Medical-GAT: Cancer Document Classification Leveraging Graph-Based Residual Network for Scenarios with Limited Data

Elias Hossain*‡, Tasfia Nuzhat†, Shamsul Masum§, Shahram Rahimi¶, Noorbakhsh Amiri Golilarz¶

*Department of Computer Science & Engineering,
Mississippi State University, Mississippi State, MS, USA

†Department of Information Technology
Universiti Tenaga Nasional, Malaysia

§School of Electrical and Mechanical Engineering,
University of Portsmouth, UK

¶Department of Computer Science,
The University of Alabama, Tuscaloosa, AL, USA

‡Corresponding author: mh3511@msstate.edu

arXiv:2410.15198v4 [cs.AI] 8 Apr 2025

Abstract—Accurate classification of cancer-related medical abstracts is crucial for healthcare management and research. However, obtaining large, labeled datasets in the medical domain is challenging due to privacy concerns and the complexity of detailed clinical data. This scarcity of annotated data impedes the development of effective machine learning models for cancer document classification. To address this challenge, we present a curated dataset of 1,874 biomedical abstracts, categorized into thyroid cancer, colon cancer, lung cancer, and generic topics. Our research focuses on leveraging this dataset to improve classification performance, particularly in data-scarce scenarios. We introduce a Residual Graph Attention Network (R-GAT) with multiple graph attention layers that effectively capture the semantic information and structural relationships within lengthy cancer-related documents. Our R-GAT model is compared with various techniques, including transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) and the Robustly Optimized BERT Pretraining Approach (RoBERTa), as well as domain-specific transformer models, i.e., Bidirectional Encoder Representations from Transformers for Biomedical Text (BioBERT) and Bio+ClinicalBERT. We also evaluated deep learning-based models (e.g., CNNs, LSTMs) and traditional machine learning models (e.g., Logistic Regression, SVM) for a comprehensive comparison. Traditional models were included as they can sometimes capture specific patterns better and offer useful performance baselines. Additionally, we explore ensemble approaches that combine multiple deep learning models to further enhance classification performance. Various feature extraction methods are assessed, including Term Frequency-Inverse Document Frequency (TF-IDF) with unigrams and bigrams, Word2Vec, and tokenizers from BERT and RoBERTa. The R-GAT model outperforms other techniques, achieving precision, recall, and F1 scores of 0.99, 0.97, and 0.98 for thyroid cancer; 0.96, 0.94, and 0.95 for colon cancer; 0.96, 0.99, and 0.97 for lung cancer; and 0.95, 0.96, and 0.95 for generic topics, respectively. Furthermore, the R-GAT model demonstrates better generalizability compared to both machine learning and transformer-based models. The dataset is available on GitHub[1] and is publicly accessible for research purposes.

Keywords—**Biomedical Text Classification, Residual Graph Attention Network (R-GAT), Cancer Abstracts, Transformer Models, and Bidirectional Encoder Representations from Transformers (BERT).**

---

[1] https://github.com/eliashossain001/MedicalAbstracts

I. INTRODUCTION

Cancer is a major global health issue affecting millions of people annually, with thyroid, colon, and lung cancers being the most prevalent [1]. Patients often experience symptoms related to the affected organs, such as neck lumps, changes in bowel habits, or a chronic cough. Recent data shows a significant increase in cancer incidence, with 238,000 reported cases of thyroid cancer in 2016 rising to 567,000 by 2018 [1]. The World Health Organization (WHO) reported in 2020 that over 1.9 million people were affected by colorectal cancer, resulting in approximately 930,000 deaths [2]. Lung cancer data surpasses other types, with 2,094,000 cases reported in 2018 [3].

With the increasing prevalence of cancer, efficient methods for documenting and evaluating medical information have become crucial. Electronic Health Records (EHRs) have emerged as essential tools for handling large volumes of medical data and identifying patterns, which medical professionals utilize to provide optimal patient care. Nevertheless, EHRs often lack structure, contain incomplete information, and have inconsistent formatting, making them difficult to use in cancer research. Furthermore, patient privacy concerns restrict access to these records, and organizations frequently withhold data for research purposes, hindering efforts to obtain the extensive datasets necessary for advancing cancer research.

Given these constraints, there is an urgent need for alternative sources of data to support cancer research. Medical abstracts, which provide insights from various publications, offer a valuable resource. These widely available abstracts present diverse perspectives on different cancer types, including thyroid, colon, and lung cancers. To address the limited availability of data, we created a dataset comprising 1,874 biomedical abstracts from PubMed, categorized into thyroid cancer, colon cancer, lung cancer, and generic topics (not specifically related to cancer). This dataset aims to fill the

gap created by EHR limitations and support automated cancer classification systems. Our objective is to enhance healthcare systems by providing a well-defined dataset that facilitates research and improves the accuracy of cancer document classification, particularly in situations where large, labeled datasets are scarce.

Utilizing this dataset, we implemented the R-GAT to classify cancer types using graph-based approaches. This represents the first labeled and well-defined dataset that includes thyroid, colon, and lung cancers in biomedical abstracts. We chose the R-GAT model because it represents connections between entities using graphs, unlike traditional machine learning or deep learning methods that treat text as linear sequences. Its attention mechanism emphasizes crucial information, enabling the model to gather structural context and enhance performance on text analysis tasks. R-GAT is particularly advantageous in scenarios with limited data, as its ability to leverage structural links improves classification performance even with smaller datasets. Incorporating R-GAT into our dataset contributes significant insights to cancer research and enhances the effectiveness of cancer treatment. The key contributions of this research are outlined as follows:

- A specialized dataset of $\approx$ 1,874 cancer-related publications has been publicly released, categorized into thyroid, colon, and lung cancers, serving as a key resource for targeted healthcare research.
- An enhanced residual graph attention network (R-GAT) has been introduced to improve feature extraction and reduce information loss in complex cancer records.
- Various ML and DL models, including ensemble and transformer-based approaches, have been assessed, with their limitations and areas for improvement identified.

The rest of the article is structured into four sections. In Section II, we provide a literature review. Section III describes the R-GAT model architecture and methodology. Section IV presents the study's findings, including analyses and insights. Finally, Section V concludes the manuscript.

II. LITERATURE REVIEW

This section provides a literature review of cutting-edge approaches used to classify cancers from various text-based data, alongside other solutions that support healthcare research.

Nguyen et al. [4] created a summarization technique utilizing Dutch radiology report information. They developed a hybrid model that integrates an encoder-decoder language model with an attention mechanism and a Breast Imaging-Reporting and Data System (BI-RADS) score classifier. For Dutch reports, the model attained a ROUGE-L F1-score of 51.5%. Another classifier achieved 83.3% accuracy, outperforming the language model's 79.1% in BI-RADS classification. Despite its strong performance, the model was infeasible for clinical application.

Tang et al. [5] utilized attention-based deep learning models incorporating BERT to classify progress notes and extract key terms. Their BERT model, fine-tuned with an attention layer, reached 97.6% accuracy. This highlights the efficacy of hybrid and attention-based models in improving clinical data classification and summarization. In another study, Hepsağ et al. [6] curated a dataset comprising 62 mammography records from Turkish patients, which a subject matter expert manually classified for breast cancer diagnosis. The study also utilized state-of-the-art models, such as pre-trained BERT and Distil-BERT, in addition to an ensemble voting method. However, findings from this study indicated that BERT produced the best results on Turkish radiology reports, achieving a 91% F1-score.

Ai et al. [7] introduced the Edge-Enhanced Minimum-Margin Graph Attention Network (EMGAN) to address short-text classification challenges, focusing on feature sparsity and weak contextual relationships through a Heterogeneous Information Graph (HIG) and a Minimum Margin Graph Attention Network (MMGAN). Another study by Wei et al. [8] proposed a graph convolutional attention network (GCAN) for enhancing remaining useful life (RUL) predictions in engineered systems, utilizing temporal convolution-aware nested residual connections and an attention mechanism to prioritize relevant features.

Song et al. [9] developed Graph Sequence Pre Training with Transformer (GSPT) for text-attributed graphs (TAGs), leveraging large language models (LLMs) to unify feature spaces across diverse graphs, emphasizing feature reconstruction to boost knowledge transferability and performance in node classification and link prediction tasks. Meanwhile, Rao et al. [10] presented a Multi-layer Residual Attention Network (MRAN) for knowledge graph completion (KGC), integrating categorical information with textual descriptions to enhance entity embeddings and improve contextual understanding between entities and relationships.

These studies explored various cancers and clinical note classification using deep learning and transformer-based approaches, with a growing emphasis on capturing semantic information through graph-based methods. However, multicancer types such as thyroid, colon, and lung remain understudied, with no publicly available datasets for their classification. Biomedical abstract-based text classification has also received limited attention, presenting an opportunity for developing solutions to support healthcare professionals in disease diagnosis and treatment planning. While graph-based techniques have been widely explored, the R-GAT model has not been previously utilized in this context. To address these gaps, we created a dataset of biomedical abstracts categorizing three cancer types and developed a pipeline effective for limited data samples. Our research contributes to the field by providing a foundation for future investigations and supporting healthcare research through innovative solutions. Our findings will help the research community discover creative solutions to healthcare issues and progress medical research.

III. METHODOLOGY

This section outlines the classification of medical documents related to thyroid cancer, colon cancer, lung cancer, and generic topics divided into four subsequent phases. In the first phase, medical abstracts related to these cancers were collected from the PubMed database. The second phase
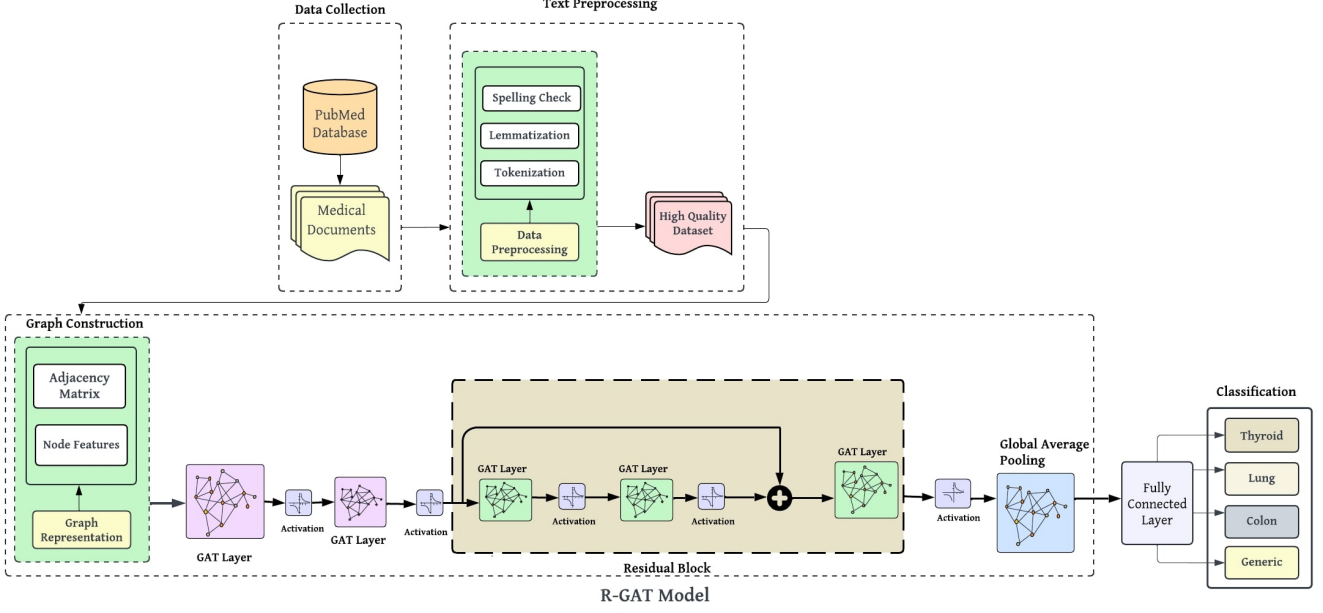
Fig. 1: Overview of the methodology for classifying medical documents. The process consists of four phases: (1) Data Collection from PubMed, (2) Text Preprocessing, (3) R-GAT Model Architecture, and (4) Classification, yielding outputs for thyroid cancer, colon cancer, lung cancer, and generic topics.

involved text preprocessing, where the raw data underwent several techniques to produce a high-quality dataset, including tokenization, spelling checks, and text normalization, such as lemmatization.

The third phase included the R-GAT model, which unfolds across four distinct steps. Initially, a graph was constructed to represent node features and the adjacency matrix in the first step. In the second step, this graph was processed through two Graph Attention Network (GAT) layers before entering to the Residual Block. The third step introduced a Residual Block, a crucial component comprising three GAT layers, each with its activation function, as depicted in Fig. 1. After that, in the forth step, a Global Average Pooling layer aggregated the features.

Ultimately, in the final phase of our workflow, a fully connected layer was used for classification. Elaborately, the first two steps are described in IV-A1 and IV-A2, while the following subsections will detail phase 3, the R-GAT model, outlining its functionalities and mathematical perspective.

A. Graph Construction

The first step is to create a graph that visually represents the cancer documents and their interconnections. In the node feature representation, each medical text document is assigned a feature vector to reflect its content. The feature matrix, denoted as $X \epsilon R^{N \times F}$, represents the document features, with N representing the number of nodes in the graph and F is the number of features per node. An adjacency matrix $A \epsilon R^{N \times N}$ is utilized to represent the relationships between documents, with A_{ij} denoting the strength of the connection between document i and document j. The edge weights can be

acquired through training or determined using domain-specific knowledge.

B. Graph Layers with Attention Mechanism

A key component of our research is the use of the Graph Attention Mechanism, which computes attention scores for surrounding documents. The attention scores for a certain node i are calculated as follows: We use Equation 1 to apply a Leaky Rectified Linear Unit (LeakyReLU) activation function to the concatenation of linear transformations of the feature vectors of nodes i and j, where nodes i and j feature representations are denoted by h_i and h_j, respectively; a is a learnable attention weight vector and W is a learnable weight matrix. The $(\top)$ represents that vector a is transposed before performing the dot product to ensure appropriate dimension alignment for matrix multiplication. In addition, a double vertical bar sign denotes concatenation.

$$e_{ij} = \text{LeakyReLU}\left(a^\top \cdot [W \cdot h_i \| W \cdot h_j]\right) \qquad (1)$$

To obtain attention coefficients, we normalize the attention scores using the SoftMax function in Equation 2. In this regard, $\mathcal{N}(i)$ represents the collection of surrounding nodes of node i.

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(e_{ik}\right)} \qquad (2)$$

C. Residual Blocks with Graph Attention Layers

To improve our model and capture complex interactions, we use a Residual Block that combines multiple GAT layers, each

followed by an activation function. The input h_i in Equation 3 is the result of two previous GAT layers before the Residual Block. Equations 3–6 mathematically describe the structure of the Block, where A represents the adjacency matrix. To be more precise, $h'_{i,1}$ and $h'_{i,2}$ represent the outputs of the first and second GAT layers, respectively; $h'_{i,3}$ is the result of adding the residual connection shown in (Equation 5). Finally, h'_i is the outcome of processing through the third GAT layer of the Residual Block.

$$h'_{i,1} = \text{GATLayer}^1(h_i, A) \tag{3}$$

$$h'_{i,2} = \text{GATLayer}^2(h'_{i,1}, A) \tag{4}$$

$$h'_{i,3} = h_i + h'_{i,2}(\text{Residual Connection}) \tag{5}$$

$$h'_i = \text{GATLayer}^3(h'_{i,3}, A) \tag{6}$$

The use of attention coefficients α_{ij} helps consolidate information from adjacent nodes, resulting in an improved representation for each particular node i.

$$h'_i = \sum_{j \in \mathcal{N}(i)} \alpha_{ij} \cdot W \cdot h_j \tag{7}$$

Simultaneously, to capture a wide range of patterns contained in the data, we employ several K independent attention heads. Each attention head, K, which operates independently, captures different aspects of the interactions between nodes in the network. These different attention heads increase the model's ability to focus on diverse patterns at the same time. Also, the non-linear activation function, denoted as σ further contributes to this process by introducing non-linearity, allowing the model to learn intricate relationships within data.

$$\vec{h}'_i = \|_{k=1}^{K} \sigma \left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k \vec{h}_j \right) \tag{8}$$

D. Global Average Pooling Layer

The final node representation is created by concatenating or averaging the results. Our network uses global average pooling, which computes the mean feature vector to represent the entire graph. Finally, the average feature vector passes through a dropout layer and then a fully connected layer, followed by a SoftMax activation function to forecast the cancer document classes. Algorithm 1 is the pseudocode of the R-GAT model, which illustrates the major steps and processes in the design.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

This section presents and analyzes the results of our experiments. We provide a detailed overview of data collection and preprocessing, followed by an in-depth explanation of the findings from several machine learning and deep learning models. We also evaluate the performance of the R-GAT model, conduct inference tests, compare our findings with existing studies, and discuss the limitations of our study.

Algorithm 1 R-GAT Model

Require: inputs
1: initialize node_feature_matrix, adjacency_matrix
2: **for** each node in node_feature_matrix **do**
3:     calculate attention_scores using Equation 1
4:     normalize attention_scores using Equation 2
5: **end for**
6: **for** each Residual Block **do**
7:     **for** each GAT Layer **do**
8:         update node feature representations using Equations 3-6
9:     **end for**
10: **end for**
11: aggregate node features using global average pooling
12: pass aggregated features through fully connected layers
13: apply Softmax activation function for classification
14: **return** classification_output

A. Experimental Setup

1) Dataset Collection: We collected 1,874 medical abstracts on thyroid, colon, and lung cancers, as well as general topics, using the open-source "Entrezpy" [11] Python library, which provides access to PubMed, a major NCBI database containing a wealth of biological literature. The data collection took place between January and March 2024. We retrieved abstracts using specific search terms and keywords such as "thyroid cancer", "colon cancer", "lung cancer", and other general topics, prioritizing studies from the past five years to ensure the dataset's relevance. Inclusion criteria were based on the abstracts' direct relevance to the specified cancer types. We excluded irrelevant, duplicate, or non-English abstracts, as well as those with insufficient detail. After retrieval and by subject matter experts, the abstracts were manually reviewed and categorized into four groups: thyroid, colon, lung, and generic. The resulting dataset consists of abstracts with an average length of 200 words, encompassing a wide range of recent research articles from various medical journals.

2) Data Preparation: Before analyzing the data, we cleaned it properly to ensure quality, so that it could be utilized for various machine learning or deep learning models. Cleanup steps include detection of missing features, tokenization, lemmatization, handling class imbalance issues, removal of redundant words, and vectorization.

Firstly, missing attributes are often present in the dataset because the data is collected from various sources and not all attribute information is available at the time of collection, leading to missing values. Hence, any missing values should be identified and addressed before using the dataset in a machine learning model. Then, tokenization helps the model capture meaningful information effectively as it breaks down a longer paragraph or sentence into smaller parts. We used the Natural Language Toolkit (NLTK) library for tokenization [12].

Next, we turn to the process of lemmatization, which is the reduction of words to their most basic or original form. By considering distinct versions of the same word as interchangeable, this technique helps models understand the links between

words. Furthermore, our dataset had class imbalance problems, which meant that the data length varied for each class. To successfully address the class imbalance issues, we employed the Synthetic Minority Oversampling Technique (SMOTE) [13], which creates synthetic data samples. Furthermore, we filtered out less significant words from our dataset, as not all parts of a sentence contribute significantly to model training. Lastly, we converted the text data into numerical vectors using a variety of vectorization techniques, including Term Frequency-Inverse Document Frequency (TF-IDF) [14], Word to Vector (Word2Vec) [15], and Bidirectional Encoder Representations from Transformers (BERT) [16] tokenizer.

3) Assessment Metrics: We employed several methodologies to examine the performance of different models, including a confusion matrix, precision, recall, and F1-score. These indicators are extremely useful in understanding actual prediction and how models recognize all relevant instances, allowing us to receive a balanced assessment of model performance. We analyze multiple models to help everyone understand how different models interact while dealing with limited medical abstracts. Without adequate validation, we may not acquire full information about a model. Hence, these indicators allowed us to understand better the effectiveness of each model utilized in this study. We provided a detailed explanation of the evaluation process in IV-B3.

B. Result and Discussion

This section presents the findings from our experiments and the conclusions drawn from the various models. The classification reports of various machine learning models is explained in IV-B1; the performance of deep learning-based models is discussed in IV-B2; the assessment of our proposed R-GAT model is illustrated in IV-B3; in IV-B4, we demonstrate how the R-GAT model performs when observing unseen data through an inference test; in IV-B5, we compare the R-GAT with existing studies; and finally, in IV-B6, we highlight the limitations of this research.

1) Classification Report of Traditional Machine Learning Models: Table I presents a classification report for several machine learning models and feature extraction strategies across four classes: thyroid, colon, lung, and generic. The table compares various models, including Decision Tree [17], Random Forest [18], K-Nearest Neighbors [19], Multinomial Naïve Bayes [20], Gradient Boosting [21], Adaptive Boosting [22], Support Vector Machine [23], Extreme Gradient Boosting (XGBoost) [24], and Logistic Regression [25], based on feature extraction approaches such as TF-IDF (Unigram and Bigram) [26] and Word2Vec [15]. Taking a closer look at Table **??**, we can observe that Random Forest and Logistic Regression performed well with TF-IDF (Unigram) for classifying the thyroid, colon, lung, and generic classes, as demonstrated by the performance metrics. It is also worth noting that TF-IDF (Unigram) proved to be suitable across various feature extraction approaches.

Turning to the TF-IDF (Bigram) and Word2Vec techniques, we notice that the K-Nearest Neighbors model performed

inconsistently. Furthermore, it is also obvious that using Word2Vec features significantly reduces the performance of Multinomial Naïve Bayes, whereas other models, e.g., Decision Trees and Gradient Boosting models, produce mixed results.

2) Performance Analysis of Deep Learning Models: Table II presents an analysis of the performance of various deep learning models on this data set, including Convolutional Neural Networks (CNN) [27], Recurrent Neural Networks (RNN) [28], Long Short-Term Memory Networks (LSTM) [29], Bidirectional Long Short-Term Memory Networks (BI-LSTM) [30], Stacked Long Short-Term Memory Networks (Stacked LSTM) [31], Stacked Bidirectional Long Short-Term Memory Networks (Stacked B-LSTM) [32], Hybrid Ensemble Models [33], Bidirectional Encoder Representations from Transformers (BERT) [16], Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) [34], Robustly optimized BERT approach (RoBERTa) [35], and Clinical BERT for Biomedical Text (Bio+ClinicalBERT) [36], using advanced feature extraction techniques.

When it comes to Keras embedding-based feature extraction, models such as CNN, RNN, LSTM, GRU, BI-LSTM, stacked LSTM, and stacked Bi-LSTM exhibit balanced performance, as demonstrated by evaluation metrics (i.e., P, R, and F1) across four distinct classes. It is noticeable that Stacked LSTM performs the worst in the case of the BERT-based tokenizer, although CNN and hybrid ensemble models outperform other models by a significant margin.

Furthermore, comparing the performance of domain-specific models such as BioBERT, Bio+ClinicalBERT, RoBERTa, and BERT, it was found that BioBERT performed best, while the BERT model turned out to be overfitted, indicating that it was unable to capture semantic information in cases where data samples are not comprehensive. To further investigate the efficacy, we applied zero-shot training on BERT and RoBERTa, and we found that they did not capture information accurately and lacked generalization. Upon closer inspection, the R-GAT model performs better across all classes and shows no signs of overfitting. This is because the model makes use of a graph-based feature representation technique, which enables it to better capture more semantic information without experiencing overfitting.

3) Model Assessment and Justification : In order to have a better understanding of the efficacy of various machine learning or deep learning-based models, we can refer to the confusion matrix as a performance evaluation technique. To put it simply, true positive, false positive, true negative, and false negative are the four values that comprise its foundation. In Fig. 2, we show the confusion matrix of the R-GAT model since we discovered its balanced performance, as evident in Table **??**. Examining Fig. 2, it is evident that the R-GAT model produces good results when correctly identifying the four distinct classes (thyroid, colon, lung, and generic) at 94%, 96%, and 97%, respectively. We may see some misclassifications even if the majority of the records are correctly classified. For instance, 3% of instances of lung cancer are mistakenly diagnosed as thyroid cancer, whereas 4% of cases of colon

TABLE I: Classification Results for Various ML Algorithms Using Different Feature Extraction Methods. P: Precision, R: Recall, F1: F1 Score.

| Model	Feature Extraction	Thyroid			Colon			Lung			Generic		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Decision Tree	TF-IDF (Unigram)	0.99	0.98	0.98	0.94	0.96	0.95	0.94	0.97	0.95	0.97	0.93	0.95
	TF-IDF (Bigram)	0.97	0.99	0.98	0.95	0.94	0.95	0.96	0.96	0.96	0.91	0.90	0.91
	Word2Vec	0.53	0.58	0.55	0.96	0.93	0.94	0.66	0.67	0.67	0.85	0.82	0.83
Random Forest	TF-IDF (Unigram)	0.99	1.00	0.99	1.00	0.98	0.99	0.99	1.00	0.99	0.99	0.99	0.99
	TF-IDF (Bigram)	1.00	0.98	0.99	0.98	0.95	0.97	0.99	0.97	0.98	0.92	0.99	0.95
	Word2Vec	0.66	0.68	0.67	0.96	0.96	0.96	0.76	0.77	0.76	0.88	0.85	0.86
K-NN	TF-IDF (Unigram)	1.00	0.15	0.27	0.58	0.66	0.62	0.95	0.19	0.31	0.39	0.99	0.56
	TF-IDF (Bigram)	0.00	0.00	0.00	0.63	0.62	0.63	1.00	0.08	0.15	0.34	1.00	0.51
	Word2Vec	0.58	0.67	0.62	0.94	0.96	0.95	0.72	0.68	0.70	0.89	0.80	0.84
Multinomial NB	TF-IDF (Unigram)	0.95	0.98	0.96	0.90	0.93	0.91	0.92	0.91	0.91	0.98	0.92	0.91
	TF-IDF (Bigram)	0.97	1.00	0.98	0.94	0.96	0.95	0.98	0.96	0.97	0.99	0.96	0.97
	Word2Vec	0.00	0.00	0.00	0.47	0.97	0.63	0.00	0.00	0.00	0.33	0.51	0.40
Gradient Boosting	TF-IDF (Unigram)	0.99	0.99	0.99	0.99	0.97	0.98	0.96	0.98	0.97	0.97	0.97	0.97
	TF-IDF (Bigram)	1.00	0.98	0.99	0.98	0.91	0.94	0.98	0.95	0.96	0.86	0.97	0.91
	Word2Vec	0.58	0.60	0.59	0.96	0.94	0.95	0.65	0.65	0.65	0.86	0.83	0.84
AdaBoost	TF-IDF (Unigram)	1.00	0.69	0.82	0.99	0.95	0.97	0.95	0.99	0.97	0.73	0.96	0.83
	TF-IDF (Bigram)	1.00	0.99	0.99	0.97	0.88	0.92	0.92	0.91	0.92	0.84	0.94	0.88
	Word2Vec	0.22	0.51	0.30	0.33	0.01	0.03	0.53	0.37	0.43	0.48	0.56	0.52
SVM	TF-IDF (Unigram)	1.00	0.94	0.97	0.88	0.98	0.93	0.99	0.94	0.96	0.96	0.97	0.96
	TF-IDF (Bigram)	0.43	0.18	0.26	0.43	0.18	0.26	0.10	0.15	0.36	0.48	0.44	0.49
	Word2Vec	0.56	0.34	0.42	0.96	0.93	0.94	0.54	0.75	0.63	0.78	0.79	0.79
XGBoost	TF-IDF (Unigram)	0.98	0.99	0.98	0.97	0.95	0.96	1.00	0.96	0.98	0.94	0.99	0.97
	TF-IDF (Bigram)	0.37	0.37	0.37	0.37	0.37	0.37	0.51	0.51	0.51	0.48	0.47	0.52
	Word2Vec	0.43	0.50	0.46	0.96	0.95	0.95	0.41	0.37	0.39	0.61	0.60	0.60
Logistic Regression	TF-IDF (Unigram)	1.00	1.00	1.00	0.99	0.96	0.97	0.99	0.99	0.99	0.95	0.98	0.96
	TF-IDF (Bigram)	0.67	0.51	0.64	0.67	0.51	0.64	0.59	0.33	0.48	0.78	0.56	0.61
	Word2Vec	0.42	0.50	0.46	0.91	0.97	0.94	0.60	0.18	0.28	0.49	0.69	0.57
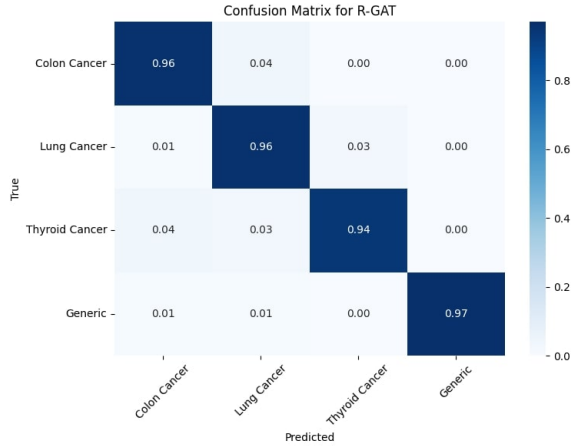
Fig. 2: Confusion Matrix Showing the Classification Performance of the R-GAT Model.

cancer are anticipated to be lung cancer. It is also incorrect to classify 4% of cases of thyroid cancer as colon cancer. Notwithstanding a few small mistakes, the model performs admirably overall.

Next, our evaluation shifted towards k-fold cross-validation, a process where we allocate some data for training and testing from the same dataset. In the case of the R-GAT model, we applied 5-fold cross-validation, and this process helped enhance both performance and generalizability.

Following that, we examine a validation loss graph to confirm the R-GAT model's performance. Looking at Fig. 3, it reveals that, in 5-fold cross-validation, all folds show a rapid decline in loss over the first few epochs, indicating quick learning. The model has been trained successfully when the validation loss steadies over all folds at roughly 10 epochs with little fluctuations. Based on the consistency of loss across folds, we can denote that the model has excellent capability to be generalized on various subsets of data. In addition, as training advances, the loss approaches zero, suggesting great performance on the validation set with no notable evidence of overfitting.

When comparing the outcomes of the various models, it is worth noting that some machine learning models, such as logistic regression, exhibited evidence of overfitting. Almost all of the traditional machine learning models were overfitted to some extent. This happens because these models tend to memorize specific patterns in training data, especially when the sample size is limited.

Deep learning ensemble models, while more resilient than classic machine learning approaches, were limited in their generalizability. Ensembles integrate numerous models to increase performance, but we observed that they fail to capture intricate relationships in the case of the limited data samples, and we believe that this approach did not carefully consider the data's underlying structure and model architecture as well. While ensembles can decrease overfitting by averaging predictions from numerous models, they frequently shortfall attention to relational structures present in more advanced models such as R-GAT.

TABLE II: Classification Results for Various Deep Learning Models Using Different Feature Extraction Methods. P: Precision, R: Recall, F1: F1 Score.

| Model	Feature Extraction	Thyroid			Colon			Lung			Generic		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
CNN	Keras Embedding	0.98	0.98	0.98	0.95	0.98	0.96	0.95	0.92	0.94	0.97	0.97	0.97
	BERT Tokenizer	0.98	0.95	0.96	0.93	0.98	0.95	0.97	0.97	0.97	1.00	0.98	0.99
RNN	Keras Embedding	0.98	0.96	0.97	0.93	0.96	0.94	0.94	0.91	0.92	0.93	0.95	0.94
	BERT Tokenizer	0.33	0.40	0.36	0.50	0.18	0.27	0.21	0.27	0.24	0.30	0.34	0.32
LSTM	Keras Embedding	0.95	0.95	0.95	0.97	0.95	0.96	0.92	0.98	0.95	0.99	0.96	0.97
	BERT Tokenizer	0.51	0.80	0.62	0.70	0.97	0.81	0.64	0.08	0.14	0.94	0.82	0.88
GRU	Keras Embedding	0.96	0.98	0.97	0.88	0.92	0.90	0.96	0.92	0.94	0.90	0.91	0.91
	BERT Tokenizer	0.85	0.47	0.61	0.76	0.80	0.78	0.56	0.85	0.67	0.99	0.92	0.95
Bi-LSTM	Keras Embedding	0.99	0.90	0.94	0.80	0.93	0.86	0.96	0.86	0.91	0.93	0.95	0.94
	BERT Tokenizer	0.93	0.89	0.91	0.96	0.94	0.95	0.89	0.94	0.92	0.97	0.98	0.97
Stacked LSTM	Keras Embedding	0.90	0.97	0.93	0.92	0.92	0.92	0.94	0.86	0.90	0.92	0.92	0.92
	BERT Tokenizer	0.00	0.00	0.00	0.00	0.00	0.00	0.24	1.00	0.38	1.00	0.01	0.02
Stacked Bi-LSTM	Keras Embedding	0.98	0.96	0.97	0.91	0.97	0.94	0.94	0.92	0.93	0.97	0.95	0.96
	BERT Tokenizer	0.91	0.91	0.91	0.94	0.94	0.94	0.95	0.88	0.91	0.90	0.97	0.93
Hybrid Ensemble	BERT Tokenizer	0.97	0.95	0.96	0.96	0.96	0.96	0.96	0.97	0.96	0.97	0.98	0.97
BioBERT	BioBERT Tokenizer	1.00	0.99	1.00	0.97	0.98	0.94	0.99	0.96	0.99	0.96	0.97	0.99
BERT	BERT Tokenizer	1.00	1.00	1.00	0.97	1.00	0.98	0.99	0.98	0.98	1.00	0.98	0.99
RoBERTa	RoBERTa Tokenizer	1.00	1.00	1.00	0.93	1.00	0.97	1.00	0.98	0.99	1.00	0.95	0.97
Bio+ClinicalBERT	Bio+ClinicalBERT Tokenizer	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	0.99	0.99
BERT (Zero-shot)	BERT Tokenizer	0.34	1.00	0.51	0.45	1.00	0.62	0.31	1.00	0.47	0.09	0.39	0.15
RoBERTa (Zero-shot)	RoBERTa Tokenizer	0.29	1.00	0.45	0.74	0.94	0.83	0.30	1.00	0.47	0.17	0.89	0.29
R-GAT	Graph Representation	0.99	0.97	0.98	0.96	0.94	0.95	0.96	0.99	0.97	0.95	0.96	0.95
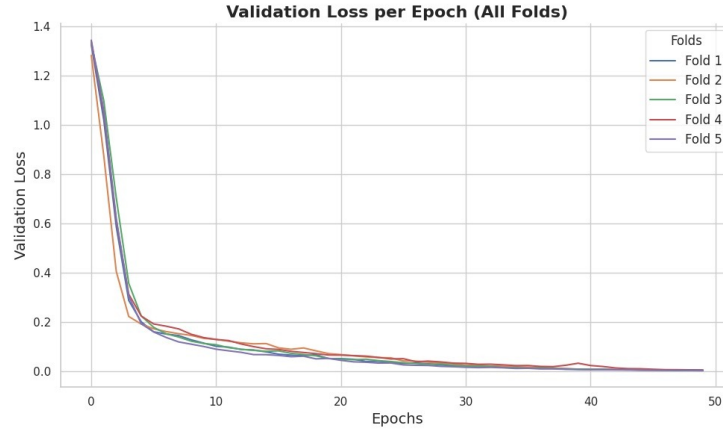
Fig. 3: Performance Evaluation of the R-GAT Model: Training and Validation Loss Using 5-Fold Cross-Validation.

Domain-specific transformer models, such as BioBERT and Bio+ClinicalBERT, are specifically designed for biomedical text and benefit from intensive pre-training on relevant collections. However, in cases with insufficient data, we found that these models overfit because of their complicated structures and huge number of parameters. This intricacy caused the models to memorize the training data rather than generalize properly, resulting in good performance on the training set but lower generalization to new, unseen data.

On the other hand, the R-GAT model performs well due to its ability to capture semantic relationships within long medical documents using the graph attention technique. This approach enables the model to focus on key areas of the data, hence boosting its generalization capabilities, particularly in data-scarce settings. The ability to properly use graph-based attention and residual connections is a major reason for its excellent performance in both the training and validation phases. This, combined with its modest loss fluctuations and low overfitting risk, makes it more generalized than other models.

4) Inference Testing: The predicted results of the R-GAT model are displayed in Fig. 4. In Fig. 4 (a), we observe the inference result of the R-GAT model, where the output class label is "Thyroid Cancer," corresponding to the input text discussing the telomere-telomerase complex in familial and sporadic cases. Similarly, in Fig. 4 (b), the R-GAT model categorizes the input text about lung cancer treatments. The text discusses using nitrosoureas, such as BCNU, CCNU, and

Raw Abstract: Telomeres are specialized structures at the ends of chromosomes, consisting of hundreds of repeated hexanucleotides (TTAGGG)n. Genetic integrity is partly maintained by the architecture of telomeres, and it is gradually lost as telomeres progressively shorten with each cell replication, due to incomplete lagging DNA strand synthesis and oxidative damage. Telomerase is a reverse transcriptase enzyme that counteracts telomere shortening by adding telomeric repeats to the G-rich strand. In the absence of telomerase or when the activity of the enzyme is low compared to the replicative erosion, apoptosis is triggered. Patients who have inherited genetic defects in telomere maintenance seem to have an increased risk of developing malignant diseases. At the somatic level, telomerase is reactivated in the majority of human carcinomas, suggesting that telomerase reactivation is a critical step for cancerogenesis. In sporadic thyroid carcinoma, telomerase activity is detectable in nearly 50% of thyroid cancer tissues. Recently a germline alteration of the telomere-telomerase complex has been identified in patients with familial papillary thyroid cancer, characterized by short telomeres and increased expression and activity of telomerase compared to patients with sporadic papillary thyroid cancer. In this report, we will review the role of the telomere-telomerase complex in sporadic and familial thyroid cancer.
Output: Thyroid Cancer

(a): Abstract related to the role of telomere-telomerase complex in familial and sporadic thyroid cancer.

Raw Abstract: BCNU, CCNU, and methyl-CCNU have undergone extensive trials in multiple drug combinations for bronchogenic carcinoma. The addition of a nitrosourea appears to be an improvement over cyclophosphamide used alone in oat cell carcinoma and over the two-drug combination of cyclophosphamide and methotrexate in both adenocarcinoma of the lung and oat cell disease. Encouraging response rates have been seen in squamous lung cancer with multiple-drug combinations of a nitrosourea, an alkylating agent, vincristine, and bleomycin with or without adriamycin. The nitrosoureas have been easily incorporated, at reduced doses, into multiple-drug regimens with cumulative myelosuppression seen only when the interval between nitrosourea doses is less than 6 weeks. Conclusions about the ultimate role of these compounds in lung cancer treatment must await (a) comparative trials of combinations with and without a nitrosourea, and (b) further exploration of new approaches to increase their therapeutic index.

Output: Lung Cancer

(b): Abstract discussing the effectiveness of nitrosoureas and other agents in treating bronchogenic carcinoma and various types of lung cancer, including oat cell carcinoma and adenocarcinoma.

Fig. 4: Analysis of cancer abstracts fed into the R-GAT model for classification: a) Thyroid Cancer—the model analyzed the abstract focused on the telomere-telomerase complex in both sporadic and familial thyroid cancer cases, emphasizing telomere shortening and telomerase activation; b) Lung Cancer—the R-GAT model processed an abstract detailing the effectiveness of nitrosoureas and other agents in treating various types of lung cancer, including oat cell carcinoma and magenta adenocarcinoma. Both abstracts were correctly classified by the R-GAT model.

methyl-CCNU, in combination with other medications to treat bronchogenic cancer. It outlines how certain combinations outperform standard therapy, particularly in squamous lung cancer. The model correctly recognizes this input as "lung cancer," proving its ability to interpret complex medical data and make precise classifications. It should be noted that once a model is developed and demonstrated to be correct, we cannot explicitly imply that it is robust. In this regard, a reasonable solution can be an inference, a process of using a trained model to classify or predict new data. The R-GAT model effectively captures complex information through its residual graph attention architecture, demonstrating strong generalization capabilities and adaptability across different cancer types. This model excels in leveraging attention mechanisms to focus on relevant features while maintaining contextual relationships, significantly enhancing classification accuracy.

5) Comparative Review of Existing Studies: Table III examines the various data formats and methods used for classifying cancers such as breast, colorectal, prostate, lung carcinoma, thyroid, colon, and lung. As can be seen from the table, the majority of the data is not publicly available, implying that it was obtained through the authorized organization for research and development purposes. Then, in terms of multi-cancer, just two studies examined several data modalities, while the remaining studies concentrated on a single form of cancer. Simply put, earlier research mostly focused on radiological and clinical reports for identifying breast cancer, while malignancies from biomedical abstracts were not considerably extracted. Medical abstracts are worth looking into, as they have not received as much attention as radiological and clinical findings in research. They include varying conclusions, and investigating them could provide interesting perspectives for cancer research in general and multicancer research in particular.

Additionally, examining Table III, it is evident that transformer-based models were employed in the earlier research, but the R-GAT model has not yet been applied to classify medical cancers. It is also discovered that previous studies were mostly concentrated on the other downstream activities indicated in Table III, and as a result, they did not address thyroid, colon, or lung cancers.

We aimed to implement a model capable of extracting semantic information from lengthy documents, such as med-

TABLE III: Summary of Studies on Cancer Classification Using Different Data Types and Methods

Study	Cancer Type(s)	Data Type	Publicly Available?	Multi-cancer Focus	R-GAT	Transformer Model
Nguyen et al. [4]	Breast Cancer	Radiology Reports	✗	✗	✗	✗
Tang et al. [37]	N/A	Clinical Progress Notes	✗	✗	✗	✓
Alachram et al. [38]	Breast Cancer and Other Cancer Datasets	PubMed Abstracts, Gene Expression Data	✗	✓	✗	✗
Uskaner et al. [6]	Breast Cancer	Mammography Radiology Reports	✗	✗	✗	✓
Jasmir et al. [39]	N/A	Clinical Trial Documents	✓	✗	✗	✗
Prabhakar et al. [40]	N/A	Clinical Patient Records	✗	✗	✗	✗
Achilonu et al. [41]	Breast Cancer, Colorectal and Prostate Cancer	Free-text Pathology Reports	✗	✓	✗	✗
Mithun et al. [42]	Lung Carcinoma	Radiology Reports	✗	✗	✗	✗
Du et al. [43]	N/A	Biomedical Literature and Clinical Notes	✗	✗	✗	✗
R-GAT Model	Thyroid, Colon and Lung	Biomedical Abstracts	✓	✓	✓	✓

ical abstracts related to thyroid, colon, and lung cancers. To validate our R-GAT model, we tested it against several transformer-based methods, including BERT, BioBERT, RoBERTa, and Bio+ClinicalBERT, as well as conventional machine learning and deep learning techniques. However, we discovered that the R-GAT is more generalizable than the other approaches we considered.

6) Study Limitations: This study created a dataset and applied several models to classify cancers such as thyroid, colon, lung, and generic subjects. Although our study achieved interesting results, it has some shortcomings. First, our dataset was quite small and may not yield comprehensive patterns; Therefore, given a complex medical abstract to classify, the models may not generalize accurately.

Second, our dataset contains abstracts related to specific types of cancer. Each abstract focuses only on its designated topic, so there is no overlap with other cancer types or unrelated topics. This means that the dataset may not capture variations that may arise from abstracts discussing multiple subjects or cancer types, potentially affecting the robustness of the data. Additionally, subject matter experts have not evaluated the effectiveness of the presented model in any practical setting.

Regarding model constraints, we introduced the R-GAT model, a graph-based approach. Although this method can be effective with limited data, its generalizability may be limited by the nature of the data used for training. As a result, we cannot confidently conclude that this model is robust to different scenarios. Therefore, in future works, we plan to address these limitations by including a larger and more diverse dataset and evaluating our R-GAT model in real-world settings.

V. CONCLUSION AND FUTURE WORK

This study introduces a labeled dataset of biomedical abstracts covering specific cancer types (thyroid, colon, and lung) and broader topics. The research aimed to evaluate the performance of cutting-edge techniques in scenarios with limited data availability. To this end, we developed an R-GAT model for categorizing cancer-related abstracts. Our R-GAT model underwent thorough testing and was found to surpass traditional machine learning (ML) and deep learning (DL) methods in terms of generalization capabilities. Although we optimized domain-specific transformer models, including BioBERT, RoBERTa, and Bio+ClinicalBERT, their ability to generalize effectively on our curated dataset was less impressive. Our results indicate that the R-GAT model is particularly adept at identifying relationships between various aspects of biomedical abstracts, such as diseases or treatments, by structuring the data as a graph. This approach allows the model to leverage correlations between these elements, enhancing its efficacy in limited data. In contrast, transformer models like BioBERT primarily consider word sequences and may not fully capture these interconnections. Furthermore, transformers often require large datasets to work effectively, and in data-scarce contexts, they are prone to overfitting, limiting their ability to generalize to new or unknown information. To this end, this study adopt a solution that can mitigate the issue of the domain specific transformer such as BIO+BERT or RoBertA.

Our future research plans involve combining the relational reasoning capabilities of R-GAT with the advanced contextual understanding provided by domain-specific transformer models such as BioBERT, RoBERTa, or Bio+ClinicalBERT. We anticipate creating a hybrid ensemble model that enhances the extraction of semantic information from the dataset, potentially leading to improved performance. Furthermore, we will explore methods to enhance the generalizability of transformer-

based models when applied to our specially curated dataset.

DECLARATION OF COMPETING INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ACKNOWLEDGMENTS

REFERENCES

[1] M. Zhai, D. Zhang, J. Long, Y. Gong, F. Ye, S. Liu, and Y. Li, "The global burden of thyroid cancer and its attributable risk factor in 195 countries and territories: A systematic analysis for the global burden of disease study," *Cancer medicine*, vol. 10, no. 13, pp. 4542–4554, 2021.

[2] W. H. Organization, "Colorectal cancer." Accessed on 10/15/2023.

[3] Y.-H. Wang, P. A. Nguyen, M. M. Islam, Y.-C. Li, and H.-C. Yang, "Development of deep learning algorithm for detection of colorectal cancer in ehr data.," *MedInfo*, vol. 264, pp. 438–441, 2019.

[4] E. Nguyen, D. Theodorakopoulos, S. Pathak, J. Geerdink, O. Vijlbrief, M. Van Keulen, and C. Seifert, "A hybrid text classification and language generation model for automated summarization of dutch breast cancer radiology reports," in *2020 IEEE second international conference on cognitive machine intelligence (CogMI)*, pp. 72–81, IEEE, 2020.

[5] M. Tang, P. Gandhi, M. A. Kabir, C. Zou, J. Blakey, and X. Luo, "Progress notes classification and keyword extraction using attention-based deep learning models with bert," *arXiv preprint arXiv:1910.05786*, 2019.

[6] P. Uskaner Hepsağ, S. A. Özel, K. Dalcı, and A. Yazıcı, "Using bert models for breast cancer diagnosis from turkish radiology reports," *Language Resources and Evaluation*, pp. 1–32, 2023.

[7] W. Ai, Y. Wei, H. Shao, Y. Shou, T. Meng, and K. Li, "Edge-enhanced minimum-margin graph attention network for short text classification," *Expert Systems with Applications*, vol. 251, p. 124069, 2024.

[8] Y. Wei, D. Wu, and J. Terpenny, "Remaining useful life prediction using graph convolutional attention networks with temporal convolution-aware nested residual connections," *Reliability Engineering & System Safety*, vol. 242, p. 109776, 2024.

[9] Y. Song, H. Mao, J. Xiao, J. Liu, Z. Chen, W. Jin, C. Yang, J. Tang, and H. Liu, "A pure transformer pretraining framework on text-attributed graphs," *arXiv preprint arXiv:2406.13873*, 2024.

[10] Q. Rao, T. Wang, X. Guo, K. Wang, and Y. Yan, "Knowledge graph completion using a pre-trained language model based on categorical information and multi-layer residual attention," *Applied Sciences*, vol. 14, no. 11, p. 4453, 2024.

[11] J. P. Buchmann and E. C. Holmes, "Entrezpy: a python library to dynamically interact with the ncbi entrez databases," *Bioinformatics*, vol. 35, no. 21, pp. 4511–4514, 2019.

[12] E. Loper and S. Bird, "Nltk: The natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, p. 321–357, June 2002.

[14] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[15] T. Mikolov, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[17] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[18] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[19] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Knn model-based approach in classification," in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pp. 986–996, Springer, 2003.

[20] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial naive bayes for text categorization revisited," in *AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17*, pp. 488–499, Springer, 2005.

[21] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[22] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[23] D. Boswell, "Introduction to support vector machines," *Departement of Computer Science and Engineering University of California San Diego*, vol. 11, pp. 16–17, 2002.

[24] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[25] J. S. Cramer, "The origins of logistic regression," 2002.

[26] K. Spärck Jones, "Idf term weighting and ir research lessons," *Journal of documentation*, vol. 60, no. 5, pp. 521–523, 2004.

[27] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proceedings of 2010 IEEE international symposium on circuits and systems*, pp. 253–256, IEEE, 2010.

[28] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[29] S. Hochreiter, "Long short-term memory," *Neural Computation MIT-Press*, 1997.

[30] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," in *International conference on artificial neural networks*, pp. 799–804, Springer, 2005.

[31] Y. Heryadi and H. L. H. S. Warnars, "Learning temporal representation of transaction amount for fraudulent transaction recognition using cnn, stacked lstm, and cnn-lstm," in *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, pp. 84–89, IEEE, 2017.

[32] Z. Ran, D. Zheng, Y. Lai, and L. Tian, "Applying stack bidirectional lstm model to intrusion detection," *CMC-Comput. Mater. Continua*, vol. 65, no. 1, pp. 309–320, 2020.

[33] V. B. Semwal, A. Gupta, and P. Lalwani, "An optimized hybrid deep learning model using ensemble learning approach for human walking activities recognition," *The Journal of Supercomputing*, vol. 77, no. 11, pp. 12256–12279, 2021.

[34] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, p. 1234–1240, Sept. 2019.

[35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[36] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly available clinical bert embeddings," 2019.

[37] M. Tang, P. Gandhi, M. A. Kabir, C. Zou, J. Blakey, and X. Luo, "Progress notes classification and keyword extraction using attention-based deep learning models with bert," *arXiv preprint arXiv:1910.05786*, 2019.

[38] H. Alachram, H. Chereda, T. Beißbarth, E. Wingender, and P. Stegmaier, "Text mining-based word representations for biomedical data analysis and protein-protein interaction networks in machine learning tasks," *PloS one*, vol. 16, no. 10, p. e0258623, 2021.

[39] J. JASMIR, S. NURMAINI, R. F. MALIK, and D. Z. ABIDIN, "Text classification of cancer clinical trials documents using deep neural network and fine grained document clustering," in *Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*, pp. 396–404, Atlantis Press, 2020.

[40] S. K. Prabhakar and D.-O. Won, "Medical text classification using hybrid deep learning models with multihead attention," *Computational intelligence and neuroscience*, vol. 2021, 2021.

[41] O. J. Achilonu, V. Olago, E. Singh, R. M. Eijkemans, G. Nimako, and E. Musenge, "A text mining approach in the classification of free-text cancer pathology reports from the south african national health laboratory services," *Information*, vol. 12, no. 11, p. 451, 2021.

[42] S. Mithun, A. K. Jha, U. B. Sherkhane, V. Jaiswar, N. C. Purandare, A. Dekker, S. Puts, I. Bermejo, V. Rangarajan, C. M. Zegers, *et al.*, "Clinical concept-based radiology reports classification pipeline for lung carcinoma," *Journal of Digital Imaging*, pp. 1–15, 2023.

[43] J. Du, Q. Chen, Y. Peng, Y. Xiang, C. Tao, and Z. Lu, "Ml-net: multi-label classification of biomedical texts with deep neural networks," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1279–1285, 2019.