BACHAREL EM CIÊNCIA DA COMPUTAÇÃO

AVALIAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NA CLASSIFICAÇÃO DO CÂNCER

DIVINO BORGES DE OLIVEIRA FILHO

Rio Verde, GO

2022

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO -
CAMPUS RIO VERDE
BACHAREL EM CIÊNCIA DA COMPUTAÇÃO

AVALIAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NA CLASSIFICAÇÃO DO CÂNCER

DIVINO BORGES DE OLIVEIRA FILHO

Trabalho de Conclusão de Curso apresentado ao Instituto Federal de Educação, Ciência e Tecnologia Goiano - Campus Rio Verde, como requisito parcial para a obtenção do Grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Heyde Francielle do Carmo França

Rio Verde, GO

Agosto, 2022

Sistema desenvolvido pelo ICMC/USP
Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas - Instituto Federal Goiano

de Oliveira Filho, Divino Borges
D618a Avaliação de Técnicas de Aprendizado de Máquina na Classificação do Câncer / Divino Borges de Oliveira Filho; orientadora Heyde Francielle do Carmo França. -- Rio Verde, 2022.
35 p.

TCC (Graduação em Ciência da Computação) -- Instituto Federal Goiano, Campus Rio Verde, 2022.

1. Aprendizado de Máquina. 2. Redes Neurais Convolucionais. 3. Classificação de Câncer. 4. Máquina de Vetores de Suporte. 5. miRNA. I. França, Heyde Francielle do Carmo, orient. II. Título.

Responsável: Johnathan Pereira Alves Diniz - Bibliotecário-Documentalista CRB-1 n°2376

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO

PARA DISPONIBILIZAR PRODUÇÕES TÉCNICO-CIENTÍFICAS NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Com base no disposto na Lei Federal nº 9.610, de 19 de fevereiro de 1998, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano a disponibilizar gratuitamente o documento em formato digital no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

IDENTIFICAÇÃO DA PRODUÇÃO TÉCNICO-CIENTÍFICA

- ☐ Tese (doutorado)
- ☐ Dissertação (mestrado)
- ☐ Monografia (especialização)
- ☑ TCC (graduação)
- ☐ Artigo científico
- ☐ Capítulo de livro
- ☐ Livro
- ☐ Trabalho apresentado em evento
- ☐ Produto técnico e educacional - Tipo:

Nome completo do autor:
Divino Borges de Oliveira Filho

Matrícula:
2017102201910030

Título do trabalho:
Avaliação de Técnicas de Aprendizado de Máquina na Classificação do Câncer

RESTRIÇÕES DE ACESSO AO DOCUMENTO

Documento confidencial: ☑ Não ☐ Sim, justifique:

Informe a data que poderá ser disponibilizado no RIIF Goiano: 22 / 08 / 2022

O documento está sujeito a registro de patente? ☐ Sim ☑ Não

O documento pode vir a ser publicado como livro? ☐ Sim ☑ Não

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O(a) referido(a) autor(a) declara:

• Que o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;

• Que obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autoria, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia Goiano os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;

• Que cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia Goiano.

Rio Verde - Goiás — Local
22 / 08 / 2022 — Data

Assinatura do autor e/ou detentor dos direitos autorais

Ciente e de acordo:

Assinatura do(a) orientador(a)

Página de assinaturas

Divino Filho
754.508.951-00
Signatário

Heyde França
006.170.711-28
Signatário

HISTÓRICO

23 ago 2022 08:52:55 — **Divino Borges de Oliveira Filho** criou este documento. (E-mail: divino100rv@gmail.com, CPF: 754.508.951-00)

23 ago 2022 08:52:58 — **Divino Borges de Oliveira Filho** *(E-mail: divino100rv@gmail.com, CPF: 754.508.951-00)* visualizou este documento por meio do IP 186.207.158.255 localizado em Rio Verde - Goias - Brazil.

23 ago 2022 08:53:01 — **Divino Borges de Oliveira Filho** *(E-mail: divino100rv@gmail.com, CPF: 754.508.951-00)* assinou este documento por meio do IP 186.207.158.255 localizado em Rio Verde - Goias - Brazil.

23 ago 2022 17:09:07 — **Heyde Francielle do Carmo França** *(E-mail: heyde.franca@ifgoiano.edu.br, CPF: 006.170.711-28)* visualizou este documento por meio do IP 177.15.85.83 localizado em Brazil.

23 ago 2022 17:11:03 — **Heyde Francielle do Carmo França** *(E-mail: heyde.franca@ifgoiano.edu.br, CPF: 006.170.711-28)* assinou este documento por meio do IP 177.15.85.83 localizado em Brazil.

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO

Ata nº 31/2022 - GEPTNM-RV/DE-RV/CMPRV/IFGOIANO

ATA DE DEFESA DE TRABALHO DE CURSO

Ao(s) 12 do mês de agosto de 2022, às 14 horas, reuniu-se a banca examinadora composta pelos docentes: Heyde Francielle do Carmo França, Heverton Barros de Macedo, Marlus Dias Silva, para examinar o Trabalho de Curso intitulado “AVALIAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NA CLASSIFICAÇÃO DO CÂNCER” do(a) estudante Divino Borges de O. Filho, Matrícula nº 2017102201910030 do Curso de Ciência da Computação do IF Goiano – Campus Rio Verde. A palavra foi concedida ao(a) estudante para a apresentação oral do TC, houve arguição do(a) candidato pelos membros da banca examinadora. Após tal etapa, a banca examinadora decidiu pela APROVAÇÃO do(a) estudante. Ao final da sessão pública de defesa foi lavrada a presente ata que segue assinada pelos membros da Banca Examinadora.

(Assinado Eletronicamente)

Heyde Francielle do Carmo França

Orientador(a)

(Assinado Eletronicamente)

Heverton Barros de Macedo

Membro

(Assinado Eletronicamente)

Marlus Dias Silva

Membro

Documento assinado eletronicamente por:

- **Heverton Barros de Macedo**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 12/08/2022 15:13:09.
- **Marlus Dias Silva**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 12/08/2022 15:12:00.
- **Heyde Francielle do Carmo Franca**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 12/08/2022 15:04:05.

Este documento foi emitido pelo SUAP em 12/08/2022. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifgoiano.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 415296
Código de Autenticação: fcc70cf0b6

INSTITUTO FEDERAL GOIANO
Campus Rio Verde
Rodovia Sul Goiana, Km 01, Zona Rural, None, None, RIO VERDE / GO, CEP 75901-970
(64) 3620-5600

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO

Declaração nº 59/2022 - GEPTNM-RV/DE-RV/CMPRV/IFGOIANO

DIVINO BORGES DE OLIVEIRA FILHO

AVALIAÇÃO DE T ÉCNICAS DE APRENDIZADO DE MÁQUINA NA CLASSIFICAÇÃO DO CÂNCER

Trabalho de curso DEFENDIDO E APROVADO em 12 de agosto de 2022, pela Banca Examinadora constituída pelos membros:

_____	_____
Dr. Heverton Barros de Macedo Instituto Federal Goiano	Marlus Dias Silva Instituto Federal Goiano

_____

Heyde Francielle do Carmo França

Orientadora

Rio Verde, GO
2022

Documento assinado eletronicamente por:

- **Marlus Dias Silva**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 19/08/2022 12:17:20.
- **Heverton Barros de Macedo**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 19/08/2022 11:30:26.
- **Heyde Francielle do Carmo Franca**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 19/08/2022 11:20:08.

Este documento foi emitido pelo SUAP em 19/08/2022. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifgoiano.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 417465
Código de Autenticação: eb57a6e964

INSTITUTO FEDERAL GOIANO
Campus Rio Verde
Rodovia Sul Goiana, Km 01, Zona Rural, None, None, RIO VERDE / GO, CEP 75901-970
(64) 3620-5600

Dedico esse trabalho as pessoas que me deram suporte e apoio para concluir essa longa jornada.

AGRADECIMENTOS

Em primeiro lugar, as pessoas que sempre acreditaram, apoiaram e incentivaram minhas escolhas. E a professora Heyde, por me acompanhar e ajudar no final desse percurso.

Você tem que ir para onde as coisas acontecem
(ROGÉRIO, Luiz, 2016).

RESUMO

BORGES, Divino.**Avaliação de técnicas de aprendizado de máquina na classificação do câncer.** Agosto, 2022. 37 f. Trabalho de Conclusão de Curso – (Curso de Bacharel em Ciência da Computação), Instituto Federal de Educação, Ciência e Tecnologia Goiano - Campus Rio Verde. Rio Verde, GO, Agosto, 2022.

O presente trabalho teve como base a importância da detecção precoce do câncer. Uma grande variedade de microRNAs (miRNAs) que indicam especificamente muitos tipos de câncer podem ser identificadas e seus perfis de expressão de miRNA analisados. Com isso, os miRNAs servem como uma ferramenta de diagnóstico de biópsia líquida não invasiva para a detecção precoce de muitos tipos de câncer. Assim sendo, foi possível avaliar técnicas de aprendizado de máquina na classificação do câncer, tendo como base um conjunto de dados obtidos do Hospital Albert Einstein. Três modelos de classificação foram experimentados, sendo dois Redes Neurais Convolucionais e um Máquina de Vetores de Suporte. Obteve-se, como resultado, que é possível treinar uma Rede Neural Convolucional com dados estruturados, porém tal conjunto precisa estar balanceado. E caso o conjunto de dados possuir apenas duas classes, modelos de classificação com Máquina de Vetores de Suporte mostram-se com melhor desempenho na classificação.

Palavras-chave: Aprendizado de Máquina. Redes Neurais Convolucionais. Classificação de câncer. Máquina de Vetores de Suporte. miRNA.

ABSTRACT

BORGES, Divino. Avaliação de técnicas de aprendizado de máquina na classificação do câncer. Agosto, 2022. 37 f. Trabalho de Conclusão de Curso – Bacharel em Ciência da Computação, Instituto Federal de Educação, Ciência e Tecnologia Goiano - Campus Rio Verde. Rio Verde, GO, Agosto, 2022.

The present work was based on the importance of early cancer detection. A wide variety of microRNAs (miRNAs) that specifically indicate many types of cancer can be identified and their miRNA expression profiles analyzed. Thus, miRNAs serve as a biopsy diagnostic tool non-invasive liquid for the early detection of many types of cancer. Therefore, it was possible to evaluate machine learning techniques in cancer classification, based on a dataset obtained from the Albert Einstein Hospital. Three classification models were tried, two Convolutional Neural Networks and one Support Vector Machine. As a result, it was possible to train a Convolutional Neural Network with structured data, but such a set needs to be balanced. And if the dataset has only two classes, classification models with Support Vector Machine show the best performance in classification.

Keywords: Machine Learning. Convolutional Neural Networks. Cancer classification. Support Vector Machine. miRNA.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANN	*Artificial Neural Network*, termo em inglês para Redes Neurais Artificiais
CNN	*Convolutional Neural Network*, termo em inglês para Rede Neural Convolucional
RNN	*Recurrent Neural Network*, termo em inglês para Rede Neural Recorrente
RNA	*Rybonucleic Acid*, termo em inglês para Ácido Ribonucleico
DNA	*Deoxyribonucleic acid*, termo em inglês para Ácido Desoxirribonucleico
API	*Application Programming Interface*, termo em inglês para Interface de Programação de Aplicações
SVM	*Support Vector Machine*, termo em inglês para Máquina de Vetores de Suporte
GCO	*Global Cancer Observatory*, termo em inglês para Observatório Global do Câncer
miRNA	microRNA
ROC	*Receiver Operating Characteristic Curve*, termo em inglês para Característica de Operação do Receptor
MCC	*Matthews Correlation Coefficient*, termo em inglês para Coeficiente de Correlação de Matthews
NLP	*Natural Language Processing*, termo em inglês para Processamento de Linguagem Natural

SUMÁRIO

1 INTRODUÇÃO

Estimativas mundiais feitas pelo *Global Cancer Observatory* (GCO) mostram que em 2020, houve o surgimento de aproximadamente 19,3 milhões de novos casos de câncer e cerca de 10 milhões de mortes ocasionadas pela doença. O estudo presume também, que a carga global de câncer em 2040, terá um aumento de 47% em comparação com 2020, chegando ao nível de 28,4 milhões de casos (SUNG et al., 2021).

De acordo com Kakushadze, Raghubanshi e Yu (2017), estima-se que, somente nos Estados Unidos, o diagnóstico precoce de câncer reduz os custos do tratamento da doença em aproximadamente 26 bilhões de dólares por ano. Estudos realçam também, que pacientes diagnosticados precocemente possuem taxa de sobrevivência entre cinco a dez vezes maiores em relação aos com a doença em processo avançado (CHO et al., 2014).

A lentidão no diagnóstico de câncer impacta no tratamento da doença. Ora ocasionada pela procura tardia do paciente, ora pelos do atendimento a saúde, a demora na detecção pode ocasionar um avanco da enfermidade, um tratamento mais agressivo e uma diminuição do índice de sobrevida (FELIPPU et al., 2016). Esse diagnóstico tardio, em se tratando de câncer oral, em que a detecção é baseado na exploração clínica completa, aumenta as taxas de mortalidade. Logo, se a Inteligência Artificial for utilizada como uma ferramenta não invasiva para a predição de câncer oral, pode melhorar o processo atual de diagnóstico precoce (GARCíA-POLA et al., 2021).

Dentre as formas de detecção de câncer, a identificação de moléculas de micro-RNA (miRNA) circulantes comporta-se como um biomarcador, capaz de antever prognóstico e diagnóstico precoce de cânceres. Os miRNAs são moléculas que atuam na regulação de vários processos fisiológicos e patológicos. Quando liberado por células cancerígenas, a identificação desses RNAs não-codantes através de fluidos corporais, colabora para a detecção precoce da doença (ALEčKOVIć; KANG, 2015).

O uso da Inteligência Artificial e Aprendizado de Máquina vem sendo temática em numerosos estudos e com aplicações em variadas áreas inclusive na medicina, indo desde aplicações na área genética (KULSKI, 2016), idenficação de doenças infecciosas (LIM; TUCKER; KUMARA, 2017), descoberta de novos medicamentos (ZHANG et al., 2017), visão computacional em diagnósticos e monitoramento de pacientes (BOHR; MEMARZADEH, 2020). Dentre as suas utilizações nesse campo, incluem a análise de imagem, oncologia de precisão, assistência em robôs cirúrgicos, descoberta de novos medicamentos, entre outros. A IA e ML são ferramentas que não substituirão os profissionais de saúde, sendo eles fundamentais para o desenvolvimento e uso de tais ferramentas (IQBAL et al., 2021).

Desta forma, o presente trabalho abrange o uso do aprendizado de máquina na detecção precoce de câncer. Posto isso, o principal objetivo é treinar modelos de

classificação/previsão, baseado em dados estruturados. Para fins de comparação, utilizamos três abordagens de aprendizado de máquina, sendo elas: a aplicação de Redes Neurais Convolucionais (CNN) com os dados unidimensionais e bidimensionais, e Máquina de Vetores de Suporte (SVM).

Portanto, ao longo desse trabalho será respondida a pergunta de qual técnica de aprendizado de máquina é mais eficiente, dado uma certa base de dados. Dessa forma, o presente trabalho apresentará uma proposta de um sistema para classificação de pacientes com câncer e saudáveis, fundamentado em modelos computacionais de Inteligência Artificial, os quais serão sustentados por dados de sequenciamento genético originários de pacientes do Hospital Albert Einstein(SP). O programa estará preparado para identificar padrões em dados estruturados e prever se um indivíduo está ou não com câncer, mostrando também a probabilidade do resultado ser confiável.

Há um considerável desequilíbrio no conjunto de dados utilizados que dificulta o aprendizado da classe minoritária, em consideração a isso, durante o desenvolvimento do modelo computacional preditivo, foram empregadas técnicas de balanceamento dos dados. Diante disso, para realizar a classificação, fez-se necessário dividir o processo em algumas etapas, dentre elas: o pré-processamento e redimensionamento da base de dados estruturada, treinamento do modelo computacional preditivo fazendo o uso das técnicas mencionadas anteriormente, realizar o treinamento do mesmo, e por fim, proceder com a avaliação e comparação da acurácia e outras métricas dos resultados gerados pelo protótipo.

Nos próximos capítulos serão apresentados a Revisão de Literatura no Capítulo 3, Material e Métodos no Capítulo 4, os Resultados e Discussão no Capítulo 5 e as Conclusões no Capítulo 6.

2 TRABALHOS RELACIONADOS

Os autores Koppad et al. (2022) evidenciaram o uso de diferentes métodos de aprendizado de máquina no estudo e identificação de assinaturas genéticas em pacientes com câncer colorretal. A pesquisa faz uso de seis métodos de ML, como classificadores, para identificar genes que podem ser assumidos como marcadores para o diagnóstico do cancer colorretal, dando assim ênfase à aplicabilidade ML para a realização de análises preditivas. Isto posto, as menções expõem a relevância de recorrer ao subcampo de Inteligência Artifical durante a investigação, estudo e diagnóstico da doença.

Na Figura 1, os autores Koppad et al. (2022) apresentam uma representação esquemática do fluxo de trabalho de identificação de biomarcadores.

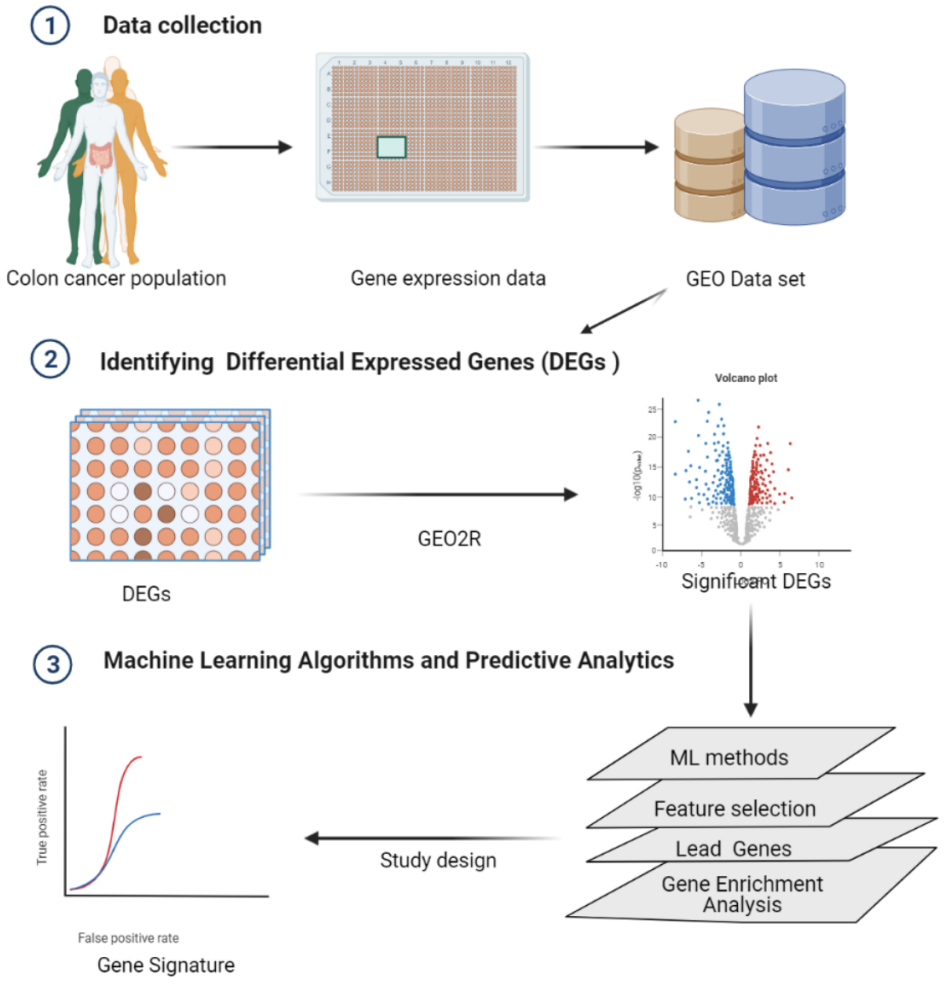

Figura 1 – Uma representação esquemática do fluxo de trabalho de identificação de biomarcadores. Fonte: (KOPPAD et al., 2022).

Os autores Koppad et al. (2022) construíram a curva ROC dos modelos que tiveram o melhor desempenho nas diferentes combinações de dados de treinamento e teste. Conforme Figura 2, nos três conjuntos de dados testados, Random Florest e Regressão Logística

obtiveram o melhor desempenho quando conjuntos de dados GSE44861 e GSE20916 foram combinados como dados de treinamento e teste, respectivamente.

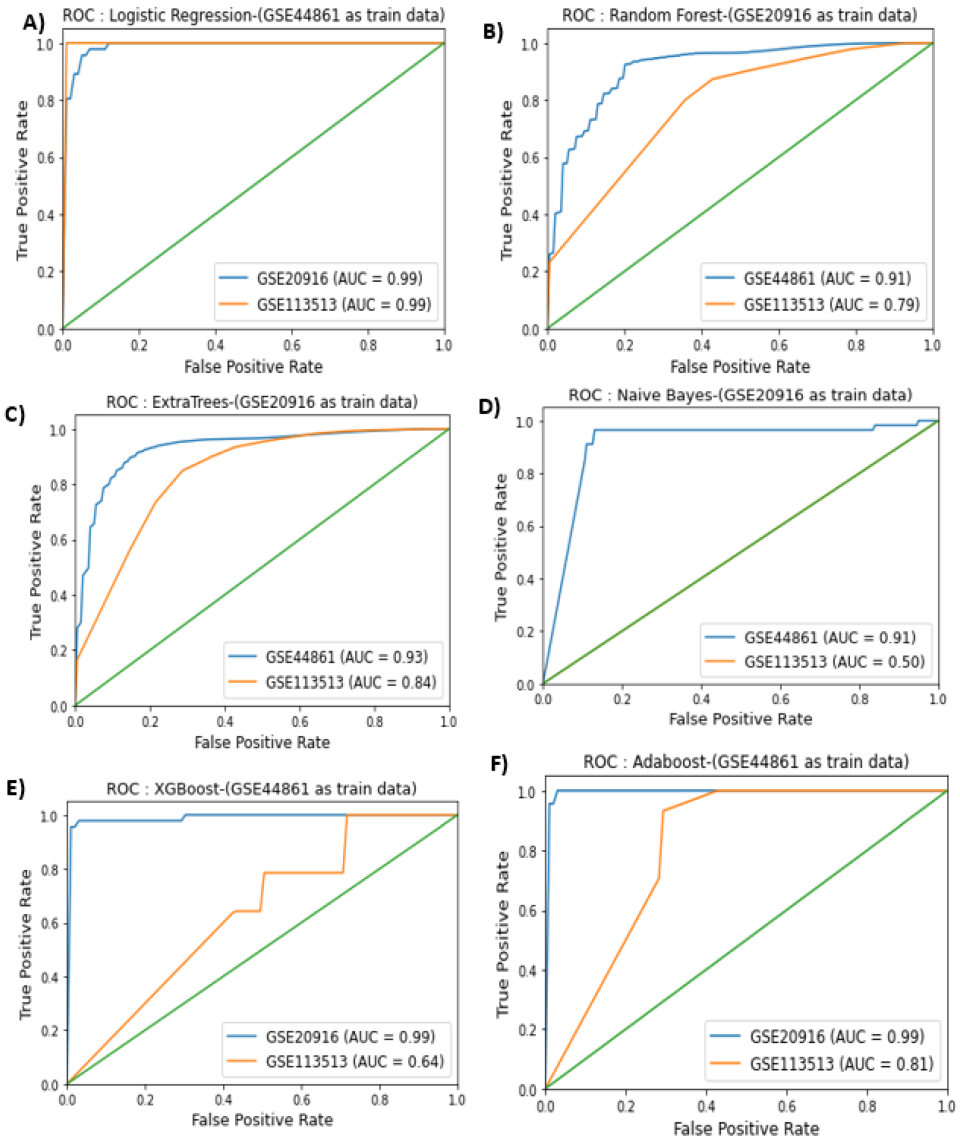

Figura 2 – Curvas ROC para os diferentes classificadores. Fonte: (KOPPAD et al., 2022).

No entanto, Koppad et al. (2022) registraram que nenhum dos classificadores avaliados obtiveram um bom desempenho utilizando o conjunto de dados GSE113513. Portanto, os modelos de Random Florest exibiram desempenho consistentemente melhor em todas as classificações modelos testados, afirmam Koppad et al. (2022).

Os autores Bohr e Memarzadeh (2020) resumem de forma simples a aplicação de Redes Neurais Convolucionais em imagens de radiologia. Conforme a Figura 3, os autores mostram que a camada de convolução, é onde uma imagem se torna essencialmente uma pilha de imagens filtradas. Explicam também, em seguida, um agrupamento é formado com todas essas imagens filtradas, em que o conjunto original de imagens torna-se uma representação menor delas mesmas e todos os valores negativos são removidos por uma unidade linear retificada (Relu).

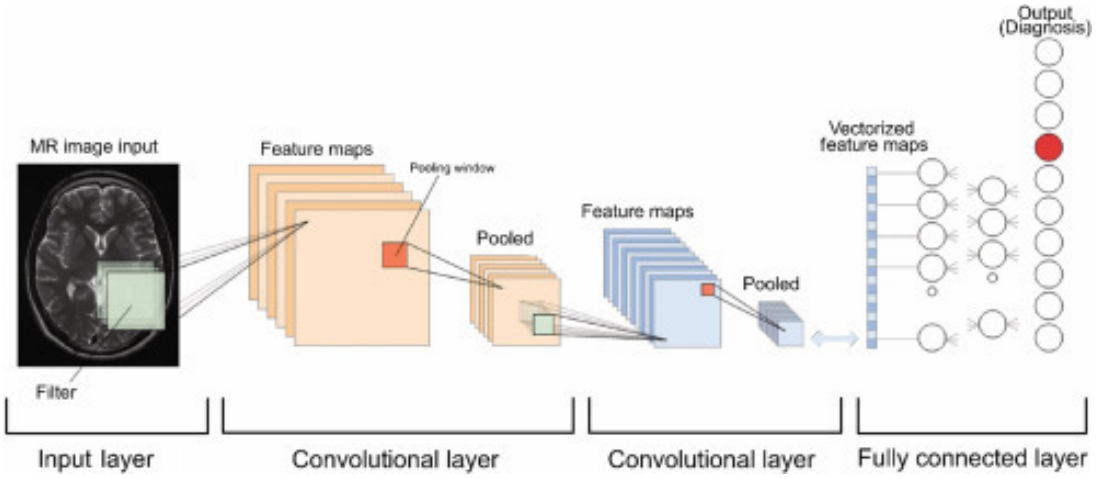

Figura 3 – Os vários estágios das Redes Neurais Convolucionais em ação. Fonte: (BOHR; MEMARZADEH, 2020)

Explicam também, que a última camada é chamada de camada totalmente conectada, em que cada valor atribuído a todas as camadas contribuirá para quais serão os resultados. Em caso do algoritmo produzir um erro na final, o gradiente descendente pode ser aplicado ajustando os valores para cima e para baixo para ver como o erro muda em relação à resposta certa de interesse. Isso pode ser alcançado por um algoritmo chamado retropropagação que aprende com os erros.

Logo, os autores Bohr e Memarzadeh (2020) finalizam com a explicação de que depois de aprender um novo recurso a partir do conjunto de dados iniciais, o modelo treinado pode ser aplicado a novas imagens e o sistema pode classificá-las na categoria correta, ação chamada de previsão ou inferência, semelhante à forma como um radiologista executa seu trabalho.

Em (ELIAS et al., 2017), os autores avaliam várias técnicas computacionais com o objetivo de identificar padrões de alterações nos níveis de expressão de miRNAs que tornem possíveis discriminar câncer de ovário de tumores benignos e salientam a dificuldade em se obter dados homogêneos de miRNA para qualificação e utilização em estudos de predição e/ou classificação.

Neste trabalho (ELIAS et al., 2017), foi realizada uma comparação de desempenho de onze ferramentas, desde as estatísticas, como LDA e regressão linear, até aquelas que incluem aprendizado de máquina, entre estes classificadores como Bayes, Random forest, SVM e o que obteve melhor resultado foi a rede neural Multi Layer perceptron, com 7 neurônios na camada oculta. Este é o primeiro estudo em câncer de ovário a combinar a tecnologia de sequenciamento de próxima geração para miRNAs séricos com técnicas de aprendizado de máquina. Mesmo está técnica sendo a que obteve melhores resultados, estes giram em torno de 75 a 90%, ainda com uma boa margem para melhoria, mas satisfatória quando comparada com o diagnóstico utilizando técnicas convencionais. Os autores corroboram a viabilidade de utilizar dados de miRNA para classificação de câncer

de ovário.

Em (KOTHANDAN; BISWAS, 2015), é apresentado um método de identificação de miRNA associados ao câncer usando *Support vector machine* (SVM). Foi realizado um estudo *in silico* envolvendo a identificação de assinaturas globais em microRNAs validados experimentalmente associados ao câncer. Subsequentemente, um classificador baseado em SVM binário foi treinado e modelado utilizando as características extraídas do estudo acima. Um total de 60 características distintas foram selecionadas e classificadas para formar o conjunto de recursos para a classificação. O modelo foi dividido em dois classificadores, miRSEQ (com os dados positivos da base de dados) e miRINT(com os dados negativos da base de dados). A ferramenta teve um desempenho satisfatório, apresentando medidas de desempenho razoavelmente boas com valores do coeficiente de correlação de Matthews (MCC) que variam de 0,72 a 0,82, os autores defendem que o desempenho da previsão deve melhorar conforme o número de dados experimentalmente validados aumentarem.

3 REVISÃO DE LITERATURA

3.1 Inteligência Artificial

A denominação de Inteligência Artificial (IA) diz respeito a tentativa de imitar a inteligência humana dentro de um sistema computacional. A capacidade de simulação engloba as principais habilidades do cognitivo humano, como por exemplo, a percepção, o raciocínio, o aprendizado, o planejamento e a previsão. De um sistema com essas habilidades, espera-se precisão em tomadas de decisões, assim como seria se fossem seres humanos (XU et al., 2021).

Em razão de sua capacidade de aprendizado e tomada de decisão, a IA tornou-se aplicável em diversas áreas. Dentre as aplicações mais populares estão: análise preditiva e inteligência nas tomadas de decisões; análise de comportamento de usuários; agricultura sustentável; reconhecimento de imagem, fala e padrões em dados; Processamento de Linguagem Natural (NLP); recomendação de produtos em lojas virtuais; cuidados de saúde; entre outros (SARKER, 2021).

Dentre as utilizações pluridisciplinares, o uso da IA destaca-se também na área da saúde. Bohr e Memarzadeh (2020) reuniram em sua publicação os principais empregos dessa tecnologia nesse setor, abrangendo utilizações com dados genéticos, visão computacional no diagnóstico de doenças e em cirurgias, aprendizado profundo no reconhecimento de imagens e monitoramento de pacientes, atendendo assim, as necessidades de médicos, pacientes e outros profissionais de saúde.

Já na descoberta de novos medicamentos, as técnicas de aprendizado de máquina e aprendizado profundo estão sendo usadas para simplificar cada etapa do processo, desde os estudos sobre o composto ativo da droga, até a predição de reação da mesma (ZHANG et al., 2017). Na maneira convencional, por ser um processo demorado devido a várias etapas e testes a serem feitos, pode levar até 15 anos entre o início da pesquisa e a homologação da droga. Além disso, o custo de todo o processo é caro, em alguns casos, ultrapassando 1 bilhão de dólares (HUGHES et al., 2011).

Em um cenário futuro de saúde pública, estima-se que as próximas gerações terão acesso ao seu sequenciamento completo do genoma, o que contribuirá para uma medicina mais precisa. Nesse cenário, todos os sistemas de saúde existentes nos dias atuais precisarão de adequações para tirar proveito dessas informações genéticas (KULSKI, 2016). Esses ajustes serão possíveis através de ferramentas de Big Data e IA, que são capazes de processar uma grande quantidade de dados e encontrar padrões nos mesmos.

3.1.1 Aprendizado de Máquina

Os autores Panch, Szolovits e Atun (2018) salientam que o Aprendizado de Máquina (ML) é o estudo de algoritmos para desenvolvimento de aplicações através da análise de exemplos de um comportamento desejado, ao invés de programá-los diretamente, como é feito na programação comum. Esse subconjunto de IA, é programado para aprender associações com base em grandes quantidades de dados brutos, e usa um conjunto amplo de técnicas estatísticas para treinamento e predição em um modelo.

O agrupamento feito por Sharma, Sharma e Jindal (2021) reune as três abordagens principais de ML, sendo elas:

- Aprendizagem supervisionada, em que é fornecido os dados de treinamento para o modelo;
- Aprendizagem não supervisionada, onde o algoritmo encontra padrões escondidos dentro da base de dados; e
- Aprendizagem por reforço, no qual os comportamentos desejados são recompensados e os indesejados são punidos.

As estratégias de aprendizado supervisionado equivale a utilizar dados categorizados para o treinamento de um modelo, e no fim, com base nesse treinamento, poderá ser utilizado para classificação de alguma informação de entrada (LIM; TUCKER, 2016).

Os autores Sidey-Gibbons e Sidey-Gibbons (2019) reiteram que essa classe de algoritmos de aprendizado de máquina recebem um conjunto de dados contendo diversas variáveis de entrada associadas a um resultado conhecido. Destacam também que tal comportamento pode representar o treinamento de um modelo para relacionar características de uma pessoa, por exemplo, altura, peso, tabagismo, a um determinado resultado como, por exemplo, diabetes. E salientam ainda, que uma vez realizado o treinamento do modelo, a partir de então ele será capaz de realizar previsões de resultados quando aplicado a novos dados. Resultados que podem ser discretos como, por exemplo, positivo ou negativo, benigno ou maligno; ou contínuo como, por exemplo, um valor qualquer entre 0 e 100.

Os algoritmos de aprendizagem não supervisionada, diferente da supervisionada, treinam a máquina para encontrar padrões e estruturas escondidas em conjunto de dados não categorizados (BASHIR; QAMAR; KHAN, 2016).

Em Lim, Tucker e Kumara (2017) popuseram um modelo de aprendizado de máquina não supervisionado com o objetivo de descobrir doenças infecciosas latentes sem usar características predeterminadas da doença. Além disso, apresentam outros exemplos de aplicações dessa técnica em áreas biomédicas, incluindo reconhecimento de dados textuais; análise de discussões sobre drogas farmacêuticas em redes sociais baseado em particularidades da respectiva ferramenta social; avaliação de risco clínico; e auxílio em diagnósticos laboratoriais.

3.2 Modelos Preditivos

3.2.1 Máquina de vetores de suporte

Uma Máquina de Vetores de Suporte (SVM) é um tipo de algoritmo de aprendizado de máquina supervisionado, frequentemente usado em problemas de classificação de duas classes (CORTES; VAPNIK, 1995).

O autor Noble (2006) declara que aplicando o conceito estatístico de regressão linear, a SVM segue a principal ideia de funcionamento criando um limite de decisão entre duas classes que possibilite a previsão de rótulos de um ou mais vetores de características, implementando o seguinte princípio de funcionamento:

Conforme Huang et al. (2018) exibem, dado um conjunto de dados para treinamento:

$$(x_1, y_1), ..., (x_n, y_n),\ x_i \in R^d \ e\ y_i \in (1, +1) \tag{1}$$

Onde x_i é uma representação de vetor de características e y_i o rótulo da classe para tais atributos em uma iteração de treinamento i.

Esse limite de decisão entre duas classe é criado atráves de um hiperplano, que em seu melhor dos casos pode ser ser definido como:

$$wx^T + b = 0 \tag{2}$$

No qual w é o vetor de peso, x o vetor de atributos de entrada e b é o *bias*.

O w e o *bias* satisfariam as seguintes desigualdades para todos os elementos do conjunto de treinamento, se:

$$w{x_i}^T + b \geq 1\ se\ y_i = 1 \tag{3}$$

$$w{x_i}^T + b \leq -1\ se\ y_i = -1 \tag{4}$$

Dado isso, o principal objetivo de treinar um modelo de SVM é encontrar o w e o *bias* para que o hiperplano separe os dados e maximize a margem $1/||w||^2$.

As definições supracitadas são apresentadas de forma gráfica na Figura 4, na qual apresenta um modelo de SVM linear.

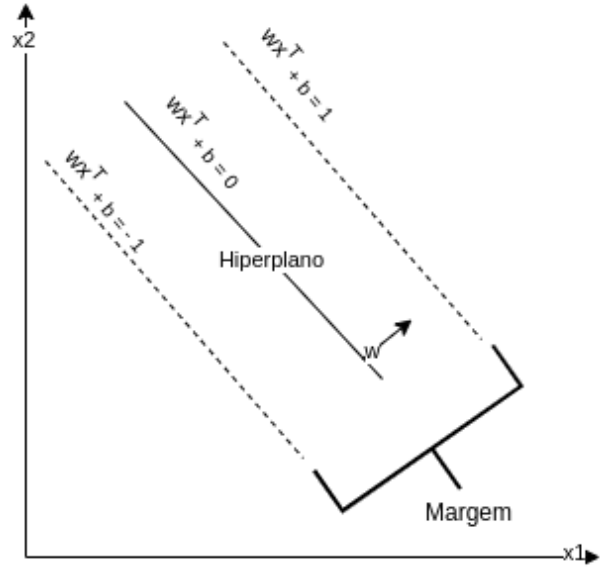

Figura 4 – Modelo de SVM linear. Fonte: Autor.

A apresentação dos Cervantes et al. (2020) esclarece que o uso do SVM em reconhecimento de padrões é de uma importância significante. Elucidam também, que esse papel significativo do emprego do SVM é em razão do bom desempenho do algoritmo aplicado a grandes conjuntos de dados, classificação múltipla e para conjuntos de dados não balanceados.

3.2.2 Redes Neurais Artificiais

Os autores Agatonovic-Kustrin e Beresford (2000) define uma Rede Neural Artificial (ANN) como um modelo computacional inspirado nos neurônios e no processamento de informações feito pelo cérebro humano. Assim como a inteligência humana, as ANNs aprendem através da experiência, reunindo conhecimento encontrados em padrões e relacionamentos entre dados.

Tal como os neurônios biológicos, o neurônios em uma ANN também recebem várias entradas, as somam e processam o resultado com uma função de ativação, podendo em seguida, transmití-los para vários outros neurônios por meio de ligações chamadas sinapses (HAN et al., 2018).

Ainda enunciado por Han et al. (2018), o processo de aprendizado de uma ANN traduz-se em atualizar os pesos das sinapses de um neurônio, no qual dado o erro entre o valor previsto e o correto, o peso é corrigido no objetivo de minimizar o erro e um nova saída próxima do valor verdadeiro seja encontrado.

As primeiras Redes Neurais concebidas são simples e de camada única, no qual implementam uma regra básica de aprendizado supervisionado, e recebem o nome de Perceptrons (WALLISCH et al., 2009). Originárias da apresentação do Rosenblatt (1958), as Redes Perceptrons, seguem o princípio de funcionamento que consiste em fornecer um

número finito de entradas ponderadas $w_i x_i$ ao nerônio e espera-se uma saída binária de 0 ou 1. Essa definição é apresentada na Figura 5, no qual mostra um modelo Perceptron de Rede Neural Artificial.

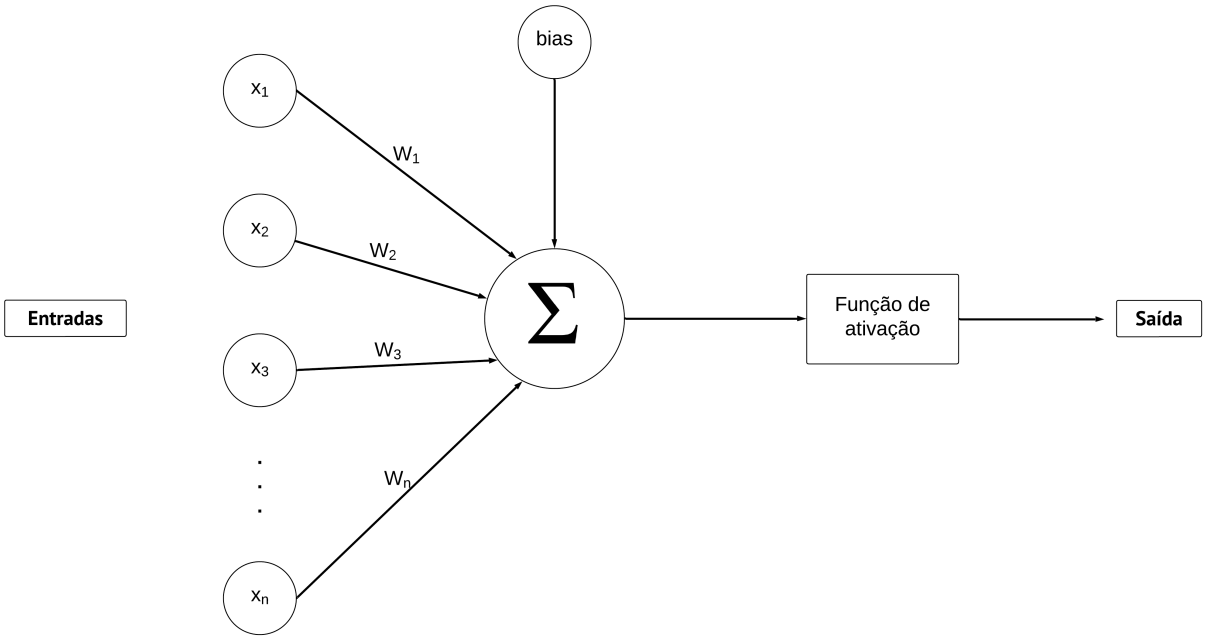

Figura 5 – Modelo Perceptron de Rede Neural Artificial. Fonte: Autor.

Considerando o *bias* uma constante que não depende de nenhum outro valor, o processo de treinamento de um modelo Perceptron traduz-se em ajustar os pesos w_i para que aproximam-se das entradas classificadas como 1 e distanciam-se dos classificados como 0 (WALLISCH et al., 2009).

Logo, o modelo matemático básico de classificação do Percepron retorna 1, caso satisfaça a inequação algébrica $\sum_{i=1}^{n} w_i x_i + bias > 0$. Caso contrario, é retornado 0.

3.2.3 Redes Neurais Feed-forward

As ANNs, como Sazli (2006) apresenta, dependendo do tipo de conexão entre neurônios, podem ser classificadas dentre: Rede Neural Feed-forward e Rede Neural Recorrente.

A denominação de Rede Neural Feed-forward é dadas àquelas Redes Neurais que não possuem um retorno, também chamado de *feedback*, com origem na Camada de Saída em direção a Camada de Entrada. De outro modo, se houver tal conexão sináptica, a Rede é nomeada como Rede Neural Recorrente. Evidencia também, que ambas as Redes Neurais são organizadas em camadas, o qual dependendo da quantidade podem ser classificadas também como Rede Neural de Camada Única ou Rede Neural de Múltiplas Camadas (SAZLI, 2006).

A Figura 6 apresenta um exemplo de Rede Neural Feed-Forward de camada única.

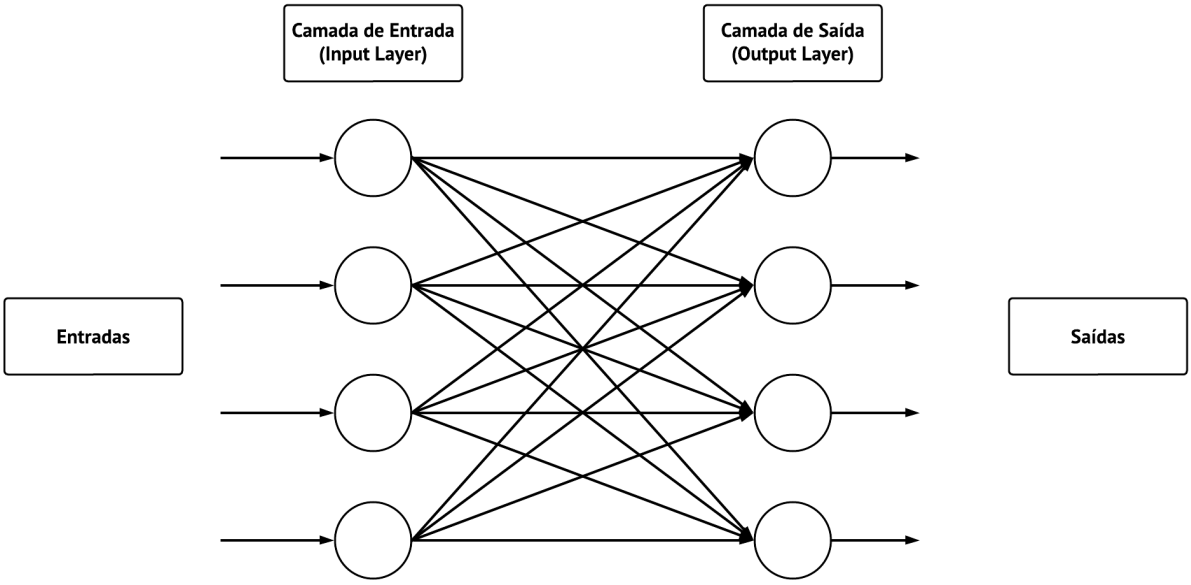

Figura 6 – Exemplo de Rede Neural Feed-Forward de camada única. Fonte: Autor.

Ainda apresentado por Sazli (2006), caso a Rede Neural Feed-Forward seja de múltiplas camadas, as camadas situadas entre a Camada de Entrada (Input Layer) e a Camada de Saída (Output Layer), são chamadas de Camadas Ocultas (Hidden Layers) e possuem a finalidade de realizar computação entre a Camada de Entrada e saída da Rede.

3.2.4 Redes Neurais Convolucionais

O principal objetivo das Redes Neurais Convolucionais (CNN), também chamada de ConvNet, é exceder as limitações dos algoritmos de aprendizado de máquina tradicionais.

Conforme Indolia et al. (2018) mostram, a CNN no qual é implementada com uma arquitetura Feed-forward, possui capacidade de aprender automaticamente com base em representações de recursos complexos, como por exemplo, uma imagem, podendo extrair informações suficientes para classificá-las, obtendo assim desemprenho melhor do que os algoritmos tradicionais.

Na Figura 7 é mostrado uma representação de imagem na qual os olhos humanos conseguem enxergar. Por outro lado, na Figura 8 é apresentada a representação da imagem anterior dentro de um sistema computacional.

Figura 7 – Como os olhos humanos enxergam. Fonte: Autor.

```
[[[  7   6   4]
  [  6   5   3]
  [  5   4   2]
  ...
  [146 125 122]
  [143 122 119]
  [139 120 116]]

 [[  5   4   2]
  [  6   5   3]
  [  5   4   2]
  ...
  [148 124 122]
  [145 124 121]
  [141 122 118]]

 [[  5   4   2]
  [  5   4   2]
  [  5   4   2]
  ...
  [145 121 121]
  [143 122 119]
  [147 126 123]]

 ...

 [[168  27 167]
  [169  28 166]
  [170  28 166]
  ...
  [155   0 122]
  [158   0 125]
  [159   0 126]]

 [[173  32 170]
  [173  32 170]
  [175  33 171]
  ...
  [155   0 124]
  [157   0 122]
  [160   2 125]]

 [[173  32 170]
  [172  31 169]
  [177  32 173]
  ...
  [155   0 124]
  [155   0 120]
  [159   1 124]]]
```

Figura 8 – Como o computador ”enxerga”. Fonte: Autor.

Muito aplicadas em soluções que envolvem imagens, a CNN pode realizar a redução de dimensionalidade de uma imagem, com a finalidade de reduzir o número de parâmetros que o algoritmo precisa computar. Ademais, uma CNN pode receber como entrada imagens bidimensionais e tridimensionais (WANG et al., 2021).

Exibido por Zhu et al. (2021), além da utilização em reconhecimento de imagens e vídeos, a ConvNet pode ser usada no processamento de linguagem natural e reconhecimento de fala. Mostram também, dentre outras aplicações estão áreas como bioinformática, finanças, entre outros, em que muitos dados dessas áreas não são imagens, mas sim dados estruturados. Nesse cenário, a Rede Neural Convolucional não pode ser aplicada com sua máxima eficiência, e nos casos de dados tabulares, tal coleção deve ser reorganizada para

um espaço bidimensional, com o objetivo de representar as relações entre os dados. Dado isso, Zhu et al. (2021) afirmam que há a motivação de transformar dados tabulares em imagens para que a CNN possa melhorar o desempenho da previsão, comparado a modelos treinados com dados tabulares.

3.3 MicroRNAs (miRNAs)

Os MicroRNAs (miRNAs) são uma classe de RNAs não codificantes que atuam na regulação da expressão gênica. Grande parte dessas moléculas são transcritas de sequências de DNA em miRNAs primários e processados em miRNAs precursores e, por último, em miRNAs maduros. (O'BRIEN et al., 2018)

Os autores Dragomir, Knutsen e Calin (2021) apontam que algumas mudanças na expressão de miRNAs são indicativos de iniciação e progressão de cânceres humanos. Afirmam também, que a localização de uma expressão diferencial desse ácido nucleico em regiões genômicas associadas a cânceres podem exemplificar, por exemplo, uma malignidade. Dentre as expressões diferenciais estão: os mecanismos epigenéticos, desregulação transcricional, modificações químicas e edição e alterações em proteínas da biogênese do miRNA.

Em Condrat et al. (2020) reforçam que as moléculas de miRNAs possuem grande importância quando estão associados a um doença, dispondo de um papel estratégico no diagnóstico, prognóstico e tratamento. Realçam também, conforme as técnicas atuais evoluem, será comum o uso dos miRNAs para criar perfis personalizados de pacientes, o que irá permitir intervenções terapêuticas direcionadas.

3.4 Pré-processamento de dados

Conforme os autores Kotsiantis, Kanellopoulos e Pintelas (2006) apresentam, um dos fatores de sucesso do Aprendizado de Máquina é a representação e qualidade da base de dados, visto que dados irrelevantes, redundantes e não confiáveis, atrapalham a fase de treinamento de um modelo de classificação.

3.4.1 Normalização e Desbalanceamento

Em concordância ao estudo feito por Singh e Singh (2020), a performance dos algoritmos de aprendizado de máquina está diretamente relacionada com a qualidade dos dados, cujo objetivo final é alcançar um modelo preditivo generalizado do problema de classificação. Enfatiza também que a normalização é de extrema importância para melhorar a qualidade dos dados e, por consequência, o desempenho dos algoritmos de aprendizado de máquina.

Os autores França et al. (2021) mostram que o Big Data potencializa recursos que organizam e catalogam esses dados, aumentando a disponibilidade de dados relevantes para a tomada de decisões utilizando algoritmos preditivos.

A normalização tem como finalidade transformar todos os dados para que tenham a mesma importância. Para isso, essa técnica redimensiona os valores dos atributos para que os mesmos fiquem numericamente no mesmo intervalo (JAVAHERI; SEPEHRI; TEIMOURPOUR, 2014).

Sendo uma maneira comum de normalizar dados, o objetivo da equação abaixo, usualmente denominada Normalização Min-Max, é que dado um conjunto de dados, o mesmo é transformado em um novo conjunto, no qual o valor mínimo pertencente ao conjunto inicial é transformado em 0, o valor máximo é transformado em 1 e todos os outros valores são transformados em um decimal entre 0 e 1.

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{5}$$

Suplementário a normalização, Yanminsun, Wong e Kamel (2011) demonstram que o desbalanceamento de dados usados no treinamento de modelos preditivos, expõe grande desvantagem em comparação a conjunto de dados balanceados.

Conforme Johnson e Khoshgoftaar (2019) apresentam, é comum os dados brutos na maioria dos *datasets* não serem balanceados, o qual significa que em sua maioria há uma diferença quantitativa entre as classes, havendo assim a divisão entre classe majoritária e minoritária. Ainda ressaltam que para lidar com esse desequilíbrio de classe faz-se necessário aplicar algumas técnicas de desbalanceamento.

Da maneira exposta por Lee e Seo (2022), o Downsample é uma técnica que visa a seleção de amostras mais significativas dentro da classe majoritária em um conjunto de dados desbalanceados. Exibem também que o critério para essa separação são métricas ótimas para a minimização do erro de generalização do modelo treinado.

Um outra alternativa ao Downsample, a técnica de NearMiss, originária da apresentação do Zhang e Mani (2003), propõe a estratégia de subamostragem com base na distância entre um exemplo da classe majoritária e minoritária. Apresentaram também, três algoritmos implementando essa técnica, sendo eles:

- NearMiss-1: cujo objetivo é selecionar exemplos negativos baseados na distância mínima de três exemplos positivos mais próximos;
- NearMiss-2: sendo o mais performático na análise feita por Zhang e Mani (2003), visa selecionar exemplos negativos que estão próximos a todos os exemplos positivos e que podem ser distribuídos entre os exemplos positivos; e
- NearMiss-3: que objetiva cada exemplo positivo ser cercado por exemplos negativos selecionados, aumentando a precisão e reduzindo o recall.

3.5 Métricas de Desempenho

As métricas de desempenho são importantes para a avaliação de um modelo de classificação, e é útil para aferir a precisão do classificador, dentre outras métricas correlacionadas (LIU et al., 2014).

A matriz de confusão, ilustrada na Figura 9, é formatada em uma tabela, no qual aponta a performance de um algoritmo de classificação (SINGH et al., 2021).

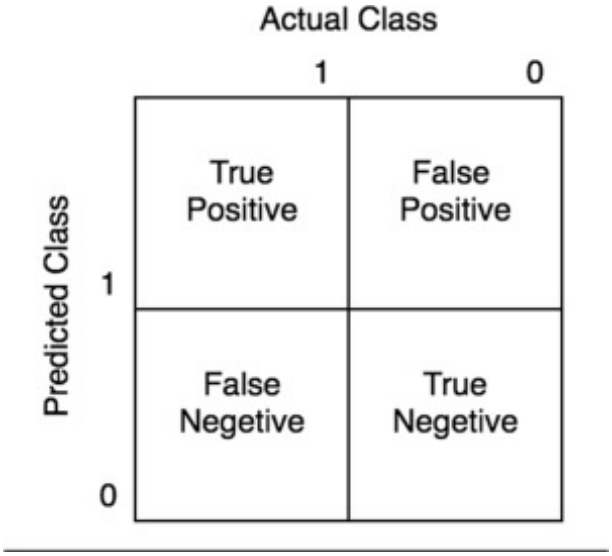

Figura 9 – Matriz de Confusão para classificação binária. Fonte: (SHARMA et al., 2022)

Na Tabela 1 são apresentadas algumas terminologias usadas na matriz de confusão.

TP	*True Positive*	Verdadeiro Positivo
TN	*True Negative*	Verdadeiro Negativo
FP	*False Positive*	Falso Positivo
FN	*False Negative*	Falso Negativo

Tabela 1 – Terminologia de matriz de confusão (SINGH et al., 2021)

A acurácia, assim como outras métricas, deviva da matriz de confusão e é a métrica mais comum. Ela expressa o quanto o modelo classificou acertadamente (KULKARNI; CHONG; BATARSEH, 2020).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

A precisão representa a proporção de casos pevistos que expressam a presença da condição ou característica avaliada (POWERS, 2008).

$$precision = \frac{TP}{TP + FP} \tag{7}$$

Em complemento, o recall representa a proporção de casos pevistos que expressam a real presença da condição ou característica avaliada (ARJARIA; RATHORE; CHERIAN, 2021).

$$recall = \frac{TP}{TP + FN} \tag{8}$$

Os autores Zou, O'Malley e Mauri (2007) revelam que outra análise de métrica importante é a da curva Característica de Operação do Receptor, comumente chamada de curva ROC, da qual sua utilidade é avaliar o desempenho e precisão de um mobelo de classificação. Referindo as suas características e conforme apresentado na Figura 10, uma curva ROC possui um gráfico de sensibilidade no eixo y contra (1-especificidade) no eixo x para valores variados do limiar t. A linha diagonal de 45° conectando as coordenadas (0,0) a (1,1) é a curva ROC correspondente ao acaso aleatório. A curva ROC para o padrão-ouro é a linha que liga (0,0) a (0,1) e (0,1) a (1,1), sendo que em geral, as curvas ROC ficam entre esses 2 extremos. A área abaixo da curva ROC, denominada AUC (do inglês, *Area Under the Curve*) é uma área que calcula a média da precisão com base em todos os valores de teste.

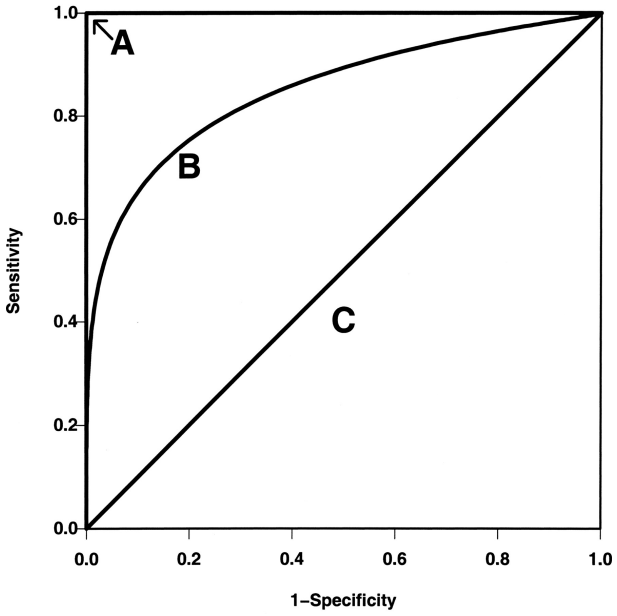

Figura 10 – Representação de três curvas ROC hipotéticas. Fonte: (ZOU; O'MALLEY; MAURI, 2007)

Na figura acima, Zou, O'Malley e Mauri (2007) ilustram as três curvas ROC hipotéticas representando a precisão do padrão-ouro (linha A; AUC=1) nos eixos superior

e esquerdo no quadro unitário, uma curva ROC típica (curva B; AUC=0,85) e uma linha diagonal correspondente a chance aleatória (linha C; AUC=0,5). À medida que a precisão do teste melhora, a curva ROC se move em direção a A e a AUC se aproxima de 1.

4 MATERIAL E MÉTODOS

4.1 Ferramentas

Como ferramenta de trabalho, utilizou-se um computador com as seguintes configurações: processador Intel Core i5-10400 2.90GHz × 12, 32GB de memória RAM e 240GB armazenamento em disco de estado sólido (SSD).

Para o desenvolvimento e treinamento dos modelos preditivos, e avaliação de métricas de desempenho, houve o emprego da linguagem Python dentro do ambiente de desenvolvimento do Google Colaboratory. Além disso, fez-se necessário o uso da biblioteca Tensorflow, da qual seu objetivo é abstrair diversos algoritmos de aprendizado de máquina. Ademais, foi feito o uso de bibliotecas auxiliares como a Sklearn e NumPy.

4.1.1 Linguagem de Programação Python

Python é uma linguagem de programação criada por Guido Van Rossum em 1989, sendo uma linguagem multiparadigma, orientada a objetos, imperativa, funcional e interpretada (ROSSUM; TEAM, 2016). Atualmente, por ser uma linguagem de fácil entendimento, python é usada em numerosos casos, sendo alguns deles: visualização e análise de dados, *Machine Learning*, automação de tarefas e desenvolvimento de sites e de APIs.

4.1.2 Google Colaboratory

O Google Colaboratory, comumente chamado de "Colab", é uma ferramenta online, com ambiente para desenvolvimento interativo, para criação de protótipos de modelos de aprendizado de máquina em servidores na nuvem do Google.

4.1.3 Biblioteca Tersorflow

Tersorflow é uma plataforma e biblioteca para várias linguagens, utilizada no desenvolvimento de algoritmos para Aprendizado de Máquina. Dentre as suas funções, encontra-se diversos algoritmos para criação e treinamento de Redes Neurais Artificiais (ANN), usados para detectar padrões e correlações. (ABRAHAMS et al., 2016)

4.1.4 Biblioteca Sklearn

Sklearn é uma biblioteca, para a linguagem de programação Python, utilizada no desenvolvimento de algoritmos para Aprendizado de Máquina. Dentre as suas funções, encontra-se diversos algoritmos de regressão, agrupamento e classificação, entre eles estão

os mais conhecidos, como por exemplo: Support Vector Machine (SVM), Random Florest, k-means. (PEDREGOSA et al., 2011)

4.1.5 Biblioteca NumPy

NumPy é uma biblioteca, para a linguagem de programação Python, com a finalidade de processar matrizes multi-dimensionais. Ela é utilizada para desenvolver algoritmos para diversos fins, como a física, química, astronomia, geociência, biologia, psicologia, ciência dos materiais, engenharia, finanças e economia. Também em campos da astronomia, o que já foi usada na descoberta de ondas gravitacionais de um buraco negro (HARRIS et al., 2020).

4.2 Construção dos modelos de classificação

Para o progresso do presente trabalho, a construção dos modelos de classificação seguiram os passos exibidos na Figura 11 e esmiuçados nas seções seguintes.

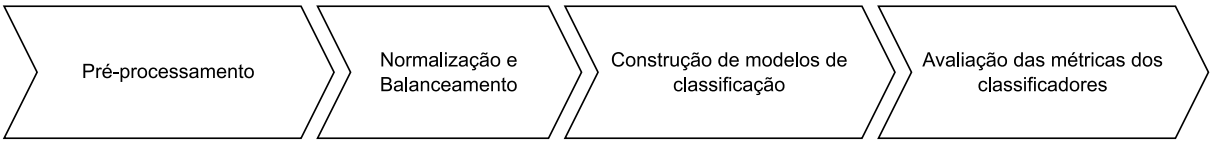

Figura 11 – Fluxo de trabalho do desenvolvimento dos modelos de classificação. Fonte: Autor.

4.3 A Base de Dados

Obtido do Hospital Albert Einstein de São Paulo, o conjuto de dados utilizados nesse estudo veio estruturado na forma de tabela, em que foram registrados 552 pacientes, no qual cada um deles tiveram 1181 miRNAs sequenciados.

Em sua estrutura inicial, a base da dados contém 1038312 registros, conforme apresentado na Tabela 2, em que cada grupo de 1181 registros representam as leituras dos miRNAs de um único indivíduo. A última coluna, denominada classe, caracteriza a presença ou ausência de câncer no paciente examinado, sendo simbolizado por 1 ou 0, respectivamente.

FILE	MIRNA	READ	READS	classe
1	1	8735	76124	1
1	2	8442	74836	1
1	3	8801	76411	1
1	4	23959	126073	1
1	5	7706	71499	1
⋮	⋮	⋮	⋮	⋮
1038312	1181	7706	71499	1

Tabela 2 – Base de dados original de registros de miRNAs em pacientes examinados

Após compreensão da base de dados, realizou-se a normalização empregando a técnica da Normalização Min-Max, descrita no Capítulo 3.

Como o conjunto de dados original esteve desbalanceado, conforme apresentado na Figura 12, contendo o número de registros de miRNAs em pacientes com câncer maior que em pacientes sadios, empregou-se as técnicas de Downsample e NearMiss, descrita no Capítulo 3, para balancear o conjunto.

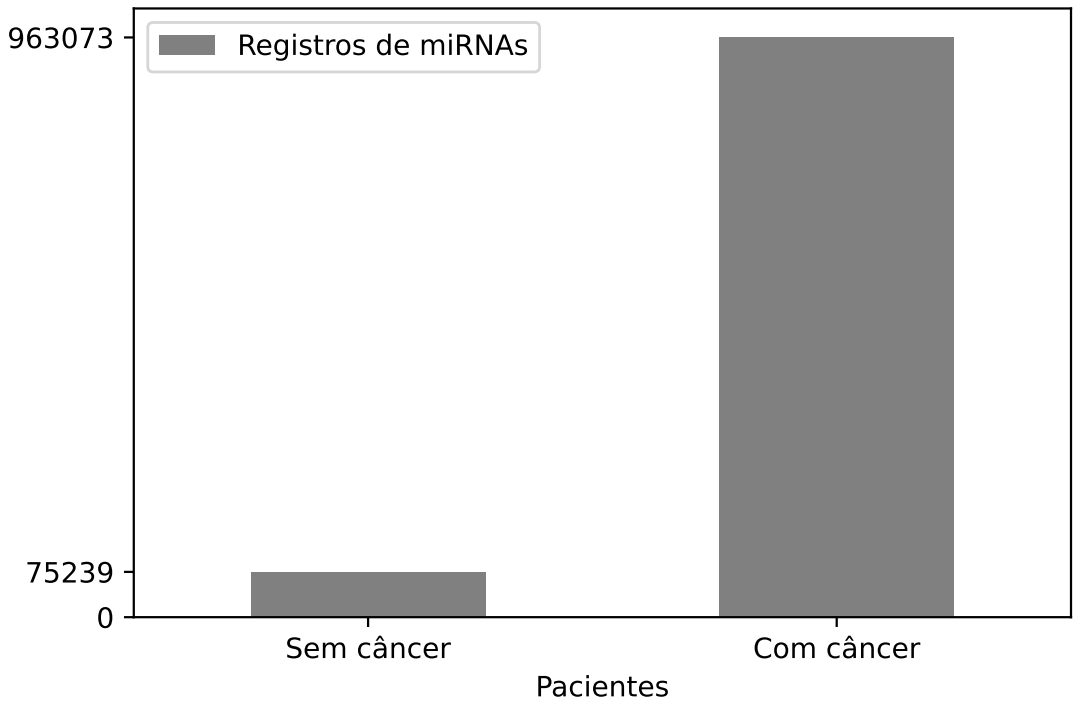

Figura 12 – Quantidade de registros de miRNAs em pacientes examinados

Após a aplicação da técnica de Downsample na classe majoritária, obteve-se um conjunto de dados com 75239 registros de cada classe, conforme apresentado da Figura 13, representando assim 40 pacientes em cada classe.

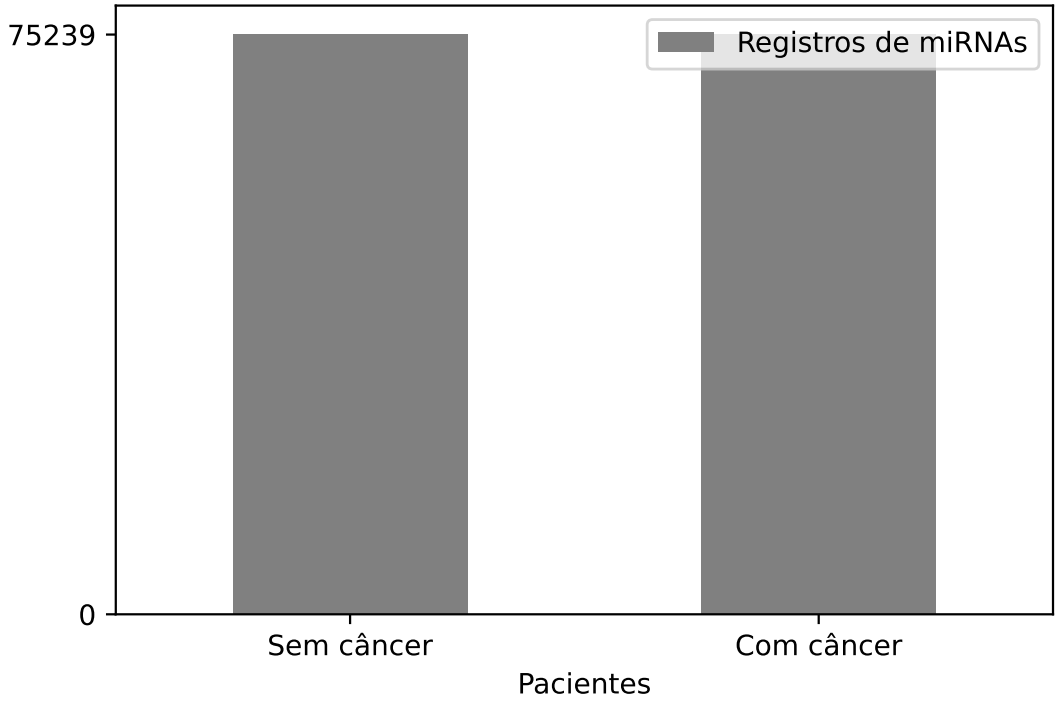

Figura 13 – Quantidade de registros de miRNAs após aplicada a técnica de Downsample

Posteriormente, houve o redimensionamento desse dataset, além da eliminação de dados inúteis, como por exemplo, dados nulos e com caracteres especiais, um vez que os registros devem ser apenas numéricos. Esse estágio foi realizado utilizando a biblioteca científica *NumPy*.

Em sequida, houve a separação de dados, para fins de treinamento, validação e teste, em 3 (três) partes. Nessa fase houve o uso da biblioteca *NumPy* com o objetivo de embaralhar os dados antes da divisão randômica do conjunto de treinamento e validação, e de 10% a 20% de conjunto de testes.

4.4 Desenvolvimento dos modelos

Conforme aprendido no Capítulo 3, de acordo com Zhu et al. (2021), a CNN pode ser usada em aplicações que fazem o uso de dados estruturados, e não apenas em imagens. Por esse motivo, há uma justificativa para a avaliação do emprego de Redes Neurais Convolucionais aplicados a base de dados de miRNAs, visto que são dados estruturados.

4.4.1 Rede Neural Convolucional 1D

No decorrer do redimensionamento do conjunto de dados para servir como entrada para a CNN 1D, utilizou-se um editor de planilhas para reestruturar os dados na forma de tabela, em que as linhas passaram a representar cada paciente e as colunas cada

registro de miRNAs. Ao final do processo, tivemos como resultado uma coleção com o dimensionamento (552, 3762).

Logo depois, durante o balanceamento do conjunto de dados, utilizou-se a técnica de Downsample, descrita no Capítulo 3, durante o desenvolvimento do modelo de CNN 1D. Além disso, houve o estudo e aplicação da técnica de NearMiss, descrita no Capítulo 3, para fins de avaliação.

Seguidamente, sucedeu o desenvolvimento do modelo de CNN 1D, iniciando pela camanda de entrada com a dimensão de (None, 3762, 1), seguido de duas camadas de convolução 1D com 64 filtros de saída e função de ativação Relu.

Em continuação, colocou-se uma camada de Pooling para reduz a dimensionalidade para 1D, com o objetivo de preparar os dados para usá-los como entrada em uma camada densa.

Feito isso, colocou-se uma camada densa cuja dimensionalidade da camada de saída possui o valor 100 e sua função de ativação é a Relu. Continuando, empregou-se uma camada Dropout com o objetivo de evitar o overfitting. Por último, como camada de saída utilizou-se uma camada densa com dimensionalidade de saída 1 e função de ativação Softmax.

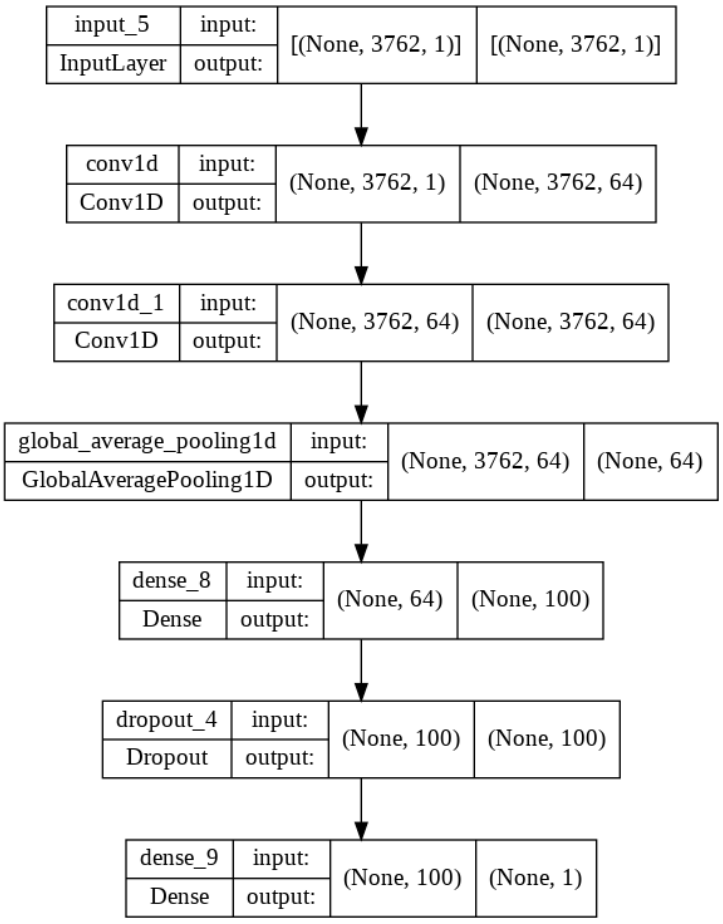

Figura 14 – Modelo de CNN unidimensional.. Fonte: Autor.

4.4.2 Rede Neural Convolucional 2D

Ao longo do desenvolvimento da CNN 2D, também precisou-se realizar o redimensionamento do conjunto de dados para servir como entrada para esse modelo. Para esse feito, foi usada a biblioteca *NumPy* para obter como resultado uma coleção com o dimensionamento (552, 1881, 4).

Assim como no modelo de CNN unidimensional, logo depois, durante o balanceamento do conjunto de dados, utilizou-se a técnica de Downsample durante o desenvolvimento desse modelo. E depois, também houve o estudo e uso da técnica de NearMiss.

Mais tarde, houve o desenvolvimento do modelo propriamente dito, começando pela camanda de entrada com a dimensão de (None, 1881, 4, 1), seguido de duas camadas de convolução 2D, sendo a primeira com 8 filtros de saída e a segunda com 16, ambas usando Relu como função de ativação.

Depois, colocou-se uma camada de Pooling para reduz a dimensionalidade para 2D, com o objetivo de preparar os dados para usá-los como entrada na próxima camada. Continuando, colocou-se uma camada densa cuja dimensionalidade de saída possui o valor 64 e Relu como função de ativação.

Após isso, também houve o emprego de uma camada Dropout com o objetivo de evitar o overfitting, assim como o modelo anterior. Por fim, como camada de saída foi usado uma camada densa com dimensionalidade de saída 1 e função de ativação Softmax.

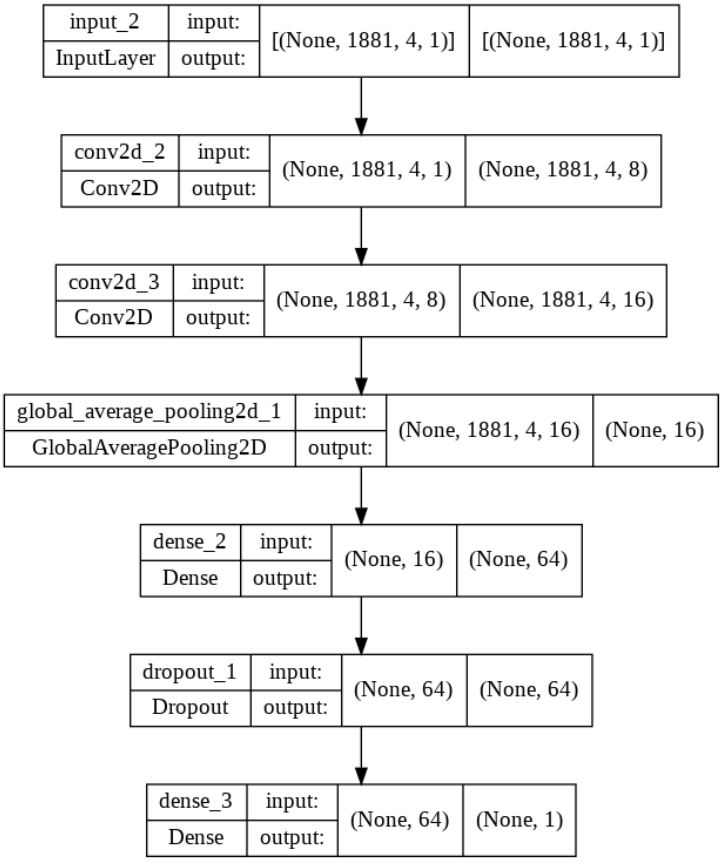

Figura 15 – Modelo de CNN bidimensional. Fonte: Autor.

4.5 Máquina de Vetores de Suporte

Para o desenvolvimento do modelo de SVM, aproveitou-se dos resultados do redimensionamento do conjunto de dados realizado durante o desenvolvimento da CNN 1D.

Houve a separação do conjuntos de dados baseado nas suas classes, resultando em um conjunto contendo os registros da classe majoritária e outro da classe minoritária, na qual foi aplicada a técnica de Downsample.

Após a etapa anterior, para corrigir o desbalanceamento do conjunto de dados, realizou-se o emprego a técnica de Downsample para reduzir o tamanho dos registros da classe majoritária. E também, como Javaheri, Sepehri e Teimourpour (2014) mostram, os SVMs produzem modelos melhores quando os dados são normalizados, logo todos os dados foram normalizados e padronizados antes da classificação utilizando a técnica de Normalização Min-Max.

E finalmente, para o construção do modelo, foi empregado a função SVC da biblioteca *Sklearn*.

5 RESULTADOS E DISCUSSÃO

A partir da avaliação e comparação das métricas entre os modelos treinados, podemos ter argumentos suficientes para apontar qual obteve os melhores resultados e que satisfaça o nosso objetivo.

5.1 Modelos desenvolvidos

Na abordagem em que foi desenvolvido o modelo de CNN 1D, o classificador atingiu uma acurácia média de 94,64%, *f1-score* de 0,96 e *recall* médio de 0,5. Todavia, avaliando individualmente cada classe, observou-se que a classe majoritária obteve os melhores resultados. No entanto, para a classe minoritária, devido ao desbalanceamento, o modelo em questão não conseguiu prever nenhum exemplo.

Na Tabela 3 são apresentadas métricas de desempenho desse classificador em questão.

Classe	Acurácia (%)	Recall	f1-score
majoritária	93	1.0	0.96
minoritária	0	-	-

Tabela 3 – Métricas de desempenho do modelo CNN 1D. Fonte: Autor

A curva ROC desse modelo, apresentada na Figura 16, ilustra previsões aleatórias.

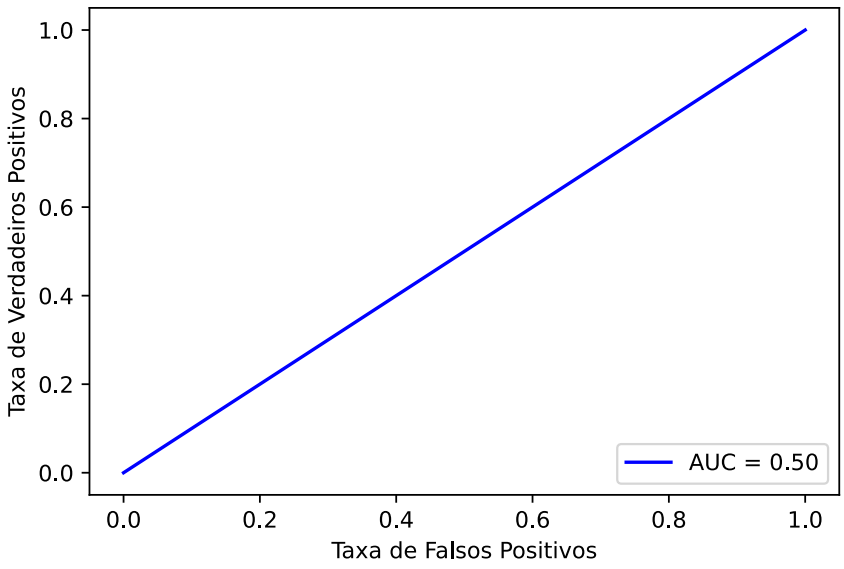

Figura 16 – Curva ROC do modelo CNN 1D

Sua matriz de confusão, ilustrada na Figura 17, evidencia que não foram previstas nenhum item da categoria 0, ou seja, pacientes sadios.

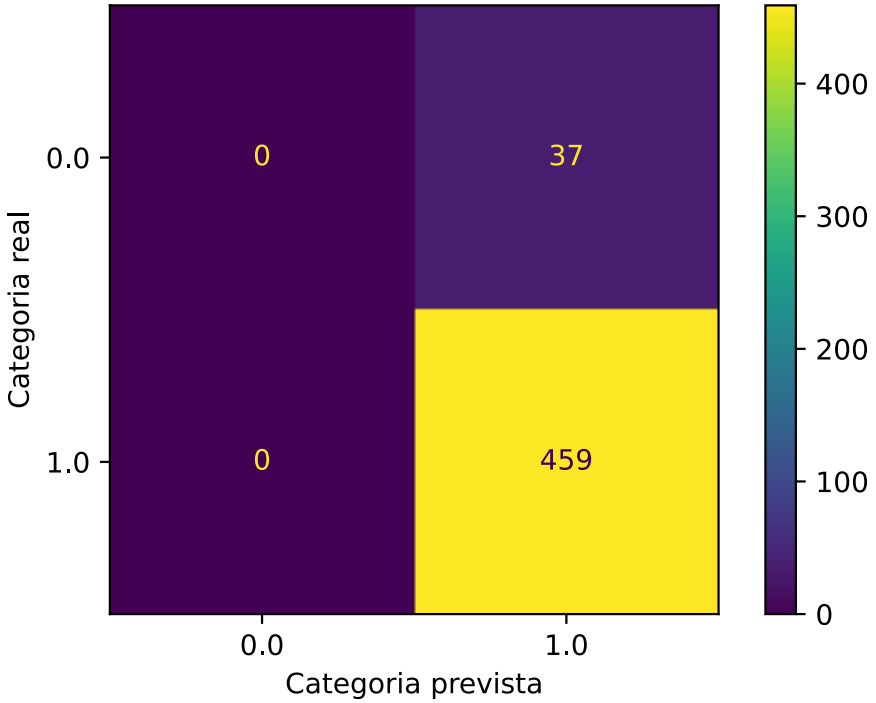

Figura 17 – Matriz de Confusão do modelo CNN 1D

Para o modelo de CNN 2D, teve-se durante o relatório de classificação, que o classificador em questão atingiu os mesmo resultados do modelo de CNN 1D, logo, obteve-se uma acurácia de 94,64%, *f1-score* de 0,96 e *recall* médio de 0,5. E seguiu o mesmo comportamento que o modelo anterior em uma perspectiva individual.

Na Tabela 4 são apresentadas métricas de desempenho desse classificador em questão.

Classe	Acurácia (%)	Recall	f1-score
majoritária	93	1.0	0.96
minoritária	0	-	-

Tabela 4 – Métricas de desempenho do modelo CNN 2D. Fonte: Autor

Assim como o modelo anterior, a curva ROC desse modelo, apresentada na Figura 18, ilustra previsões aleatórias.

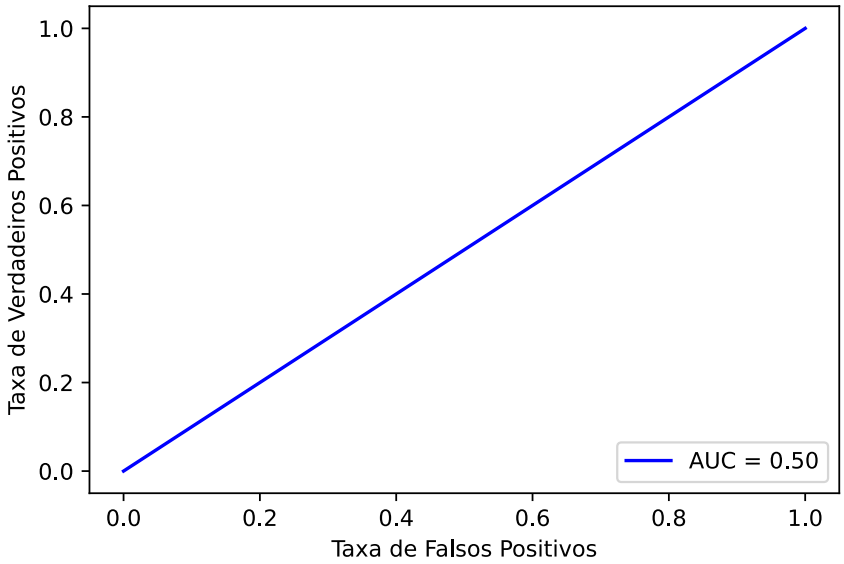

Figura 18 – Curva ROC do modelo CNN 2D

Sua matriz de confusão, ilustrada na Figura 19, também evidencia que não foram previstas nenhum item da categoria 0, ou seja, pacientes sem câncer.

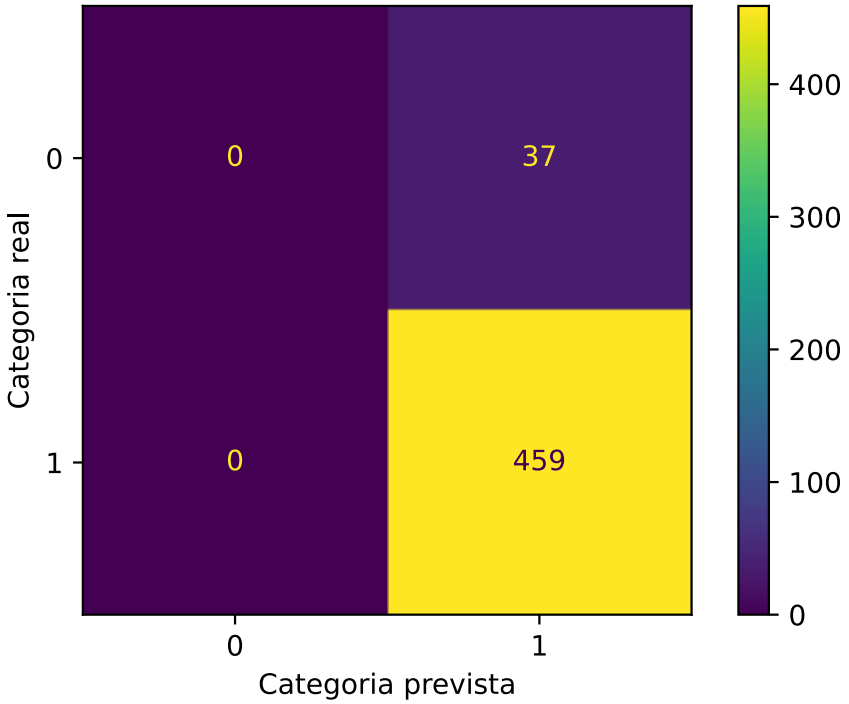

Figura 19 – Matriz de Confusão do modelo CNN 2D

Tendo em vista que os modelos convolucionais não atingiram taxas de acerto suficientes para alcançarem o objetivo proposto, após o desenvolvimento de um modelo de

SVM, que conforme apresentado no Capítulo 3, se propõe a ter melhores resultados em classificação de problemas com duas classes, obteve-se uma acurácia de 75%, precisão de 63,63% e *recall* de 1.0.

Em uma consideração singular, nesse modelo a classe majoritária continuou apresentando os melhores resultados, e o presente classificador conseguiu prever itens da classe minoritária, sendo que os classificadores anteriores não previram, logo, foi possível obter suas métricas para esse cenário.

Na Tabela 5 são apresentadas métricas de desempenho desse classificador em questão.

Classe	Precisão (%)	Recall	f1-score
majoritária	93	1.0	0.78
minoritária	64	0.56	0.71

Tabela 5 – Métricas de desempenho do modelo SVM. Fonte: Autor

A cruva ROC desse modelo, apresentada na Figura 20, ilustra o desempenho do classificador, mostrando-se melhor que os anteriores.

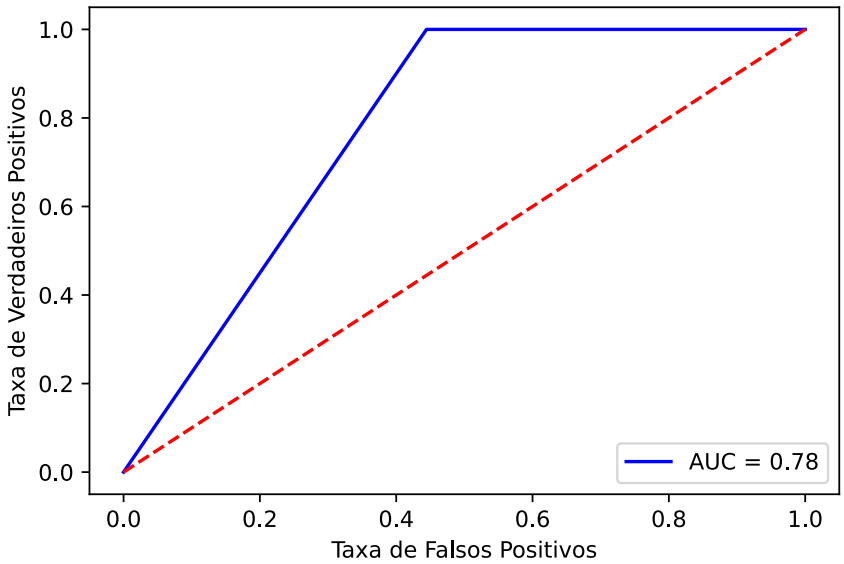

Figura 20 – Curva ROC do Modelo SVM

A matriz de confusão do modelo de SVM, apresentada na Figura 21, diferente dos modelos anteriores, mostra que foram previstos items das duas categorias.

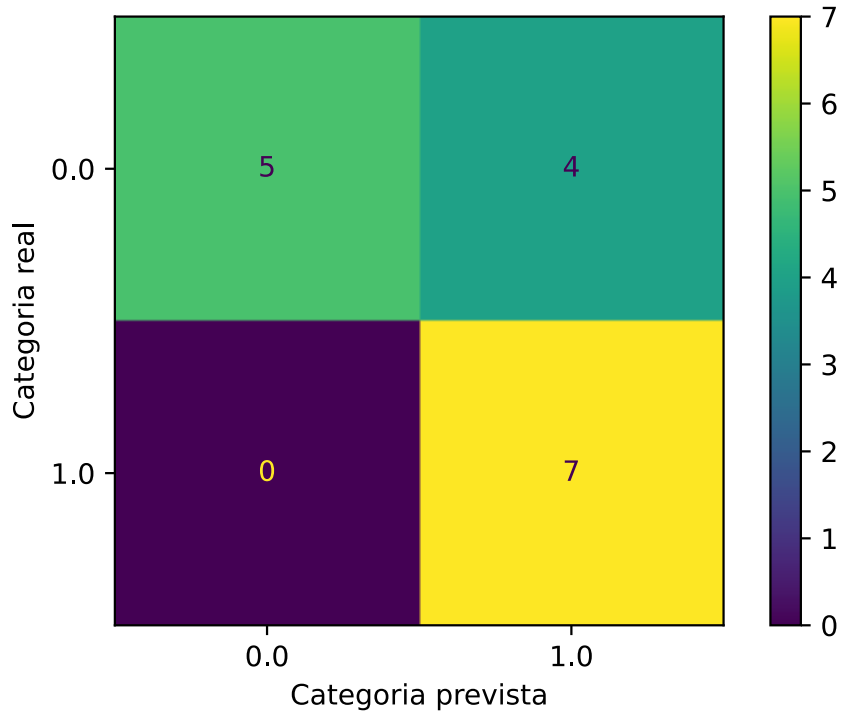

Figura 21 – Matriz de Confusão do Modelo SVM

5.2 Avaliando as diferentes abordagens

Em um comparativo, no primeiro momento em uma visão macro, os modelos de CNN apresentaram-se ilusoriamente como os melhores classificadores devido as suas métricas gerais. Porém, essa reputação é descartada quando é feita uma análise individual das classes, o que mostra que os modelos de CNN não classificaram itens da classe minoritária.

Na Tabela 6 é apresentado um comparativo das métricas entre todos os modelos avaliados.

	Acurácia (%)	Precisão (%)	Recall	f1-score
CNN 1D	93	50	0,5	0,48
CNN 2D	93	50	0,5	0,48
SVM	75	82	0,78	0,75

Tabela 6 – Métricas de desempenho dos modelos de classificação. Fonte: Autor

Dando importância a isso, o modelo a ser considerável é o de SVM, que mesmo com métricas razoáveis, consiguiu prever casos dentro das duas classes.

6 CONCLUSÕES

Verificou-se ser possível treinar uma Rede Neural Convolucional com dados estruturados que não sejam apenas imagens. Entretando, devido ao desbalanceamento das classes do nosso dataset, os resultados dos modelos de CNN 1D e CNN 2D não foram satisfatórias para a classe minoritária. Em razão do desbalanceamento de classes, criou-se um modelo de SVM, levando em consideração o seu desempenho para lidar com problemas de duas classes. Após a avaliação das métricas de performance, a aplicação do SVM mostrou-se com um melhor desempenho, em comparação às implementações de CNN, em razão de que todas as classes foram possíveis serem previstas, diferentemente dos modelos de CNN que previram apenas a classe majoritária.

Portanto, conclui-se que quando se têm um conjunto de dados balanceados, pode-se criar um classificador utilizando uma Rede Neural Convolucional. Contudo, se o conjunto de dados possuir apenas duas classes, modelos de classificação com SVM mostram-se melhores nesse quesito.

6.1 Trabalhos Futuros

Para a continuação desse trabalho, sugere-se testar outras técnicas e aumentar os registros da base de dados com mais incidências de pacientes saudáveis, o que representa a classe minoritária. Além disso, será importante melhorar o desempenho da classificação para que o modelo seja mais preciso em sua previsão.

Referências

ABRAHAMS, S. et al. *TensorFlow for Machine Intelligence: A Hands-On Introduction to Learning Algorithms*. [S.l.]: Bleeding Edge Press, 2016. ISBN 1939902452. Citado na página 19.

AGATONOVIC-KUSTRIN, S.; BERESFORD, R. Basic concepts of artificial neural network (ann) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, v. 22, n. 5, p. 717–727, 2000. ISSN 0731-7085. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0731708599002721>. Citado na página 10.

ALEčKOVIć, M.; KANG, Y. Regulation of cancer metastasis by cell-free mirnas. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, v. 1855, n. 1, p. 24–42, 2015. ISSN 0304-419X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0304419X14000948>. Citado na página 1.

ARJARIA, S. K.; RATHORE, A. S.; CHERIAN, J. S. Chapter 13 - kidney disease prediction using a machine learning approach: A comparative and comprehensive analysis. In: N, P.; KAUTISH, S.; PENG, S.-L. (Ed.). *Demystifying Big Data, Machine Learning, and Deep Learning for Healthcare Analytics*. Academic Press, 2021. p. 307–333. ISBN 978-0-12-821633-0. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128216330000064>. Citado na página 17.

BASHIR, S.; QAMAR, U.; KHAN, F. H. Intellihealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework. *Journal of Biomedical Informatics*, v. 59, p. 185–200, 2016. ISSN 1532-0464. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1532046415002816>. Citado na página 8.

BOHR, A.; MEMARZADEH, K. The rise of artificial intelligence in healthcare applications. *Artificial Intelligence in Healthcare*, p. 25–60, 2020. PMC7325854[pmcid]. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325854/>. Citado 5 vezes nas páginas , 1, 4, 5 e 7.

CERVANTES, J. et al. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, v. 408, p. 189–215, 2020. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231220307153>. Citado na página 10.

CHO, H. et al. When Do Changes in Cancer Survival Mean Progress? The Insight From Population Incidence and Mortality. *JNCI Monographs*, v. 2014, n. 49, p. 187–197, 11 2014. ISSN 1052-6773. Disponível em: <https://doi.org/10.1093/jncimonographs/lgu014>. Citado na página 1.

CONDRAT, C. E. et al. miRNAs as biomarkers in disease: Latest findings regarding their role in diagnosis and prognosis. *Cells*, v. 9, n. 2, jan. 2020. Citado na página 14.

CORTES, C.; VAPNIK, V. Support-vector networks. *Mach. Learn.*, Kluwer Academic Publishers, USA, v. 20, n. 3, p. 273–297, sep 1995. ISSN 0885-6125. Disponível em: <https://doi.org/10.1023/A:1022627411411>. Citado na página 9.

DRAGOMIR, M. P.; KNUTSEN, E.; CALIN, G. A. Classical and noncanonical functions of mirnas in cancers. *Trends in Genetics*, 2021. ISSN 0168-9525. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0168952521002869>. Citado na página 14.

ELIAS, K. M. et al. Diagnostic potential for a serum mirna neural network for detection of ovarian cancer. *eLife*, eLife Sciences Publications, Ltd, v. 6, p. e28932, oct 2017. ISSN 2050-084X. Disponível em: <https://doi.org/10.7554/eLife.28932>. Citado na página 5.

FELIPPU, A. W. D. et al. Impacto da demora no diagnóstico e tratamento no câncer de cabeça e pescoço. *Brazilian Journal of Otorhinolaryngology*, v. 82, n. 2, p. 140–143, 2016. ISSN 25300539. Disponível em: </25300539/0000008200000002/v0_201702231038/X253005391650688X/v0_201702231039/pt/main.assets>. Citado na página 1.

FRANçA, R. P. et al. Chapter 3 - an overview of deep learning in big data, image, and signal processing in the modern digital age. In: PIURI, V. et al. (Ed.). *Trends in Deep Learning Methodologies*. Academic Press, 2021, (Hybrid Computational Intelligence for Pattern Analysis). p. 63–87. ISBN 978-0-12-822226-3. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128222263000039>. Citado na página 15.

GARCíA-POLA, M. et al. Role of artificial intelligence in the early diagnosis of oral cancer. a scoping review. *Cancers*, v. 13, n. 18, 2021. ISSN 2072-6694. Disponível em: <https://www.mdpi.com/2072-6694/13/18/4600>. Citado na página 1.

HAN, S.-H. et al. Artificial neural network: Understanding the basic concepts without mathematics. *Dement Neurocogn Disord*, v. 17, n. 3, p. 83–89, dez. 2018. Citado na página 10.

HARRIS, C. R. et al. Array programming with numpy. *Nature*, v. 585, n. 7825, p. 357–362, Sep 2020. ISSN 1476-4687. Disponível em: <https://doi.org/10.1038/s41586-020-2649-2>. Citado na página 20.

HUANG, S. et al. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics*, v. 15, n. 1, p. 41–51, jan. 2018. Citado na página 9.

HUGHES, J. P. et al. Principles of early drug discovery. *British journal of pharmacology*, Blackwell Science Inc, v. 162, n. 6, p. 1239–1249, Mar 2011. ISSN 1476-5381. 21091654[pmid]. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/21091654>. Citado na página 7.

INDOLIA, S. et al. Conceptual understanding of convolutional neural network- a deep learning approach. *Procedia Computer Science*, v. 132, p. 679–688, 2018. ISSN 1877-0509. International Conference on Computational Intelligence and Data Science. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050918308019>. Citado na página 12.

IQBAL, M. J. et al. Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future. *Cancer Cell International*, v. 21, n. 1, p. 270, May 2021. ISSN 1475-2867. Disponível em: <https://doi.org/10.1186/s12935-021-01981-1>. Citado na página 1.

JAVAHERI, S. H.; SEPEHRI, M. M.; TEIMOURPOUR, B. Chapter 6 - response modeling in direct marketing: A data mining-based approach for target selection. In: ZHAO, Y.; CEN, Y. (Ed.). *Data Mining Applications with R*. Boston: Academic Press, 2014. p. 153–180. ISBN 978-0-12-411511-8. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780124115118000062>. Citado 2 vezes nas páginas 15 e 25.

JOHNSON, J. M.; KHOSHGOFTAAR, T. M. Survey on deep learning with class imbalance. *Journal of Big Data*, v. 6, n. 1, p. 27, Mar 2019. ISSN 2196-1115. Disponível em: <https://doi.org/10.1186/s40537-019-0192-5>. Citado na página 15.

KAKUSHADZE, Z.; RAGHUBANSHI, R.; YU, W. Estimating cost savings from early cancer diagnosis. *Data*, v. 2, n. 3, 2017. ISSN 2306-5729. Disponível em: <https://www.mdpi.com/2306-5729/2/3/30>. Citado na página 1.

KOPPAD, S. et al. Machine learning-based identification of colon cancer candidate diagnostics genes. *Biology*, v. 11, n. 3, 2022. ISSN 2079-7737. Disponível em: <https://www.mdpi.com/2079-7737/11/3/365>. Citado 3 vezes nas páginas , 3 e 4.

KOTHANDAN, R.; BISWAS, S. Identifying micrornas involved in cancer pathway using support vector machines. *Computational Biology and Chemistry*, v. 55, p. 31 – 36, 2015. ISSN 1476-9271. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1476927115000171>. Citado na página 6.

KOTSIANTIS, S.; KANELLOPOULOS, D.; PINTELAS, P. Data preprocessing for supervised learning. *International Journal of Computer Science*, v. 1, p. 111–117, 01 2006. Citado na página 14.

KULKARNI, A.; CHONG, D.; BATARSEH, F. A. 5 - foundations of data imbalance and solutions for a data democracy. In: BATARSEH, F. A.; YANG, R. (Ed.). *Data Democracy*. Academic Press, 2020. p. 83–106. ISBN 978-0-12-818366-3. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128183663000058>. Citado na página 16.

KULSKI, J. K. Next-generation sequencing — an overview of the history, tools, and "omic" applications. In: KULSKI, J. K. (Ed.). *Next Generation Sequencing*. Rijeka: IntechOpen, 2016. cap. 1. Disponível em: <https://doi.org/10.5772/61964>. Citado 2 vezes nas páginas 1 e 7.

LEE, W.; SEO, K. Downsampling for binary classification with a highly imbalanced dataset using active learning. *Big Data Research*, v. 28, p. 100314, 2022. ISSN 2214-5796. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2214579622000089>. Citado na página 15.

LIM, S.; TUCKER, C. S. A bayesian sampling method for product feature extraction from large-scale textual data. *Journal of Mechanical Design*, American Society of Mechanical Engineers Digital Collection, v. 138, n. 6, 2016. Citado na página 8.

LIM, S.; TUCKER, C. S.; KUMARA, S. An unsupervised machine learning model for discovering latent infectious diseases using social media data. *Journal of Biomedical Informatics*, v. 66, p. 82–94, 2017. ISSN 1532-0464. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1532046416301812>. Citado 2 vezes nas páginas 1 e 8.

LIU, Y. et al. A strategy on selecting performance metrics for classifier evaluation. *International Journal of Mobile Computing and Multimedia Communications*, v. 6, p. 20–35, 10 2014. Citado na página 16.

NOBLE, W. S. What is a support vector machine? *Nat Biotechnol*, United States, v. 24, n. 12, p. 1565–1567, dez. 2006. Citado na página 9.

O'BRIEN, J. et al. Overview of microrna biogenesis, mechanisms of actions, and circulation. *Frontiers in Endocrinology*, v. 9, p. 402, 2018. ISSN 1664-2392. Disponível em: <https://www.frontiersin.org/article/10.3389/fendo.2018.00402>. Citado na página 14.

PANCH, T.; SZOLOVITS, P.; ATUN, R. Artificial intelligence, machine learning and health systems. *Journal of global health*, Edinburgh University Global Health Society, v. 8, n. 2, p. 020303–020303, Dec 2018. ISSN 2047-2986. 30405904[pmid]. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/30405904>. Citado na página 8.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, JMLR.org, v. 12, n. null, p. 2825–2830, nov. 2011. ISSN 1532-4435. Citado na página 20.

POWERS, D. Evaluation: From precision, recall and f-factor to roc, informedness, markedness correlation. *Mach. Learn. Technol.*, v. 2, 01 2008. Citado na página 17.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, v. 65, n. 6, p. 386–408, 1958. Citado na página 10.

ROSSUM, G. van; TEAM, P. D. *Python 3.6 Language Reference*. London, GBR: Samurai Media Limited, 2016. ISBN 9789888406883. Citado na página 19.

SARKER, I. H. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, v. 2, n. 3, p. 160, Mar 2021. ISSN 2661-8907. Disponível em: <https://doi.org/10.1007/s42979-021-00592-x>. Citado na página 7.

SAZLI, M. A brief review of feed-forward neural networks. *Communications, Faculty Of Science, University of Ankara*, v. 50, p. 11–17, 01 2006. Citado 2 vezes nas páginas 11 e 12.

SHARMA, D. K. et al. 3 - deep learning applications for disease diagnosis. In: GUPTA, D. et al. (Ed.). *Deep Learning for Medical Applications with Unique Data*. Academic Press, 2022. p. 31–51. ISBN 978-0-12-824145-5. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128241455000058>. Citado 2 vezes nas páginas e 16.

SHARMA, N.; SHARMA, R.; JINDAL, N. Machine learning and deep learning applications-a vision. *Global Transitions Proceedings*, v. 2, n. 1, p. 24–28, 2021. ISSN 2666-285X. 1st International Conference on Advances in Information, Computing and Trends in Data Engineering (AICDE - 2020). Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666285X21000042>. Citado na página 8.

SIDEY-GIBBONS, J. A. M.; SIDEY-GIBBONS, C. J. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, v. 19, n. 1, p. 64, Mar 2019. ISSN 1471-2288. Disponível em: <https://doi.org/10.1186/s12874-019-0681-4>. Citado na página 8.

SINGH, D.; SINGH, B. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, v. 97, p. 105524, 2020. ISSN 1568-4946. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1568494619302947>. Citado na página 14.

SINGH, P. et al. Chapter 5 - diagnosing of disease using machine learning. In: SINGH, K. K. et al. (Ed.). *Machine Learning and the Internet of Medical Things in Healthcare*. Academic Press, 2021. p. 89–111. ISBN 978-0-12-821229-5. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128212295000033>. Citado 2 vezes nas páginas e 16.

SUNG, H. et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, United States, v. 71, n. 3, p. 209–249, fev. 2021. Citado na página 1.

WALLISCH, P. et al. Chapter 29 - neural network part ii: Supervised learning. In: WALLISCH, P. et al. (Ed.). *Matlab for Neuroscientists*. London: Academic Press, 2009. p. 319–337. ISBN 978-0-12-374551-4. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780123745514000294>. Citado 2 vezes nas páginas 10 e 11.

WANG, Z. et al. A convolutional neural network-based classification and decision-making model for visible defect identification of high-speed train images. *Journal of Sensors*, Hindawi, v. 2021, p. 5554920, Mar 2021. ISSN 1687-725X. Disponível em: <https://doi.org/10.1155/2021/5554920>. Citado na página 13.

XU, Y. et al. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, v. 2, n. 4, p. 100179, 2021. ISSN 2666-6758. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666675821001041>. Citado na página 7.

YANMINSUN; WONG, A.; KAMEL, M. S. Classification of imbalanced data: a review. *International Journal of Pattern Recognition and Artificial Intelligence*, v. 23, 11 2011. Citado na página 15.

ZHANG, J.; MANI, I. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. In: *Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets*. [S.l.: s.n.], 2003. Citado na página 15.

ZHANG, L. et al. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discov Today*, England, v. 22, n. 11, p. 1680–1685, set. 2017. Citado 2 vezes nas páginas 1 e 7.

ZHU, Y. et al. Converting tabular data into images for deep learning with convolutional neural networks. *Scientific Reports*, v. 11, n. 1, p. 11325, May 2021. ISSN 2045-2322. Disponível em: <https://doi.org/10.1038/s41598-021-90923-y>. Citado 3 vezes nas páginas 13, 14 e 22.

ZOU, K. H.; O'MALLEY, A. J.; MAURI, L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, United States, v. 115, n. 5, p. 654–657, fev. 2007. Citado 2 vezes nas páginas e 17.