REVISTA TÓPICOS
ISSN 2965-6672 | Qualis A2

# APLICAÇÃO DA CLASSIFICAÇÃO DE VETORES DE SUPORTE E REGRESSÃO LOGÍSTICA NA DETECÇÃO DE CÂNCER DE PULMÃO, CÓLON E TIREOIDE A PARTIR DE DADOS TEXTUAIS

**APPLICATION OF SUPPORT VECTOR CLASSIFICATION AND LOGISTIC REGRESSION FOR THE DETECTION OF LUNG, COLON, AND THYROID CANCERS FROM TEXTUAL DATA**

Engenharias, Ciências da Saúde • 08/05/2026

REGISTRO DOI: 10.70773/revistatopicos/778110043

Bryan Neves Pinto
Francisco Ponciano Maciel Neto
Johnson Tavares Pinto
Juíle Yoshie Sarkis Hanada
Maria Giovanna Gonçalves Sales
Hylbert Bentes Rocha Rodrigues
Lilian Kalinka da Silva Carvalho
Ailon dos Santos Teixeira

**RESUMO**

Este estudo avalia a aplicação de algoritmos de aprendizado de máquina no diagnóstico de câncer com base na classificação de textos biomédicos. Foram analisados três tipos de câncer, o pulmão, cólon e tireoide, utilizando o conjunto de dados *Biomedical Text Publication Classification.* Dois modelos de aprendizado de máquina foram testados: Classificação por Vetores de Suporte (CVS) com kernel RBF e Regressão Logística, ambos treinados com embeddings gerados pelo Sentence-BERT. A performance foi medida por acurácia, precisão, recall e F1-score. Os resultados demonstram que a CVS superou a Regressão Logística em todas as métricas, alcançando 95% de acurácia, em textos clínicos da base de treinamento. Os achados destacam o potencial da inteligência artificial no apoio ao diagnóstico precoce de câncer por meio de textos clínicos

**Palavras-chave:** Aprendizado de máquina; Classificação por Vetores de Suporte; Regressão Logística; Inteligência Artificial.

**ABSTRACT**

This study evaluates the application of machine learning algorithms to cancer diagnosis based on the classification of biomedical texts. Three types of cancer—lung, colon, and thyroid—were analyzed using the Biomedical Text Publication Classification dataset. Two machine learning models were tested: Support Vector Classification (SVC) with an RBF kernel and Logistic Regression, both trained with embeddings generated by Sentence-BERT. Performance was measured using accuracy, precision, recall, and F1-score. The results demonstrate that SVC outperformed Logistic Regression across all metrics, achieving 95% accuracy on clinical texts from the training dataset. The findings highlight the potential of artificial intelligence to support the early diagnosis of cancer through clinical text

analysis.

**Keywords:** Machine Learning; Support Vector Classification; Logistic Regression; Artificial Intelligence.

## 1. INTRODUÇÃO

O câncer é uma das principais causas de morte no século XXI, com impacto significativo na saúde pública global. Estima-se que seja responsável por 16,8% das mortes gerais e 22,8% das mortes causadas por doenças não transmissíveis [Bray et al., 2022].

De acordo com Chehade et al. [2022], os cânceres de pulmão e cólon estão entre os mais letais. O câncer de pulmão é responsável por 18,4% das mortes relacionadas ao câncer, enquanto o de cólon responde por 9,2% dos óbitos atribuídos à doença em escala global.

O câncer de tireoide, por sua vez, ocupa a nona posição entre os tipos de câncer com maior incidência mundial. Embora possa acometer ambos os sexos, sua ocorrência é significativamente mais comum em mulheres, que representam cerca de 75% dos casos diagnosticados [Chen et al., 2023].

Esses dados indicam a necessidade de estratégias eficazes para diagnóstico precoce e tratamento direcionado. Essas estratégias devem ser personalizadas conforme o tipo de câncer, a idade do paciente e o estágio do tumor. A escolha do tratamento adequado, o uso de terapias inovadoras e o acompanhamento especializado são fundamentais para maximizar as chances de remissão.

No tratamento do câncer de cólon, Nair et al. [2024] ressaltam a importância da quimioterapia neoadjuvante e de uma abordagem multidisciplinar que inclui testes precoces de mismatch Repair

(MMR). Adicionalmente, para o câncer de tireoide com metástase pulmonar, Zhang et al. [2024] sugerem a terapia com radioiodo (RAI) no pós-operatório, especialmente para pacientes com menos de 55 anos de idade.

O diagnóstico precoce é difundido como uma estratégia também muito importante. Os autores Pashayan and Pharoah [2020] afirmam que as chances de sobrevivência de um paciente com câncer aumentam substancialmente se a doença for diagnosticada e tratada em um estágio clínico precoce, pois permite uma intervenção médica antecipada com o objetivo de retardar ou prevenir o desenvolvimento e a letalidade da doença.

No entanto, existem os desafios e riscos do diagnóstico precoce. Crosby et al. [2022] destacam o sobrediagnóstico e o subtratamento, que podem gerar prejuízos a indivíduos que não manifestaram uma malignidade clinicamente evidente. O excesso de intervenções, como procedimentos invasivos, terapias agressivas sem uma justificativa clinicamente, podem acarretar consequências físicas e psicológicas desnecessárias.

Nesse contexto, ferramentas inovadoras, como as baseadas em Inteligência Artificial (IA), emergem como aliadas para acelerar e refinar o processo diagnóstico. Uma subárea da IA que tem destaque são os algoritmos de aprendizado de máquina (AM).

Os algoritmos de AM utilizam grandes conjuntos de dados para tarefas específicas, permitindo que aprendam a partir da experiência com os dados, sem a exigência de programação explícita, como destacam Chehade et al. [2022]. Essa abordagem é aplicada em

contextos biomédicos para predição e classificação de diversos tipos de dados.

Este estudo se dedica à avaliação da aplicabilidade de algoritmos de AM no contexto do diagnóstico oncológico. A investigação foca especificamente no desempenho da Classificação de Vetores de Suporte (CVS) e da Regressão Logística.

O objetivo central desta pesquisa é analisar a capacidade desses modelos no reconhecimento e classificação de cânceres de pulmão, cólon e tireoide utilizando as métricas de precisão, recall, F1-score e acurácia. Para tal, a metodologia emprega dados rotulados e informações de sintomas de pacientes em formato textual, servindo como base para o treinamento e teste dos algoritmos propostos.

## 2. FUNDAMENTAÇÃO TEÓRICA OU REVISÃO DA LITERATURA

Esta seção apresenta uma revisão de estudos que empregam modelos de AM rasos (*shallow learning*, em inglês) na medicina diagnóstica, estabelecendo paralelos com o tema central deste artigo.

As pesquisas mencionadas entendem-se do câncer às condições como malária e doenças cardiovasculares. Essa abrangência se justifica pela busca por diagnósticos rápidos e precisos, pela viabilidade de tempo para a pesquisa e pela oportunidade de comparar modelos de AM rasos em diversos cenários clínicos.

A pesquisa de Atoyebi and Uwazie [2024] apresenta a comparação de três *shallow learning* aplicados no diagnóstico de malária, que são: Multinomial Naive Bayes (MNB), Gaussian Naive Bayes (GNB) e Random Forest (RF), visando desenvolver um sistema confiável para

prever se um paciente tem malária grave com base nas características apresentadas utilizando *prompt*.

Os modelos foram treinados num *dataset* que contém 2121 instâncias com os seguintes atributos: fatores demográficos da região, tratamento de água e risco em malária, bem como dados do paciente (o objetivo principal do estudo é ter um controle sobre os casos de malária na região (o local não será mencionado no artigo) e visualizar um futuro livre da malária [Atoyebi and Uwazie, 2024].

Ainda na pesquisa de Atoyebi and Uwazie, 2024], os classificadores de Naive Bayes foram avaliados como modelos robustos e que categorizam muito bem casos sérios de malária com base nos dados fornecidos, levando em consideração que a acurácia do MNB foi medida em 97% e do GNB e RF foram de 100% para dados de teste e validação.

Vale ressaltar que os autores não mencionaram uma possibilidade de sobreajuste nos modelos GNB e RF, pois os dados foram testados em base de teste/validação; porém, seria prudente testar modelos em outras amostras de dados.

A pesquisa de [Debnath et al., 2023] propõe uma abordagem baseada em algoritmos de aprendizado de máquina para prever a ocorrência de doenças cardiovasculares (DCVs), que são uma das principais causas de mortalidade global.

[Debnath et al., 2023] utilizaram um conjunto de dados com 14 atributos extraídos do repositório Kaggle, contendo informações clínicas de pacientes. Foram aplicados tanto algoritmos individuais (como SVM, GNB, Regressão Logística, Árvore de Decisão e RF)

quanto métodos de ensemble (como Bagging, AdaBoosting e Gradient Boosting).

Após o pré-processamento de dados e seleção de atributos, os modelos foram avaliados com base em métricas como precisão, *recall*, F1-score e acurácia. Os melhores desempenhos foram obtidos com os classificadores *Bagging* e *Gradient Boosting*, ambos atingindo uma acurácia de 81,3%. O estudo concluiu que técnicas de *ensemble* são mais eficazes na predição de DCVs em comparação com classificadores isolados.

[Geeitha et al., 2023] apresenta uma abordagem baseada em aprendizado de máquina para prever a sobrevida livre de doença em casos de câncer cervical recorrente. Utilizando o algoritmo GNB, os autores buscaram analisar fatores clínicos relevantes e prever a probabilidade de recorrência, auxiliando na escolha de tratamentos mais eficazes.

A pesquisa emprega um conjunto de dados clínicos, selecionando as dez características mais importantes por meio de técnicas de *feature selection*. Foram testados diferentes classificadores, incluindo *Naive Bayes*, RF, *Decision Stump*, *Decision Table* e *Lazy IBK*.

O modelo baseado em *Naive Bayes* obteve o melhor desempenho, atingindo 85% de acurácia. O estudo reforça o potencial da IA na personalização do tratamento e no suporte à decisão médica em oncologia ginecológica.

Na próxima seção serão apresentados: o protocolo experimental, a base de dados, uma descrição dos algoritmos de AM utilizados e os experimentos

## 3. METODOLOGIA CIENTÍFICA

Nesta seção, é apresentada a metodologia de pesquisa adotada, uma visão geral dos algoritmos AM utilizados, o conjunto de dados empregado e as etapas de pré-processamento aplicadas.

### 3.1. Dataset

Nesta pesquisa, foi utilizado o conjunto de dados denominado *Biomedical Text Publication Classification* disponibilizado na plataforma Kaggle, que possui as seguintes informações:

- **index:** Atua como o identificador único de cada linha, e serve como chave primária que permite referenciar uma entrada específica sem ambiguidade.

- **Disease:** Campo categórico que contém a identificação da doença como câncer de tireoide, pulmão ou cólon, representa a variável alvo que se deseja prever a partir de um texto descritivo.

- **Text:** Campo textual descritivo que indica o diagnóstico do câncer, é o recurso principal para classificar ou identificar a doença correspondente na coluna “Disease”.

A escolha do dataset *Biomedical Text Publication Classification* se justifica por conter a estrutura, qualidade de descrições clínicas relevantes para classificação de tipos cancerígenos a partir de dados textuais, conforme ilustra a Figura 1.

**Figura 1:** Dataset Biomedical text publication classification.

```
df.head()
```

| | Index | Disease | Text |
|---|---|---|---|
| 0 | 0 | Thyroid_Cancer | Thyroid surgery in children in a single insti... |
| 1 | 1 | Thyroid_Cancer | " The adopted strategy was the same as that us... |
| 2 | 2 | Thyroid_Cancer | coronary arterybypass grafting thrombosis ï¬^b... |
| 3 | 3 | Thyroid_Cancer | Solitary plasmacytoma SP of the skull is an u... |
| 4 | 4 | Thyroid_Cancer | This study aimed to investigate serum matrix ... |

Fonte: Próprio Autores, 2026

A seguir será apresentada como as técnicas de Processamento da Linguagem Natural foram utilizadas, aplicando-se o *framework* Sentence-BERT.

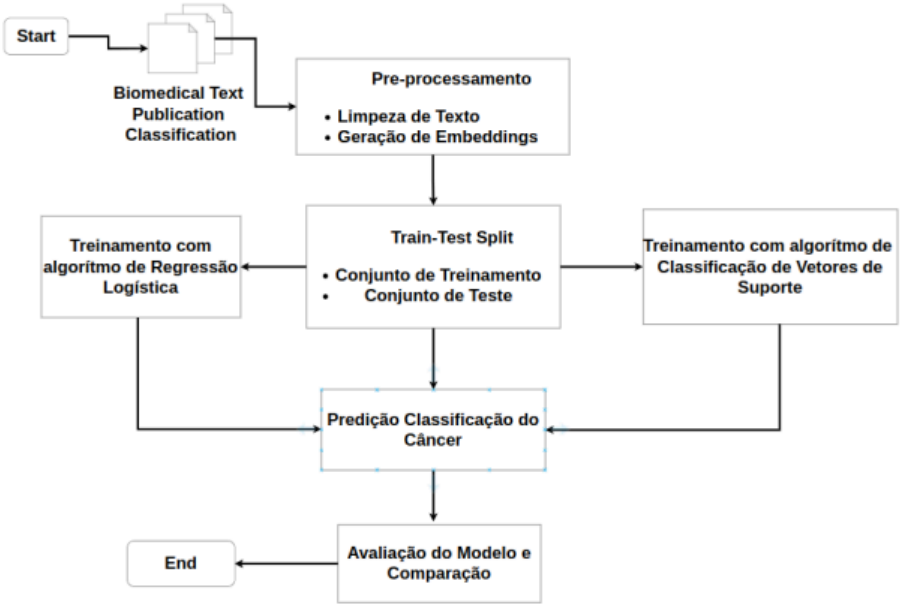
### 3.2. Engenharia de Atributos

A etapa de engenharia de atributos é crucial para transformar os dados brutos em um formato adequado e otimizado para os modelos de AM. No contexto da classificação de textos biomédicos, esta fase envolve a extração e representação de características significativas a partir do campo textual.

O processo de classificação de textos biomédicos começa com a aquisição e pré-processamento de dados. Essa etapa inicial é crucial e inclui a limpeza textual, garantindo que os dados estejam prontos para as fases seguintes.

Após a limpeza dos dados, o *Sentence-BERT* foi utilizado para gerar *embeddings* (vetores) densos, otimizando comparação por similaridade de cosseno e carga computacional [Reimers and Gurevych, 2019]. Esses vetores alimentam os modelos de AM para predição de câncer, conforme ilustra a Figura 2.

**Figura 2:** Diagrama do pipeline de classificação de textos biomédicos




**Fonte:** Próprio Autores, 2026

Após o pré-processamento, os dados foram divididos em conjuntos de 80% para treino e 20% para teste, foi utilizada uma divisão estratificada para lidar com o desbalanceamento de classes, preservando as proporções e garantindo uma avaliação mais eficaz dos modelos.

### 3.3. Modelos Utilizados

Esta seção detalha os modelos de aprendizado de máquina escolhidos para a classificação de textos biomédicos (Classificação de Vetores de Suporte e Regressão Logística). A escolha dos algoritmos foi baseada na sua adequação ao processamento de linguagem natural e na capacidade de gerenciar dados textuais complexos e volumosos.

#### 3.3.1. Classificação de Vetores de Suporte

A Classificação de Vetores de Suporte (CVS) é uma técnica de aprendizado de máquina amplamente utilizada para classificar dados em categorias distintas. De acordo com Cervantes et al. [2023], as CVS são um dos algoritmos de classificação e regressão mais poderosos e robustos em diversos campos de aplicação e classificação binária, semelhante à regressão logística.

O modelo CVS, baseado em funções de *kernel*, é especialmente eficaz para tarefas de classificação em dados vetoriais de alta dimensionalidade, como os gerados a partir de textos clínicos. Sua capacidade de encontrar fronteiras de decisão não lineares e maximizar a separação entre as classes torna particularmente adequado para o dataset *Biomedical Text Publication Classification* [Cervantes et al., 2023].

### 3.3.2. Regressão Logística

A Regressão Logística é comumente utilizada para dados categóricos, realizando classificações de novos pontos de dados. Sua principal vantagem reside na simplicidade de execução e interpretação, o que a torna um algoritmo ideal para esta pesquisa, especialmente ao lidar com as representações vetoriais dos textos biomédicos geradas a partir do nosso conjunto de dados [Zabor et al., 2021].

No campo de AM, a Regressão Logística destaca-se pela sua eficiência e ampla aplicação, inclusive em contextos médicos [Miranda et al., 2021]. Especificamente na pesquisa oncológica, este algoritmo tem sido essencial para analisar a relação entre diversos fatores de risco e o desenvolvimento do câncer [Kumar and Gota, 2023].

### 3.4. Métricas Adotadas

Para avaliar os modelos de classificação, foram utilizadas métricas conhecidas na literatura de AM, que possibilitam medir tanto a acurácia quanto a qualidade das previsões, especialmente em contextos com desequilíbrio entre classes.

- **Acurácia:** mede a proporção de previsões corretas em relação ao total de amostras avaliadas;
- **Precisão:** indica a proporção de verdadeiros positivos entre todas as previsões positivas realizadas pelo modelo;
- **Recall (ou Sensibilidade):** representa a proporção de verdadeiros positivos identificados entre os casos positivos reais;
- **F1-Score:** é a média harmônica entre precisão e recall, oferece uma medida balanceada entre esses dois aspectos.

A seguir é apresentado o experimento realizado, assim como a comparação entre os modelos.

## 4. EXPERIMENTO

Após o treinamento e otimização dos classificadores, a etapa final da metodologia consistiu na comparação do desempenho dos modelos investigados (Classificação de Vetores de Suporte e Regressão Logística). Esta análise comparativa é essencial para identificar qual abordagem demonstra maior eficácia na classificação de textos biomédicos relacionados a diferentes tipos de câncer.

A comparação foi realizada com base nas métricas de desempenho definidas na Seção 3.4. Para cada modelo, esses indicadores foram apresentados a partir do conjunto de testes, garantindo uma avaliação imparcial sobre dados não vistos durante o treinamento.

A Figura 3 mostra o relatório de classificação do modelo CVS, detalhando *precision*, *recall*, *f1-score* e *support* para cada classe. O modelo teve alto desempenho em todas as categorias: *Colon Cancer*, *Lung Cancer* e *Thyroid Cancer*. Destaca-se a classe *Lung Cancer*, que obteve pontuação de 100% em todas as métricas, indicando classificação perfeita.

**Figura 3:** Relatório de classificação do CVS

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Colon_Cancer | 0.89 | 0.97 | 0.92 | 517 |
| Lung_Cancer | 1.00 | 1.00 | 1.00 | 407 |
| Thyroid_Cancer | 0.97 | 0.89 | 0.93 | 590 |
| accuracy | | | 0.95 | 1514 |
| macro avg | 0.95 | 0.95 | 0.95 | 1514 |
| weighted avg | 0.95 | 0.95 | 0.95 | 1514 |

**Fonte:** Próprio Autores, 2026

O modelo CVS atingiu uma acurácia de 0,95 em 1514 amostras, mostrando alta precisão. As médias macro e ponderada também foram 0,95, indicando equilíbrio entre as classes. Isso demonstra que o modelo lida bem com variações na quantidade de exemplos por categoria. Portanto, o modelo é robusto e eficaz na classificação dos três tipos de câncer.

**Figura 4:** Relatório de classificação do algoritmo de Regressão Logística

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Colon_Cancer | 0.78 | 0.82 | 0.80 | 517 |
| Lung_Cancer | 0.96 | 0.96 | 0.96 | 407 |
| Thyroid_Cancer | 0.82 | 0.78 | 0.80 | 590 |
| accuracy | | | 0.84 | 1514 |
| macro avg | 0.85 | 0.85 | 0.85 | 1514 |
| weighted avg | 0.84 | 0.84 | 0.84 | 1514 |

**Fonte:** Próprio Autores, 2026

A Tabela 1 apresenta os resultados comparativos entre os modelos de CVS com kernel RBF e Regressão Logística. Observa-se que o CVS superou a Regressão Logística em todas as métricas avaliadas, com destaque para a classe *Lung Cancer*, na qual atingiu um ótimo desempenho, cujas métricas têm score igual a 1,00. A acurácia geral obtida pelo CVS foi de 95%, enquanto a Regressão Logística alcançou 84%.

**Tabela 1:** Comparação de desempenho entre os modelos CVS (RBF) e Regressão Logística

| Classe | Métrica | CVS (RBF) | Regressão Logística |
|---|---|---|---|
| Colon_Cancer | Precisão | 0.89 | 0.78 |
| | Revocação | 0.97 | 0.82 |
| | F1-Score | 0.92 | 0.80 |
| Lung_Cancer | Precisão | 1.00 | 0.96 |
| | Revocação | 1.00 | 0.96 |
| | F1-Score | 1.00 | 0.96 |
| Thyroid_Cancer | Precisão | 0.97 | 0.82 |
| | Revocação | 0.89 | 0.78 |
| | F1-Score | 0.93 | 0.80 |

| | | | |
|---|---|---|---|
| Geral | Acurácia | 0.95 | 0.84 |
| | Média Macro | 0.95 | 0.85 |
| | Média Ponderada | 0.95 | 0.84 |

**Fonte:** Próprio Autores, 2026

Esses resultados indicam que a utilização de um modelo não linear, como o CVS com kernel RBF, se mostra mais eficaz na tarefa de classificação dos *embeddings* gerados. O desempenho superior do CVS sugere uma maior capacidade de modelar fronteiras de decisão mais complexas, aspecto relevante ao lidar com dados textuais de alta dimensionalidade e variabilidade semântica.

## 5. CONCLUSÃO

Este estudo demonstrou a viabilidade da aplicação de modelos de AM na classificação de textos biomédicos voltados ao diagnóstico de câncer de pulmão, cólon e tireoide. Por meio da vetorização dos dados textuais utilizando *embeddings* do *Sentence-BERT*, foi possível estruturar uma representação semântica densa e contextualizada dos textos clínicos, o que potencializou o desempenho dos classificadores empregados.

A comparação entre os modelos de Classificação por Vetores de Suporte (CVS) com kernel RBF e Regressão Logística evidenciou a superioridade do primeiro em todas as métricas avaliadas, com destaque para a acurácia global de 95%. Os resultados obtidos ressaltam a capacidade dos algoritmos de aprendizado supervisionado em lidar com tarefas de comparação em ambientes clínicos complexos, mesmo utilizando modelos rasos.

Em síntese, os resultados reforçam o potencial da IA no diagnóstico precoce de câncer a partir da análise automatizada de textos clínicos. Como continuidade, propõe-se explorar modelos mais avançados, como redes neurais profundas e *transformers* biomédicos, além de expandir o estudo para outras doenças e tipos de dados.

## REFERÊNCIAS BIBLIOGRÁFICAS


Atoyebi and Uwazie. Comparison of multinodal naive bayes (mnb), gaussian naive bayes (gnb) and random forest (rf) algorithm in malaria disease diagnosis. *International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*, 2024.

Bray, Laversanne, Sung, Ferlay, Siegel, Soerjomataram, and Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. 2022.

DW. Chen, BH. Lang, DSA. McLeod, K. Newbold, and MR. Haymart. Thyroid cancer. 2023.

D. Crosby, S. Balata, K. Brindle, L. Coussons, C. Dive, and M. Emberton. Early detection of cancer. *Nature Reviews Cancer*, 2022.

Emily C. Zabor, Chandana A. Reddy, Rahul D. Tendulkar, and Sujata Patil. Logistic regression in clinical studies. *International Journal of Radiation Oncology, Biology, Physics*, Volume 112, Issue 2, 271–277, 2021.

Eka Mardiana, Fajri M. Bhatti, Medianta Surya, and Charles Bernardo. Intelligent computational model for early heart disease prediction

using logistic regression and stochastic gradient descent (a preliminary study). 2021.

H. Chehade, A. Abdallah, N. Maroun, and J.M. et al. Lung and colon cancer classification using medical imaging: a feature engineering approach. In *Phys Eng Sci Med*, 2022.

Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodriguez-Mazahua, and Asdrubal Lopez. A comprehensive survey on support vector machine classification. *Applications, challenges and trends*. *Neurocomputing*, 2023.

K. Nair, S. Kamath, N. Chowattukunnel, and S. Krishnamurthi. Preoperative strategies for locally advanced colon cancer. 2024.

N. Pashayan and P. Pharoah. The challenge of early detection in cancer. 2020.

N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using siamese bert-networks. *Conference on Empirical Methods in Natural Language Processing*, 2019.

S. K. Denath, G. Kaur S. Malik, S. Bagchi, A. M. Soomro, and A. Naeem. 10th ieee uttar pradesh section international conference on electrical, electronics and computer engineering (upcon). 2023.

S. Geeitha, P. Renuka, K. Poongothai, S. Ananth, K. Sinduja, and K. R. Sangeetha. 10th ieee uttar pradesh section international conference on electrical, electronics and computer engineering (upcon). 2023.

Sharath Kumar and Vikram Gota. Cancer research: statistics, and treatment.

S. Zhang, M. Zhu, H. Zhang, H. Fan Liu, X. Zhang, and F. Yu. Preoperative strategies for locally advanced colon cancer. 2024.