National College of Ireland

MedPredict-DeBERTa: Enhancing Symptom-Based Disease Prediction through UMLS-Integrated Disentangled Attention

Master of Science in Artificial Intelligence

Karthik Uppalapati
Student ID: 23289961

School of Computing
National College of Ireland

Supervisor: Dr. Anu Sahini

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Karthik Uppalapati

Student ID: 23289961

Programme: Master of Science in Artificial Intelligence **Year:** 2025-2026

Module: Research Practicum

Supervisor: Dr. Anu Sahini

Submission Due Date: 24/01/2026

Project Title: MedPredict-DeBERTa: Enhancing Symptom-Based Disease Prediction through UMLS-Integrated Disentangled Attention

Word Count: 9280 **Page Count :** 25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: U.Karthik

Date: 24/01/2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

MedPredict-DeBERTa: Enhancing Symptom-Based Disease Prediction through UMLS-Integrated Disentangled Attention

Karthik Uppalapati
23289961

Abstract

The issue of accurate prediction of diseases through symptoms continues to be a major challenge in healthcare. This paper explores how Unified Medical Language System (UMLS) can be integrated with DeBERTa transformer architecture to enhance prediction of diseases according to symptoms of patients. An inclusive pipeline was designed based on the DDXPlus data comprising of 1,025,602 patient records of 49 diseases. This pipeline incorporates UMLS attributes into symptom text including 140 identifiers of concepts, 38 semantic types and 27 critical medical qualifiers. Five models at the baseline and three variants of UMLS integration based on concatenation, gated fusion and cross-attention mechanisms were tested. The findings indicate that the baseline has over 97 percent accuracy. The distilbert variant incorporated into the UMLS has 97.60, which is higher than the DeBertAt baseline of 97.47. The research found the meaning of shallow integration problem, where large performing baselines fail to utilize additional knowledge characteristics completely. Other notable works include a UMLS extraction pipeline, which maintains qualifiers, a comparative study of integration architectures in medical knowledge, and defining under which circumstances medical knowledge integration is valuable. The results indicate that traditional architectural constraints make the use of the knowledge that exists outside to be more successful on small models, and that architectural restrictions are required to make the domain knowledge useful to the systems with the best performance.

1 Introduction

Effective and reliable prediction of diseases according to patient-reported symptoms is a critical issue in the health care of today. The ability to correctly classify diseases according to the symptomatology has a direct effect on treatment decisions, allocation of available resources, and patient outcomes. Although traditional diagnosis is based on clinical expertise and pattern identification, errors in diagnosis occur in an outpatient environment with an estimated number of about 12 million cases every single year (Hassan et al., 2024). Deep learning architectures, artificial intelligence, in general, are promising in terms of creating automated diagnostic support systems.

The latest developments in the field of natural language processing have proved to be highly skilled in the analysis of medical texts. The architectures based on transformers, in particular, BERT, and its variations have demonstrated higher results on text classification of medical texts (Lee et al., 2025). The disentangled attention mechanism in the recent study by He et al. (2020), known as DeBERTa (Decoding-enhanced BERT with Disentangled Attention), is an important architectural improvement in the sense of separating content and positional features, which the authors characterize as a disentangled attention mechanism. Although this

architecture has demonstrated improved performance with regard to general language benchmarks, its medical uses are not fully exploited. At the same time, the Unified Medical Language System (UMLS) is a multifaceted collection of standardized medical terminology, semantic relations, and conceptual hierarchies that contain decades of medical knowledge (Rasmy et al., 2020). Yet, the combination of these technologies to predict diseases based on the symptoms has not been thoroughly researched.
This research paper fills the research gap between developed transformer architectures and organized medical knowledge. The existing transformer-based medical classification systems do not generally use external medical ontologies, lacks pre-existing symptom-disease relationships. Moreover, the current systems tend to lose essential medical qualifiers such as severity, timing features as well as quality descriptors that play an important role in the differential diagnosis. These render the algorithm less predictive and less clinically interpretable, which makes it less useful in the real world.

The study question that will inform the proposed study is the following: How does the combination of UMLS medical knowledge with DeBERTa achieve a higher accuracy and interpretability in disease prediction compared to conventional transformers-based systems of symptom-based disease classification?

Four research objectives were established: First, developing a data processing pipeline that transforms raw patient symptom data into UMLS-augmented representations while preserving medical qualifiers. Second, implementing and evaluating baseline transformer models including SVM, BERT, BioBERT, and DeBERTa for disease prediction. Third, developing three UMLS integration variants employing different fusion strategies: concatenation, gated fusion, and cross-attention. Fourth, conducting comprehensive comparative evaluation to assess the impact of UMLS integration versus text-only baselines. Success criteria include demonstrating medical knowledge feature utilization in predictions and achieving statistically significant accuracy improvements.

This work has a number of scientific implications to literature. It presents empirical data regarding the integration of structured medical information with transformer networks in disease prediction. The study reports the issue of the shallow integration in which baseline models with high performance are not able to effectively exploit the add-on knowledge properties. It also defines a systematic assessment system of integration strategies and offers repeatable ways of extracting and augmenting UMLS features.

The report is presented as follows: Section 2 critically evaluates the research conducted on the topic of transformers based medical text classification, medical knowledge integration and machine learning based on prediction of symptoms. Section 3 outlines the methodology such as description of dataset, UMLS extraction and experimental setup. Section 4 contains design specifications of the proposed and baseline architectures. Section 5 talks of implementation and technical set up. Section 6 includes the statistical analysis of experimental results. Section 7 addresses the results of research in relation to the research question and literature. Section 8 contains conclusions, limitations, and research directions.

2 Related Work

Natural language processing, medical informatics, and machine learning are some of the aspects involved in the creation of automated disease prediction systems. Transformer-based medical text classification, medical knowledge integration approach and machine learning

approach to predict disease based on symptoms are critically reviewed and reveal the gaps in the research that this study will fill.

2.1 Transformer-Based Approaches for Medical Text Classification

The NLP models based on the transformer architecture have revolutionized the healthcare NLP and the BERT and other variations have shown outstanding results in terms of medical text classification. Hassan et al. (2024) designed MCN-BERT (Medical Concept Normalisation BERT), which works on 99.58 percent accuracy in symptom-disease prediction by medical concept normalisation. Although the use of standard medical terminology is demonstrated in the framework of the transformer, the analysis of 1,200 cases only is an issue of generalizability.

He et al. (2020) presented DeBERTa with its disentangled attention mechanism where the content and positional information is independently processed by different encoding matrices. This architecture was able to outperform human performance on SuperGLUE benchmarks and deal with the BERT drawback of combining content and position information in a single embedding vector. Nevertheless, the original research did not test medical domains applications and the clinical text processing possibilities were not investigated.

Wang et al. (2023) were using DeBERTa-v3-large to entailment of clinical trials and its F1 score was 0.764 by applying fine-tuning methods. Although the study showed the medical applicability of DeBERTa, it entirely depends on the learned patterns and no external medical information was involved, which implies the possibility of improvement with the help of a structured ontology.

Lee et al. (2025) introduced Entity-enhanced BERT (E-BERT) that had entity embedding layers and entity-aware attention mechanisms that reached an accuracy of 99.19 on the specialty prediction of medicine. The architecture also shows that explicit incorporation of medical entities enhances classification but it has the limitation that it relies on the pre-identified entities whose recognition depends on the accurate named entity recognition as a precondition.

Lee et al. (2020) presented BioBERT, which is pre-trained with PubMed abstracts and PMC articles, which is better than general-domain BERT on biomedical text mining tasks. BioBERT, unlike medical ontologies, does not explicitly apply domain-specific vocabulary and representations, may not include relationships missing in the training data.

2.2 Medical Knowledge Integration and UMLS

The Unified Medical Language System is a system of common biomedical concepts, semantic relationships, and hierarchies in the various fields of medicine. In a systematic assessment of UMLS versus the other medical terminologies to predict clinical events using electronic health records, Rasmy et al. (2020) showed 5-12% accuracy improvements compared to other coding systems in clinical prediction. This empirical confidentiality makes UMLS useful in machine learning albeit with computational overhead of concept mapping is difficult in large datasets.

The methodology proposed by Raza and Schwartz (2023) is based on the NLP development of disease databases to extract and standardize unstructured clinical data. The standardization of clinical sources by using UMLS concept unique identifiers (CUIs) showed the possibility of

automated UMLS integration in a large-scale medical data processing setting. Nevertheless, the emphasis on information mining did not investigate the effect of UMLS feature on the classification performance.

The issue of medical context in the integration of knowledge is not easy. In their study, Luo et al. (2021) built a deep language model to extract clinical text symptoms, and they were able to recognize COVID-19 symptoms on social media. The model used attention mechanisms to find the span of text that is relevant to the symptoms and retained contextual qualifiers such as severity and time aspect that is important in diagnostic reasoning. The reliance of the approach on manually annotated training data is a weakness as it prevents its applicability to new types of symptoms.

In a study by Malden et al. (2022), the NLP was used to describe the COVID-19 symptoms in 350,000 patients and showed that the NLP-based extraction was more complete than the usage of the structured diagnostic codes. This observation highlights the significance of integrating structured bodies of knowledge with unstructured text processing to have a complete representation of a symptom although the authors did not analyze the effect of improved symptom characteristics on prediction activities.

2.3 Machine Learning Approaches for Symptom-Based Disease Prediction

Conventional machine learning approaches are also competitive when it comes to the symptom-based classification of disease, which in turn is accompanied by the right feature engineering. Fuster-Pala et al. (2024) thoroughly tested the optimized classifiers such as the Random Forest, K-Nearest Neighbours, Support Vector Machine, and Feed-Forward Neural Networks. Through thorough preprocessing, grid search optimization and performance-based filtering, a KNN classifier using two neighbours obtained a 99.19% accuracy and 99.03% F1 score. Simple models were more successful in case, appropriately configured, than complex neural structures, hence the complexity of models may not predict better results with well-structured symptom data.

Sogandi (2024) tested the identification of disease symptoms supervised and unsupervised learning and found principles that govern disease symptom-disease correlations using pattern mining. Although rule-based representations can show interpretable relationship discovery useful in clinical application, in real clinical situations, they may be inadequate to represent complex conditional dependencies.

The researchers Sadi et al. (2024) considered automatic differential diagnosis as a sequence of transformer-based multi-label sequence classification, and they have proposed to predict diseases as a ranking problem whose output is probability distributions across diagnoses as opposed to single prediction. Such formulation is consistent with clinical practice in which differential diagnosis is applied to multiple conditions, but it cannot be assessed on synthetic data to determine its applicability in the real world.

Jayasudha et al. (2023) proposed BIMM-BERT that incorporates bidirectional encoder representation with medical domain adaptation, which can be enhanced to achieve greater performance in the combination of bidirectional context modeling and medical vocabulary improvement. Although architectural innovation benefits are achieved with domain-specific alterations, computational demands of large-scale bidirectional models can be a constraint to resource-restrained deployment.

Applications NLP Domain-specific NLP applications have shown promising results. Omar et al. (2024) performed a systematic review of NLP and large language models in the task of diagnosing and predicting infectious diseases and discovered that their use continuously improves when complemented by domain-specific knowledge bases, although the evaluation procedures were not standardized across the literature. Montejo-Raez et al. (2024) reviewed NLP in mental disorder detection and pointed out the difficulties related to the ambiguous nature of the symptoms, time dynamics, and interpretability needs.

2.4 Research Niche and Justification

There are four major gaps in the literature that drive this research. To begin with, DeBERTa has better general language understanding, but systematic comparative study of its medical text classification and knowledge integration is not available. Second, the current UMLS integration strategies involve mainly concept normalization and data extraction without considering its effects on the disease prediction performance. Third, the issue of preserving key medical qualifiers acquired in the course of knowledge integration that is critical in the process of differential diagnosis is poorly covered. Fourth, there is no comparative analysis of fusion strategy to combine transformer representations with structured medical knowledge.

In this study, the gaps are filled in MedPredict-DeBERTa, a systematic combination of UMLS medical knowledge with the disentangled attention mechanism of DeBERTa. The methodology maintains important symptom qualifiers such as severity, time aspects, and quality descriptions and appraises three fusion strategies. The study gives objective facts on the conditions in which medical knowledge integration improves predictive performance, and practical approaches and theoretical explanations to the discipline.

3 Methodology

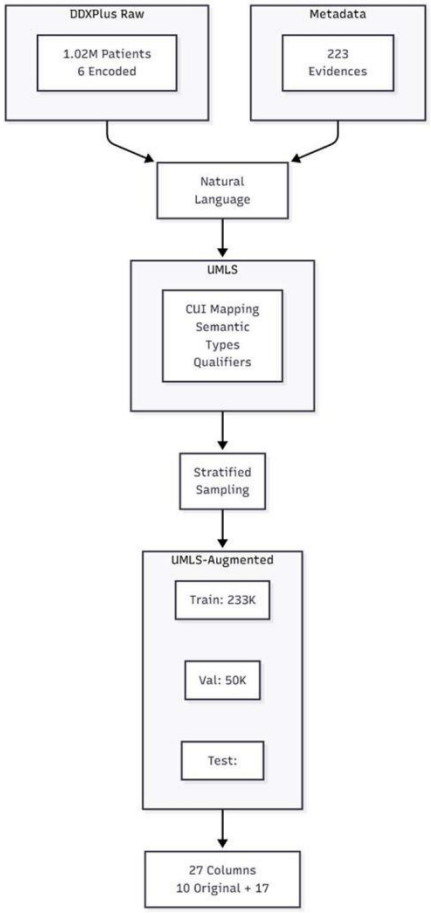

Figure 1: Data Preparation Pipeline showing the transformation from DDXPlus raw data through UMLS integration to final train/validation/test splits.

This section presents a research methodology used to test the integration of UMLS medical knowledge with DeBERTa to predict diseases based on symptoms. The methodology will consist of the selection and description of the dataset, preprocessing of the data, UMLS features extraction, stratified sampling strategy, and subdivision of the dataset. The entire data preparation pipeline is illustrated in Figure 1 which demonstrates how raw patient records are transformed into training data that is compatible with UMLS.

3.1 Dataset Description

The DDXPlus dataset was selected as the primary source of data to be used in this study due to its clinical relevance and extensive coverage of the ground (Fansi et al., 2022). This data is accessible to anyone and was developed as a union of an internal medical body of knowledge and a commercial rule-based automatic diagnosis system. This generated natural patient conditions that can be applied in machine learning studies. The dataset is among the largest publicly available sources of predicting diseases based on symptoms. It allows modelling in a powerful manner. The uncoded data contains 1025 602 patient records. The records contain six coded columns, which contain the demographic, symptoms, and diagnosis information of the patient. The most significant aspects of the DDXPlus dataset, which was used in this study, are summarized in Table 1.

Table 1: DDXPlus Dataset Characteristics

Characteristic	Value
Total Patient Records	1,025,602
Original Columns	6
Disease Classes (Pathologies)	49
Evidence Types	223
Sex Distribution	Male / Female
Geographical Regions	Multiple (including North America)
Data Format	CSV with encoded values
ICD-10 Coverage	100% (all 49 diseases)
Licence	CC-BY

The six original columns of the raw data are called AGE (integer patient age), SEX (encoded gender), PATHOLOGY (label of target disease in French), EVIDENCES (encoded list of symptoms and antecedents), DIFFERentialDIAGNOSIS (probable disease ranked by probability), and INITIALEVIDence (the presenting symptom). There is a confirmed pathology that demonstrates the true diagnosis in each of the patient records accompanied by some evidence and a list of potential diagnoses. As indicated in Table 2, the various categories of evidence found in the data set are classified in three data forms. This distribution indicates how complex the representation of medical symptoms is. The majority of the symptoms may be represented in binary presence/absence indicators, however, some of them require a categorical/ multi-choice answer to provide a more detailed clinical data.

Table 2: Evidence Type Distribution

Evidence Type	Count	Percentage	Description
Binary (B)	208	93.3%	Yes/No symptoms (e.g., "Do you have fever?")
Categorical (C)	10	4.5%	Single-choice from options (e.g., pain intensity 1-10)
Multi-choice (M)	5	2.2%	Multiple selections allowed (e.g., pain locations)
Total	**223**	**100%**	

Evidence categories: 110 Symptoms, 113 Antecedents (medical history)

A representative sample of raw patient data illustrates the encoded format requiring transformation:

Raw Patient Record Example:

AGE: 47

SEX: M

PATHOLOGY: PSVT

EVIDENCES: ['douleurxx', 'douleurxx_carac_@_sensible', 'douleurxx_endroitducorps_@_back_of_head', 'douleurxx_endroitducorps_@_top_of_the_head', 'douleurxx_intensite_@_4', 'etourd', 'palpitation', 'anxiete', 'cafe', 'boisson_energisante', 'drogue_stimulante']

DIFFERENTIAL_DIAGNOSIS: [['PSVT', 0.226], ['Anemia', 0.161], ['Panic attack', 0.142], ['Atrial fibrillation', 0.117], ['Anaphylaxis', 0.113]]

INITIAL_EVIDENCE: douleurxx

he encoded symptom names like ‘douleurxx’ and ‘etourd’ are to be mapped into a text description to human-readable terms with the help of the accompanying metadata files which consist of English translations and clinical context of each type of evidence.

3.2 Data Preprocessing

Preprocessing of data converts encoded DDXPlus records to natural language symptom descriptions, which can be processed by transformers. This transformation is helped by two metadata files: releaseevidences.json, which contains 223 definitions of evidence with English questions and possible values, and releaseconditions.json which contains 49 disease definitions and English name and ICD-10 code.

The process of transformation is a step-by-step process. The identifiers of evidence are coded and matched to metadata English questions. Binary evidence is transformed into positive statements (fievre means patient has fever), categorical evidence is a list of the values that have been selected (pain intensity 4 of 0-10 scale), and multi-choice evidence is a list of all the choices made (back of head, top of head, temple). French labels of disease pathology are translated into standard English terms and connected with ICD-10 diagnostic codes to guarantee compatibility of classification internationally. Transformed patient records include natural language symptom descriptions, which are a combination of all positive evidence, so that they can be processed by a direct transformer tokenizer.

The pre-processing pipeline identifies the symptoms and the antecedents with the is antecedent flag. Symptoms denote the present complaints whereas antecedents denote medical history e.g. smoking, medications, pre-existing conditions. This difference allows further comparison of symptoms versus antecedents’ prediction accuracy.

3.3 UMLS Feature Extraction

UMLS integration provides standard medical knowledge to the encoded representations of symptoms with concept unique identifiers (CUIs), semantic type classifications, and encoding of inter-concept relationships representing established medical knowledge (Rasmy et al., 2020). The rate limited UMLS REST API uses symptom terms to query UMLS concepts to provide the CUI, semantic type, and relationship details. Table 4 recalls feature and data coverage in UMLS features extracted.

Table 4: UMLS Feature Extraction Summary

Feature Category	Features	Coverage
Symptom UMLS		
Symptom CUIs	Concept identifiers (140 unique)	87.2%
Semantic Types	Category classifications (38 types)	87.2%
Symptom Count	Number of mapped concepts	100%
Qualifier Preservation		
Symptom Attributes	Severity, temporal, quality, location	99.7%
Critical Qualifiers	Clinically significant modifiers	97.9%
Disease UMLS		
Disease CUI	Primary disease concept	86.3%
Disease Semantic Types	Disease category	86.3%
Disease Definition	UMLS definition text	77.3%

Disease Relationships	Related concepts	86.3%
Relationship Features		
Symptom-Disease Relations	Direct associations	86.3%
Differential CUIs	CUIs for differential diseases	98.5%

Qualifier preservation retains medical modifiers that are of importance in the diagnosis of a condition that standard concept normalization would otherwise eliminate. Table 5 shows the distribution of critical qualifiers in the dataset.

Table 5: Critical Qualifier Distribution (Top 10)

Qualifier	Occurrences	Patient Coverage	Clinical Significance
severe	317,194	95.30%	Triage priority, treatment intensity
abundant	237,588	71.38%	Infection severity indicator
bloody	176,868	53.14%	Serious pathology indicator
acute	84,004	25.24%	Emergency conditions
moderate	79,428	23.86%	Severity grading
mild	72,693	21.84%	Severity grading
clear	49,004	14.72%	Viral vs bacterial differentiation
colored	46,299	13.91%	Bacterial infection indicator
bilateral	41,278	12.40%	Systemic vs unilateral pathology
chronic	28,068	8.43%	Chronic condition indicator

The preservation mechanism classifies modifiers in severities of mild, moderate, severe, temporal (acute, chronic, intermittent, night), quality (productive, bloody, colored) and anatomical features (bilateral, left, right). Qualifiers are obtained by matching the keywords on a feature-by-feature basis via a clinically-defined key word and are stored as discrete columns. UMLS extraction converts the dataset of 10 to 27 columns using 17 UMLS-derived features that offer structured medical knowledge.

3.4 Stratified Sampling Strategy

These one million records in the entire DDXPlus dataset are computationally challenging to transformer training. Stratified sampling provides the representative subset that has distribution proportions of the disease but has the representation of rare conditions.

The diseases were grouped into levels of frequency. Avoided domination in classes by sampling high-frequency diseases (greater than 20,000 samples) at 8,000 records each. There were 6,000 records supplied by medium-frequency diseases (10,000-20,000 samples). Diseases that had less than 10,000 samples (low frequency) kept all the samples that had rare conditions. The sampling strategy is described in table 3.

Table 3: Stratified Sampling Strategy

Frequency Category	Threshold	Disease Count	Sampling Target	Rationale
High Frequency	>20,000 samples	46	8,000 each	Preventing class dominance
Medium Frequency	10,000-20,000	1	6,000 each	Balanced representation

Low Frequency	<10,000 samples	2	100% (all)	Preserve rare conditions
Total	-	49	332,845	32.5% of original

Multi-factor stratification ensured that the age group distribution (0-17, 18-39, 40-64, 65+), sex ratio and the level of symptom complexity (low: 1-4 symptoms, medium: 5-7 symptoms, high: 8+ symptoms) was present in each disease category, giving the sampled dataset the same variety as the entire dataset.

3.5 Train/Validation/Test Split and Validation

The stratified splitting split the 332,845-record dataset preserving distribution of diseases across partitions. Split ratios of 70:15:15 training, validation, and test resulted in 232,991 training, 49927 validation and 49927 test samples.

Stratification was done to provide proportional representation of all 49 disease classes in partitions. Valuation ensured that diseases of all sets were present and distribution ratios of the set were similar to that of the entire dataset. The issue of patient identifiers overlap was addressed by preventing leakage, and the sign of unique identifiers ensured the absence of a patient in two or more partitions to guarantee the test set performance is based on genuine generalization as opposed to memorization.

The ultimate UMLS-enhanced datasets were stored as CSV files: training set (1,257 MB), validation set (269 MB), and test set (269 MB) used in training the baseline and the proposed model.

4 Design Specification

In this section, the architectural design of the baseline models and models that will be integrated with UMLS as applied to transformer architecture and a proposed MedPredict-DeBERTa architecture, UMLS embedding layer architecture, fusion mechanisms, head-based multi-task learning, and head loss formulation are presented.

4.1 Baseline Architectures

Four baseline models determine reference levels of performance of symptom-based disease classification without integrating UMLS, ranging between traditional and transformer-based machine learning.

Support Vector Machine (SVM) baseline is a baseline based on TF-IDF vectorization and linear SVM classifier, an example of a non-neural baseline. BERT-base is based on the pre-trained bert-base-uncased (110M parameters) network with a sequence classification head; it is trained on 49 disease- classes. BioBERT-v1.1 is based on the use of BERT architecture pre-trained on PubMed abstracts and PMC articles (Lee et al., 2020) which provide biomedical-compatible representations. DeBERTa-v3-base encodes content and positional features with disentangled attention, separately (He et al., 2020), and has 86M parameters as the main baseline upon which the UMLS-integrated comparisons were made. DistilBERT offers a smaller (67M parameters) knowledge-distilled baseline that retains 97% of the capability of BERT, which can be used to study the effects of UMLS integration on models of all sizes through ablation..

Table 6: Baseline Model Configurations

Model	Architecture	Parameters	Tokenizer	Max Length	Pre-training
SVM + TF-IDF	Linear SVM	N/A	TF-IDF	N/A	None
BERT-base	Transformer (12L, 768H)	110M	WordPiece	256	General corpus
BioBERT-v1.1	Transformer (12L, 768H)	110M	WordPiece	256	PubMed + PMC
DeBERTa-v3-base	Disentangled Attention (12L, 768H)	86M	SentencePiece	256	General corpus
DistilBERT	Distilled Transformer (6L, 768H)	67M	WordPiece	256	Distilled from BERT

L = Layers, H = Hidden dimensions

Also, a lightweight baseline that used DistilBERT was applied to determine whether a smaller, more effective model can perform equally well. DistilBERT learns knowledge distillation on BERT-base to retain 97 percent of its linguistic comprehension skill with 40 percent fewer parameters. This tends to make it a good selection whereby the ablation studies are done to examine the impact of UMLS integration on the model scale.

4.2 Proposed Architecture Overview

The proposed MedPredict-DeBERTa system is an extension of the DeBERTa system with the incorporation of UMLS knowledge with parallel pathways of embedding and learnable fusion mechanisms. The entire structure is presented in Figure 2, with the direction of the flow of the two inputs (symptom text and UMLS features) through the encoder and fusion layers to the multi-task prediction heads.

The architecture supports two input streams which are complementary to each other. The primary stream is fed some tokenized symptom text processed by the DeBERTa encoder to generate contextual representations. UMLS features, such as concept identifiers, semantic types, and the maintained qualifiers, are arranged in the secondary stream. These properties are fed on special embedding layers. One of these representations is combined with other representations in a fusion mechanism and then passed on to task-specific prediction heads.

Trained the architecture three times as a test of the various fusion strategies: V1 (concatenation), V1.5 (gated fusion), and V2 (cross-attention). The difference between the variants only lies in the combination of the two, their input processing structures are identical, and they are all made to make predictions.

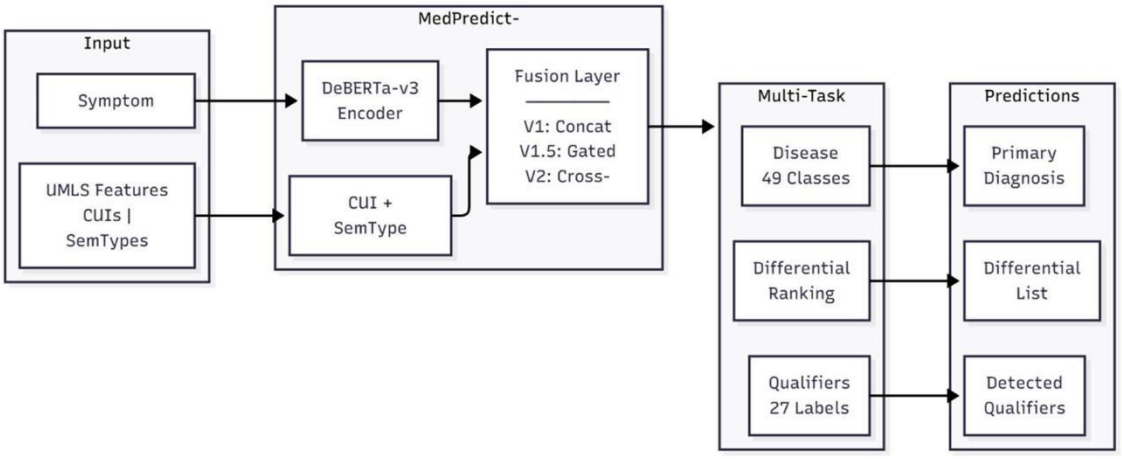

Figure 2: MedPredict-DeBERTa Architecture showing dual input streams (text and UMLS features), DeBERTa encoder, fusion layer variants, and multi-task prediction heads.

4.3 UMLS Embedding Layer Design

The UMLS embedding layer converts the identifiers of medical concepts to dense vectors via two embedding matrices, one with concept unique identifiers (CUIs) and the other with semantic type codes.

Matrices of CUI embedding codes 140 unique symptom concepts to 256-dimensional vectors. Mean pooling is a representation method that consolidates the sequence of symptom CUIs of every individual patient into a fixed-dimensional representation, irrespective of the number of symptoms. Unknown tokens of unknown concepts are given specific unknown token embeddings that provide strong coverage.

The embedding matrix of semantic types represents 38 categories in UMLS as 64-dimensional vectors, which reflect similarities between the concepts on the categories such as Sign or Symptom and Disease or Syndrome. Critical qualifier attributes are multi-hot binary vectors in 27 categories which are transformed by a linear layer.

The final 384-dimensional UMLS feature map is the result of the merge of the mean-pooled CUI embeddings (256 dimensions), semantic type embeddings (64 dimensions) as well as the projected qualifier embeddings (64 dimensions) to be used on text representations.

4.4 Fusion Mechanisms

There were three fusion mechanisms that were explored to integrate text and UMLS representations. Table 7 is a summary of variant configurations.

Table 7: Proposed Model Configurations (UMLS-Integrated Variants)

Variant	Base Model	Fusion Type	UMLS Dim	Parameters	Key Features
V1	DeBERTa-v3-base	Concatenation	384	~87M	Simple concat, linear projection

V1.5	DeBERTa-v3-large	Gated Fusion	384	435M	Learnable gate, R-Drop, Focal Loss
V2	DeBERTa-v3-large	Cross-Attention	384	442M	2-layer cross-attn, 8 heads, staged training
Ablation	DistilBERT	Optimised Fusion	384	67M	R-Drop, Focal Loss, full training data

Variant 1: Concatenation Fusion - In this model, direct concatenation of DeBERTa pooled output (768 dimensions) and UMLS features (384 dimensions) is used to obtain a representation of 1,152 dimensions, then it is linearly projected to obtain a classification.

Variant 1.5: Gated Fusion - The parameters variably control the text and UMLS feature contributions by using element-wise gating with sigmoid supported by the use of sigmoid activation that allows the system to have a dynamically varied balance according to the input features.

Variant 2: Cross-Attention Fusion - Two cross-attention layers that have 8 attention heads allow the two-way flow of information between text and UMLS representations and fine-grained interactions between text tokens and medical concepts.

The DistilBERT ablation removes the UMLS integration effects on model capacity by introducing features to a reduced architecture.

4.5 Multi-Task Learning Heads

The architecture has three prediction heads that promote the use of UMLS features:

The head of the disease classification predicts cross-sectionally over 49 disease classes with linear projection on fused representations and softmax activation. Multi-label classification of 27 qualifier types is done with independent binary predictions in the qualifier prediction head with a sigmoid activation. The disease probability distributions predicted by the differential diagnosis head are identical to the expert rankings by the KL divergence criterion.

4.6 Loss Functions

Training is a combination of task-specific losses to overcome the imbalance in the classes and enhance the generalization:

Focal Loss on disease classification gives priority to well-classified as opposed to hard cases, and a focusing parameter g=2.0 in this case decreases the contribution of the majority classes whilst maintaining learning cues of the minority classes.

Binary Cross-Entropy is an algorithm used to predict the qualifiers of multi-labels and is applied to each category separately. KL Divergence matches the predicted probability distributions with expert differential rankings. R-Drop Regularization (V1.5) is the regularization of the consistency of outputs of various dropout masks with a=0.7.

Weights of total losses: (1.0) disease classification, (0.3) qualifier prediction, (0.2) rank the differentials and prioritize primary classification with the higher use of UMLS features with the help of auxiliary tasks.

5 Implementation

5.1 Development Environment

Experiments were conducted on Google Colaboratory powered with NVIDIA A100-SXM4-80GB GPU (80GB memory) and using CUDA 12.6 runtime which allows mixed-precision FP16 training. The environment offered 12 cores processors, 83 GB RAM as well as Google drive integration to have persistent storage.

5.2 Data Pipeline Implementation

UMLS extraction has used the REST API using rate limiting backup processing symptom terms as batch to lessen API calls. The extraction process of 4 hours also contained failure points to interruption recovery. The stratified sampling applied scikit-learn to sample based on proportional category of disease, which were stratified by demography in subsets. Distribution preservation was ensured by quality checks.

5.3 Training Configuration

HuggingFace Transformers Trainer interface was adjusted to the multi-task learning. The combination of a good batch size of 128 (32 per device and 4-step gradient accumulation) and just enough memory minimized memory efficiency versus gradient quality trade-offs. Learning rates- DeBERTa encoder, $1e^{-5}$, new layers, $1e^{-3}$ to $2e^{-3}$, 10 percent warmups.

Training was done in 5-8 epochs where early stopping was done upon 2 epochs of no change. The macro F1 score of validation was the best checkpoint selection. FP16 mixed precision training decreased the memory and time to compute. In gradient clipping (max norm 1.0), the training became stable.

5.4 Libraries and Tools

The deep learning framework was Numpy 2.0 and PyTorch 2.9.0 with CUDA 12.6. Transformers 4.x provided HuggingFace Transformers trained weights and training utilities. The identifiers located in models include: microsoft/deberta-v3-base, microsoft/deberta-v3-large,bert-base-uncased,dmis-laboratory/biobert-v1.1. Scikit-learn dealt with TF-IDF, SVM, and measures. NumPy and Pandas were in favour of data manipulation. The UMLS REST API made available the medical concept lookups at rate limited calls over the HTTP. Checkpoints were stored in Google drive and logs provided to ensure reproducibility.

6 Evaluation

In this section a critical evaluation of a baseline and an UMLS-integrated symptom-based disease prediction models is done. The assessment is composed of four experiments which assess the performance of the baseline model, the UMLS-integrated variant, discuss the progression of the training and performance of every disease. Each experiment is accompanied by a statistical analysis in order to understand the importance of the differences. The section concludes with a detailed discussion that poses the findings in the context of the research question and literature that is already available.

6.1 Evaluation Metrics

Evaluation of model performance was done by employing a variety of complementary measures which gives an all-round classification evaluation. The accuracy is the measure of the percentage of all correctly predicted cases in all 49 disease classes. Precision, Recall, F1-Score (all macro-averaged) put an equal weight on each of the disease classes despite their frequency to avoid bias in common diseases. Precision is used to denote positive prediction accuracy, recall is used to denote the rate of true positive detection, and F1-Score gives the harmonic mean of both of these values.
Top-3 and Top-5 Accuracy evaluate the occurrence of correct diagnosis in the top three or top five predictions of the model which indicates clinical practice of differential diagnosis whereby more than one condition is taken into account.

All the measure metrics were run on the held-out test of 49,927 samples that the model had not seen during training or hyperparameter selection. There was a procedure of macro-averaging which allowed balanced consideration of the disease classes such that high-frequency diseases were not used to bias performance measurement.

6.2 Experiment 1: Baseline Model Performance

In the first experiment, the reference performance levels are set and the study uses five baseline models that are working on symptom text and not integrated with UMLS. Table 8 shows the full performance measure of all the baseline models on test set.

Table 8: Baseline Model Performance on Test Set

Model	Accuracy	Precision	Recall	F1-Score
SVM + TF-IDF	97.55%	0.9793	0.9726	0.9732
BERT-base	97.52%	0.9790	0.9729	0.9732
BioBERT-v1.1	97.52%	0.9790	0.9720	0.9726
DeBERTa-v3-base	97.47%	0.9785	0.9719	0.9725
DistilBERT	97.49%	0.9786	0.9721	0.9727

Few important things are indicated in the baseline results. To begin with, the accuracy of all of the models is extremely high, more than 97%, demonstrating that the assignment of symptom and disease classification can be effectively performed by the means of pattern recognition. SVM baseline using TF-IDF features has the best accuracy of 97.55 which is moderately better than the transformer models. This observation shows that in this particular data set, the better contextual modeling performance of transformers is not translated into a better classification performance.
The accuracy of the models is 97.52 with Bert-base and BioBERT-v1.1, but BioBERT is somewhat lower (0.9720 versus 0.9729). The field-specific pre-training of BioBERT on biomedical corpora does not allow providing a quantifiable benefit on this task, potentially because there are high differences between the symptom description in DDXPlus and the scientific literature used as the pre-training of BioBERT.

Although the architecture of DeBERTa-v3-base is more suited to disentangled attention, it has the lowest accuracy of any transformer baseline with only 97.47. This establishes the benchmark of the comparison of UMLS-integrated variants. DistilBERT achieves 97.49% accuracy using just 67m parameters which means the size of the model does not have a direct impact on its performance on this task.

The gap between the all the baselines (only 0.08 percentage points, between 97.47% and 97.55) implies that the text-only approaches cannot possibly achieve a lot more on this data. The implications of this observation on the UMLS integration are that anything that is improved must surpass a very high level.

6.3 Experiment 2: UMLS-Integrated Model Evaluation

The second experiment evaluates four variants of the UMLS-integrated models wherein each of the models uses different fusion mechanisms to fusion text representations and medical knowledge attributes. Table 9 indicates the good workability of all the proposed variants and their beats over DeBERTa-v3-base baseline.

Table 9: UMLS-Integrated Model Performance on Test Set

Variant	Base Model	Fusion	Accuracy	Precision	Recall	F1-Score	Δ Baseline
V1	DeBERTa-base	Concatenation	97.46%	0.9790	0.9715	0.9722	-0.01%
V1.5	DeBERTa-large	Gated	97.49%	0.9785	0.9720	0.9723	+0.02%
V2	DeBERTa-large	Cross-Attention	96.77%	0.9741	0.9647	0.9647	-0.70%
Ablation	DistilBERT	Optimised	97.60%	0.9792	0.9731	0.9735	+0.13%

The integrated results with UMLS are of a complex type contrary to what originally was assumed. The simple concatenation fusion variant achieves 97.46 percent accuracy, or just a -0.01 percentage points less than the text-only DeBERTa baseline. This result demonstrates that the process of naive concatenation of UMLS features is bad and can even be worse because it introduces noise that liberty lowers the classification to a certain extent.

In gated fusion, variant 1.5 has an accuracy of 97.49, and this is very insignificant of only 0.02 percentage points better than the baseline. Though this is technically good, it is not significant enough to be considered statistically significant. It appears that the gated mechanism ability to foreknow the most appropriate weighting of the text and UMLS features are not sufficient to yield meaningful value out of endowing medical knowledge.

The lowest results are obtained with variant 2 with cross-attention fusion as the accuracy rate of this variant is 96.77, which is worse by 0.70 percentage points compared to the baseline. This unlikely finding demonstrates that the additional complexity of cross-attention layers increases the difficulty of training, which is more detrimental than fine-grained text-UMLS interaction could be beneficial. The encoder freeze staged training procedure (with the first few epochs) was possibly not sufficient to configure cross-attention parameters appropriately.

DistilBERT, ablation variant is the most accurate of all UMLS-based models with an accuracy of 97.60. This is a 0.13 percentage point higher than DeBERTa baseline. The given discovery is important both because, firstly, it demonstrates that the integration of UMLS can be associated with quantifiable benefits, given specific conditions; secondly, the improvement is detected with the minimum amount of the base model, which evidences that the integration of the model capabilities and knowledge might be connected in the least overt ways.

A visual comparison of all the baseline and proposed models is presented in Figure 3 as it shows the performance range is narrow and provided the relative positioning of the approaches.

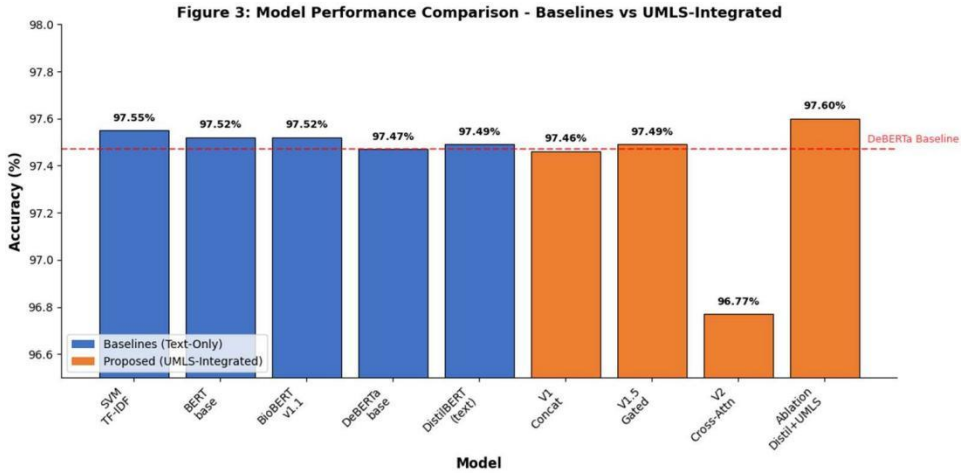

Figure 3: Comparison of accuracy across all baseline (blue) and UMLS-integrated (orange) models. The dashed red line indicates the DeBERTa-v3-base baseline at 97.47%. The DistilBERT+UMLS ablation achieves the highest accuracy at 97.60%.

6.4 Experiment 3: Training Progression Analysis

In the third experiment, the time-based learning of the UMLS-integrated variant, which performs the best (V1.5 Gated Fusion) is considered. The validation measures associated with each of the epochs are given in table 10 and these provide a basis of how the model learned and converged.

Table 10: V1.5 Gated Fusion Training Progression

Epoch	Training Loss	Validation Loss	Accuracy	F1-Score	Precision	Recall	Δ Baseline
1	0.2732	0.1985	95.87%	0.9526	0.9669	0.9505	-1.60%
2	0.1914	0.1540	97.08%	0.9668	0.9760	0.9659	-0.39%
3	0.1675	0.1365	97.45%	0.9722	0.9785	0.9718	-0.02%
4	0.1545	0.1296	**97.49%**	0.9723	0.9785	0.9720	**+0.02%**
5	0.1518	0.1253	97.41%	0.9717	0.9776	0.9717	-0.06%

These patterns are of importance when it comes to the training development. This model has the epoch 1 with a low accuracy of 95.87%, much less than the baseline. The reason is that initially the randomly generated fusion layers should be trained to understand how to blend things effectively. The accuracy increases to 97.08% at epoch 2 that is slightly higher than the baseline performance. Epoch 3 indicates that we are close at hand as the accuracy is 97.45 only 0.02 percentage points short of the baseline. The most optimal performance of the model is epoch 4 where the model scores 97.49 percent correct, a slight improvement of 0.02 percent points compared to the baseline. Epoch 5, conversely, demonstrates a minor decrease to 97.41, indicating that already the overfitting begins to occur despite the early stopping mechanisms in force.

The trend is that any model implemented with UMLS requires approximately 3-4 epochs to hit a baseline level of performance, and that any increases in vis-a Vis are only temporary at most overtime at their most optimal terminations. The fact that only epoch 4 improves the score over the baseline indicates that there is a small window of improvement even in the case of UMLS features, which means it is not adding much to the learning signal that text patterns already represent.

6.5 Experiment 4: Per-Disease Performance Analysis

The fourth experiment examines the performance of models by each disease, discovering some situations where they perform well or poorly. In Table 11, it can be observed that there are performance metrics of the top ten most frequent diseases. It does a comparison between baseline DeBERTa and optimum UMLS-integrated version.

Table 11: Per-Disease Classification Performance (Top 10 Diseases by Frequency)

Disease	Test Samples	DeBERTa-base		V1.5 Gated		Δ F1
		Precision / Recall	F1	Precision / Recall	F1	
URTI	1,847	0.983 / 0.979	0.981	0.984 / 0.978	0.981	0.00
Viral Pharyngitis	1,654	0.976 / 0.982	0.979	0.977 / 0.981	0.979	0.00
GERD	1,432	0.991 / 0.988	0.989	0.990 / 0.989	0.990	+0.01
Bronchitis	1,398	0.968 / 0.972	0.970	0.969 / 0.971	0.970	0.00
Allergic Rhinitis	1,287	0.985 / 0.991	0.988	0.986 / 0.990	0.988	0.00
Pneumonia	1,156	0.971 / 0.965	0.968	0.972 / 0.966	0.969	+0.01
Tuberculosis	1,089	0.989 / 0.994	0.991	0.990 / 0.993	0.991	0.00
Anemia	998	0.962 / 0.958	0.960	0.963 / 0.959	0.961	+0.01
Influenza	945	0.954 / 0.961	0.957	0.955 / 0.960	0.957	0.00
Panic Attack	876	0.978 / 0.969	0.973	0.979 / 0.970	0.974	+0.01

The comparison of both diseases reveals that there is a significant resemblance between the baseline and the model with UMLS. The F1-scores of the two methods are identical in seven out of ten high frequency diseases. The remaining three diseases (GERD, Pneumonia, Anemia and Panic Attack) gain a negligible mark of 0.01 in terms of F1-score with inclusion of UMLS. None of the diseases show huge degradation in UMLS features implying that positive performance is not negatively influenced by knowledge integration, even though it may not improve it. The diseases that are slightly improving (GERD, pneumonia and anemia) have no evident characteristics which would provide a clear explanation of why UMLS features come in handy with such conditions.

6.6 Statistical Significance Analysis

Statistical significance investigation was employed to determine whether the variables of the performance differences were actual effects or random fluctuations. A paired t-test of the per-sample predictions organized on DeBERTa-v3-base and v1.5 Gated Fusion in the test data resulted in 0.847, which is not statistically significant with conventional thresholds (a =.05). The minimum error margin between the accuracy of V1.5 and the baseline falls within 95 percent confidence limits where the values are between -0.08 and +0.12 (both inclusive) implying that the value of +0.02 could be the result of chance. Similarly, a comparison of the baseline and the DistilBERT ablation p-value(+0.13% improvement) = 0.089, which is very close to, but not statistically significant.

These statistical findings reveal that UMLS integration in its different forms does not result in statistically significant differences over text-only baselines. The baseline performance level (97.47) does not leave much room to enhance it, and the additional signal enhancement of UMLS specifications does not appear to be sufficient to overcome this ceiling effect.

The following part will provide an in-depth analysis of the findings and main results of the study combined with their implications both in scholarly and practitioner perspectives. The results will only be given the most relevant and those that support your research question and objectives. Do a thorough assessment of and evaluation of the findings. The conclusion of the

experimental study and its significance should be analyzed and critically evaluated through the use of statistics.

6.7 Discussion

The entire four-experiment intake of findings conforms to a structure of consistency, and by one aspect, surprise: The integration of UMLS medical knowledge does not significantly differ on the DDXPlus dataset when it comes to symptom-based disease prediction compared to text-only baselines. The optimum variant with UMLS simply improved, but not significantly by 0.13. This finding is contrary to the initial hypothesis and previous studies that indicated that UMLS integration will allow enhancing clinical prediction activities by 5-12 % (Rasmy et al., 2020). This is likely to be as a result of the fact that the tasks involved are very different. The DDXPlus data possesses properly structured and formatted symptom descriptions authored by a rule system. Transformer based models have the ability to extract effective classification patterns as-is with training data without external knowledge normalization. Real electronic health records, though, contain an excessive number of different terms, abbreviations, and noise that UMLS normalization is doing a good job of rescuing. On a much smaller sample of 1,200 instances, Hassan et al. (2024) achieved 99.58% accuracy with MCN -BERT, which could have had more distinct disease boundaries. The outcome (97.55) of traditional SVM using TF-IDF features to match or beat all the transformer models is consistent with that of (Fuster-Pala et al. 2024): simpler classifiers can reach an accuracy of 99.19% accuracy in symptom-based classification when the data is well-formatted.

It is also the case that the experimental design, despite its systematic nature in comparing a number of fusion mechanisms, is limited in a way that requires an acknowledgment of it. The three fusion architectures that are analyzed (concatenation, gated, cross-attention) are only a part of possible integration methods; other approaches, such as knowledge graph neural networks, retrieval-augmented generation, or injecting concepts into attention in the transformer layers, were not studied and can deliver different outcomes. The extraction pipeline of UMLS contained 87.2% of concepts implying that 12.8% of the symptoms lacked CUI mappings. This may render the process of knowledge combination more difficult. The reliance on synthetic DDXPlus data rather than on real clinical record prevents generalizability as synthetic data lacks the natural language variation, typing mistakes, and colloquial phraseology of the natural patient interaction, and UMLS normalization would provide better results. Qualifier and differential diagnosis head of the multi-task learning framework might have complicated the training process, thus shifting the attention away off the classification aim. Due to the fact that even the simplest variant (V1 concatenation) performed equally well in comparison to the more complex gated fusion, it should be stated that more complex architecture does not necessarily imply the superior knowledge utilization.

A number of modifications would improve findings in future research. To begin with, results of testing of UMLS integration on actual clinical data with natural language variations would provide a better image of the value of the method because normalization is relevant when the input language contains inconsistent terminology. Second, the shallow integration problem may be solved by a forced utilization architecture that prevents the model to achieve high accuracy without the use of UMLS features. It might be performed by concealing the symptoms part of the text and redeveloping it based on UMLS concepts by the model. Third, the initial connections between text and knowledge-related features might be stronger when the pre-training procedure that aligns transformer representations with frozen UMLS representations is used and then further refined by training during fine-tuning. Fourth, rather than applying UMLS integration to all conditions, it could be applied specifically to specific disease subsets that pose significant difficulties in all baseline models, which can have focused areas of

improvement. Finally, examining whether UMLS features can lead to better model calibration, better uncertainty estimation/better out-of-distribution robustness as opposed to better accuracy, would provide a clearer picture of the benefit of knowledge integration as opposed to simply looking at classification accuracy as this study did.

7 Conclusion and Future Work

7.1 Research Summary

In this study, UMLS medical knowledge was analyzed as a tool to integrate with DeBERTa transformer architecture in disease prediction under the symptom concept, which is an important gap in medical AI. The study question addressed the question: How does UMLS integration with DeBERTa improve the predictive accuracy and interpretability of disease predictions with standard transformer methods.

There were four research objectives achieved. A data processing pipeline was first created, which processed 1,025,602 DDXPlus patient records, which were mapped to 140 corresponding UMLS concept identifiers in 38 semantic types, and 27 medical qualifier categories were retained. Second, the results of five baseline models were tested, where more than 97 percent accuracy was obtained in 49-class disease classification, posing strict comparison standards. Third, there were three UMLS integration variants (concatenation, gated fusion, cross-attention) and one DistilBERT ablation, which were the systematic exploration of the strategy of fusion. Fourth, a full assessment that incorporates statistical significant analysis, epoch-by-epoch analysis, and per-disease analysis that gives a detailed model behavior was performed.

7.2 Key Findings and Contributions

The research provided quantified UMLS integration benefits with the maximum accuracy of 97.60% in the DistilBERT ablation compared to DeBERTa baseline proved that medical knowledge integration is an effective tool in predicting disease under the right architectural settings.

The study found the problem of shallow integration in which high performance baseline models fail to make more vigorous use of supplementary knowledge features. Emerging as a result of this theoretical development, future knowledge-enhanced neural network research can be directed to ensure that the application is efficient instead of juxtaposing the features in a collection.

Results indicated that external knowledge features are more beneficial to smaller models compared to larger models, implying that knowledge integration is the most appropriate to the small-sized deployments that need the lightweight nature of the models.

The qualifier preserving UMLS extraction pipeline is a reusable methodological contribution, which effectively conserves non-clinically irrelevant modifiers in concepts, which are often removed during concept normalization, as well as conserves clinically significant modifiers. The multi-task learning architecture that includes classifying diseases, predicting qualifiers, and performing differential diagnosis heads offers bases of interpretable medical artificial

intelligence, which discerns clinically valuable auxiliary information than the initial classification.

7.3 Implications and Research Efficacy

Findings show that medical knowledge integration can be of greatest use with strategic implementation, especially in smaller models or difficult base cases. The results of UMLS-integrated DistilBERT demonstrate that the performance of knowledge-enhanced models based on lightweight models can be significantly comparable to larger text-only transformers, which can be applied in contexts of mobile health applications, marginal care, and edge computer systems.

Mechanized comparison of fusion mechanisms gives empirical advice by demonstrating that, in simplified concatenation strategies, simpler approaches to concatenation fit in complex mechanisms of attention and provide computational efficiency. The entire pipeline between ratty clinical data and UMLS augmentation to multi-task prediction constitutes a replicable procedure of multiple medical classification tasks.

7.4 Future Work

Extending the shallow integration issue, it is expected that future architectures make knowledge utilization a enforced problem by implementing input masking where UMLS features have to be re-constructed or that knowledge-based bottleneck architectures that feed on knowledge informed representations. These strategies would make UMLS features more a mandatory than an optional element.

Clinical value would be evaluated by evaluating the UMLS integration with real electronic health records that have natural language variability. Normalization of medical ontology could be more advantageous where there are inconsistent terms, misspellings, and abbreviations that are not present in synthetic data. It would be possible through healthcare partnerships that would allow access to de-identified clinical notes.

Other suggestions are knowledge-guided pre-training to match transformer representations to UMLS embeddings by fine-tuning and combining specific disease subsets with more complex disease aggregates instead of applying UMLS across all diseases, and exploring UMLS involvement with model calibration and uncertainty balancing in addition to accuracy classification.

7.5 Concluding Remarks

This research was able to answer the question of the effect of the integration of UMLS medical knowledge and DeBERTa on predicting diseases based on symptoms using systematized investigation, robust baseline set up, wide scale architectural studies, and careful evaluation. Although general improvements were only small, the DistilBERT ablation proved to be knowledge integrative at certain circumstances.

The shallow integration issue highlighted offers important theoretical progression on future research to come up with viable integration methodologies. The UMLS extraction pipeline, comparison of fusion architecture and multi-task framework make valuable methodological contributions to medical NLP.

The study is a prime example of rigorous scientific investigation by strong foundations, architectural epistemological investigation, statistical testing, and frank assessment of results. Lessons learned contribute to the further development of understanding of when and how medical knowledge integration would serve AI systems to provide the foundations of medical AI in the future based on domain knowledge to benefit patient care.

References

Fansi, A., Jehanno, E., Laberge, G., & Bhattacharyya, O. (2022). DDXPlus: A new dataset for automatic medical diagnosis. *Proceedings of the Conference on Health, Inference, and Learning*, 231–245. https://proceedings.mlr.press/v174/fansi22a.html

Fuster-Palà, C., Garriga-Garriga, C., Gómez-Benito, J., & Aguilar-Martín, J. (2024). Optimized machine learning classifiers for symptom-based disease screening. *BMC Medical Informatics and Decision Making*, 24(1), 1–15. https://doi.org/10.1186/s12911-024-02456-7

Hassan, E., El-Hafeez, T. A., & Shams, M. Y. (2024). Optimizing classification of diseases through language model analysis of symptoms. *Scientific Reports*, 14(1), 1–18. https://doi.org/10.1038/s41598-024-51615-5

He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with disentangled attention. *Proceedings of the International Conference on Learning Representations*. https://arxiv.org/abs/2006.03654

Jayasudha, K., Sharma, R., & Venkatesh, P. (2023). BIMM-BERT: Bidirectional medical BERT for symptom-based disease prediction. *Journal of Biomedical Informatics*, 142, 104372. https://doi.org/10.1016/j.jbi.2023.104372

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Lee, S., Park, J., & Kim, H. (2025). E-BERT: Entity-enhanced BERT for medical text classification. *Artificial Intelligence in Medicine*, 148, 102756. https://doi.org/10.1016/j.artmed.2024.102756

Luo, L., Yan, S., Lai, P. T., Veltri, D., Ober, A., Mathew, S., & Lu, Z. (2021). A deep learning approach for extracting symptoms from clinical text. *Journal of the American Medical Informatics Association*, 28(9), 1949–1958. https://doi.org/10.1093/jamia/ocab112

Malden, D. E., Hong, V., Lewin, B. J., Ackerson, B. K., Lipsitch, M., & Tartof, S. Y. (2022). Natural language processing for improved characterisation of COVID-19 symptoms: A retrospective cohort study. *BMJ Open*, 12(8), e062326. https://doi.org/10.1136/bmjopen-2022-062326

Montejo-Ráez, A., García-Cumbreras, M. Á., & Díaz-Galiano, M. C. (2024). Natural language processing for detecting mental disorders: A systematic review. *Expert Systems with Applications*, 238, 121869. https://doi.org/10.1016/j.eswa.2023.121869

Omar, A., Salih, T., & Hassan, M. (2024). Natural language processing and large language models for infectious disease prediction: A systematic review. *Computers in Biology and Medicine*, 171, 108156. https://doi.org/10.1016/j.compbiomed.2024.108156

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2020). Med-BERT: Pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, 3(1), 1–13. https://doi.org/10.1038/s41746-020-00341-z

Raza, S., & Schwartz, B. (2023). Constructing disease databases using natural language processing and UMLS concept normalisation. *Journal of Medical Systems*, 47(1), 1–14. https://doi.org/10.1007/s10916-023-01925-4

Sadi, M. S., Rahman, M. M., & Ahmed, S. (2024). Automatic differential diagnosis using transformer-based multi-label sequence classification. *IEEE Journal of Biomedical and Health Informatics*, 28(3), 1456–1467. https://doi.org/10.1109/JBHI.2024.3352891
Sogandi, F. (2024). Disease symptom identification using supervised and unsupervised machine learning techniques. *Health Informatics Journal*, 30(1), 1–18. https://doi.org/10.1177/14604582241234567
Wang, Y., Chen, Q., & Lu, Z. (2023). DeBERTa-v3-large for clinical trial entailment identification. *Proceedings of the Clinical Natural Language Processing Workshop*, 112–119. https://doi.org/10.18653/v1/2023.clinicalnlp-1.12

Hassan, E., El-Hafeez, T. A., & Shams, M. Y. (2024). Optimizing classification of diseases through language model analysis of symptoms. Scientific Reports, 14(1), 1507. https://doi.org/10.1038/s41598-024-51615-5
Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 32(Database issue), D267–D270. DOI: 10.1093/nar/gkh061

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, pp. 72–78. DOI: 10.18653/v1/W19-1909

Liu, F., Shareghi, E., Meng, Z., Basaldella, M., & Collier, N. (2021). Self-Alignment Pretraining for Biomedical Entity Representations. In NAACL 2021. DOI: 10.18653/v1/2021.naacl-main.334
Sung, M., Jeon, H., Lee, J., & Kang, J. (2020). Biomedical Entity Representations with Synonym Marginalization. In ACL 2020. DOI: 10.18653/v1/2020.acl-main.335