Performance Comparison of Pre-trained and Fine-tuned BERT-Based Transformer Language Models in Health Domain

Payel Patra, (3rd year PhD student)
Payel.patra@graduate.univaq.it

Università degli Studi dell'Aquila , Italy

Joint work with Prof. Antinisca Di Marco, Prof. Phuong T. Nguyen

This work is supported by European Union — NextGenerationEU — National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) — Project: "SoBigData.it — Strengthening the Italian RI for Social Mining and Big Data Analytics" — Prot. IR0000013 — Avviso n. 3264 del 28/12/2021.

Outline:

Introduction

- Motivation
- challenges
- Research Gap
- Contribution

Related Work

Transformer Models Concept (BERT)

- Descriptions of BERT-based Models

Entity Extraction of Medical text (Cancer) data

- Dataset and Selected BERT models
- Methodology

Evaluation of research Questions

Result Analysis

Conclusion and future work

Motivation

- Cancer data is one of the most common issue now-a-day, so it is selected due to its significance in research.
- Early detection and precise treatment improve patient outcomes.
- Clinical records contain essential medical insights, and it is also unstructured and difficult to analyze
- Automated entity extraction improves efficiency and accuracy.
- Transformer-based models enhance clinical text processing. It supports AI-driven expert systems for better decision-making.
- Transformer-based models enhances efficiency, accuracy, and reliability in processing clinical text data.

Challenges in Clinical Text Processing

MEDICAL TEXT IS UNSTRUCTURED AND COMPLEX.

LIMITED AVAILABILITY OF ANNOTATED DATASETS.

TRADITIONAL RULE-BASED OR MACHINE LEARNING METHODS LACK.

SCALABILITY ISSUES.

Research Gap

LACK OF BENCHMARKING

LIMITED AI BASED MODEL COMPARISONS

ABSENCE OF REAL-WORLD IMPLICATIONS

Contributions

a transformer-based approach cancer clinical data

fine-tune five BERT-based models for cancer entity extraction

Conduct a comprehensive performance evaluation

contributes to the development **AI-driven expert system cancer diagnosis and treatment planning (Breast Cancer)**

Listing: The Query for Systematic Review using advanced string search

TITLE-ABS (("Breast Cancer" OR "Clinical Text Data" OR "Breast Cancer Data" OR "Cancer Text Data") AND ("*Entity Extraction" OR "Extracting") AND ("Natural Language Processing" OR "NLP" OR "Large Language Model" OR "LLM*"OR "Artificial Intelligence" OR "AI" OR "BERT" OR "*BERT" OR "Transformer Model*")) AND PUBYEAR >2020 AND PUBYEAR <2025

Table of total number of papers related to the topics

	Cancer (Breast)	Entity Extraction	NLP/LLM/BERT
Cancer (Breast)	48,548		
Entity Extraction	282	46,757	
NLP/LLM/BERT	1,737	1,073	45,884

Selection Criteria:

We have selected 39 papers which matches any one keyword from these three keyword groups.

Keywords for this advance string search operation:

(i) **Cancer(Breast)** — “Breast Cancer” OR “Clinical Text Data” OR “Breast Cancer Data” OR “Cancer Text Data”;
(ii) **Entity Extraction** — “*Entity Extraction” OR “Extracting”;
(iii) **NLP/LLM/BERT** — “Natural Language Processing” OR “NLP” OR "Large Language Model” OR “LLM*” OR “Artificial Intelligence” OR “AI” OR “BERT” OR “*BERT” OR “Transformer Model*”.

Overview of BERT Models, Their Pretraining Datasets, and Exclusion Criteria.

Model Name	Description	Pretraining Dataset(s)	Exclusion Criteria
BERT	Original BERT model by Google for general NLP tasks.	BookCorpus + English Wikipedia	Not optimized for domain-specific tasks like biomedical or legal text.
BioBERT	Specialized for biomedical text.	PubMed abstracts + PMC full-text articles	Selected (Designed specifically for biomedical applications.)
ClinicalBERT	Tailored for clinical data processing.	MIMIC-III (clinical records)	Clinical records, lacks broader biomedical corpus.
SciBERT	Designed for scientific and research text.	Semantic Scholar corpus	Not fine-tuned for clinical or biomedical applications.
BlueBERT	Combines biomedical and clinical data for broad use.	PubMed abstracts + MIMIC-III	Selected (Optimized for biomedical corpus)
RoBERTa	Enhanced version of BERT by Facebook for better results.	BookCorpus + English Wikipedia + CC-News + OpenWebText	Selected (Strong general NLP performance makes it suitable for fine-tuning on medical corpora.)
DistilBERT	Lightweight, faster version of BERT.	Same as BERT (BookCorpus + English Wikipedia)	Reduced model size leads to lower accuracy in domain-specific tasks.
ALBERT	Efficient, smaller variant of BERT.	BookCorpus + English Wikipedia	Not medical domain data.
BioClinical BERT	Advanced ClinicalBERT for better medical NLP.	MIMIC-III + PubMed abstracts	Selected (Focused only on clinical and biomedical data, limiting generalizability.)
COVID-Twitter-BERT	Analyzes COVID-19 related tweets.	COVID-19 related tweets	Limited to social media and COVID-related discussions.
FinBERT	Focused on financial text analysis.	Financial reports, news articles, company filings	Not designed for medical or scientific applications.
LegalBERT	Designed for legal text processing.	Legal documents and court cases	Unsuitable for biomedical and general NLP applications.
CamemBERT	BERT model for French language.	French Common Crawl (OSCAR)	Only supports French, not useful for English medical NLP.
M-BERT	Supports 104 languages for multilingual NLP.	Wikipedia in multiple languages	Not optimized for domain-specific NLP like biomedical or legal text.
XLM-RoBERTa	Cross-lingual version for multiple languages.	Common Crawl in 100 languages	Focuses on multilingual tasks, lacks domain specialization.
PubMedBERT	Fully trained on biomedical literature for better domain adaptation.	Full PubMed abstracts	Selected (Clinical records (e.g., MIMIC-III), making it less effective for real-world clinical applications.)

Background in Selecting BERT-based models and Annotation Tools

We have Selected 5 BERT-based Models:

1. BioBERT
2. BioClinicalBERT
3. PubMedBERT
4. BlueBERT
5. RoBERTa

Background

Tool Name	Available Links	Functionality	Exclusion Criteria from Others
BRAT	https://brat.nlplab.org/	Web-based text annotation and validation	Requires external setup, lacks built-in ML support
INcepTION	https://inception-project.github.io/	Collaborative annotation with machine learning assistance	Complex for beginners, requires ML expertise
MetaMap	https://metamap.nlm.nih.gov/	Maps biomedical text to UMLS concepts	Limited to UMLS concepts, lacks broader NLP capabilities
MedLEE	https://www.dbmi.columbia.edu/research-projects/medlee/	Structuring clinical narratives	Designed for clinical narratives, less flexible for general medical text
CLAMP	https://clamp.uth.edu/	Named entity recognition and relation extraction	Focuses mainly on named entity recognition, lacks deep inference
MedNLI	https://physionet.org/content/mednli/1.0.0/	Medical language inference and analysis	Limited to medical natural language inference tasks
BioMedGPT	https://github.com/stanford-crfm/BioMedGPT	Deep medical language processing	Primarily designed for deep learning-based medical text generation
MedspaCy	https://github.com/medspacy/medspacy	Extends spaCy for clinical text processing	Requires spaCy knowledge, limited pre-built clinical models
Doccano	https://github.com/doccano/doccano	User-friendly text annotation with JSON export support	Selected for its ease of use, efficient annotation workflow, and JSON export support

Selection of Annotation Tool.

We have selected Doccano Tool among these open-source tools

Background

MODEL	PRE-TRAINING DATA	BUILT BY	NSP (NEXT SENTENCE PREDICTION)	MASKING STRATEGY	PRIMARY APPLICATIONS
BioBERT	PubMed & PMC articles	Korea University & Clova AI	Yes	Dynamic masking	Entity recognition, relation extraction, QA in biomedicine
BioClinicalBERT	MIMIC-III clinical notes	Google Research	Yes	Dynamic masking	Clinical NER, patient record analysis
RoBERTa	General NLP corpus	Facebook AI	No	Full-sentence & full-word masking	Text classification, QA, sentiment analysis
BlueBERT	PubMed abstracts & clinical notes	Azure AI & Microsoft Research	Yes	Dynamic masking	Medical text classification, entity recognition
PubMedBERT	PubMed abstracts	Microsoft Research & NIH	No	Whole-word masking	Biomedical NLP tasks, domain-specific text analysis

Context: What is Transformer Models?

- A transformer model is a type of deep learning model that was introduced in 2017. These models have quickly become fundamental in natural language processing (NLP) and have been applied to a wide range of tasks in machine learning and artificial intelligence.

Transformer Models introduce two key innovations:

- **Positional Encoding**, which assigns unique numerical positions to tokens to capture sequence order.
- **Self-Attention**, which calculates the relationships between tokens to understand context and importance. These mechanisms allow transformers to process data in parallel while effectively capturing patterns and relationships in sequential data.

Transformer models are used for tasks like **natural language processing**, **computer vision**, and **speech recognition** due to their ability to **capture context**, **understand complex relationships in data**, and **generate high-quality outputs** using self-attention mechanisms and positional encoding.

Context: Types of Transformer models?

1. Encoder-Only Models
Focus: Tasks requiring understanding, such as classification and regression.
Examples: **BERT**, **RoBERTa**, **ALBERT**, **DistilBERT**.

2. Decoder-Only Models
Focus: Generative tasks, such as text generation and summarization.
Examples: **GPT (e.g., GPT-2, GPT-3, GPT-4), OPT, LLaMA.**

3. Encoder-Decoder Models
Focus: Sequence-to-sequence tasks, such as translation and summarization.
Examples: **T5**, **BART**, **MarianMT**, **Pegasus**.

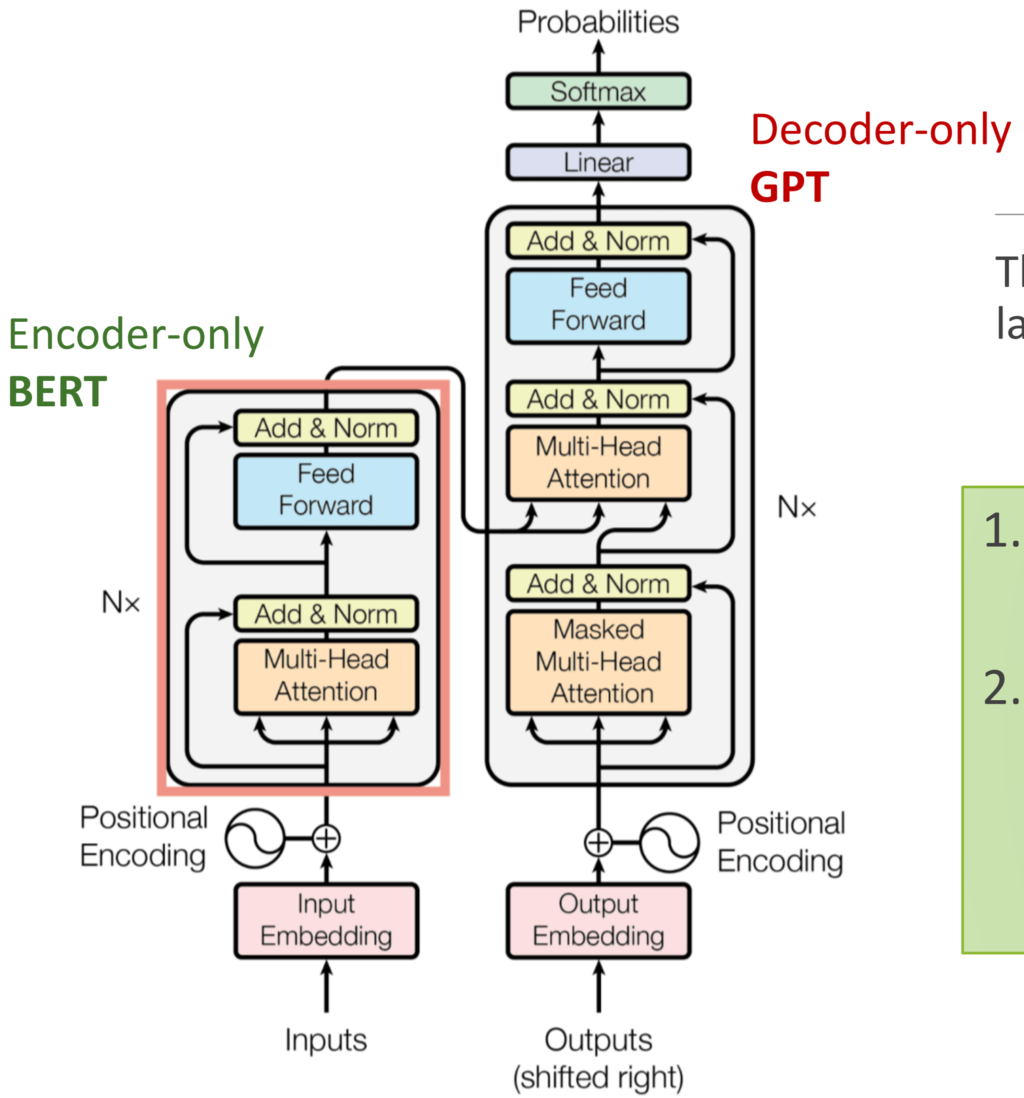
4. Vision Transformers (ViTs)
Focus: Image-related tasks like classification, object detection, and segmentation.
Examples: **ViT, DeiT, Swin Transformer.**

5. Multimodal Transformers
Focus: Integrating multiple types of data (e.g., text, image, video).
Examples: **CLIP, DALL·E, Florence, Flamingo.**

6. Specialized Transformers
Designed for specific domains or tasks(based on Encoder):
BioBERT, ClinicalBERT: Biomedical and clinical tasks.
CodeBERT, Codex: Programming language understanding and code generation.
Graph Transformers: Graph-based data tasks.
(we are using this type of Transformer Model)

Transformer Architecture: focused on Encoder

The structure of the Encoder of Transformer Model



The encoder consists of a stack of N = 6 identical layers, where each layer is composed of two sublayers:

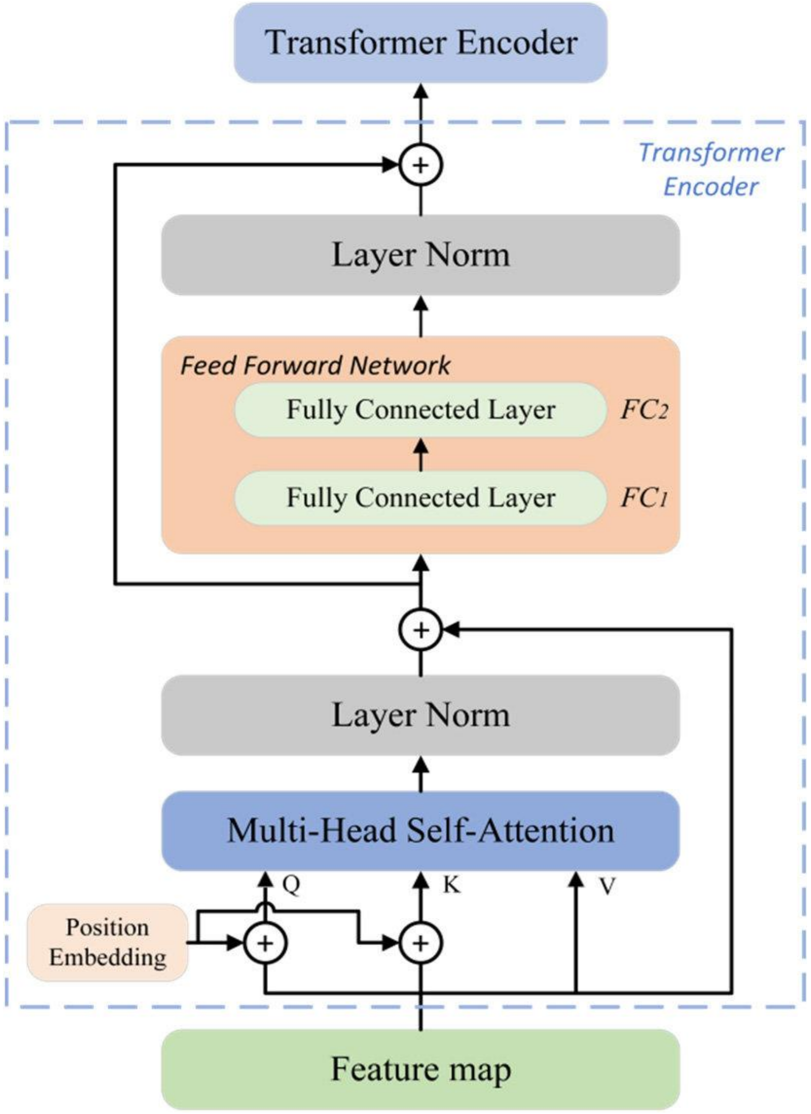
1. The first sublayer implements **a multi-head self-attention mechanism**.
2. The second sublayer is **a fully connected feed-forward network** consisting of two linear transformations with Rectified Linear Unit (ReLU) activation.

$$\mathrm{FFN}(x) = \mathrm{ReLU}(W_1 x + b_1) W_2 + b_2$$

Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017). https://arxiv.org/abs/1706.03762

Transformer Architecture: focused on Encoder

The inside structure of the Encoder of Transformer Model



1 **Self-Attention:** Enables each word to **attend to all other words** in a sentence, capturing contextual relationships effectively.

2 **Positional Encoding:** Adds position-based values to word embeddings to **retain word order information**, as Transformers process data in parallel.

3 **Multi-Head Attention:** Uses multiple attention heads to **analyze different aspects of the text simultaneously**, improving contextual understanding.

4 **Feed-Forward Layers:** Applies **fully connected layers** to refine word representations and **enhance feature extraction**.

5 **Layer Normalization & Residual Connections: Stabilizes training** by preventing vanishing/exploding gradients and improving model convergence.

6 **Stacked Encoder Layers:** Repeated layers process text at multiple levels, **allowing deeper feature learning for complex NLP tasks**.

Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017). https://arxiv.org/abs/1706.03762

BioBERT Model: Pre-Training dataset.

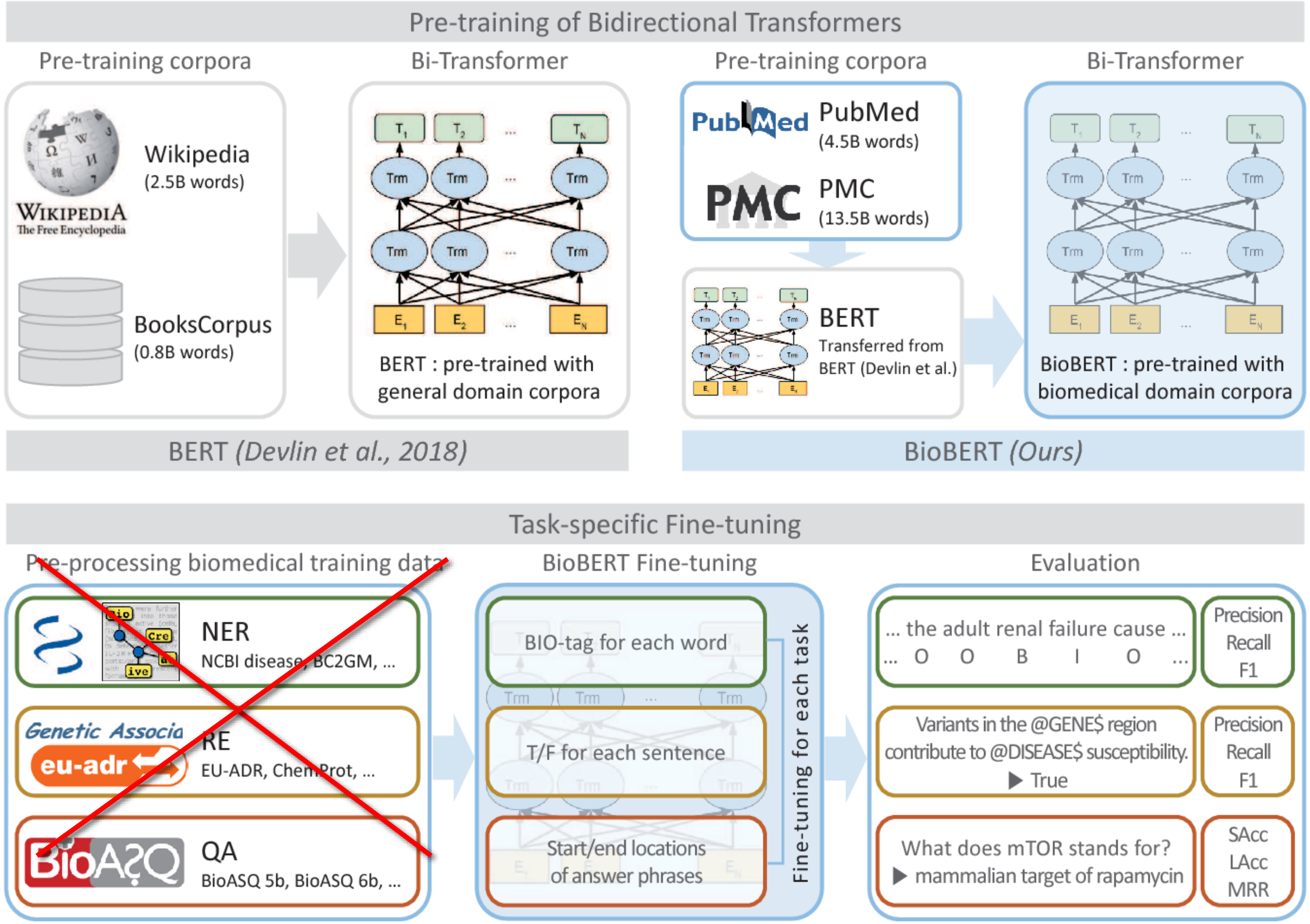
BioBERT (**Biomedical Bidirectional Encoder Representations from Transformers**) is a specialized language model built upon Google's BERT architecture to address the challenges of understanding biomedical texts.

It is pre-trained on a variety of huge general and biomedical text corpora, such as Books Corpus, English Wikipedia, PubMed abstracts (4.5 billion words), and full-text articles from PubMed Central (PMC) (13.5 billion words) shown below.

Model	Pre-Training Corpus	Number of Words	Domain
BERT (Baseline)	Wiki + Books	3.3 billion	General
BioBERT (+ PubMed)	Wiki + Books + PubMed	7.8 billion	Biomedical
BioBERT (+ PMC)	Wiki + Books + PMC	16.8 billion	Biomedical
BioBERT (+ PubMed + PMC)	Wiki + Books + PubMed + PMC	20.3 billion	Biomedical

BioBERT Architecture : (Pre-Training and Fine-Tuning)

BioBERT architecture of Pre-training and Fine-Tuning phase



Here for training data, we are using our manually annotated medical corpus using Doccano Open-source Tool.

Data Description:

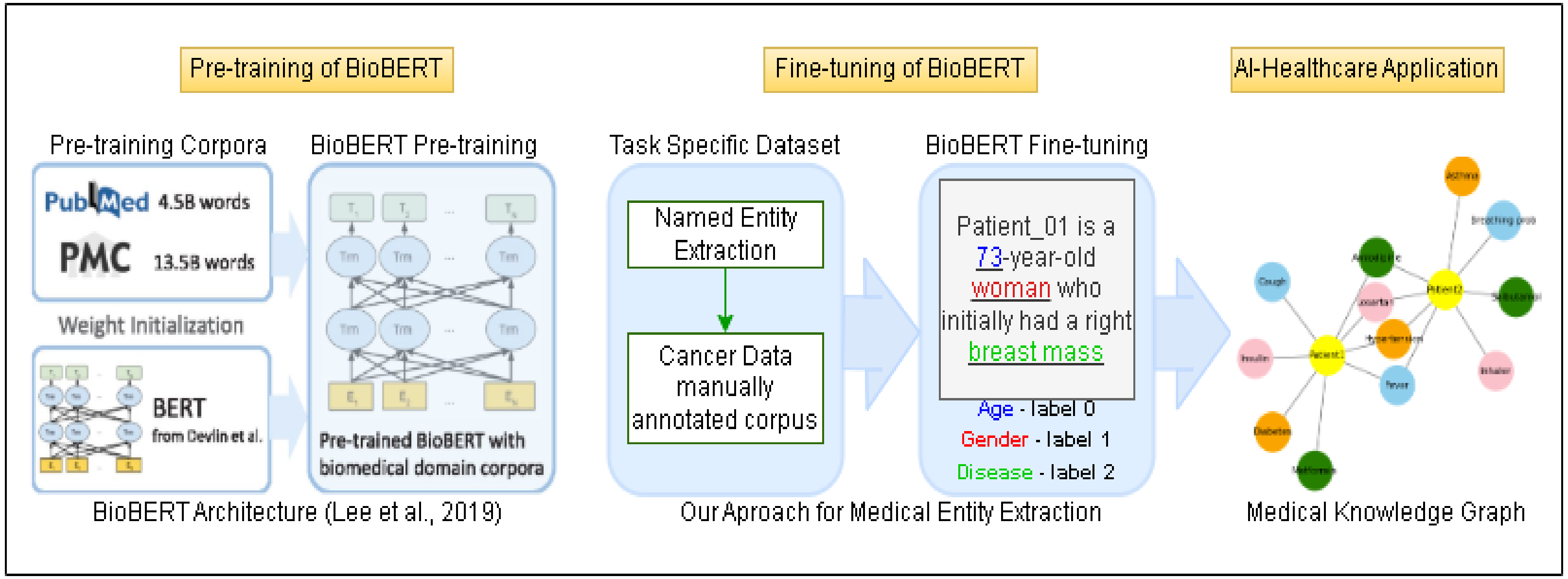
The dataset is consisting of cancer patient data samples was curated by Sushil et al. (Sushil et al., 2024) from the University of California, San Francisco (UCSF) Information Commons in the period from 2012 to 2022, which is available on this website: https://physionet.org/content/curated-oncology-reports/1.0/

M. Sushil, V. E. Kennedy, D. Mandair, B. Y. Miao, T. Zack, and A. J. Butte.1018Coral: expert-curated oncology reports to advance language model inference.1019NEJM AI, 1(4):AIdbp2300110, 2024.1020

Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* 36.4 (2020): 1234-1240.

BioBERT Architecture : (Pre-Training and Fine-Tuning)

BioBERT architecture of Pre-training and our Fine-Tuning phase



Here for training data, we are using our manually annotated medical corpus using Doccano Open-source Tool.

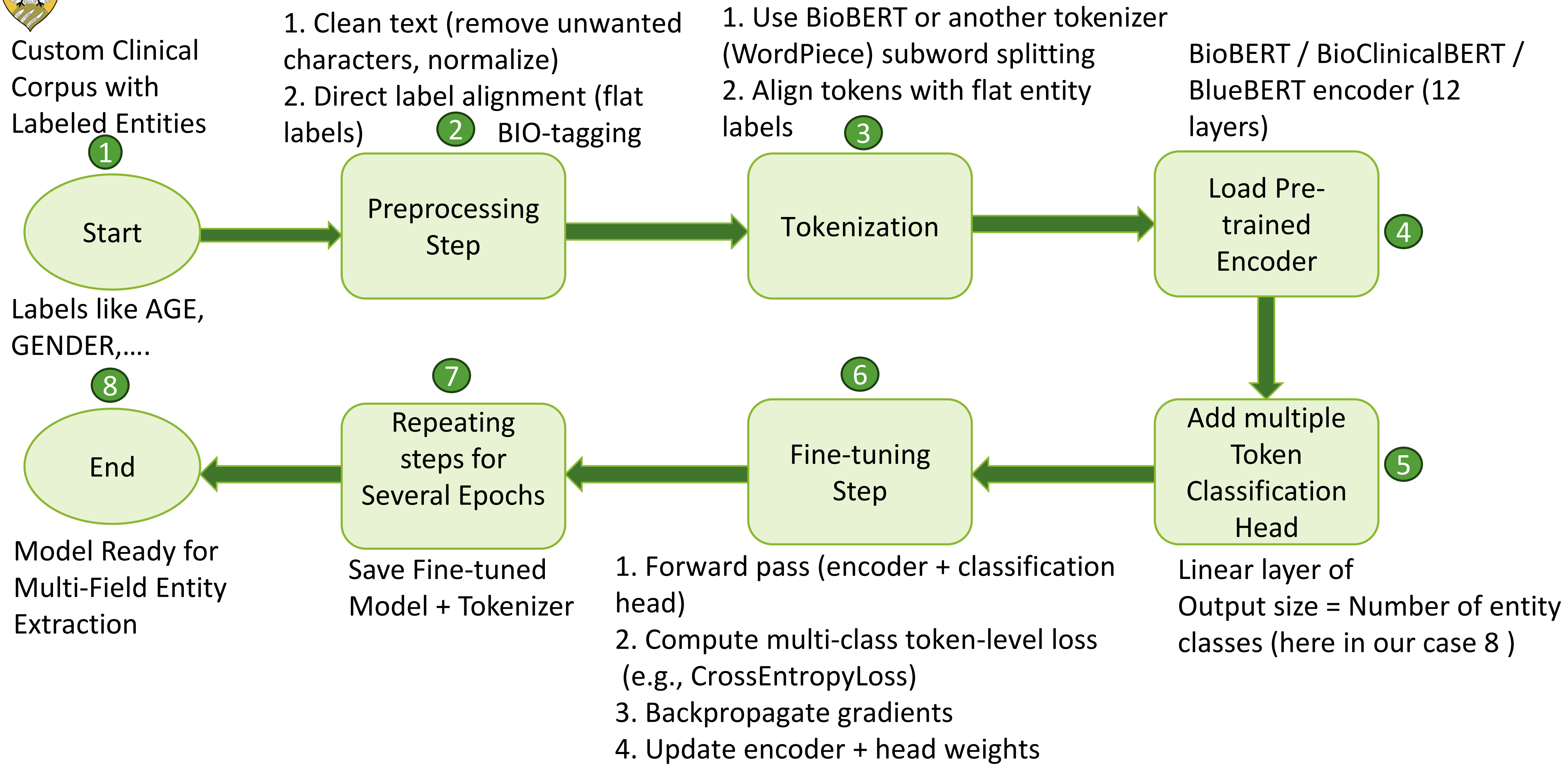
Data Description:

The dataset is consisting of cancer patient data samples was curated by Sushil et al. (Sushil et al., 2024) from the University of California, San Francisco (UCSF) Information Commons in the period from 2012 to 2022, which is available on this website: https://physionet.org/content/curated-oncology-reports/1.0/

M. Sushil, V. E. Kennedy, D. Mandair, B. Y. Miao, T. Zack, and A. J. Butte.1018Coral: expert-curated oncology reports to advance language model inference.1019NEJM AI, 1(4):AIdbp2300110, 2024.1020

Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." ***Bioinformatics*** **36.4 (2020): 1234-1240.**

Flowchart: Fine-tuning BioBERT (or similar) on Custom Clinical Entity Labels



Goal of our research work:

The research focuses on **medical entity extraction** using pre-trained and fine-tuned BERT models to identify cancer-related entities such as diseases, symptoms, medications, and medical history from unstructured clinical text.

A **knowledge graph** is then constructed to map relationships between these entities, enabling pattern identification and insights from patient data. Finally, the knowledge graph is used to **provide medication and dosage suggestions** based on patterns from past cases, supporting personalized cancer treatment strategies. **(future work)**

The main focus of the Research work is:

- **Understanding Pretrained variant of BERT Models and Domain-Specific Data.**
- **Cleaning Noisy Cancer Clinical Text Data.**
- **Preparing Manually Annotated Corpora.**
- **Ensuring Ground Truth Reliability.**
- **Fine-Tuning Pretrained BERT Models.**
- **Evaluating Fine-Tuned Models for Entity Extraction.**

We did an evaluation of 5 specialized Transformer Models as shown before.

Clinical Text Data selection:

In this work, we primarily focus on the technical aspects of processing and analyzing clinical text data to enhance language model inference. Using the **CORAL: Expert-Curated Medical Oncology Reports to Advance Language Model Inference** dataset, focusing on the **unannotated raw clinical text data** available within the "coral" folder. This dataset comprises **unstructured clinical text notes** that lack manual annotations, making them a rich source of natural language data for computational processing. These notes include detailed narratives on patient medical histories, symptoms, diagnoses, medications, and demographic information, written by healthcare professionals.

The unannotated data is divided into two subfolders, each containing notes related to **breast cancer** and **pancreatic cancer**. We took both data sets for building the annotated corpus for ground truth.

Inter-Annotator Agreement (IAA) scores and Cohen's Kappa for various annotation tasks.

Cohen Kappa's value= $\kappa = \dfrac{P_o - P_e}{1 - P_e}$

Po = the observed agreement

Pe = the expected agreement

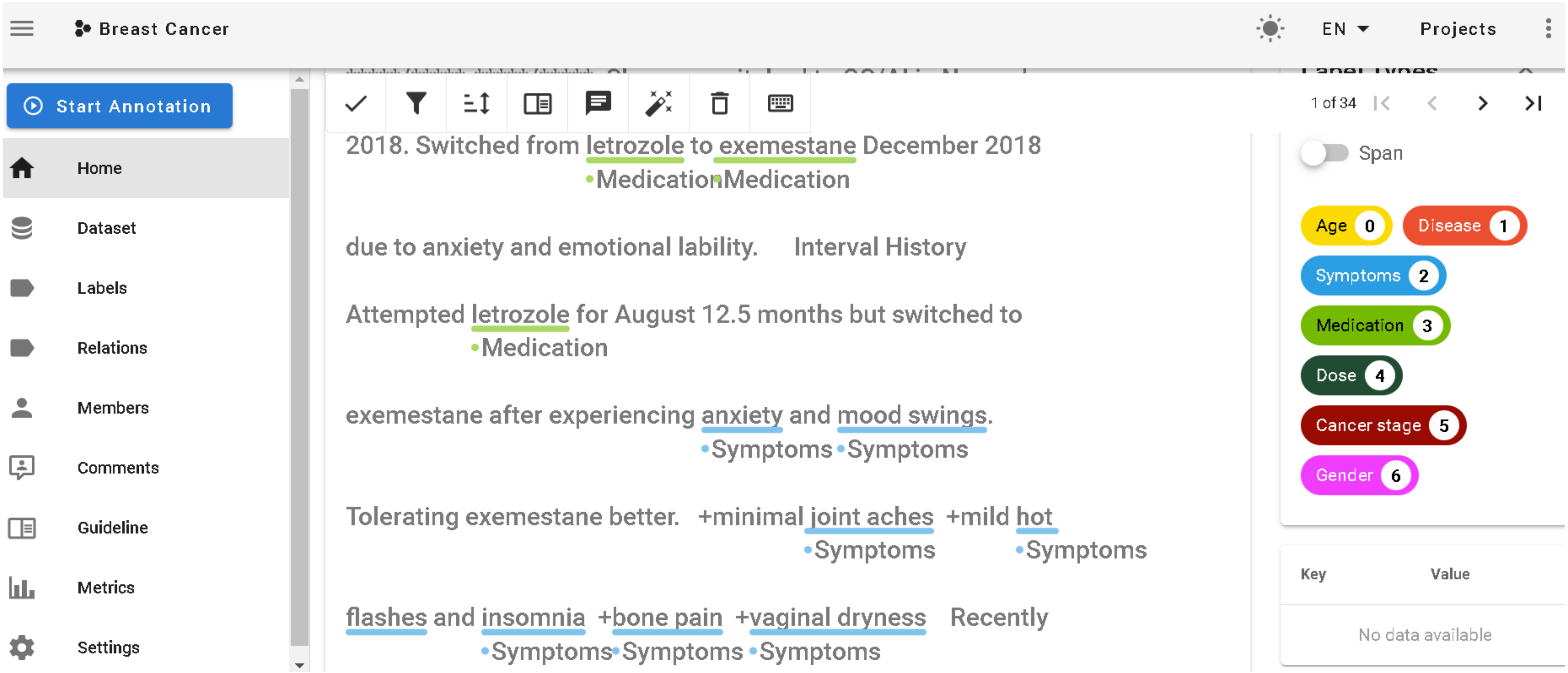
$$\text{Accuracy} = \frac{\text{Correct Annotations by Annotator}}{\text{Total Annotations}} \times 100$$

Annotation Task	Annotator 1 (%)	Annotator 2 (%)	Cohen's Kappa	Interpretation
Age Annotation	94%	93%	0.85	Almost Perfect Agreement
Gender Annotation	96%	95%	0.90	Almost Perfect Agreement
Disease Annotation	95%	94%	0.89	Almost Perfect Agreement
Symptom Annotation	89%	91%	0.76	Substantial Agreement
Medication Annotation	92%	88%	0.64	Substantial Agreement
Dose Annotation	85%	87%	0.57	Moderate Agreement
Medical History Annotation	84%	86%	0.56	Moderate Agreement
Cancer Stage Annotation	90%	92%	0.80	Substantial Agreement

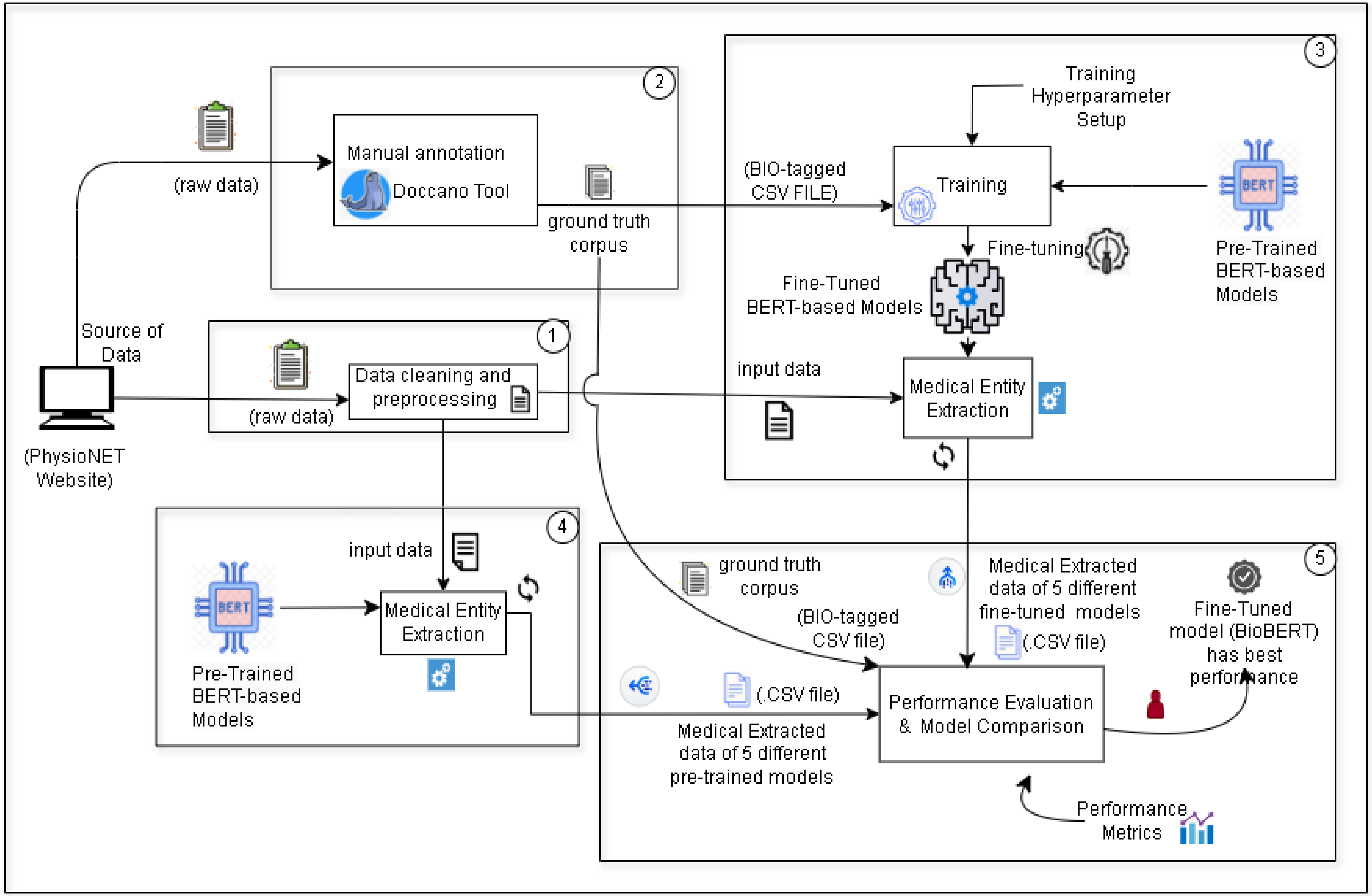
$$\text{Agreement Percentage} = \frac{\text{Matching Annotations between Annotator 1 and Annotator 2}}{\text{Total Annotations}} \times 100$$

M. Sushil, V. E. Kennedy, D. Mandair, B. Y. Miao, T. Zack, and A. J. Butte.1018Coral: expert-curated oncology reports to advance language model inference.1019NEJM AI, 1(4):AIdbp2300110, 2024.1020

Open-source Annotation Tool (Doccano):



The architecture of our Proposed methodology



Proposed Methodology

- Receiving Data from Source
- Preprocessing Data
- Manual Annotation
- Manually Annotated Medical Corpus for Fine-Tuning
- Fine-Tuning BERT-Based Models
- Medical Entity Extraction Using Pre-Trained Models
- Medical Entity Extraction Using Fine-Tuned Models
- Comparing Results Using Evaluation Metrics

Research Questions

RQ1: How can transformer BERT-based models be used to extract medical entities as knowledge from raw clinical text data?

RQ2: What criteria should be considered when selecting the best-performing BERT-based models for medical entity extraction?

RQ3: Which BERT-based approach, based on selected criteria, is the best for knowledge extraction in the health domain?

Answer to RQ1:

RQ1: How can transformer BERT-based models be used to extract medical entities as knowledge from raw clinical text data?

1. Investigate **transformer BERT-based models** for **Named Entity Recognition (NER)** in clinical text.
2. **Contextual embeddings** to extract **medical entities** using **(bidirectional context** and **self-attention mechanisms**).
3. **Convert unstructured text into structured data.**
4. Identify Essential medical terms include **diseases, symptoms, medications with doses, and medical history** based on contextual understanding.

Experimental Settings of our approach

Component	Description
Data Source	Research-based (Sushil et al., 2024) clinical text datasets (e.g., BC5CDR, MIMIC-III).
Preprocessing	Tokenization, stopword removal, abbreviation expansion, annotation using Doccano tool.
Annotation Tool	Doccano for manual medical entity labeling.
Models Used	BioBERT, ClinicalBERT, PubMedBERT, BlueBERT, RoBERTa.
Training Strategy	Fine-tuning pre-trained models using annotated medical corpus.
Evaluation Metrics	Precision, Recall, F1-score, Accuracy.
Comparison	Pre-trained vs. fine-tuned models for medical entity extraction.
Implementation Tools	Python, SpaCy, TensorFlow/PyTorch, SciSpacy, Hugging Face Transformers.
Data Split Strategy	70% for training, 15% for validation, and 15% for testing.
Annotation Guidelines	Custom-labeled medical corpus with 8 entity labels for multi-head entity extraction.
Pre-training Corpora	PubMed abstracts, PMC full-text articles, MIMIC-III clinical notes.

Hyperparameters used for training the models.

Hyperparameter	Value	Hyperparameter	Value
Seed	42	Batch size buffer	256
Epochs	10	Discard oversize batches	True
Dropout	0.1	Learning rate scheduler	Warmup-linear
Optimizer	AdamW	Initial learning rate	5e-5
GPU allocator	PyTorch	Total training steps	1,500 (approx)
Batch size	16-32	Warmup steps	1200

Experimental Settings for our research

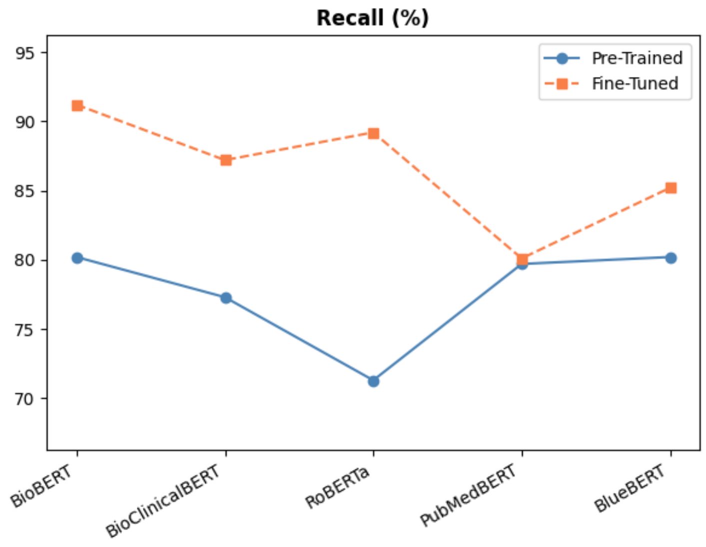
Answer to RQ2:

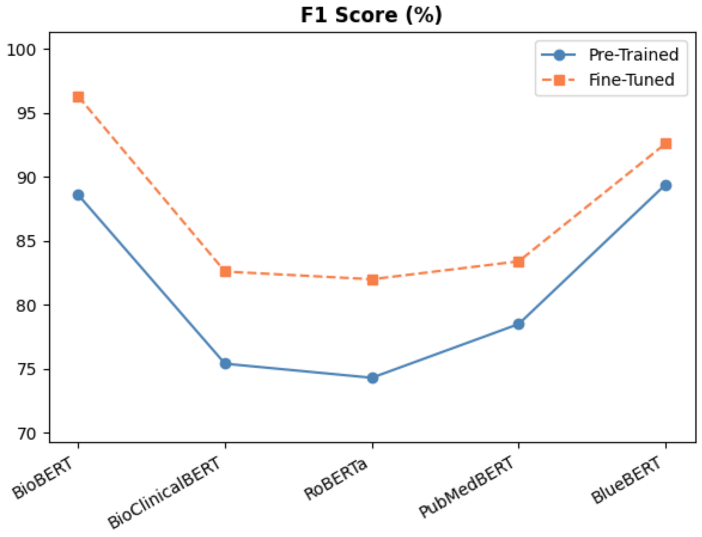
RQ2: What criteria should be considered when selecting the best-performing BERT-based models for medical entity extraction?

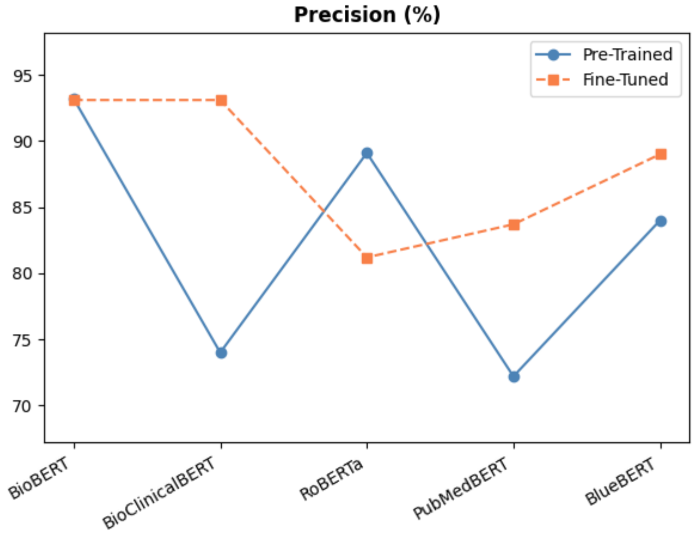
1. **Effectiveness Criteria:** The evaluation focused on **contextual embedding quality, NER accuracy, and domain adaptability** to medical texts.

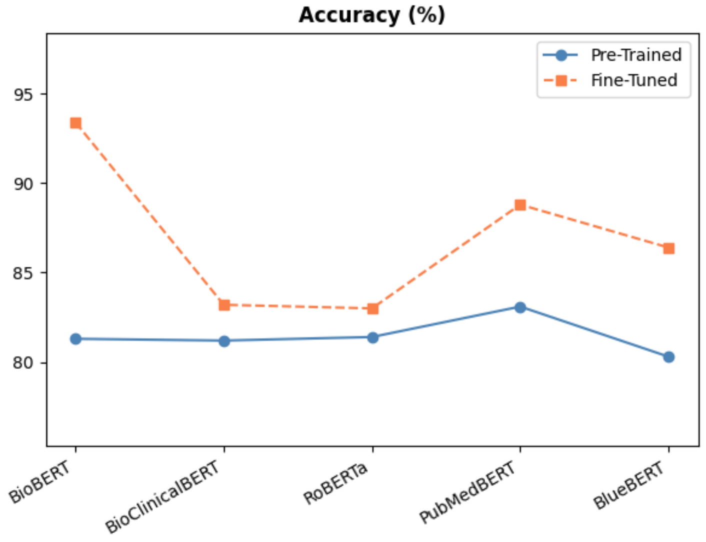
2. **Performance Evaluation:** accuracy, recall, F1-score, and precision.

Evaluation Metrices for all BERT-based models









Answer to RQ3:

RQ3: Which BERT-based approach, based on selected criteria, is the best for knowledge extraction in the health domain?

1. **Comparison Approach:** Accuracy / F1-Score
2. **Key Findings: PubMedBERT** showed strong accuracy among pre-trained models also pre-trained **BlueBERT** has good F1-Score than **PubMedBERT**, while **BioBERT** outperformed other fine-tuned models.
3. **Conclusion: BioBERT** was selected as the best model for future research on **constructing a knowledge graph for pattern matching**.

The result of our experiment

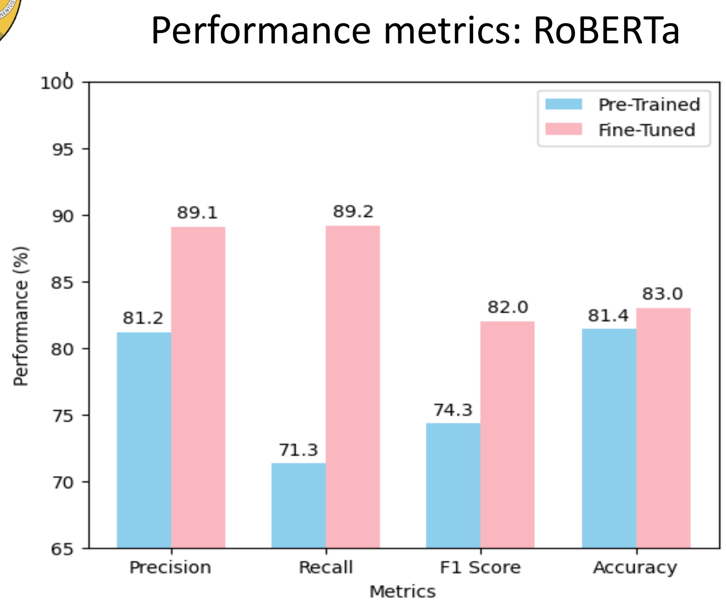
The overall performance matrix of pre-trained BERT models

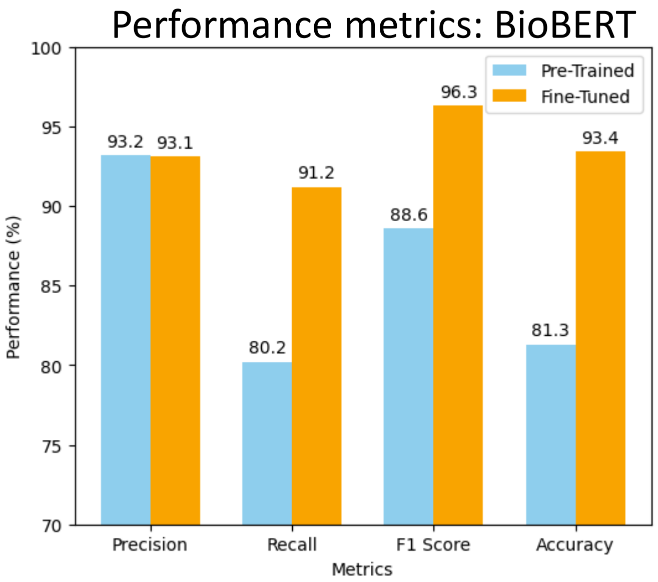
Model Names	Recall (%)	F1 Score (%)	Precision (%)	Accuracy (%)
BioBERT	80.2	88.6	93.2	81.3
BioClinicalBERT	77.3	75.4	74.0	81.2
RoBERTa	71.3	74.3	81.2	81.4
PubMedBERT	79.7	78.5	77.2	83.1
BlueBERT	80.2	89.4	84.0	80.3

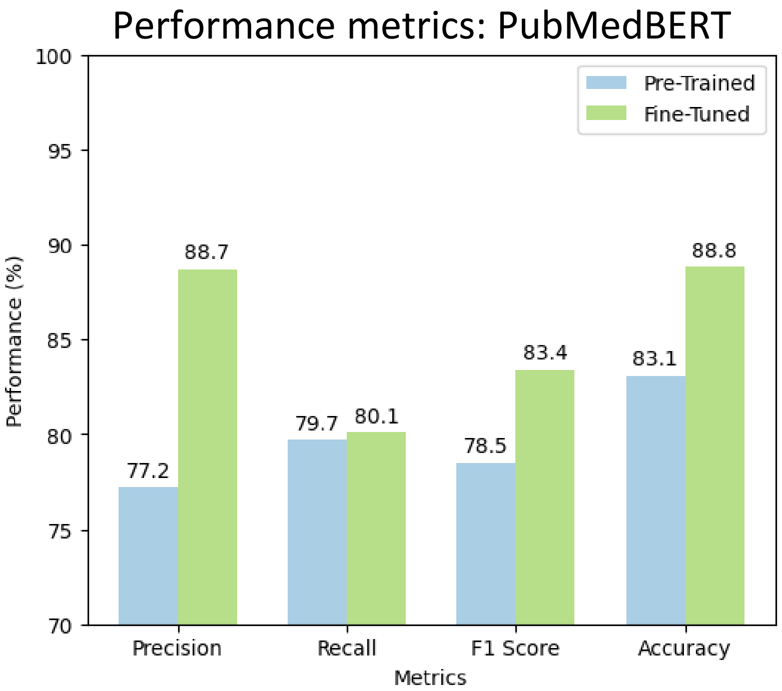
The overall performance matrix of fine-tuned BERT models.

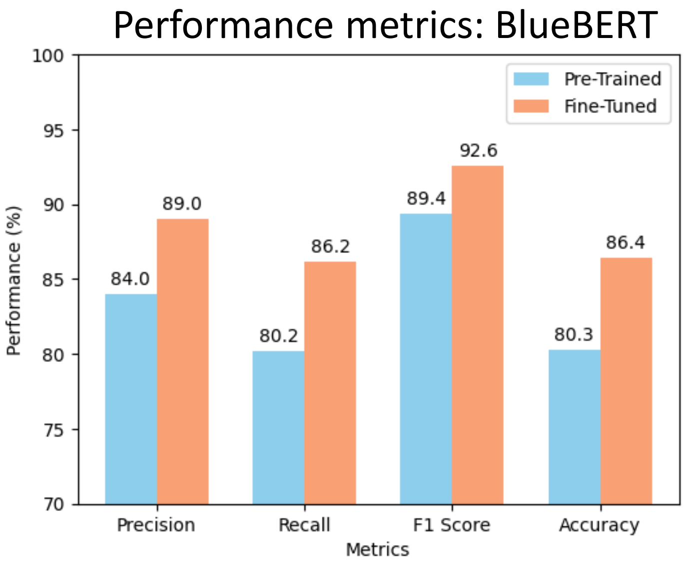
Model Names	Recall (%)	F1 Score (%)	Precision (%)	Accuracy (%)
BioBERT	91.2	96.3	93.1	93.4
BioClinicalBERT	87.2	82.6	93.1	88.2
RoBERTa	89.2	82.0	89.1	83.0
PubMedBERT	80.1	83.4	88.7	88.8
BlueBERT	86.2	92.6	89.0	86.4

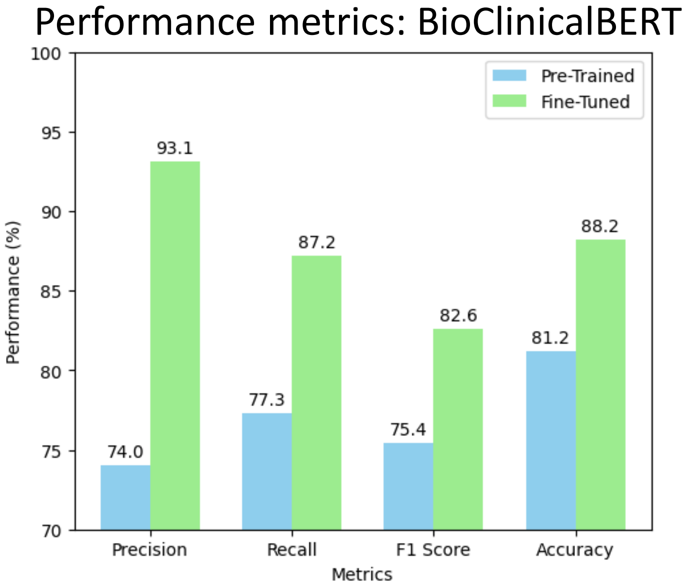
Comparison of Five Different Models Based on Performance Metrics



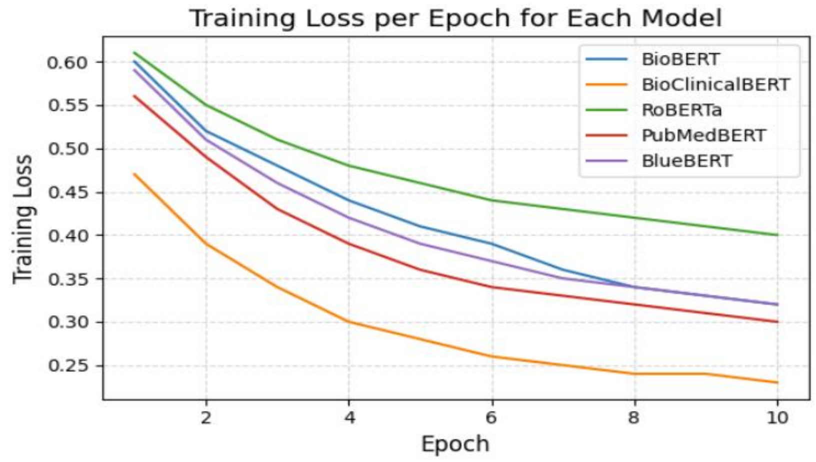


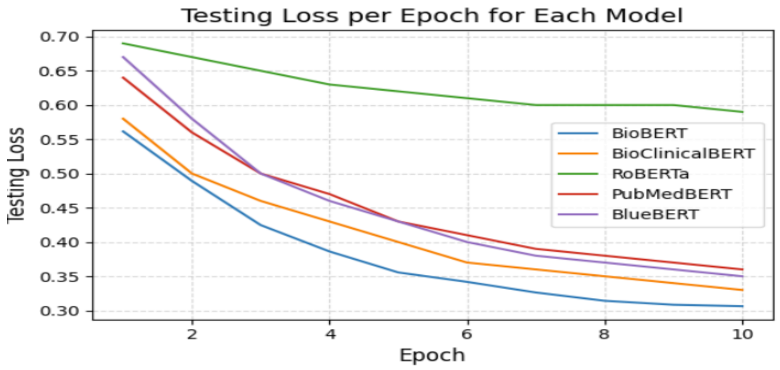


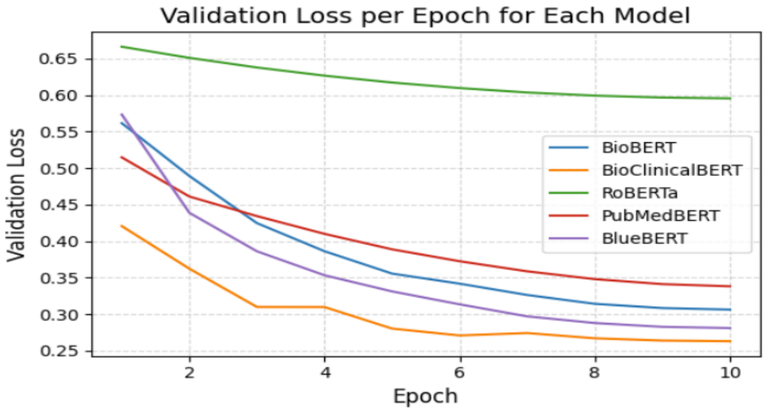




Line graph of Training, Validation and Testing loss of Pre-trained Models while training



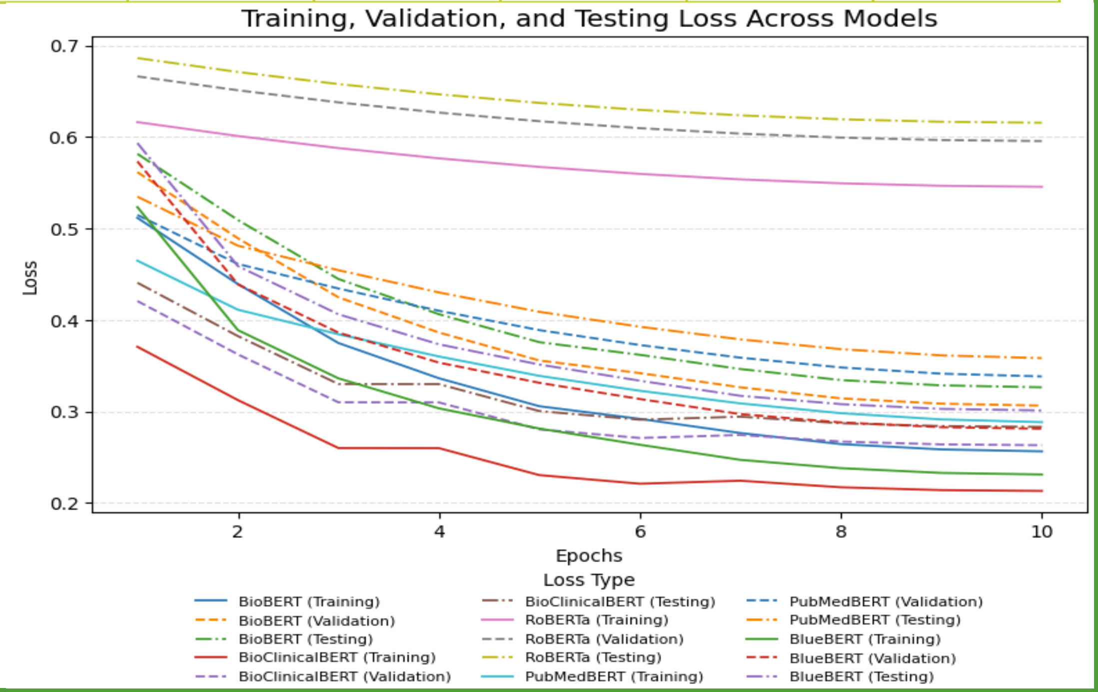




The line graph illustrates the loss variation across epochs

Model	(Epoch 1)	Epoch 2	Epoch 3	Epoch 4	Epoch 5	Epoch 6	Epoch 7	Epoch 8	Epoch 9	Epoch 10
BioBERT	0.5615	0.4892	0.4247	0.3862	0.3556	0.3418	0.3263	0.3142	0.3084	0.3063
BioClinicalBERT	0.4206	0.3623	0.3099	0.3098	0.2803	0.2709	0.2742	0.2670	0.2639	0.2630
RoBERTa	0.6662	0.6510	0.6378	0.6266	0.6172	0.6096	0.6036	0.5993	0.5966	0.5955
PubMedBERT	0.5147	0.4611	0.4343	0.4100	0.3888	0.3725	0.3587	0.3480	0.3412	0.3382
BlueBERT	0.5733	0.4388	0.3861	0.3533	0.3312	0.3135	0.2970	0.2879	0.2827	0.2810

This table shows the decreased values of validation loss in each epochs .
the validation loss was used to detect potential overfitting and to select the best-performing model checkpoint.



Research Impact and Novelty:

- Enhanced Clinical Decision-Making:
- Focused on Breast Cancer with other cancer data:
- Integration of Knowledge Graphs for Pattern Recognition: (Therapy Suggestions)

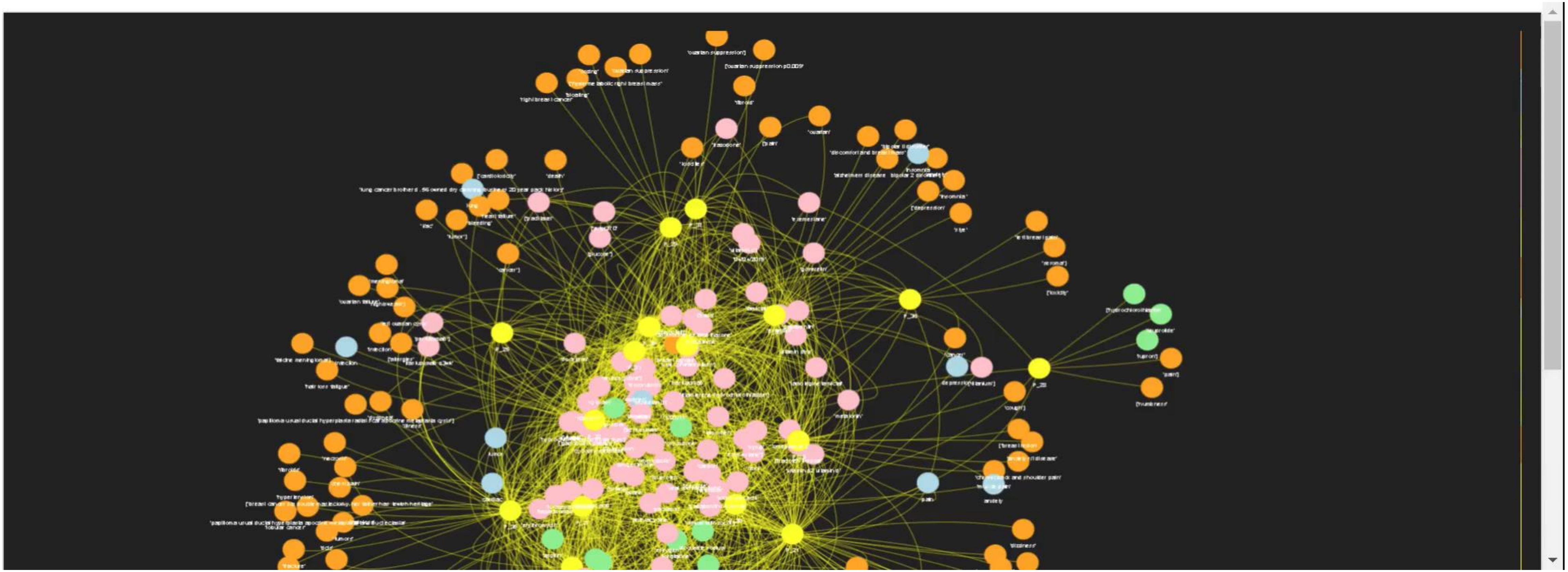
Conclusion:

This study represents a meaningful advancement in extraction knowledge of breast cancer by transforming unstructured clinical text data into well-organized, accessible information using AI-powered techniques. By extracting key details—such as age, gender, diseases, symptoms, medications, dosages, medical histories, and cancer stages—we've made it easier for healthcare professionals to access and use this data effectively. This approach supports better decision-making, enabling care that is more precise and patient-focused.

Future work:

Looking to the future, our work will focus on building medical knowledge graphs and using pattern-matching techniques to explore the relationships between patient histories, including diseases, symptoms, and medications with their dosages.

Knowledge Graph Represent for our future work (Therapy Suggestion)



PatientID = Yellow nodes, Disease = Orange nodes, Symptoms = Sky Blue nodes, Medication = Green nodes, Suggested Medication = Pink nodes.

Acknowledgement

This work is supported by European Union — NextGenerationEU — National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Re silienza, PNRR) — Project: “SoBigData.it — Strengthening the Italian RI for Social Mining and Big Data Analytics” — Prot. IR0000013 — Avviso n. 3264 del 28/12/2021 and by LIFEMAP-Dalla patologia pediatrica alle malattie cardiovascolari e neoplastiche nell’adulto: mappatura genomica per la medicina e prevenzione personalizzata Traiettoria 3 “Medicina rigenerativa, predittiva e personalizzata” - Linea di azione 3.1 “Creazione di un programma di medicina di precisione per la mappatura del genoma umano su scala nazionale” of the Ministry of Health.

References

[1] O. Solarte-Pabón, O. Montenegro, A. García-Barragán, M. Torrente, M. Provencio, E. Menasalvas, and V. Robles. Transformers for extracting breast cancer information from spanish clinical narratives. Artificial Intelligence in Medicine, 143:102625, 2023.

[2] N. Srinivashini and R. Lavanya. False positive reduction in mammographic mass detection using image representations for textural analysis. In 2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP), pages 1–6. IEEE, 2021.

[3] M. Sushil, V. E. Kennedy, D. Mandair, B. Y. Miao, T. Zack, and A. J. Butte. Coral: expert-curated oncology reports to advance language model inference. NEJM AI, 1(4):AIdbp2300110, 2024.

[4] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim. Roberta-lstm: a hybrid model for sentiment analysis with transformer and recurrent neural network. IEEE Access, 10:21517–21525, 2022.

[5] C. Zhang and X. Cao. Biological gene extraction path based on knowledge graph and natural language processing. Frontiers in Genetics, 13:1086379,5902023.

[6] D. Mamakas, P. Tsotsi, I. Androutsopoulos, and I. Chalkidis. Processing long legal documents with pre-trained transformers: Modding legalbert and longformer. arXiv preprint arXiv:2211.00974, 2022.

[7] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de La Clergerie, D. Seddah, and B. Sagot. Camembert: a tasty French language model. arXiv preprint arXiv:1911.03894, 2019.

[8] J. Mozafari, A. Fatemi, and P. Moradi. A method for answer selection using distilbert and important words. In 2020 6th International Conference on Web Research (ICWR), pages 72–76. IEEE, 2020.

[9] F. W. Mutinda, K. Liew, S. Yada, S. Wakamiya, and E. Aramaki. Automatic data extraction to support meta-analysis statistical analysis: a case study on breast cancer. BMC Medical Informatics and Decision Making,54622(1):158, 2022.

[10] A. Basaad, S. Basurra, E. Vakaj, A. K. Eldaly, and M. M. Abdel samea. A bert-gnn approach for metastatic breast cancer prediction using histopathology reports. Diagnostics, 14(13):1365, 2024.

Thank you for your attention