Medical Text Classification using Bio Clinical BERT

Mahnoor
fmahnoor1@umassd.edu
University of Massachusetts
Dartmouth
Massachusetts, USA

Mahmuda Akter Keya
mkeya@umassd.edu
University of Massachusetts
Dartmouth
Massachusetts, USA

Shuwen Wang
swang5@umassd.edu
University of Massachusetts
Dartmouth
Massachusetts, USA

Abstract

Medical text classification is a critical task in healthcare, where accurate identification of disease categories from clinical narratives can significantly enhance diagnostic and treatment processes. In this study, we propose leveraging Bio_ClinicalBERT, a BERT-based transformer model pretrained on large-scale clinical and biomedical corpora, in order to improve the classification performance for medical abstracts extracted from doctor's prescription. Our approach focuses on leveraging domain-specific pretraining to capture medical terminologies and long-range semantic relationships within clinical texts. Evaluation metrics including accuracy, precision, recall and F1-score demonstrated that Bio_ClinicalBERT outperforms traditional machine learning methods, showcasing the efficacy of transformer-based models in medical text classification tasks. The finetuned model offers significant improvements in identifying disease categories.

Keywords

Natural Language Processing (NLP), Medical Text Classification, AI in Health Care

ACM Reference Format:
Mahnoor, Mahmuda Akter Keya, and Shuwen Wang. 2025. Medical Text Classification using Bio Clinical BERT. In . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The rapid growth of medical data, particularly clinical abstracts and physician notes, has necessitated the development of robust natural language processing (NLP) systems for categorizing and analyzing textual information. Accurate classification of medical text into disease categories is critical for downstream applications such as clinical decision support, automated triage, and medical research analytics. Traditional machine learning approaches often rely on handcrafted features and shallow embeddings such as TF-IDF [9] for text, SIFT or HOG [6] for images etc. These representations when used in models like Support Vector Machines [4], Random Forests [2] have shown promises due to their simplicity and interpretability. However, these methods frequently struggle to capture the complex semantic relationships inherent in biomedical language. Recent advances in deep learning, particularly transformer-based models, offer powerful alternatives by learning contextualized representations of text.

In this work, we aim to classify doctor's prescription into disease categories by comparing the traditional machine learning classifiers with a domain-specific transformer model, **Bio_ClinicalBERT**. We evaluate model's performance using standard classification metrics and demonstrate the superiority of contextualized, domain-aware language models in clinical NLP tasks.

2 Related Works

The research in medical text classification has steadily evolved from traditional machine learning approaches to the adoption of deep contextualized language models. Early studies commonly replied on manually crafted features and shallow embeddings. For instance, logistic regression and support vector machines have been effectively applied to clinical text classification using TF-IDF and Word2Vec features [5]. These models offer interpretability and low computational costs, making them suitable for resource-constrained settings.

Subsequent research introduced more robust embedding techniques. Le and Mikolov (2014) proposed Doc2Vec for capturing paragraph-level semantics, while other studies experimented with weighting Word2Vec vectors to enhance the performance in domain-specific tasks. Such feature engineering proved effective in medium-sized datasets, especially those with limited contextual depth.

In recent years, transformer-based models have become dominant paradigm in biomedical NLP. Notably, **Bio_ClinicalBERT** [1], pretrained on MIMIC-III clinical notes, has demonstrated superior performance from large-scale pretraining on domain-specific corpora and minimal preprocessing, enabling them to preserve the subtle nuance of clinical language.

Prabhankar et al. [7] proposes a novel approach for medical text classification by utilizing two hybrid deep learning models: the **Quad Channel Hybrid Long Short-Term Memory (QC-LSTM)** model and the **Hybrid BiDirectional Gated Recurrent Unit (Bi-GRU) with Multihead Attention** model. The QC-LSTM model is designed to improve classification performance by integrating four input channels, which allows for a richer and more diverse set of features to be extracted from the text data. The hybrid BiGRU model, on the other hand, leverages multihead attention, which enables the model to attend different parts of input sequence simultaneously, thereby capturing more complex relationships in the text. Both models are validared on 2 medical text datasets, demonstrating superior performance in classification accuracy. Notably, the QC-LSTM model achieved an accuracy of **96.72**%, showcasing the effectiveness of hybrid architectures in handling the intricate and high-dimensional nature of medical text data.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

On the other hand, Qing et al. [8] present an innovative approach to medical text classification by proposing a unified neural network method that addresses the challenges of high-dimensionality and data sparsity in the medical domain. The authors utilize a Bidirectional Gated Recurrent Unit (BiGRU) combined with attention mechanism to extract meaningful sentence level features from medical texts. In the proposed method, the convolutional layer first captures the local features within sentences, and the BiGRU then processes these features in both forward and backward directions to capture the full contextual information. The attention mechanism helps highlight important words and sentences, enabling the model to focus on the most relevant parts of the text. This method is validated on four different medical texts datasets (including both medical records and literature), where it demonstrates effective classification capabilities and robust handling of the complexities inherent in medical text data.

Building on these advancements, recent studies have further explored the use of Convolutional Neural Networks for sentence level medical text classification. Hughes et al. [3] presented a novel deep learning approach for the semantic classification of clinical text. The authors utilize Word2Vec to represent words in the clinical text as dense vectors, capturing semantic relationships between words. The CNN model is designed to automatically learn complex features representations from these embeddings, which are then classified into predefined medical categories. The model architecture includes multiple convolutional layers followed by max-pooling and dropout layers to prevent overfitting. This approach is evaluated against other text classification techniques such as sentence embeddings and bag-of-words (BOW), and the results show that the CNN-based method outperforms traditional models by **15**% in terms if classification accuracy. This work demonstrates the power of CNNs in extracting rich, context-dependent features from medical text, which could be particularly valuable for large-scale text classification tasks.

3 Methods

In this section, we outline the methodologies employed for classifying medical abstracts extracted from doctor's prescriptions. This includes a detailed a description of the dataset, preprocessing steps, and the approach used for model training and evaluation.

3.1 Data Description

The dataset used in this study consists of **medical abstracts** extracted from doctor's prescriptions, categorized into five distinct disease groups. These categories represent a wide range of medical conditions, allowing the model to generalize across various diagnoses. The dataset is structures as follows:

- **Class 1: Neoplasm**: Abstract related to cancerous growths and related medical conditions.
- **Class 2: Digestive System Diseases**: Medical text describing disorders related to the digestive tract.
- **Class 3: Nervous System Diseases**: Includes conditions affecting the nervous system, such as neurological disorder.
- **Class 4: Cardiovascular Diseases**: Focusing on diseases affecting the heart and blood vessels.
- **Class 5: General Pathological Conditions**: Miscellaneous diseases or conditions that do not fall under the above categories but are relevant for broader medical classification.

The dataset consists of **11,550 instances**, where each instance is composed of a short clinical test excerpt paired with its corresponding disease label. The diversity and size of the dataset make it suitable for both traditional machine learning and deep learning approaches.

3.2 Preprocessing

Initial preprocessing steps were carried out to prepare the data for model training. These steps were aimed at normalizing the text data while preserving domain-specific medical terms critical for accurate classification. The processing pipeline included the following stages:

- **Lowercasing**:All text data was converted to lowercase to ensure uniformity and reduce dimensionality.
- **Tokenization**: Texts were split into individual tokens (words) to make the text suitable for vectorization.
- **Removal of Special Characters**: Non-essential characters (e.g., punctuation marks, symbols) were removed to ensure the model focuses on the relevant words.
- **Lemmatization**: Words were lemmatization to their base or root forms, which helped reduce inflectional forms and make the features more consistent.

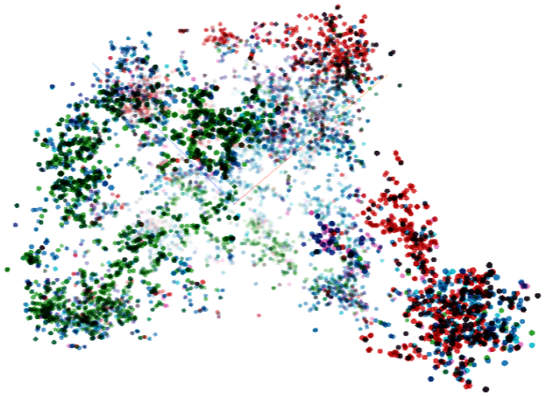
After preprocessing, the text data was vectorized using several strategies to represent the words and sentences as numerical features. These strategies included:

- **Word2Vec Mean Embeddings**: A word embedding technique where each word in the text is represented by a dense vector. The mean embedding of the words in a sentence is used to represent the entire sentence. Let $W = [w_1, w_2, \ldots, w_n]$ be a sequence of words in a sentence, and each word w_i is represented by a word vector $v(w_i)$. The mean embedding for the sentence A is computed as:

$$\mathbf{v}(S) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{v}(w_i)$$

Where:
– $v(w_i)$ is the vector representation of the i-th word.
– n is the total number of words in the sentence S
The equation computes the average of the word vectors in the sentence, which is then used as the sentence-level embedding.

Figure 1: Word2Vec Embedding Plot Using T-SNE



Note: The Word2Vec model was trained on a medical corpus using the CBOW architecture with parameters: `vector_size=100`, `window=5`, `min_count=2`, `workers=4`, and `sg=0` (CBOW).

- **Doc2Vec Embeddings**: Doc2Vec extends Word2Vec by introducing a document-level vectors and the document vector. Let D represent a document and $W = [w_1, w_2, \ldots, w_n]$ represent the words in the document. Each word w_i has a corresponding word vector $v(w_i)$, and there is also a document vector v(D). The document vector is learned in a way that helps predict the next word in a sliding window during training, similar to how Word2Vec works. The Doc2Vec embedding for a document D is computed as:

$$\mathbf{v}(D) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{v}(w_i) + d(D)$$

  Where:
  - $v(w_i)$ is the vector for the i-th word
  - d(D) is the document vector for the document D
  - n is the number of words in the document

  The final embedding is a combination of word vectors and the document vector, where the document vector d(D) provides context for the whole document.
- **Weighted Word2Vec Embeddings**: In the Weighted Word2Vec embeddings, each word vector is weighted by it's importance in the context of the sentence, such as a term frequency (TF) or other factors that reflect the importance of certain terms. Let $W = [w_1, w_2, \ldots, w_n]$ represent the words in a sentence, $v(w_i)$ the corresponding word vectors. Each word has a weight $w(w_i)$, which can be based on factors like term frequency (TF) or TF-IDF (Term Frequency-Inverse Document Frequency). The weigthted mean embedding for the sentence is:

$$\mathbf{v}(S) = \frac{\sum_{i=1}^{n}\omega(w_i)\cdot\mathbf{v}(w_i)}{\sum_{i=1}^{n}\omega(w_i)}$$

  Where:
  - $w(w_i)$ is the weight of the i-th word
  - $v(w_i)$ is the vector for the i-th word
  - The denominator normalizes the weigjts so that they sum to 1, ensuring that the average embedding remains properly scaled.

  In this method, most important words (according to the weights) contribute more to the final sentence embedding.

3.3 Approach

The primary objective of this study is to classify medical abstracts into one of five disease categories. This task is formulated as a **single-label, multiclass text classification problem**, where each abstract described a patient's condition or diagnosis, and the goal is to predict the corresponding disease label. Given the complexity and variability of the medical language, the model must be able to understand domain-specific terminology, abbreviations, and the contextual cues inherent in clinical narratives.

We applied deep learning-based approach, using transformer-based model pretrained on clinical text (Bio_ClinicalBERT). This model is particularly well-suited for handling the complexities of medical language, as it leverages the transformer architecture's self-attention mechanism to capture long-range dependencies and context within the text.

Performance is evaluated using standard classification metrics such as **Accuracy, Precision, Reall** and **F1-score**, along with **Macro** and **Weighted averages** to account for potential class imbalance in the dataset.

4 Model

4.1 Traditional Machine Learning Baselines

To establish robust baseline performances, we implemented three classical machine learning models:

- Logistic Regression
- Random Forest Classifier
- Support Vector Classifier (SVC)

Each model was trained using different text representation strategies:

- **Word2Vec Mean Embeddings:** Averaging word vectors to obtain document-level features.
- **Doc2Vec Embeddings:** Document-specific embeddings capturing holistic semantics.
- **Weighted Word2Vec Embeddings:** Averaging word embeddings weighted by TF-IDF scores.

These models are chosen for their interpretability, low computational requirements, and effectiveness in non-deep learning natural language processing pipelines.

4.2 Proposed Model: Bio_ClinicalBERT

In this study, we enhance medical text classification performance by proposing Bio_ClinicalBERT, a BERT-based transformer model specifically pretrained on clinical notes and biomedical literature.

The model was fine-tuned by adding a fully connected classification head on top of pretrained embeddings, aimed at improving domain-specific understanding and predictive accuracy.

Key advantages of Bio_ClinicalBERT include:

- **Domain-Specific Pretraining**: Extensive exposure to clinical language improves understanding of specialized terminology.
- **Context-Aware Representations**|: Ability to model long-range dependencies and nuanced semantic relations in medical text.
- **Transfer Learning Benefits**: Reduced training data requirements due to knowledge transfer from large-scale biomedical corpora.

The fine-tuning process was carried out using a cross-entropy loss function, optimized with AdamW optimizer, to ensure efficient learning and model convergence. Additionally, early stopping was implemented to prevent overfitting, ensuring that the model's performance remains robust even when trained on smaller or imbalanced datasets.

4.3 Fine-Tuning Strategy

The fine-tuning process for Bio_ClinicalBERT involved several key components to optimize the model for medical text classification.

- **Model Architecture**: The Bio_ClinicalBERT model was initialized with pretrained weights from the transformer model, which was pretrained on clinical notes and biomedical literature. The model architecture consists of a **Transformer Encoder** layer with self-attention mechanisms, followed by a **fully connected classification head**. This head was specifically added to adapt the model for our medical text classification task, predicting one of the disease categories based on the text input.
- **Hyperparameter Tuning**: A range of hyperparameters were explored during the fine-tuning process, including:
  - **Learning Rate**: A learning rate of **2e-5** was chosen after several experiments. This is a typical starting point for finetuning BERT-like models, ensuring gradual and stable convergence.
  - **Batch Size**: A batch size of 16 was used to fit the model efficiently into memory while balancing the trade-off between training speed and model performance.
  - **Epochs**: The model was trained for maximum of 5 epochs, based on preliminary results indicating that the model converged relatively quickly. This is also due to small sclae of the dataset and the use of pretrained weights.
- **Optimizer**: AdamW was chosen for its ability to adaptively adjust the learning rate and its proven effectiveness in fine-tuning transformer-based models.

The fine-tuning procedure ensures that Bio_ClinicalBERT is optimized specifically for medical text classification tasks, providing domain-specific understanding of clinical texts and significantly improving classification accuracy.

5 Results and Discussion

Early studies on medical text classification primarily relied on traditional machine learning algorithms. Johnson et al. [1] demonstrated the effectiveness of SVMs with TF-IDF features for categorizing clinical discharge summaries. Similarly, Joachims [2] explored the application of support vector machines for text categorization tasks, highlighting the model's robustness across various domains.

Recent advances have shifted toward transformer-based models. Lee et al. [3] introduced BioBERT, a BERT model pretrained on biomedical corpora, showing substantial improvements over general-purpose BERT in biomedical named entity recognition and question answering tasks. Alsentzer et al. [4] further fine-tuned BioBERT on clinical notes from the MIMIC-III database, resulting in Bio_ClinicalBERT, which demonstrated state-of-the-art performance on several clinical NLP benchmarks.

These studies emphasize the value of domain-specific pretraining and contextual embeddings in improving medical text classification, motivating the selection of Bio_ClinicalBERT for our task.

5.1 Evaluation Metrics

Model performance was assessed using the following metrics:

- **Accuracy:** Overall proportion of correctly classified samples.
- **Precision:** Correct positive predictions over all positive predictions made.
- **Recall:** Correct positive predictions over all actual positives.
- **F1-Score:** Harmonic mean of Precision and Recall.
- **Macro Average:** Metric averaged equally across all classes.
- **Weighted Average:** Metric averaged proportionally to class support.

The below tables shows the evaluation metrics using different embedding methods

Table 1: Word2Vec Embeddings - Performance Metrics

Model	Class	Precision	Recall	F1-score
Logistic Regression	0	0.422	0.420	0.421
Logistic Regression	1	0.333	0.022	0.042
Logistic Regression	2	0.333	0.007	0.013
Logistic Regression	3	0.492	0.315	0.385
Logistic Regression	4	0.364	0.679	0.474
Random Forest	0	0.321	0.319	0.320
Random Forest	1	0.070	0.031	0.043
Random Forest	2	0.038	0.017	0.023
Random Forest	3	0.320	0.256	0.284
Random Forest	4	0.277	0.415	0.332
SVC	0	0.425	0.290	0.345
SVC	1	0.000	0.000	0.000
SVC	2	0.000	0.000	0.000
SVC	3	0.538	0.083	0.143
SVC	4	0.349	0.840	0.494

Table 2: Word2Vec Embeddings - Accuracy Metrics

Model	Class	Precision	Recall	F1-score
Logistic Regression	macro avg	0.361	0.289	0.265
Logistic Regression	weighted avg	0.384	0.396	0.343
Random Forest	macro avg	0.199	0.198	0.192
Random Forest	weighted avg	0.239	0.261	0.241
SVC	macro avg	0.272	0.243	0.197
SVC	weighted avg	0.340	0.367	0.273

Table 3: Doc2Vec Embeddings - Performance Metrics

Model	Class	Precision	Recall	F1-score
Logistic Regression	0	0.612	0.724	0.663
Logistic Regression	1	0.494	0.366	0.421
Logistic Regression	2	0.536	0.349	0.423
Logistic Regression	3	0.626	0.583	0.604
Logistic Regression	4	0.464	0.528	0.494
Random Forest	0	0.514	0.474	0.493
Random Forest	1	0.067	0.022	0.033
Random Forest	2	0.071	0.024	0.036
Random Forest	3	0.449	0.329	0.380
Random Forest	4	0.320	0.532	0.400
SVC	0	0.632	0.683	0.657
SVC	1	0.494	0.362	0.418
SVC	2	0.537	0.322	0.403
SVC	3	0.642	0.535	0.583
SVC	4	0.443	0.569	0.498

Table 4: Doc2Vec Embeddings - Accuracy Metrics

Model	Class	Precision	Recall	F1-score
Logistic Regression	macro avg	0.562	0.520	0.533
Logistic Regression	weighted avg	0.556	0.555	0.550
Random Forest	macro avg	0.279	0.269	0.261
Random Forest	weighted avg	0.326	0.352	0.324
SVC	macro avg	0.562	0.503	0.521
SVC	weighted avg	0.555	0.545	0.541

5.2 Baseline Model Performance

The Bio_ClinicalBERT model outperforms all traditional baselines across all evaluation metrics, confirming the advantage of domain-specific contextualized language modeling for clinical text classification tasks.

5.3 Error Analysis

Most misclassifications occurred between Neoplasm and General Pathological Conditions, likely due to semantic overlaps in descriptions. Further fine-tuning or hierarchical classification strategies could help address this challenge.

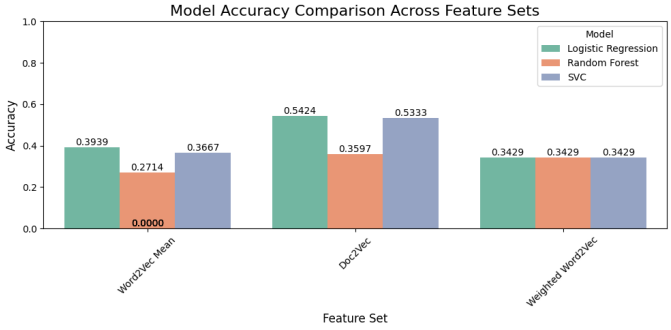
Table 5: Weighted Word2Vec Embeddings - Performance Metrics

Model	Class	Precision	Recall	F1-score
Logistic Regression	0	0.000	0.000	0.000
Logistic Regression	1	0.000	0.000	0.000
Logistic Regression	2	0.000	0.000	0.000
Logistic Regression	3	0.000	0.000	0.000
Logistic Regression	4	0.343	1.000	0.511
Random Forest	0	0.000	0.000	0.000
Random Forest	1	0.000	0.000	0.000
Random Forest	2	0.000	0.000	0.000
Random Forest	3	0.000	0.000	0.000
Random Forest	4	0.343	1.000	0.511
SVC	0	0.000	0.000	0.000
SVC	1	0.000	0.000	0.000
SVC	2	0.000	0.000	0.000
SVC	3	0.000	0.000	0.000
SVC	4	0.343	1.000	0.511

Table 6: Weighted Word2Vec Embeddings - Accuracy Metrics

Model	Class	Precision	Recall	F1-score
Logistic Regression	macro avg	0.069	0.200	0.102
Logistic Regression	weighted avg	0.118	0.343	0.175
Random Forest	macro avg	0.069	0.200	0.102
Random Forest	weighted avg	0.118	0.343	0.175
SVC	macro avg	0.069	0.200	0.102
SVC	weighted avg	0.118	0.343	0.175

Figure 2: Model Accuracy - Different Embedding Methods



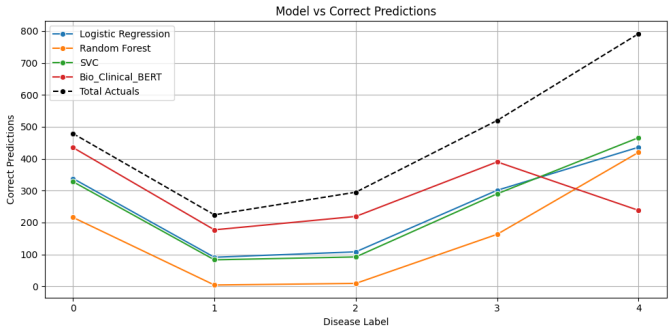
Note: Logistic Regression and SVC models perform best with the Doc2Vec features, achieving accuracies of 0.5424 and 0.5333 respectively. Overall, Doc2Vec yields higher accuracies for most models compared to other feature embeddings.

6 Conclusion

References

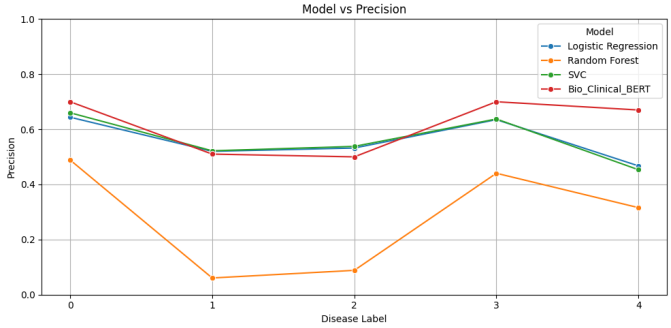
[1] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323* (2019).

Figure 3: Comparison Correct Predictions - Doc2Vec



Note: This plot compares the number of correct predictions made by various models (Logistic Regression, Random Forest, SVC, and Bio_Clinical_BERT) against the actual class distribution. Bio_Clinical_BERT outperforms traditional models in most classes but shows underperformance on class 4. This is due to overlapping embeddings of "general pathology conditions" with other conditions, as Bio_Clinical_BERT is pretrained on a broad dataset. The dotted line represents the total number of actual instances per class.

Figure 4: Comparison Models Precisions - Doc2Vec

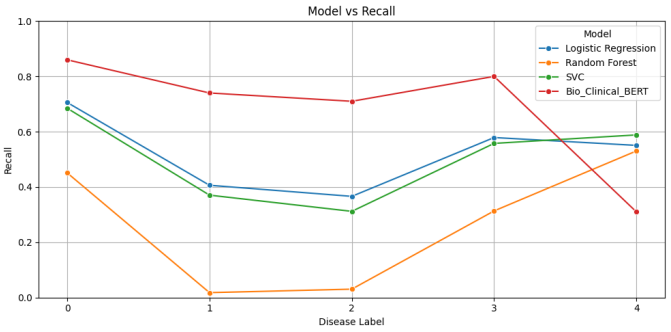


Note: This plot compares the precision of various models (Logistic Regression, Random Forest, SVC, and Bio_Clinical_BERT) across disease classes. Bio_Clinical_BERT consistently achieves the highest precision in all classes except 2, outperforming traditional models. This highlights the effectiveness of large-scale pretraining for capturing discriminative features across diverse pathology conditions.

Table 7: Performance Metrics using Bio_Clinical BERT

Model	Class	Precision	Recall	F1-score
Bio_clinicalBERT	0	0.70	0.86	0.77
Bio_clinicalBERT	1	0.51	0.74	0.61
Bio_clinicalBERT	2	0.50	0.71	0.59
Bio_clinicalBERT	3	0.70	0.80	0.74
Bio_clinicalBERT	4	0.67	0.31	0.42

Figure 5: Comparison Models Recall - Doc2Vec



Note: Bio_Clinical_BERT consistently outperforms traditional machine learning models (Logistic Regression, Random Forest, SVC), especially for disease labels 0 to 3.

Table 8: Macro and Weighted Average Metrics - Bio_Clinical BERT

Model	Class	Precision	Recall	F1-score
Bio_clinicalBERT	macro avg	0.62	0.68	0.63
Bio_clinicalBERT	weighted avg	0.64	0.63	0.61

[2] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.
[3] Mark Hughes, Irene Li, Spyros Kotoulas, and Toyotaro Suzumura. 2017. Medical text classification using convolutional neural networks. In *Informatics for health: connected citizen-led wellness and population health*. IOS Press, 246–250.
[4] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*. Springer, 137–142.
[5] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* (2016).
[6] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60 (2004), 91–110.
[7] Sunil Kumar Prabhakar and Dong-Ok Won. 2021. Medical text classification using hybrid deep learning models with multihead attention. *Computational intelligence and neuroscience* 2021, 1 (2021), 9425655.
[8] Li Qing, Weng Linhong, and Ding Xuehai. 2019. A novel neural network-based method for medical text classification. *Future Internet* 11, 12 (2019), 255.
[9] Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. 2003. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)*. 616–623.