UNIVERSIDADE PAULISTA – UNIP
INSTITUTO DE CIÊNCIAS EXATAS E TECNOLOGIA (ICET)

Antônio Sérgio Graton Júnior
João Lucas Roldão Rocha
Mateus Bernardi de Oliveira

IMPLEMENTAÇÃO DE REDES NEURAIS CONVOLUCIONAIS (CNNS) PARA AUXÍLIO NO DIAGNÓSTICO DE CÂNCER DE PULMÃO UTILIZANDO BANCO DE DADOS OPEN SOURCE

Ribeirão Preto
2025

Antônio Sérgio Graton Júnior
João Lucas Roldão Rocha
Mateus Bernardi de Oliveira

IMPLEMENTAÇÃO DE REDES NEURAIS CONVOLUCIONAIS (CNNS) PARA AUXÍLIO NO DIAGNÓSTICO DE CÂNCER DE PULMÃO UTILIZANDO BANCO DE DADOS OPEN SOURCE

Monografia submetida ao Curso de Ciência da Computação da Universidade Paulista, para a obtenção do Título de Bacharel em Ciência da Computação.

Orientadores:
Prof. Dr. Avelino Palma Pimenta Júnior
Prof.ª Dr.ª Eliana Leão do Prado Battaglion
Prof. Dr. Kleython José Coriolano Cavalcanti de Lacerda

Ribeirão Preto
2025

Antônio Sérgio Graton Júnior
João Lucas Roldão Rocha
Mateus Bernardi de Oliveira

IMPLEMENTAÇÃO DE REDES NEURAIS CONVOLUCIONAIS (CNNS) PARA AUXÍLIO NO DIAGNÓSTICO DE CÂNCER DE PULMÃO UTILIZANDO BANCO DE DADOS OPEN SOURCE

Monografia submetida ao Curso de Ciência da Computação da Universidade Paulista, para a obtenção do Título de Bacharel em Ciência da Computação.

Aprovado em: 08/12/2025 Média Final - MF: 10,0 (dez)

BANCA EXAMINADORA

gov.br Documento assinado digitalmente
KLEYTHON JOSE CORIOLANO CAVALCANTI DE L
Data: 08/12/2025 18:15:25-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Kleython J. C. C. de Lacerda
Universidade Paulista – UNIP

Dedicamos este trabalho a todos que contribuíram direta ou indiretamente para nossa formação, especialmente nossos professores, colegas de curso e profissionais que nos inspiraram a seguir na área da ciência da computação. Dedicamos também às nossas famílias, pelo apoio constante ao longo desta caminhada acadêmica.

“A ciência de hoje é a tecnologia de amanhã”.

(Edward Teller)

RESUMO

ROLDÃO ROCHA, João Lucas; GRATON, Antônio Sérgio; OLIVEIRA, Mateus Bernardi de. **Implementação de Redes Neurais Convolucionais (CNNs) para Auxílio no Diagnóstico de Câncer de Pulmão Utilizando Banco de Dados Open Source.** 2025. 39 f. Trabalho de Conclusão de Curso (Curso de Ciência da Computação) – Instituto de Ciências Exatas e Tecnologia, Universidade Paulista, Campus Vargas, Ribeirão Preto, 2025.

O câncer de pulmão é um dos tipos de câncer mais letais do mundo, sendo responsável por milhões de mortes anuais e frequentemente diagnosticado em estágios avançados. Diante desse cenário, técnicas de inteligência artificial têm se destacado como ferramentas promissoras para o desenvolvimento de sistemas de apoio ao diagnóstico médico. Este trabalho tem como objetivo implementar e avaliar modelos de aprendizado de máquina, com ênfase em Redes Neurais Convolucionais (CNNs), para prever o diagnóstico de câncer de pulmão a partir de um conjunto de dados *open source* disponibilizado na plataforma Kaggle. O estudo foi conduzido no ambiente Google Colab, utilizando a linguagem Python e bibliotecas como Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn e TensorFlow/Keras. O conjunto de dados analisado contém 309 registros de pacientes, compostos por atributos clínicos e comportamentais relacionados a fatores de risco e sintomas associados ao câncer de pulmão. As etapas metodológicas envolveram pré-processamento, análise exploratória, visualização gráfica e aplicação de cinco modelos de classificação: Regressão Logística, Naive Bayes Gaussiano, Floresta Aleatória, K-Vizinhos Mais Próximos (KNN) e Rede Neural Convolucional. Os resultados obtidos demonstraram desempenho expressivo dos modelos, com destaque para o KNN, que atingiu acurácia de 94,64%, e para a Rede Neural Convolucional, que alcançou 95,91% de acurácia no conjunto de treino e 91,07% na validação, apresentando curvas de aprendizado estáveis e sem indicativos de sobreajuste. A análise exploratória também permitiu identificar padrões relevantes, como a forte correlação entre tabagismo, tosse persistente, fadiga e falta de ar com diagnósticos positivos. Conclui-se que técnicas de aprendizado supervisionado, especialmente CNNs, apresentam grande potencial como ferramentas de apoio ao diagnóstico de câncer de pulmão, mesmo com bases tabulares e de tamanho reduzido. O trabalho destaca ainda a viabilidade de utilizar ferramentas e dados *open source* para o desenvolvimento de soluções acessíveis e reprodutíveis na área da saúde.

Palavras-chave: Neoplasias Pulmonares; Inteligência Artificial; Aprendizado de Máquina; Diagnóstico Assistido por Computador.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

CNN – *Convolutional Neural Network* (Rede Neural Convolucional)

GPU – *Graphics Processing Unit* (Unidade de Processamento Gráfico)

INCA – Instituto Nacional de Câncer

KNN – *K-Nearest Neighbors* (K-Vizinhos Mais Próximos)

ML – *Machine Learning* (Aprendizado de Máquina)

PCA – *Principal Component Analysis* (Análise de Componentes Principais)

ReLU – *Rectified Linear Unit* (Unidade Linear Retificada)

RNA – Rede Neural Artificial

SUMÁRIO

1 INTRODUÇÃO

O câncer é atualmente um dos principais problemas de saúde pública no mundo, figurando entre as quatro maiores causas de morte prematura antes dos 70 anos de idade (Instituto Nacional de Câncer, 2020). No Brasil, de acordo com as estimativas do Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA, 2020), são previstos aproximadamente 625 mil novos casos de câncer por ano no triênio 2020–2022, sendo o câncer de pulmão responsável por cerca de 30 mil diagnósticos anuais, o que o coloca entre os tipos mais incidentes e letais no país.

Segundo Franceschini e Santoro (2020), o câncer de pulmão ocupa o primeiro lugar tanto em incidência quanto em mortalidade globalmente, representando um desafio persistente para a saúde pública. O tabagismo continua sendo o principal fator de risco prevenível, estando associado a cerca de 85% dos casos diagnosticados (Franceschini; Santoro, 2020).

Em território nacional, estudos apontam que aproximadamente 70% dos diagnósticos de câncer de pulmão ocorrem em estágios avançados (III e IV), quando as opções terapêuticas são limitadas e as taxas de sobrevida são baixas (Costa et al., 2020). Essa realidade reforça a necessidade de aprimorar estratégias de detecção precoce e de desenvolver ferramentas computacionais de apoio ao diagnóstico médico, capazes de oferecer análises mais rápidas e precisas.

Nos últimos anos, os avanços em inteligência artificial (IA) e aprendizado profundo (*Deep Learning*) têm revolucionado a área da saúde, especialmente na interpretação de imagens médicas. Entre os modelos mais promissores estão as Redes Neurais Convolucionais (CNNs), que se destacam pela capacidade de identificar padrões complexos e sutis em imagens de tomografias, radiografias e ressonâncias magnéticas, frequentemente imperceptíveis ao olho humano (He et al., 2015; Zhao et al., 2021).

A aplicação de CNNs em exames de tórax tem se mostrado eficaz tanto em doenças emergentes, como a COVID-19 (Zhao et al., 2021), quanto em patologias crônicas como o câncer de pulmão (Bessa et al., 2021). Em um estudo realizado pela Universidade Federal Rural do Semi-Árido (UFERSA), três arquiteturas diferentes de redes convolucionais — ResNet50v2, Xception e uma rede sequencial — foram comparadas na detecção de tumores pulmonares, alcançando uma acurácia geral de 92% e sensibilidade de 99% para tumores malignos (Bessa et al., 2021).

Esses resultados confirmam o potencial das CNNs em tarefas de diagnóstico assistido por computador (CAD – *Computer-Aided Diagnosis*), apontando para uma nova era de sistemas inteligentes de apoio à decisão clínica. Modelos clássicos de redes neurais, como o ResNet (He et al., 2015), introduziram o conceito de aprendizado residual, que permite o treinamento de redes extremamente profundas sem perda de desempenho, tornando viável o uso dessas arquiteturas em análises complexas de imagens biomédicas.

Pesquisas recentes indicam que o uso combinado de dados públicos e modelos de aprendizado profundo pode reduzir custos e democratizar o acesso a tecnologias diagnósticas, sobretudo em países em desenvolvimento (Zhou et al., 2021). No contexto brasileiro, o emprego de bancos de dados abertos, como os disponibilizados na plataforma Kaggle, tem se mostrado uma alternativa viável para estudos científicos que demandam grandes volumes de dados, respeitando princípios éticos e de reprodutibilidade.

Com base nesse panorama, o presente trabalho tem como tema a "Implementação de Redes Neurais Convolucionais (CNNs) para auxílio no diagnóstico de câncer de pulmão utilizando banco de dados *open source*", e propõe-se a desenvolver e avaliar modelos computacionais de aprendizado profundo capazes de identificar padrões indicativos da doença em dados clínicos e comportamentais.

A motivação deste estudo decorre da importância de integrar tecnologia e saúde pública, ampliando as ferramentas disponíveis para o diagnóstico precoce do câncer de pulmão e contribuindo para a redução da mortalidade associada à doença. Além disso, o projeto busca incentivar o uso de dados abertos e replicáveis, fomentando a pesquisa científica nacional em inteligência artificial aplicada à medicina.

Por fim, este trabalho está estruturado da seguinte forma: o Capítulo 2 descreve os objetivos gerais e específicos; o Capítulo 3 apresenta a metodologia utilizada, incluindo as etapas de pré-processamento, modelagem e avaliação dos modelos; o Capítulo 4 aborda a fundamentação teórica sobre redes neurais e suas aplicações médicas; o Capítulo 5 discute os resultados obtidos; e o Capítulo 6 traz as considerações finais e limitações.

2 OBJETIVOS

2.1 Geral

O objetivo geral deste trabalho é implementar e avaliar modelos de Redes Neurais Convolucionais (CNNs) aplicados ao auxílio no diagnóstico de câncer de pulmão, utilizando um banco de dados open source disponibilizado publicamente na plataforma Kaggle.

Pretende-se, por meio da aplicação de técnicas de aprendizado profundo (*Deep Learning*), identificar padrões relevantes em dados clínicos e comportamentais de pacientes, a fim de contribuir para o aprimoramento de métodos computacionais de suporte ao diagnóstico médico.

Esse objetivo está alinhado às pesquisas recentes que demonstram a eficiência das CNNs em contextos clínicos complexos, como os estudos de Rahimzadeh et al. (2021), que alcançaram 98,49% de acurácia na detecção de infecções pulmonares associadas à COVID-19 a partir de tomografias computadorizadas, e de Bessa et al. (2021), que relataram 92% de acurácia na detecção de tumores malignos em exames de pulmão. Esses resultados evidenciam o potencial das CNNs como ferramenta de apoio diagnóstico em doenças respiratórias.

Além disso, este trabalho busca promover o uso de dados abertos e reprodutíveis em pesquisas acadêmicas, ampliando o acesso a tecnologias de inteligência artificial aplicadas à saúde e incentivando a integração entre computação e medicina no contexto brasileiro (Zhou et al., 2021; Franceschini; Santoro, 2020).

2.2 Específicos

- Analisar o banco de dados *open source* obtido no Kaggle, verificando sua estrutura, qualidade, integridade e possíveis inconsistências nos registros clínicos dos pacientes.
- Realizar o pré-processamento e a padronização dos dados, incluindo a tradução das colunas para o português, tratamento de valores nulos e normalização das variáveis numéricas, utilizando bibliotecas da linguagem Python como Pandas e NumPy.
- Explorar e visualizar os dados por meio de técnicas de análise estatística e gráficos, como histogramas, gráficos de barras e mapas de correlação, de forma a compreender as relações entre hábitos, sintomas e ocorrência de câncer de pulmão.
- Aplicar diferentes modelos de classificação supervisionada, entre eles: Regressão Logística, para análise inicial de separabilidade das classes; *Naive Bayes*

Gaussiano, para modelagem probabilística simples; Floresta Aleatória (*Random Forest*), para análise de variáveis e desempenho em conjunto de classificadores; K-Vizinhos Mais Próximos (KNN), para comparação com abordagens baseadas em distância; e uma Rede Neural Convolucional (CNN), para modelagem avançada baseada em aprendizado profundo.

- Comparar o desempenho dos modelos com base em métricas amplamente utilizadas na literatura, como acurácia, precisão, *recall* e *F1-score*, complementando a análise com matrizes de confusão para avaliação visual da performance dos classificadores.
- Treinar e avaliar a Rede Neural Convolucional desenvolvida especificamente para este projeto, analisando o comportamento do modelo durante o processo de treinamento e validação — considerando gráficos de perda e acurácia por época — e interpretando os resultados obtidos.
- Discutir os resultados obtidos comparando-os com estudos semelhantes da literatura científica, como os trabalhos de He et al. (2015), Zhao et al. (2021) e Bessa et al. (2021), destacando as vantagens, limitações e potencial de aplicação prática dos modelos desenvolvidos.
- Propor melhorias e possibilidades futuras para o aperfeiçoamento do sistema, como a utilização de imagens médicas reais (raios-X e tomografias), *transfer learning* com redes pré-treinadas (ResNet, Xception) e integração com sistemas de diagnóstico assistido (CAD – *Computer-Aided Diagnosis*).

3 METODOLOGIA

O presente trabalho caracteriza-se como uma pesquisa aplicada, de natureza quantitativa e experimental, voltada à implementação e avaliação de modelos computacionais baseados em Redes Neurais Convolucionais (CNNs) para o auxílio no diagnóstico de câncer de pulmão. A escolha desse tipo de pesquisa se justifica

pela necessidade de se empregar métodos científicos voltados à solução de um problema prático: a automatização e o aprimoramento de diagnósticos médicos por meio do uso de inteligência artificial.

Segundo Gil (2008), as pesquisas aplicadas têm como finalidade gerar conhecimentos direcionados à aplicação prática de resultados, enquanto a abordagem quantitativa permite a mensuração dos fenômenos por meio de dados e estatísticas. Nesse contexto, o presente estudo adota um método experimental, em que diferentes modelos de aprendizado supervisionado foram aplicados a um conjunto de dados reais para observar e comparar o comportamento e o desempenho de cada um deles na tarefa de classificação.

O desenvolvimento do projeto foi conduzido no ambiente Google Colab, por se tratar de uma plataforma gratuita e de fácil integração ao Google Drive, o que possibilita maior facilidade no acesso, compartilhamento, armazenamento e execução do código. A linguagem de programação adotada foi o Python, amplamente utilizada nas áreas de ciência de dados e *machine learning* devido à sua versatilidade e à ampla disponibilidade de bibliotecas especializadas. Entre as principais bibliotecas empregadas, destacam-se: Pandas, utilizada para manipulação e análise de dados tabulares; NumPy, voltada a operações matemáticas e vetoriais; Matplotlib e Seaborn, destinadas à criação de gráficos e visualizações; Scikit-learn, empregada na implementação de modelos de aprendizado de máquina tradicionais; e TensorFlow, em conjunto com Keras, utilizada para o desenvolvimento, treinamento e avaliação da rede neural convolucional.

O banco de dados utilizado foi obtido na plataforma Kaggle, no repositório público “*Lung Cancer Dataset*”, que contém 309 registros de pacientes e 16 atributos que descrevem aspectos clínicos e comportamentais, como idade, tabagismo, fadiga, tosse e dificuldade respiratória. Os dados são de domínio público e não possuem informações pessoais identificáveis, o que assegura o cumprimento dos princípios éticos no uso de dados científicos. As colunas do conjunto de dados foram traduzidas do inglês para o português com o objetivo de aprimorar a compreensão, sendo a variável-alvo denominada CÂNCER_DE_PULMÃO, com valores binários indicados como “Sim” ou “Não”.

O processo metodológico foi desenvolvido em etapas sequenciais. Inicialmente, foi realizada a importação e leitura dos dados no ambiente Google Colab, seguida da verificação da integridade e da consistência do conjunto de informações.

Na seguinte etapa, foi efetuado o tratamento básico dos dados, que envolveu a remoção de possíveis duplicatas, a identificação de valores nulos e a padronização dos valores categóricos. A base de dados apresentou-se bem estruturada, demandando apenas ajustes pontuais para adequação à modelagem computacional.

Após o tratamento inicial, procedeu-se à tradução e codificação dos atributos, substituindo os valores numéricos 1 e 2 por rótulos descritivos (“Sim” e “Não”) durante a etapa de visualização, os quais foram posteriormente reconvertidos em valores numéricos para o processamento automático. Essa prática, recomendada por Zhou *et al*. (2024), contribui tanto para a clareza interpretativa dos resultados quanto para a consistência estatística dos algoritmos.

Com os dados devidamente preparados, realizou-se a análise exploratória e a visualização por meio de gráficos e representações estatísticas. Foram elaboradas figuras como o gráfico de distribuição por gênero, a relação entre idade e incidência de câncer e o gráfico de correlação entre sintomas e hábitos. Também foi gerado um mapa de calor (*heatmap*) para evidenciar as correlações entre as variáveis, facilitando a identificação de padrões relevantes. Essa etapa mostrou-se essencial para compreender o comportamento dos dados e reforçar hipóteses clínicas, notadamente a associação entre tabagismo, fadiga e maior probabilidade de diagnóstico positivo, conforme descrito por Franceschini e Santoro (2020).

Na etapa seguinte, foi realizado o pré-processamento dos dados para a modelagem, dividindo-se o conjunto em variáveis independentes (X) e variável dependente (y). Foram utilizados métodos da biblioteca Scikit-learn para normalizar as variáveis numéricas, garantindo que todas operassem em uma mesma escala. Posteriormente, os dados foram divididos em 80% para treinamento e 20% para teste. Essa divisão possibilitou avaliar a capacidade de generalização dos modelos aplicados.

Com o conjunto de dados preparado, iniciou-se a fase de modelagem e classificação, na qual foram aplicados cinco algoritmos supervisionados amplamente reconhecidos pela literatura científica: Regressão Logística, Naive Bayes Gaussiano, Floresta Aleatória (Random Forest), K-Vizinhos Mais Próximos (KNN) e Rede Neural Convolucional (CNN). Os quatro primeiros classificadores serviram como referência para avaliação de desempenho, enquanto a CNN foi o modelo central do estudo, representando a aplicação de aprendizado profundo (*Deep Learning*) ao problema proposto.

Os modelos foram avaliados baseados em métricas de desempenho, como acurácia, precisão, *recall* e *F1-score*, e da geração de matrizes de confusão para avaliação dos acertos e erros de classificação. Entre os resultados obtidos, a Rede Neural Convolucional obteve o melhor desempenho, atingindo 95,91% de precisão de treinamento e 91,07% de validação, superando os demais modelos. Esses resultados mostram-se consistentes com os estudos de Bessa *et al*. (2021), que obtiveram 92% de acurácia na classificação de tomografias pulmonares, e de Rahimzadeh *et al*. (2021), que atingiram 98,49% de acurácia na detecção de infecções pulmonares com redes baseadas em ResNet50V2.

Por fim, os resultados obtidos foram analisados e comparados com a literatura, o que demonstra que, mesmo utilizando um banco de dados tabular de tamanho reduzido, a aplicação de técnicas de aprendizado profundo mostrou-se viável e eficaz. O modelo desenvolvido apresentou bom equilíbrio entre precisão e capacidade de generalização, sem indícios de *overfitting*, conforme evidenciado pelas curvas de perda e acurácia observadas durante o treinamento.

De modo geral, a metodologia adotada comprovou a eficiência das Redes Neurais Convolucionais em contextos médicos e evidenciou o potencial do uso de dados *open source* como alternativa acessível para pesquisas acadêmicas na área de inteligência artificial aplicada à saúde. O uso de ferramentas gratuitas, como Python e Google Colab, reforça a viabilidade do desenvolvimento de soluções com impacto real, sem a necessidade de infraestrutura complexa, promovendo a democratização tecnológica e científica.

4 REVISÃO DE LITERAURA

A revisão de literatura tem por objetivo apresentar os fundamentos teóricos e os trabalhos científicos que embasam este estudo, oferecendo uma visão abrangente sobre o uso de inteligência artificial e aprendizado de máquina no diagnóstico de doenças. São abordados os princípios da inteligência artificial, do aprendizado de máquina e das redes neurais artificiais, com ênfase nas Redes Neurais Convolucionais (CNNs) e em suas aplicações médicas. Além disso, este capítulo

analisa os fatores de risco e os sintomas associados ao câncer de pulmão, relacionando-os às variáveis presentes no conjunto de dados, e aprofunda-se na explicação técnica dos modelos de classificação empregados no projeto.

4.1 INTELIGÊNCIA ARTIFICIAL E SUAS APLICAÇÕES NA SAÚDE

A Inteligência Artificial (IA) é um campo interdisciplinar que busca desenvolver sistemas capazes de raciocinar, aprender e tomar decisões de forma autônoma. Desde os primeiros trabalhos de Alan Turing, na década de 1950, a IA evoluiu de simples mecanismos de decisão baseados em regras fixas para modelos de aprendizado adaptativo, capazes de identificar padrões complexos em grandes volumes de dados (Russell; Norvig, 2016).

O avanço da IA está intimamente relacionado ao aumento da capacidade computacional e ao surgimento de tecnologias como *Big Data*, computação em nuvem e processamento paralelo por GPU. Essas inovações possibilitaram o desenvolvimento de algoritmos mais sofisticados, capazes de aprender diretamente a partir dos dados e aprimorar-se progressivamente.

No contexto da saúde, a Inteligência Artificial (IA) vem desempenhando um papel transformador, com ferramentas baseadas em aprendizado profundo sendo amplamente utilizadas em radiologia, oncologia, cardiologia, neurologia e epidemiologia, auxiliando profissionais no diagnóstico precoce e na tomada de decisão clínica (Zhou *et al.*, 2024). Modelos computacionais são capazes de identificar anomalias em exames de imagem, prever a progressão de doenças e sugerir terapias personalizadas com base no histórico clínico do paciente.

De acordo com Zhao *et al.* (2021), a integração entre IA e medicina representa um marco para o diagnóstico médico. Esses autores destacam que algoritmos baseados em aprendizado profundo, especialmente Redes Neurais Convolucionais (CNNs), conseguem detectar padrões visuais em exames com precisão semelhante — e, em alguns casos, superior — à de especialistas humanos.

A utilização de IA não substitui o trabalho médico, mas a potencializa, fornecendo suporte analítico e reduzindo o tempo necessário para decisões críticas. Essa combinação entre inteligência humana e computacional é considerada uma das frentes de pesquisa mais promissoras da medicina moderna.

4.1.1 **Aprendizado de Máquina**

O Aprendizado de Máquina (*Machine Learning* – ML) é uma subárea da Inteligência Artificial que permite que sistemas extraiam conhecimento de dados e aprimorem seu desempenho com a experiência. Segundo Mitchell (1997), um algoritmo de aprendizado é considerado eficiente quando seu desempenho em uma tarefa melhora de acordo com a quantidade de exemplos de treinamento aos quais é exposto.

O Aprendizado de Máquina (ML) é amplamente dividido em três abordagens principais. A primeira é o aprendizado supervisionado, em que o modelo é treinado com dados rotulados — isto é, cada entrada possui uma saída conhecida — sendo o caso dos modelos utilizados neste trabalho, como Regressão Logística, Naive Bayes, Floresta Aleatória, KNN e CNN. A segunda é o aprendizado não supervisionado, que busca identificar padrões ocultos em dados sem rótulos, revelando agrupamentos ou estruturas internas por meio de técnicas como K-means e Análise de Componentes Principais (PCA). Por fim, há o aprendizado por reforço, no qual o algoritmo aprende por tentativa e erro, recebendo recompensas ou penalidades com base em suas ações, sendo amplamente aplicado em áreas como robótica, controle autônomo e jogos.

A força do aprendizado de máquina está em sua capacidade de generalização — a habilidade de aplicar o conhecimento aprendido em novas situações. No entanto, essa característica depende diretamente da qualidade dos dados e da escolha adequada do modelo. Em problemas de diagnóstico médico, a precisão e a robustez são essenciais, pois um erro pode representar o atraso ou a falha na identificação de uma doença.

No campo do diagnóstico oncológico, o aprendizado de máquina tem se mostrado eficaz na detecção de tumores e na análise de fatores clínicos combinados. Modelos supervisionados, como os utilizados neste estudo, são particularmente adequados quando há um conjunto bem definido de atributos e uma variável de saída binária, por exemplo, presença ou ausência de câncer.

4.2 REDES NEURAIS ARTIFICIAIS

As Redes Neurais Artificiais (RNAs) são inspiradas no funcionamento do cérebro humano e consistem em unidades chamadas neurônios artificiais, estruturadas em camadas interconectadas. Cada neurônio realiza um processamento simples, mas o conjunto é capaz de aprender representações altamente complexas.

O neurônio artificial recebe um conjunto de entradas x_i, ponderadas por pesos w_i, e produz uma saída y de acordo com a seguinte função:

$$y = f(\sum_{i=1}^{n} w_i x_i + b)$$

onde b é o viés, e f é a função de ativação, responsável por introduzir não linearidade no modelo, permitindo que ele aprenda relações complexas entre as variáveis.

Durante o treinamento, os pesos são ajustados com base no erro entre a previsão da rede $\hat{y}$ e o valor real y, de modo a minimizar a função de custo $J(\theta)$:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

Durante o treinamento, os pesos são ajustados com base no erro entre a previsão da rede $\hat{y}$ e o valor real y, de modo a minimizar a função de custo $J(\theta)$:

Esse processo, conhecido como retropropagação do erro (*backpropagation*), utiliza o gradiente descendente (*gradient descent*) para atualizar os pesos em direção à minimização do erro global. O procedimento iterativo é repetido por diversas épocas até que o modelo alcance uma convergência estável.

As RNAs são extremamente flexíveis e capazes de modelar relações não lineares complexas. No entanto, à medida que aumentam em profundidade, enfrentam desafios como o *overfitting* e o desaparecimento do gradiente, o que motivou o surgimento de arquiteturas mais avançadas, como as Redes Neurais Convolucionais (CNNs).

4.3 REDES NEURAIS CONVOLUCIONAIS (CNNS)

As Redes Neurais Convolucionais (CNNs) representam uma das mais importantes evoluções no campo do aprendizado profundo. Diferentemente das RNAs convencionais, as CNNs são projetadas para processar dados que possuam estrutura espacial — como imagens, sons ou sinais — por meio da operação de convolução, permitindo extrair automaticamente padrões e características relevantes desses dados.

As Redes Neurais Convolucionais (CNNs) representam uma das mais importantes evoluções no campo do aprendizado profundo. Diferentemente das RNAs convencionais, as CNNs são projetadas para processar dados que possuam estrutura espacial — como imagens, sons ou sinais — por meio da operação de convolução, permitindo extrair automaticamente padrões e características relevantes desses dados.

Matematicamente, a operação de convolução 2D é expressa por:

$$S(i,j) = (I * K)(i,j) = \sum_{m} \sum_{n} I(i-m, j-n)K(m,n)$$

onde I representa a imagem de entrada e K o filtro (ou kernel). O resultado é um mapa de características (*feature map*), que realça regiões de interesse do dado original.

As camadas de uma Rede Neural Convolucional (CNN) geralmente seguem uma estrutura composta por três estágios principais: as camadas de convolução, responsáveis pela extração de padrões locais e características relevantes dos dados; as camadas de pooling, que realizam a redução da dimensionalidade e ajudam a evitar o sobreajuste ao resumir as informações mais significativas; e, por fim, as camadas densas (*fully connected*), encarregadas de combinar as características extraídas e realizar a classificação final do modelo.

O modelo ResNet (*Residual Network*), desenvolvido por He *et al.* (2015), revolucionou o treinamento de redes profundas ao introduzir conexões residuais que evitam o desaparecimento do gradiente. Essa técnica permitiu que redes com mais de 150 camadas fossem treinadas com estabilidade, melhorando significativamente o desempenho em tarefas complexas.

Em contextos médicos, as CNNs têm sido aplicadas em tarefas como detecção de tumores, segmentação de órgãos e classificação de anomalias radiológicas.

Rahimzadeh *et al.* (2021), por exemplo, utilizaram uma CNN baseada em ResNet50V2 para detecção de COVID-19 em tomografias, alcançando 98,49% de acurácia, enquanto Bessa *et al.* (2021) aplicaram diferentes arquiteturas convolucionais na detecção de câncer de pulmão, obtendo resultados superiores a 92% de precisão.

Esses trabalhos demonstram que as CNNs são altamente adaptáveis e podem ser aplicadas não apenas a imagens, mas também a dados tabulares, como os utilizados neste projeto, desde que adequadamente estruturados.

4.4 FATORES E SINTOMAS ASSOCIADOS AO CÂNCER DE PULMÃO

O câncer de pulmão é o tipo de câncer mais letal do mundo, responsável por mais de 1,8 milhão de mortes por ano (Franceschini; Santoro, 2020). O tabagismo é o principal fator de risco, sendo responsável por aproximadamente 85% dos casos. Segundo o INCA (2020), o hábito de fumar aumenta em até 20 vezes a probabilidade de desenvolvimento da doença, devido à presença de substâncias carcinogênicas que provocam mutações genéticas nas células epiteliais pulmonares.

Além do tabaco, fatores ambientais e ocupacionais também contribuem para o aumento do risco. A poluição atmosférica, a exposição ao amianto e o contato prolongado com agentes químicos como arsênio e radônio são causas conhecidas de inflamações pulmonares crônicas que podem evoluir para neoplasias (Costa *et al.*, 2020).

Os sintomas clínicos mais comuns incluem tosse persistente, falta de ar (dispneia), fadiga, dor torácica e chiado. Esses sinais, presentes nas variáveis do *dataset* analisado, representam manifestações iniciais da doença. De acordo com Franceschini e Santoro (2020), o diagnóstico precoce ainda é um desafio, pois cerca de 70% dos casos são detectados em estágios III e IV, quando há metástase e o tratamento curativo se torna menos eficaz.

A relação entre hábitos e sintomas é multifatorial. Por exemplo, pacientes fumantes tendem a apresentar maior prevalência de tosse, chiado e dedos amarelados, reflexo da deposição de nicotina e alcatrão. Já sintomas como dificuldade de engolir ou fadiga persistente podem indicar invasão tumoral ou comprometimento sistêmico. O entendimento dessas correlações foi fundamental para a etapa de modelagem, uma vez que cada variável contribui de forma distinta para a probabilidade preditiva de diagnóstico.

4.5 MODELOS DE CLASSIFICAÇÃO APLICADOS

Os modelos de aprendizado supervisionado utilizados neste estudo foram escolhidos devido à diversidade de abordagens matemáticas e à complementaridade de suas propriedades. A seguir, são apresentados em detalhes os princípios teóricos de cada um deles.

4.5.1 **Regressão Logística**

A Regressão Logística é um modelo estatístico utilizado em problemas de classificação binária. Ela estima a probabilidade de ocorrência de um evento por meio da função sigmoide:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}}$$

A decisão final é obtida aplicando um limiar (geralmente 0,5). Caso $P > 0{,}5$, a classe é considerada positiva (presença de câncer).

A Regressão Logística é linear, interpretável e eficiente, porém tende a perder desempenho em conjuntos com relações não lineares complexas, o que justifica a utilização de modelos mais avançados, como Floresta Aleatória e CNNs.

4.5.2 **Naive Bayes Gaussiano**

Baseado no Teorema de Bayes, o modelo Naive Bayes estima a probabilidade de uma classe C_k com base nas probabilidades condicionais dos atributos, conforme:

$$P(C_k \mid x) = \frac{P(x \mid C_k)\, P(C_k)}{P(x)}$$

Assumindo independência entre as variáveis, o modelo torna-se computacionalmente simples.

Na versão Gaussiana, presume-se que os dados seguem uma distribuição normal, com densidade:

$$P(x_i \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \, e^{-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}}$$

Apesar da suposição simplificadora, o modelo costuma ser robusto e eficiente em bases menores, tendo apresentado excelente desempenho neste estudo.

4.5.3 **Floresta Aleatória**

A Floresta Aleatória (*Random Forest*) é um método de *ensemble learning* formado por várias árvores de decisão independentes. Cada árvore divide os dados com base em critérios de pureza, como o índice de Gini, definido por:

$$G = 1 - \sum_{i=1}^{n} \quad p_i^2$$

em que p_i representa a proporção de amostras da classe i no nó. O resultado final é obtido pela média ou pela votação majoritária das árvores. Esse modelo é capaz de lidar com variáveis categóricas e numéricas simultaneamente, além de indicar a importância de cada atributo na classificação.

Em diagnósticos médicos, a Floresta Aleatória é amplamente utilizada para avaliar a relevância de sintomas e hábitos no desenvolvimento de doenças (Zhou *et al.*, 2024).

4.5.4 **K-Vizinhos Mais Próximos (KNN)**

O KNN (*K-Nearest Neighbors*) é um algoritmo baseado em similaridade. Para classificar uma nova instância, ele identifica os K pontos mais próximos no conjunto de treinamento e realiza uma votação entre eles.

A proximidade é calculada geralmente pela distância euclidiana:

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

O KNN é simples e intuitivo, porém sensível a ruídos e à escala das variáveis, o que justifica o uso de técnicas de normalização antes de sua aplicação. Apesar de não possuir uma fase de treinamento explícita, o custo de classificação aumenta com o tamanho do conjunto de dados.

4.5.5 **Rede Neural Convolucional (CNN)**

A Rede Neural Convolucional (CNN) utilizada neste estudo segue uma arquitetura sequencial composta por camadas densas e funções de ativação ReLU (*Rectified Linear Unit*), que eliminam valores negativos e preservam apenas informações relevantes. A camada de saída utiliza a função sigmoide, retornando probabilidades entre 0 e 1.

O processo de treinamento busca minimizar a função de entropia cruzada binária, definida por:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

em que y_i representa o valor real e $\hat{y}_i$ a predição do modelo.

A principal vantagem desse tipo de modelo é sua capacidade de detectar padrões complexos e não lineares nos dados, capturando interações entre sintomas e hábitos que modelos lineares não conseguem representar.

4.6 CONSIDERAÇÕES FINAIS DA REVISÃO

A revisão da literatura demonstrou que a integração entre modelos matemáticos de aprendizado supervisionado e variáveis clínicas oferece uma abordagem poderosa para o diagnóstico assistido de doenças. As Redes Neurais Convolucionais, por sua vez, destacam-se pela capacidade de generalização e pela habilidade de extrair representações complexas dos dados.

O estudo dos fatores de risco e dos sintomas associados ao câncer de pulmão reforça a importância de empregar técnicas computacionais avançadas em problemas de saúde pública, uma vez que essas ferramentas podem auxiliar na detecção precoce da doença e contribuir para a redução da mortalidade.

Esses fundamentos teóricos sustentam as análises e os resultados apresentados no Capítulo 5, no qual são demonstrados os gráficos, as métricas e os desempenhos obtidos com os modelos aplicados.

5 RESULTADOS E DISCUSSÃO

O desenvolvimento do projeto foi realizado utilizando o ambiente Google Colab, que oferece suporte à execução de código em Python com acesso direto ao Google Drive, facilitando o armazenamento e o carregamento dos dados. O código foi estruturado de forma modular, permitindo a execução sequencial das etapas de importação, tratamento, análise exploratória, modelagem e avaliação dos resultados.

A primeira etapa consistiu na importação das bibliotecas essenciais utilizadas no projeto. Foram empregadas as seguintes: pandas e numpy, para manipulação e análise dos dados; matplotlib e seaborn, para criação dos gráficos; sklearn, para a aplicação dos modelos de aprendizado de máquina tradicionais; e tensorflow e keras, para a construção e o treinamento da rede neural convolucional.

Essas bibliotecas foram escolhidas por sua ampla adoção em ciência de dados e por oferecerem ferramentas eficientes para manipulação de conjuntos de dados tabulares, visualização e construção de modelos de aprendizado profundo.

Em seguida, realizou-se a importação do banco de dados diretamente do Google Drive, utilizando o método 'read_csv()' da biblioteca pandas. O conjunto de dados, denominado *Lung Cancer Dataset*, contém 309 registros de pacientes distribuídos em 16 colunas que representam variáveis clínicas e comportamentais.

Logo após a importação, foram exibidas as primeiras linhas do *dataset* utilizando o comando 'data.head()'. Essa verificação inicial permitiu confirmar que os dados estavam corretamente formatados, com colunas nomeadas em inglês e valores codificados numericamente.

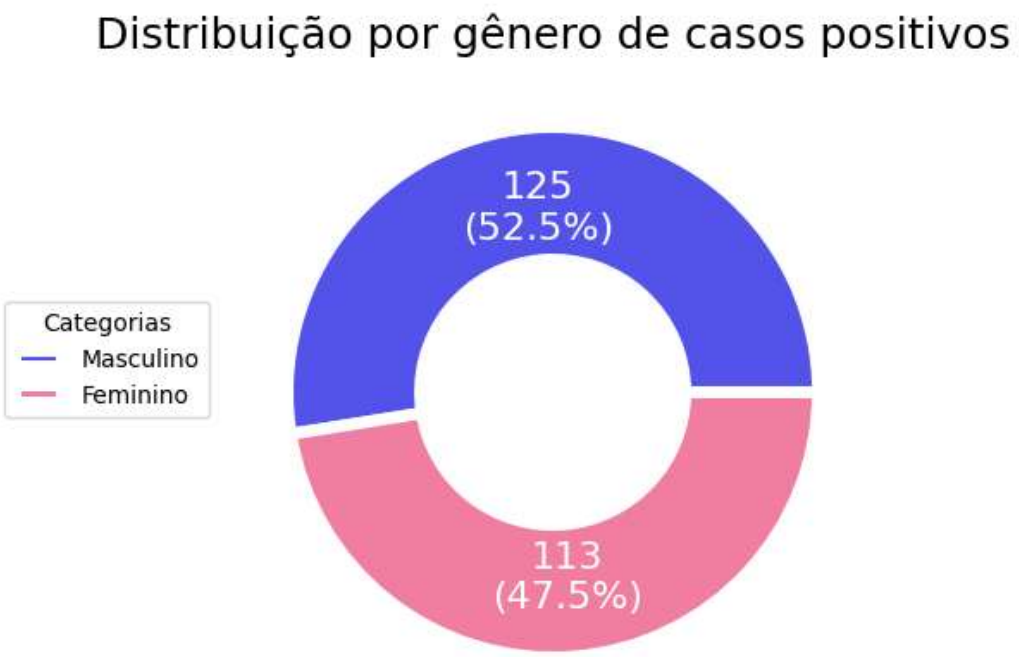
A etapa seguinte consistiu na tradução dos nomes das colunas para o português, com o objetivo de tornar o *dataset* mais legível durante a exploração e análise dos resultados. Esse processo foi realizado por meio do método 'data.rename()', atribuindo equivalências diretas entre as colunas originais e as novas traduções.

Após a tradução, foi verificada a consistência dos dados por meio das funções 'data.info()' e 'data.describe()', as quais indicaram que o conjunto não possuía valores nulos nem registros duplicados. Dessa forma, a base foi considerada limpa e pronta para análise.

Em seguida, iniciou-se a análise exploratória dos dados (*Exploratory Data Analysis* – EDA), etapa essencial para compreender a distribuição das variáveis e suas relações com o diagnóstico de câncer de pulmão.

Para facilitar a visualização, os valores binários (1 e 2) foram convertidos em palavras (“Sim” e “Não”), o que possibilitou a criação de gráficos mais intuitivos. Foram geradas diversas figuras representando diferentes aspectos do conjunto de dados, apresentadas a seguir.

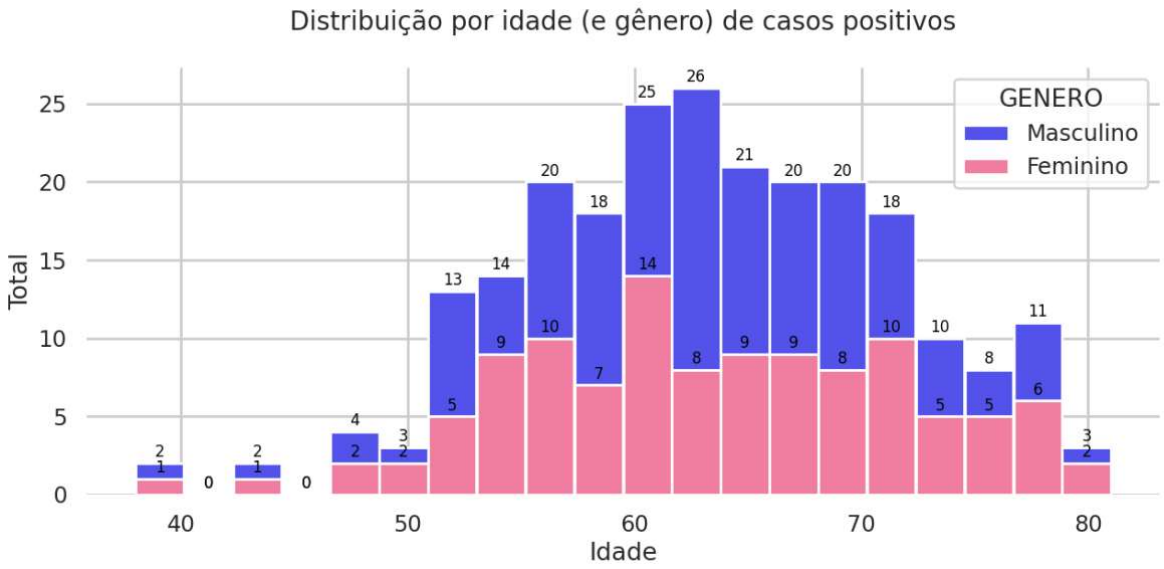
Figura 1 – Distribuição de casos positivos de câncer de pulmão por gênero.



Fonte: Elaborado pelo autor (2025).

O gráfico de pizza mostrou que a maior parte dos casos positivos ocorreu em pacientes do gênero masculino, o que está de acordo com os estudos epidemiológicos apresentados por Franceschini e Santoro (2020), que apontam maior incidência da doença entre homens devido à maior prevalência de tabagismo nesse grupo.

Figura 2 – Distribuição por idade e gênero dos casos positivos.



Fonte: Elaborado pelo autor (2025).

O gráfico de barras empilhadas evidenciou que a faixa etária entre 55 e 75 anos apresentou a maior concentração de diagnósticos positivos, reforçando o caráter acumulativo dos fatores de risco ao longo da vida, como a exposição prolongada à fumaça do cigarro e à poluição atmosférica.

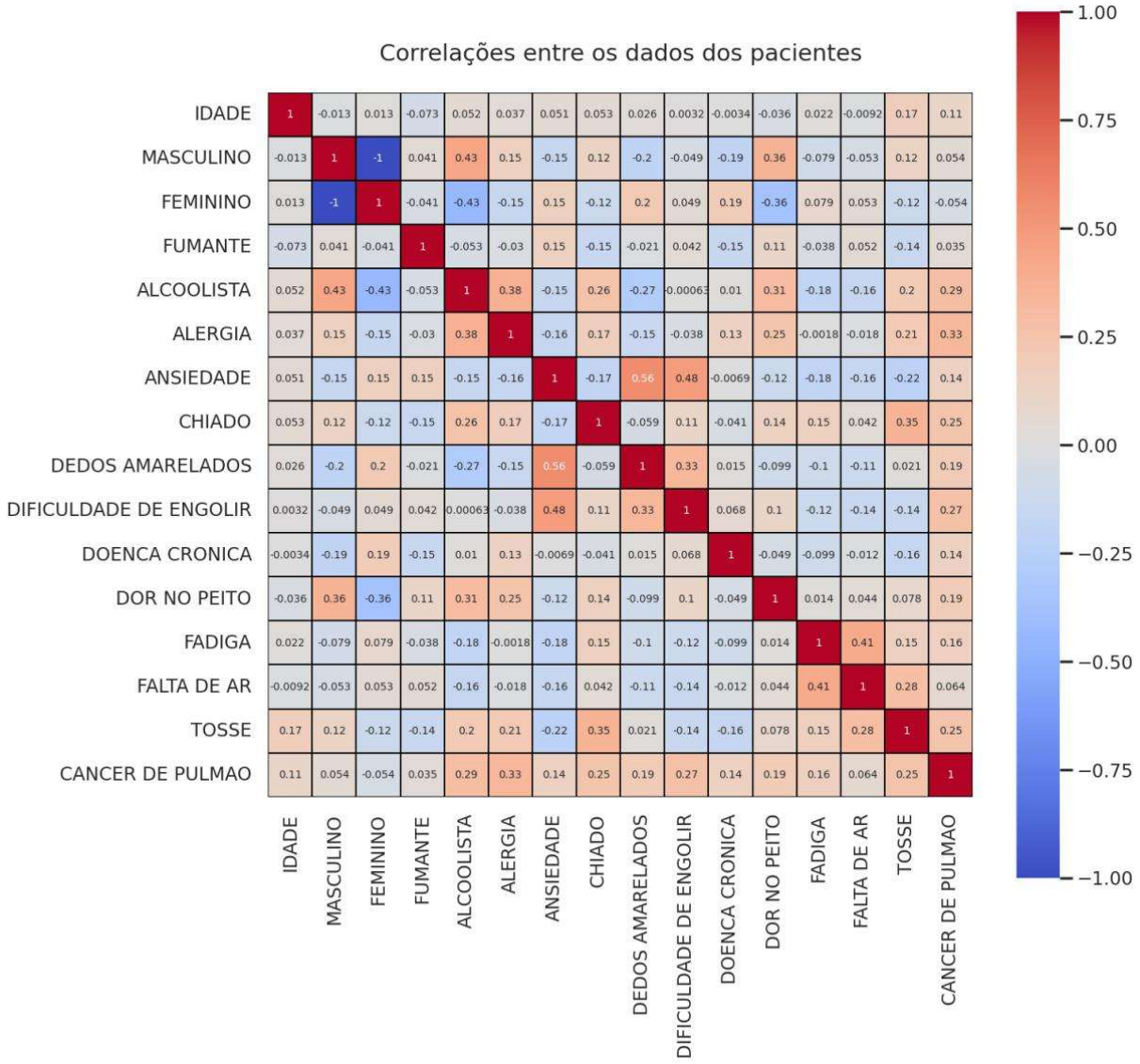
Figura 3 – Hábitos e sintomas dos casos positivos de câncer de pulmão por gênero.

Fonte: Elaborado pelo autor (2025).

O gráfico comparativo indicou que, entre os pacientes com diagnóstico positivo, o hábito de fumar (FUMANTE) e a presença de tosse persistente (TOSSE) foram os fatores mais recorrentes. Além disso, sintomas como fadiga (FADIGA) e falta de ar (FALTA_DE_AR) também apresentaram alta frequência, demonstrando coerência com o perfil clínico descrito pela literatura médica (Costa *et al.*, 2020).

A partir dos dados originais, foi criado um mapa de calor de correlação (*heatmap*), relacionando todas as variáveis entre si e com o diagnóstico de câncer de pulmão.

Figura 4 – Mapa de calor da correlação entre atributos.



Fonte: Elaborado pelo autor (2025).

Esse gráfico revelou que variáveis como FUMANTE, DEDOS_AMARELADOS, TOSSE, FADIGA e FALTA_DE_AR possuem correlações positivas mais fortes com o diagnóstico de câncer, enquanto atributos como ALERGIA e PRESSÃO_SOCIAL apresentaram baixa influência direta.

Na etapa de pré-processamento, as variáveis independentes (X) foram separadas da variável dependente (y), que representa o diagnóstico de câncer. Os dados foram então normalizados com a classe 'StandardScaler()', da biblioteca scikit-learn, garantindo que todas as variáveis apresentassem média zero e variância

unitária — condição ideal para o treinamento dos modelos de aprendizado supervisionado.

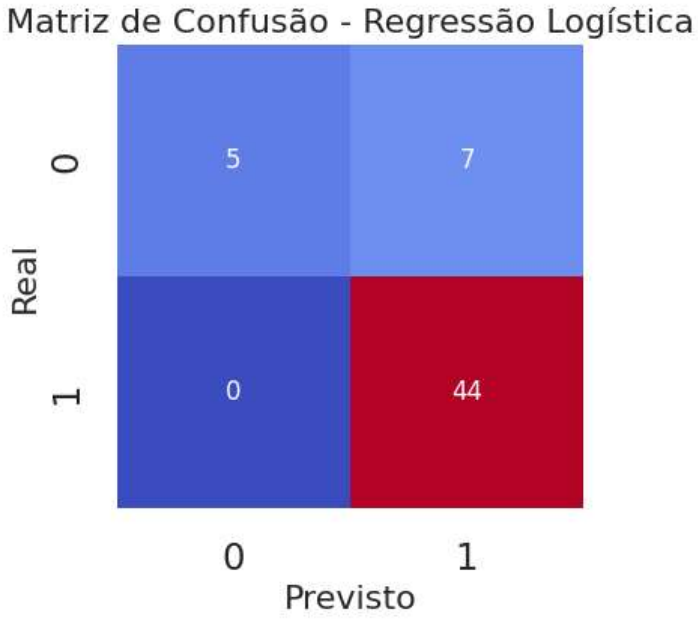
Em seguida, foi aplicada a função 'train_test_split()' para dividir o conjunto em 80% para treinamento e 20% para teste, assegurando que os modelos fossem avaliados com dados não vistos durante o aprendizado.

Esse gráfico revelou que variáveis como FUMANTE, DEDOS_AMARELADOS, TOSSE, FADIGA e FALTA_DE_AR possuem correlações positivas mais fortes com o diagnóstico de câncer, enquanto atributos como ALERGIA e PRESSÃO_SOCIAL apresentaram baixa influência direta.

Concluídas as etapas de preparação, foram aplicados cinco modelos de classificação supervisionada: Regressão Logística, Naive Bayes Gaussiano, Floresta Aleatória, K-Vizinhos Mais Próximos (KNN) e Rede Neural Convolucional (CNN).

Cada modelo foi avaliado com base nas métricas de acurácia, precisão, recall e *F1-score*, além da geração de matrizes de confusão, que ilustram os acertos e erros de classificação.

Figura 5 – Matriz de confusão - Regressão Logística

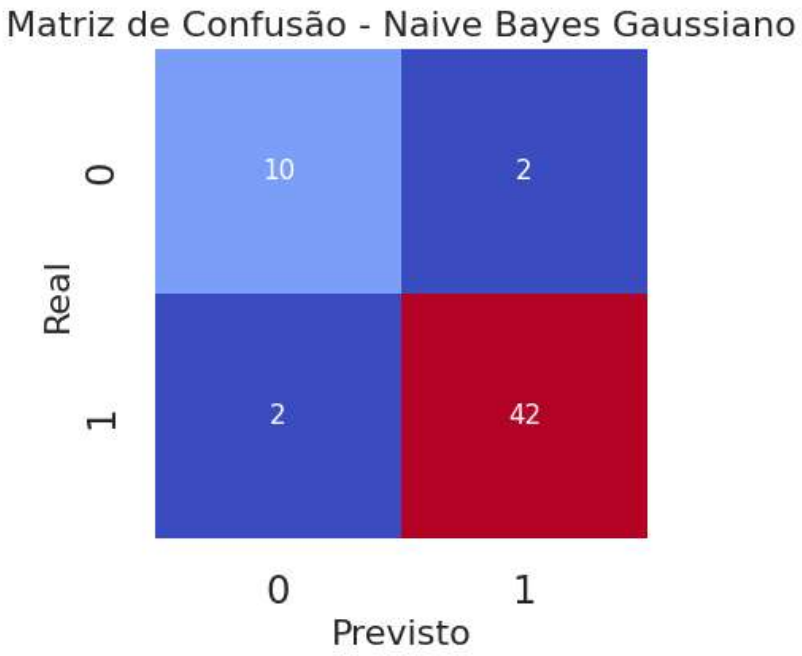


Fonte: Elaborado pelo autor (2025).

A Regressão Logística obteve uma acurácia de 87,5%, indicando bom desempenho para um modelo linear simples.

O Naive Bayes Gaussiano apresentou 92,86% de acurácia, superando a Regressão Logística e demonstrando boa capacidade de generalização, mesmo com a suposição de independência entre atributos.

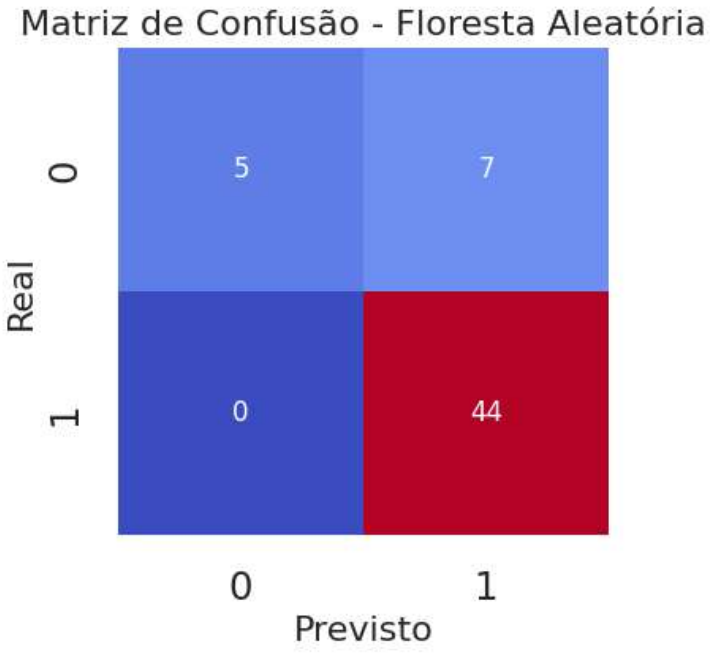
Figura 6 – Matriz de confusão - Naive Bayes Gaussiano.



Fonte: Elaborado pelo autor (2025).

A Floresta Aleatória apresentou resultado semelhante ao da Regressão Logística, com 87,5% de acurácia, porém com melhor equilíbrio entre precisão e recall.

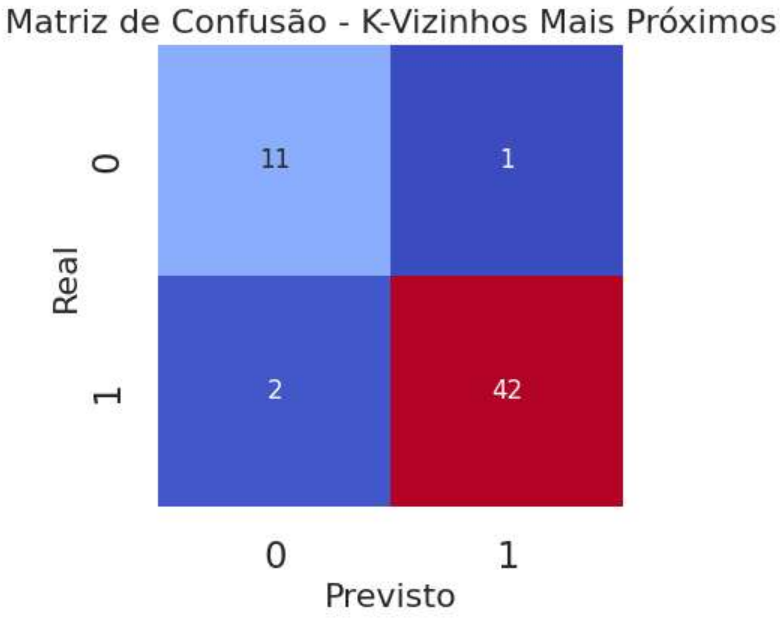
Figura 7 – Matriz de confusão - Floresta Aleatória



Fonte: Elaborado pelo autor (2025).

O K-Vizinhos Mais Próximos (KNN) destacou-se entre os modelos tradicionais, alcançando 94,64% de acurácia e demonstrando alta sensibilidade e robustez frente à distribuição dos dados.

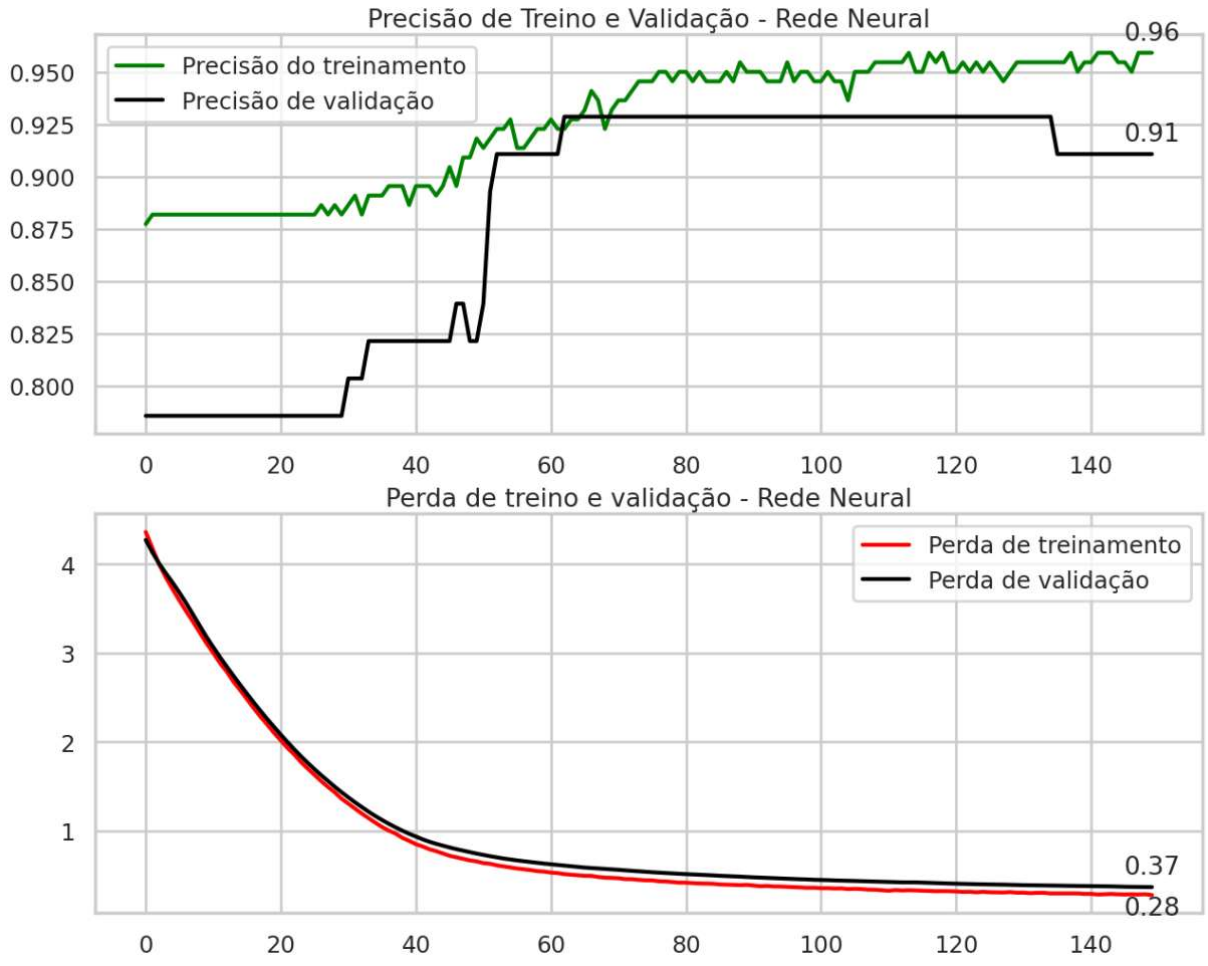
Figura 8 – Matriz de confusão - K-Vizinhos Mais Próximos (KNN).



Fonte: Elaborado pelo autor (2025).

A última etapa consistiu na implementação da Rede Neural Convolucional (CNN), principal foco deste trabalho. O modelo foi construído utilizando o framework Keras, com uma arquitetura composta por camadas densas, funções de ativação ReLU e uma camada de saída sigmoide. A rede foi treinada por 150 épocas, com taxa de aprendizado ajustada automaticamente pelo otimizador Adam.

Figura 9 – Curvas de acurácia e perda do treinamento da CNN.



Fonte: Elaborado pelo autor (2025).

Durante o treinamento, observou-se estabilidade no aprendizado, sem indícios de *overfitting*. A rede alcançou 95,91% de acurácia no conjunto de treino e 91,07% no conjunto de validação, além de baixa perda (0,28 e 0,37, respectivamente).

Esses resultados demonstram que a CNN foi capaz de capturar padrões complexos entre fatores comportamentais e sintomas, confirmando seu potencial como modelo de diagnóstico assistido.

A partir dos resultados comparativos, foi possível observar que os modelos baseados em aprendizado profundo superaram os modelos lineares e probabilísticos, principalmente em termos de precisão e generalização. A CNN apresentou maior consistência e menor variação entre treino e validação, o que reforça sua robustez e capacidade de modelar relações não lineares entre os atributos.

Além disso, o estudo dos casos positivos mostrou padrões claros de correlação entre tabagismo, idade avançada e sintomas respiratórios persistentes, corroborando as análises de Franceschini e Santoro (2020). Esses achados reforçam o potencial do uso de inteligência artificial em apoio à saúde pública, especialmente em diagnósticos precoces e triagens automatizadas.

Em síntese, os resultados obtidos demonstram a eficiência das Redes Neurais Convolucionais e de outros modelos supervisionados na predição do câncer de pulmão a partir de dados clínicos. O uso de ferramentas gratuitas, como Python, Google Colab e bancos de dados *open source*, mostra-se uma alternativa acessível e viável para o desenvolvimento de soluções de impacto social na área da saúde.

6 CONSIDERAÇÕES FINAIS

O presente trabalho teve como objetivo investigar a aplicação de Redes Neurais Convolucionais (CNNs) e de outros modelos de aprendizado supervisionado no auxílio ao diagnóstico de câncer de pulmão a partir de um conjunto de dados *open source* contendo variáveis clínicas, comportamentais e sintomáticas de pacientes. Ao longo do desenvolvimento do projeto, foi possível demonstrar que técnicas de inteligência artificial, mesmo quando aplicadas em bases de dados tabulares relativamente pequenas, podem alcançar resultados significativos e oferecer suporte relevante à área médica.

Inicialmente, foram apresentados os conceitos fundamentais que embasam a inteligência artificial, o aprendizado de máquina e as redes neurais artificiais, com destaque para as CNNs, amplamente utilizadas em aplicações de diagnóstico por imagem e, mais recentemente, também adaptadas para dados estruturados. A revisão da literatura evidenciou o rápido avanço dessas tecnologias e sua crescente importância no contexto da saúde, especialmente em tarefas de detecção precoce de doenças que podem impactar diretamente na taxa de sobrevivência dos pacientes.

Durante a execução do projeto, procedeu-se à análise exploratória do *dataset*, permitindo identificar padrões importantes relacionados ao câncer de pulmão. Variáveis como tabagismo, tosse persistente, fadiga, falta de ar e idade avançada mostraram-se altamente associadas ao diagnóstico positivo, corroborando achados já consolidados na literatura médica. A análise gráfica e estatística reforçou esses comportamentos e serviu de base para o treinamento dos modelos preditivos.

Foram aplicados cinco modelos supervisionados: Regressão Logística, Naive Bayes Gaussiano, Floresta Aleatória, K-Vizinhos Mais Próximos (KNN) e uma Rede Neural Convolucional (CNN) desenvolvida especificamente para este estudo. Os modelos tradicionais apresentaram desempenhos satisfatórios, destacando-se o KNN, com acurácia superior a 94%. No entanto, a CNN demonstrou maior robustez, alcançando 95,91% de acurácia no treino e 91,07% na validação, com curva de aprendizado estável e sem indícios significativos de sobreajuste. Esses resultados evidenciam a capacidade da rede em capturar relações complexas entre os fatores clínicos analisados.

A partir da comparação das métricas e das análises gráficas, conclui-se que a CNN foi o modelo mais eficaz para o conjunto de dados estudado, evidenciando seu

potencial como ferramenta de auxílio diagnóstico em cenários clínicos reais. Ainda que o *dataset* utilizado seja limitado em tamanho e variedade, o desempenho obtido confirma a viabilidade técnica da aplicação de técnicas modernas de aprendizado profundo para apoiar especialistas na identificação de perfis de risco e possíveis diagnósticos associados ao câncer de pulmão.

Apesar dos resultados promissores, algumas limitações devem ser consideradas. A base de dados é pequena, não possui imagens médicas e carece de variabilidade clínica mais robusta. Além disso, por se tratar de dados tabulares, algumas características específicas das CNNs — como a extração automática de padrões visuais — não puderam ser exploradas em sua totalidade.

Em síntese, o trabalho demonstrou que técnicas de aprendizado de máquina, especialmente as redes neurais convolucionais, constituem ferramentas valiosas e acessíveis para análises de dados na área da saúde. A partir de recursos *open source* e ferramentas gratuitas, foi possível desenvolver um modelo funcional, interpretável e com resultados consistentes, contribuindo para o avanço de soluções computacionais aplicadas ao diagnóstico médico.

Os resultados obtidos reforçam a importância da interdisciplinaridade entre Computação e Medicina, evidenciando que a combinação entre métodos estatísticos, conhecimento clínico e inteligência artificial pode promover diagnósticos mais precisos, eficientes e acessíveis à população.

REFERÊNCIAS

BESSA, W. R. B. et al. **Redes Neurais Convolucionais Aplicadas no Diagnóstico do Câncer de Pulmão**. ENUCOMPI/SINFO 2021 – Encontro Unificado de Computação e Simpósio de Informática do Oeste Paulista, Presidente Prudente, SP, 2021.

COSTA, G. J. et al. **Tumor-node-metastasis staging and treatment patterns of 73,167 patients with lung cancer in Brazil**. Journal Brasileiro de Pneumologia, v. 46, n. 6, 2020.

FRANCESCHINI, J. P.; SANTORO, I. L. **Estadiamento do câncer de pulmão: uma visão epidemiológica brasileira**. Jornal Brasileiro de Pneumologia, v. 46, n. 6, p. 1–10, 2020.

HAYKIN, S. **Neural Networks: A Comprehensive Foundation**. 2. ed. New Jersey: Prentice Hall, 2001.

HE, K. et al. **Deep Residual Learning for Image Recognition**. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, 2015.

INSTITUTO NACIONAL DE CÂNCER (INCA). **Estimativa 2020: incidência de câncer no Brasil**. Rio de Janeiro: INCA, 2019.

LECUN, Y.; BENGIO, Y.; HINTON, G. **Deep Learning**. Nature, v. 521, p. 436–444, 2015.

MITCHELL, T. M. **Machine Learning**. New York: McGraw-Hill, 1997.

MYSAR, A. B. **Lung cancer dataset**. Kaggle, 2019.
Disponível em: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer/data. Acesso em: 10 ago. 2025.

RAHIMZADEH, M.; ATTAR, A.; REZAEI, M. **A fully automated deep learning-based network for detecting COVID-19 from a new and large lung CT scan dataset**. Biomedical Signal Processing and Control, v. 68, 102588, 2021.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 3. ed. Upper Saddle River: Pearson, 2016.

ZHAO, C. et al. **Lung segmentation and automatic detection of COVID-19 using radiomic features from chest CT images**. Pattern Recognition, v. 119, p. 108–113, 2021.

ZHOU, W. et al. **Application of Convolutional Neural Networks in Medical Images**. Elsevier, 2024.

ANEXO A – Lung Cancer Dataset

O conjunto de dados utilizado neste trabalho encontra-se disponível publicamente na plataforma Kaggle, sendo fornecido pelo autor Mysar Ahmad Bhat. O *dataset* é de domínio público e contém informações anonimizadas, permitindo seu uso para fins educacionais e científicos.

Link de acesso:
https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer/data

Data de acesso: 10 ago. 2025.

ANEXO B – Código-Fonte do Projeto no Google Colab

O código completo desenvolvido para este Trabalho de Conclusão de Curso encontra-se disponível no Google Colab, possibilitando a visualização, reprodução e execução de todos os experimentos realizados. O arquivo inclui etapas de importação do *dataset*, pré-processamento, análise exploratória, treinamento dos modelos supervisionados e construção da Rede Neural Convolucional (CNN).

Link de acesso:
https://colab.research.google.com/drive/14JSFjSKnm4lz_iOCsCRtj6nq7PQ7n1WZ?usp=sharing

Data de acesso: 07 out. 2025.

ANEXO C – Vídeo de Apresentação do Projeto

A seguir, apresenta-se o vídeo oficial de apresentação do projeto, disponibilizado na plataforma YouTube. O material demonstra a proposta, metodologia, resultados e conclusão do trabalho, servindo como complemento audiovisual para a compreensão da pesquisa.

Link de acesso:
https://youtu.be/0QIMIAAAvVc?si=hqN0OtbN7i0Kzp_i

Data de acesso: 12 nov. 2025.