National College of Ireland

Symptom-Treatment Knowledge Graphs for Contagious and Chronic Disease Profiling

MSc Research Project
MSc Data Analytics

Muralidhar Kukkala
Student ID: 23346795

School of Computing
National College of Ireland

Supervisor: Shubham Subhnil

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Muralidhar Kukkala
Student ID:	23346795
Programme:	MSc. Data Analytics **Year:** 2025-2026
Module:	MSc. Research Project
Supervisor:	Shubham Subhnil
Submission Due Date:	11-12-2025
Project Title:	Symptom-Treatment Knowledge Graphs for Contagious and Chronic Disease Profiling
Word Count:	**8587 Page Count: 21**

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: Muralidhar Kukkala

Date: 11-12-2025

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):	

Symptom-Treatment Knowledge Graphs for Contagious and Chronic Disease Profiling

Muralidhar Kukkala
23346795

Abstract

This paper provides an elucidated, KG, model of disease classification in accordance with the level of chronicity and contagiousness, through the integration of symptom and treatment relationship as obtained through biomedical texts. With the recent research on biomedical NLP and graph learning, the proposed framework addresses the problem of the unaddressed graph structure and explainability problems, using entity canonicalization that is conducted with the help of the SciSpaCy model, TF-IDF and BioBERT embeddings, community extraction, learning of the proposed model, and supervised learning, and GNN. Empirical evaluations have indicated that relational learning usage has a significant effect of improvement of efficacy in disease prediction since XGBoost provides a prediction score of up to 0.976 and GNN provides a prediction score of up to 0.890 when predicting disease according to the extent of chronicity. Knowledge graph analytics are used to find insights into structured communities of diseases and gaps in their treatment and deploy model-level and graph-level explainers, SHAP and GNNExplainer, to the comprehensible representation of model decisions. These results indicate the importance of an efficient integration of meaning based embeddings and application of graph-based learning towards the achievement of strong predictive and biomedical interpretability.

Keywords: Biomedical NLP, Knowledge Graphs, Disease Classification, Graph Neural Networks, Explainable AI, BioBERT, SHAP, GNNExplainer

1 Introduction

1.1 Background and Context

Healthcare industry is gradually generating large amounts of text data, including electronic health records, clinical summaries, and biomedical literatures. Important data on symptoms, disease courses, and modes of therapy are contained in these records, but in many cases, due to their unstructured format, they are less subject to computation and analysis. The more customary approach of disease profiling and documentation primarily applies organized taxonomies such as the International Classification of Diseases as well as SNOMED CT. Even though standardized, in many cases, these taxonomies cannot adequately reflect the dynamic nature of semantics and overlapping associations between diseases, symptoms and treatment (Galluzzo, 2024).

Over the last several years, the biomedical text processing and analysis have advanced due to the creation of Natural Language Processing (NLP) and Machine Learning (ML) platforms. Specialized transformer-based models, such as BioBERT and SciSpacy, are able to identify more complicated contextual links between medical concepts. This, in turn, makes it feasible to produce semantic embeddings that represent biomedical meaning. These embeddings, if

applied for the development of knowledge graphs (KGs), help represent the linkage between symptoms, diseases, and treatments. This makes it feasible to develop an effective disease profiling tool that can be applied for disease diagnosis, disease surveillance, and drug repositioning, as discussed by Vithanage et al. (2024).

1.2 Problem Statement

Although there are considerable breakthroughs achieved by biomedical natural language processing and learning graph representation, it is noticeable that the state-of-the-art methods present considerable limitations concerning the semantic processing and computational comprehension of disease characteristics, like the nature of the disease as chronic and contagious. Present approaches tend to represent diseases as strictly separate entities, without paying attention to the fact that some inherent links exist between symptoms, treatments, and kinds of diseases. As such, the related state-of-the-art solutions neglect the possibility of therapy overlap (Fecho et al., 2021).

Furthermore, although the Graph Neural Networks (GNN) and its modifications have demonstrated impressive prediction performance on a variety of biomedical problems, the problem of interpretability in GNN type of solutions, particularly in healthcare-related issues, continues to be an issue. Actually, most of the advancement that has been made towards GNN and other AI/DA methods of disease profiling prediction has centered on entity extraction and relationship identification, but small consideration is given to the architecture of a combined solution between semantic embedding, graph learning, and XAI.

1.3 Research Aim and Question

This research aims at developing an explicable, semantic graph of knowledge representing the connections between symptoms, diseases, and therapies to enhance disease classification, and identify the gaps in treatment.

This work is guided by the following research question:

"How can NLP-generated semantic embeddings be used to construct knowledge graphs that accurately classify diseases based on their chronicity and contagiousness while identifying potential treatment gaps?"

To handle such an objective, the suggested solution is comprised of the implementation of the following pipeline that integrates the biomedical text embeddings, BioBERT, the multi-relational graph model, and different GNN learning algorithms to the categorization of diseases, as well as community identification. Conversely, the model gives interpretability to the learning process by using SHAP and GNNExplainer.

1.4 Research Motivation and Significance

The reasons for conducting the research are based on three ongoing developments.

Firstly, the openness of biomedical databases, such as the Disease Symptoms and Treatments Dataset on the Kaggle platform, provides an opportunity to train and test the model on marked biomedical relationships.

Second, the biomedical transformer model, such as BioBERT, had high capability to represent the text semantically, which enabled the exact definition of the biomedical words.

On the third point, graph-based machine learning allows for the analysis of relational data and discovers patterns that are difficult to uncover from tabular data.

Through the integration of the above elements, the research makes progress in the development of an interpretable and data-driven method for disease knowledge. The potential impacts of the research are significant and provide promise for assisting and making progress in decision support, drug development, and disease control.

1.5 Research Contribution

The research has important implications to health informatics and applied machine learning by introducing a hybrid solution combining biomedical NLP, graph-based analytical tools and interpretability processes in the same architecture. It shows how clustering and classification based on graphs can be applied to classify diseases and detect possible mismakes in treatment and, thus, facilitate more sophisticated clinical knowledge. The work also incorporates explainable AI methods to increase the level of transparency and confidence in medical decision-support systems. Also, it suggests a generalisable approach to constructing and analysing medical knowledge graphs that can be deployed on various healthcare datasets.

1.6 Structure of the Report

The structure of this report is organised as follows:

- Section 2 – Literature Review: Reviews existing studies on biomedical NLP, knowledge graphs, and graph learning, identifying key research gaps.
- Section 3 – Research Methodology: Describes the data resources, embedding generation, and experimental design.
- Section 4 – Design Specification: Outlines the proposed framework and architectural components.
- Section 5 – Implementation: Explains the technical development process and graph modelling steps.
- Section 6 – Evaluation: Presents model performance, clustering coherence, and interpretability results.
- Section 7 – Conclusion and Future Work: Summarises findings and suggests future enhancements, including ontology integration and federated learning.

2 Related Work

2.1 Overview and Scope

This work positions the current study as an advancement of the state of the art intersection of biomedical natural language processing (BioNLP), knowledge graphs (KGs), and graph machine learning (GML) for disease classification. The combination of the three research areas represents advancement from text-based entity extraction to knowledge-driven reasoning architectures applicable for intricate healthcare analytics. However, despite the progress achieved using transformer architectures for biomedical NLP and GML, little work focuses on the three-way relationship concerning symptoms, diseases, and treatments for disease profiling based on their characteristics of either being contagious or chronic.
The review critically evaluates three core domains:

- Biomedical NLP for disease representation,
- Knowledge graphs in healthcare,
- Disease classification via graph-based learning, and
- The integration of explainable AI (XAI) in medical KGs.

Each subsection highlights numerical outcomes from empirical studies, critiques their limitations, and identifies persistent methodological gaps that motivate the current research.

2.2 Biomedical NLP for Disease Representation

Biomedical NLP applications experienced a shift from rule-based extractive systems to transformer architectures that are capable of representing the semantics of the clinical domain. Although the rule-based approach to biomedical NLP was accurate, it had the limitation of scalability. On the other hand, transformer architectures, like BioBERT, ClinicalBERT, and SciSpacy, generalize well from the semantics but are faced with the problems of interpretability and disambiguation.

Houssein, Mohamed, and Ali (2021) conducted an extensive analysis regarding the use of machine learning and deep learning algorithms and their applications to biomedical NLP, specifically for secondary use and the prediction of chronic diseases from electronic health records. They observed that the use of deep learning algorithms, particularly based on recurrent neural networks and convolutional neural networks, had performed significantly well compared to conventional algorithms for entity extraction as well as relation classification but faced severe limitations while applied to free-text narratives where the implicit information played an important role. They concluded that despite the potential improvement of NLP tools for the early prediction of chronic diseases, the lack of relational modeling hindered the accuracy of predictions. The F1 scores for entity extraction were quantitatively observed to range from 0.80 to 0.90, but decreased for relation extraction.

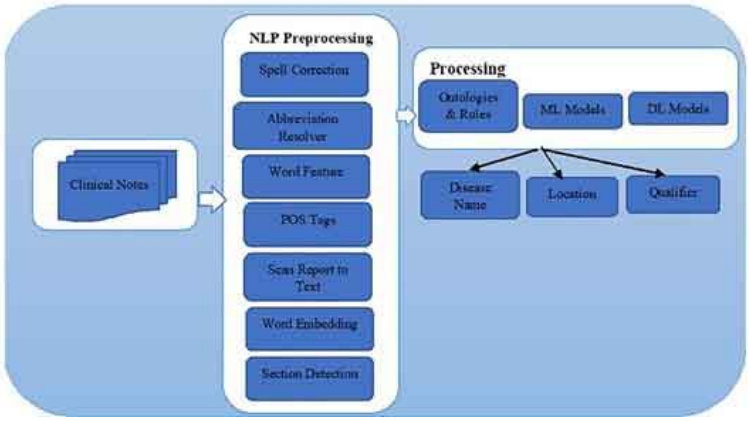

Figure 1: A graphical abstract for Machine Learning Techniques for Biomedical Natural Language Processing

In an empirical study, Ben Abdessalem Karaa et al. (2021) tackled the drug–disease relation extraction problem and adapted Support Vector Machines (SVMs) trained on UMLS ontology-based features. Their system reached an outstanding F1-score of 98.19% for the "cure" relation on the MEDLINE corpus, thus beating state-of-the-art NLP techniques by over 6%. In their study, the authors were able to show that their method, which incorporated morphological, syntactical, and semantic knowledge, performed well but relied on human-engineered features and ontologies. An important limitation of the study was that, despite the capability to extract explicit binary relations (such as drug–cures–disease), the model and approach were not generalizable to jobs involving implicit and multiple relations, for example, symptom–treatment.

This area saw advancement by Yang et al. (2021) as they compared the performance of three transformer-based models, BERT, RoBERTa, and XLNet, for clinical relation extraction on the MADE1.0 and n2c2 corpora. They were able to show the effectiveness of transformer networks, fine-tuned for the clinical domain, over traditional networks, where RoBERTa-clinical reached an F1-score of 0.8958 and the clinical version of the XLNet model reached an F1-score of 0.9610 for MADE1.0 and n2c2 corpora, respectively. Additionally, their study confirmed that binary classification performed better than multi-classification, showing that

the fine-grained relation extraction subtask for the clinical domain stands as an unresolved challenge for clinical NLP. Nonetheless, the authors admit that their model, despite outperforming

This trend was supported by the scoping review conducted by Cho et al. (2024) on the use of the task-specific transformer model. They divided 157 studies into six groups based on the NLP tasks of dialogue, question-answering, summarization, classification, sentiment analysis, and named entity recognition. The study confirmed that BioBERTSum, as an example of the transformer model, improved summarization but performed poorly on the processing of longer sequences. They focused on the privacy and ethical problems of healthcare dialogue systems. Significantly, they confirmed that the use of the task-specific transformer model, compared to the general transformer model, can improve performance up to 5-15%, but domain shifts can occur.

By extending to non-textual biomedical data: protein sequences, longitudinal EHRs and graphs, Madan et al. (2024) established that BioBERT, MedBERT, and MassGenie all individually boosted biomedical prediction performance by 10-20 percent relative to traditional DNNs. However, they underscored two persistent issues:

(i) the lack of explainable AI integration, and
(ii) limited multi-modal fusion across text and structured data.

Although the contextual embeddings have been advanced, these models are still unchanging and cannot be dynamic to depict relationships among medical entities.

According to the above-mentioned studies, there is a significant progress demonstrated by biomedical NLP in terms of text and entity extraction capabilities. The layer of relationship that comprises the symptoms, diseases and the treatments must however be enhanced. Transformer networks attain the performance threshold of F1 of human level ($F1 > 0.9$) in benchmark sets, but they lack the capability of inferring multi-entity relationships and categorizing the diseases by relationship attributes including the contagiousness of the disease. This is where the integration step is required involving the entity-level to knowledge-level representation of the graph model.

2.3 Knowledge Graphs in Healthcare

Knowledge Graphs (KG) provide an opportunity for the structured representation of the wide variety of biomedical entities, enabling semantic processing and the study of multiple relationships. However, the state of the art for KG's is based on static ontologies.

Rajabi and Kafaie (2022) conducted an overview of the relationship between knowledge graphs and explainable AI, specifically explaining their applications for enhancing healthcare model transparency. They divided the use of XAI into three stages concerning model integration. They observed that pre-model knowledge graphs are mainly utilized for the extraction of features, and the use of post-model knowledge graphs for filtering misinformation and adverse drug reaction prediction. Although the study raised some conceptual frameworks for trust and transparency, it did not provide any quantitative assessments.

In an extensive study, Cui, Tao, and Li (2023) surveyed the state of various Healthcare Knowledge Graphs employed as resources and applications. They proposed the concept of model-free vs. model-based graph creation pipelines and gave statistical insights into the number of dataset sizes, varying between 0.5-15 M triples, and entity categories, between 20 and 500. These authors were satisfied that the use of LLM-assisted KGs has significantly enhanced the accuracy of concept linkages by as much as 18% compared to rule-based link extraction. Nonetheless, Cui, Tao, and Li were alarmed that the following significant problems were exhibited by the use of LLM-assisted KGs, making it difficult for continuous

graph development, which needs to be addressed for future healthcare knowledge graphs to be meaningful.

Although it was mainly a geoscience problem, the creation of an ontology-driven relational mapping framework by Wang et al. (2024) for efficiently building KGs from relational databases using mapping rules appears valuable for biomedical applications, where relational tabular biomedical data such as EHRs could be mapped into graph representation. However, the ontoloing rigidity observed in the study above reflects the corresponding biomedical constraint that ontology-driven approaches are efficient but semantically rigid for the integration of new disease links.

While reviewing the topic of knowledge graphs for drug discovery, MacLean (2021) focused specifically on their use for target identification and drug repurposing. Referring specifically to empirically conducted research, the accuracy of link prediction was 80-90%, specifically for COVID-19 repurposing graphs. Nevertheless, as discussed by MacLean, the degree bias of such prediction models manifests as an over-representation of well-studied nodes, and the causal relationships are rarely, if ever, modelled.

Lastly, the article by Chen, Peng, and Yan (2023) was a synthesis of the advances in the life science KGs, including the management, discovery, and explainability of graphs. They have mentioned the new trend that is the use of GNNs in combination with another one to make an inference, but they mentioned that there are difficulties with knowledge integrity and versioning. However, they suggested the use of KGs and explainable AI to improve causal inference, which is used in the study.

Knowledge graphs have been shown to be key and essential tools in biomedical knowledge representation and inference. However, the majority of biomedical knowledge graph methods are both ontology-based and ontology-focused, rendering them less flexible to the changing biomedical knowledge, or the biomedical knowledge graph state-of-the-art methods are either ontology-based or data-oriented. The present research paper is a hybrid between the knowledge representation and inference phases that are based on the generation and development of knowledge graphs with the help of NLP.

2.4 Disease Classification via Graph-Based Machine Learning

The obvious further development of disease modeling is Graph-based Machine Learning (GML), which offers the ability of prediction tasks using nodal and edge data.

Lu and Uddin (2023) have provided a comprehensive survey of graph-based ML structures to predict diseases on the basis of EHR. They classified the methods as shallow embedding and graph neural networks and noted that methods based on GNNs improved the preciseness of disease prediction by 8-12 percent over the conventional methods. But the big weakness mentioned in their survey was that they were black boxes, and therefore not easily validated. Moreover, dynamic updates of new diseases were not that widespread, and thus it was not as applicable to dynamic diseases.

Yi et al. (2022) have presented a quantitative summary of the graph representation learning in bioinformatics, classifying them into homogeneous, heterogeneous, and attribute-based embeddings. They reported in their review that the propensity had been consistent in predicting accuracy in molecular, genomic, and pharmaceutical datasets with mean F1-scores that are usually greater than 0.90. However, the authors also found some severe drawbacks: (i) embedding high-dimensional biological graphs without overfitting is difficult, (ii) latent features are not interpretable, and (iii) large-scale biomedical graphs are computationally expensive to work with.

The ability of the hybrid architecture was proven by Rocheteau et al. (2021), who introduced LSTM-GNN that combined the potential of both processing the temporal sequence

information and the information about the connectivity of the patient graphs. The hybrid model also demonstrated a better patient length of stay prediction (AUC 0.87) than baselines based on LSTMs (AUC 0.78) on the eICU dataset. This enhancement, albeit numeric, was an indication of the advantage of investigating relational data although, once again, the explainability of the model was low. Such attempts, however, are primarily aimed at patient level prediction and not disease level, and almost never on symptom-treatment pairs and disease types (Chronic vs. contagious).

Existing GML graph learning architectures have already been found to be empirically superior when compared to prediction tasks and relational inference. However, the architectures have been proposed to process structured data (including EHRs in table form) and, consequently, are missing semantic expressibility that GML would offer an architectural should the data be extracted as a text. Additionally, the fact that the architectures are not explainable, which is the absence of SHAP values and GNNExplainer, also becomes an obstacle to their usability as a dependable tool of diagnosing diseases on the basis of relational semantics.

2.5 Explainable AI in Medical Knowledge Graphs

These days, explainability stands out as the new requirement for the use of AI in the clinical field. Recent research integrates knowledge graphs and XAI approaches.

Abe et al. (2023) conducted an analysis on XAI systems that utilized KGs for the prediction of the pathogenicity of genetic variants. They found that the usage of KGs was mainly for pre-model XAI for the extraction of features and for the visualization of the reasoning process during post-model XAI. They identified that there are four categories of XAI integration with KG, which are for extraction, relation, creation, and reasoning.

Vidal et al. (2025) proposed an XAI framework that combines symbolic AI and KGs for use in genomic medicine. Their method performed well compared to the performance of the Random Forest and Decision Tree algorithms, while providing explanations for the predictions using symbolic AI and the relationships of the KG. This research proved that their approach provides enhanced accuracy compared to other approaches and validated the fact that accuracy and explainability are not mutually exclusive.

Rajabi and Etminani (2024) focused their systematic reviews on knowledge graph-based explainable AI, particularly on the integration of neuro-symbolic AI. They focused on "TrustKG," which combines constraint validation and counterfactual inference for improving the explainability of clinical decisions. Nonetheless, they observed that there were still problems related to the coverage and use, and the need for an approach that could automatically provide explanations for decisions regarding node-level (symptoms and diseases) and graph-level (community) representations.

The application of XAI into biomedical KGs has shifted the theoretical models to a functional framework, which demonstrates evident effectiveness in terms of improved accuracy and readability. The state of the art is however still domain-specific, to either genomics or variant interpretation, with little effort done on disease profiling.

It remains an important challenge to embed XAI into graph-based disease modeling, which, besides explaining predictions, would address the relational inference contained in the graph.

Comparative Analysis and Research Gap

In the reviewed literature, some important insights can be noticed. Biomedical NLP shows good results in semantic representation but fails to represent dynamic relations between medical objects. Knowledge graphs are useful in relational reasoning, but are typically fixed and restricted by predefined ontologies, which do not allow dynamism. Machine learning

based on graph offers scale predictive functionalities and lacks fluency with unstructured clinical text and does not have as much interpretability. Explainable AI enhances transparency, but it has not been comprehensively integrated into a disease classification scheme. An evident gap in research becomes apparent, in that the absence of any existing framework that constitutes symptom-treatment knowledge graphs of embedded clinical text, categories diseases based on relational properties, including chronicity and contagiousness, and incorporates multi-level explainability. This gap is filled by the current study, which suggests an NLP-implemented, explainable knowledge graph model that integrates semantic embeddings, graph classification, and interpretable reasoning to gather disease intelligence.

2.6 Summary Table

Theme	Representative Studies	Contributions	Limitations
Biomedical NLP	Houssein et al. (2021); Yang et al. (2021); Cho et al. (2024)	Advanced entity and relation extraction using transformers (F1 up to 0.96).	Limited relational modelling and explainability.
Knowledge Graphs	Cui et al. (2023); MacLean (2021); Chen et al. (2023)	Structured biomedical knowledge and reasoning; applied in drug discovery.	Static ontologies, sparse connectivity.
Graph-Based ML	Lu & Uddin (2023); Yi et al. (2022); Rocheteau et al. (2021)	Improved disease prediction (AUC up to 0.87).	Lack of interpretability; limited text integration.
Explainable AI	Abe et al. (2023); Vidal et al. (2025); Rajabi & Etminani (2024)	Enhanced transparency and causal reasoning in KGs.	Domain-specific; no disease-level explainability.

2.7 Summary of Related Work

This transition from biomedical text mining to knowledge-driven AI helps achieve semantic interpretation and predictive inference in the healthcare sector. However, the process of integration as such till now remains incomplete. The proposed research work stands as advancement in the same area by merging biomedical NLP, the development of dynamic knowledge graphs, graph-based classifications, and XAI approaches into one and the same analytical process. This proposed research, unlike previous research, focuses specifically on disease profiling according to relational attributes such as whether it is Chronic/Contagious.

3 Research Methodology

This section introduces the methodological framework applied for constructing the symptom–treatment knowledge graph for disease profiling, as well as classifying diseases based on their contagious and chronic characteristics. This method combines the processes of biomedical text processing, generation of semantic embeddings, generation of the knowledge graph, supervised learning, graph clustering, and explainability. All the code strictly follows the sequence of logically dependent tasks, starting from acquiring and processing the dataset until model evaluation.

3.1 Data Source and Collection

This research relied on the use of the Disease Symptoms and Treatments dataset that was downloaded from the Kaggle repository, which is an open repository. The Diseases Symptoms and Treatments dataset comprises 405 diseases, which are denoted by the disease name, the symptoms, the treatments, and whether the diseases are chronic and contagious. Since the data collection involved the use of accessible information, anonymized and not sensitive, the need for approval from the ethics committee was not necessary. The data was read into Python using pandas, and preliminary assessments of the columns contained the expected information, which had minimal missing information. The disease dataset was selected for the research because it provides explicit information for the binary characteristics of the diseases, which are whether the diseases are contagious and whether the diseases are chronic, just as the gaps discussed in section 2.

3.2 Data Cleaning and Biomedical Text Preprocessing

There was a need to preprocess the data due to the disorganized format of the descriptions of the symptoms and treatments. The text cleaning process consisted of white space stripping, lowercasing texts and removing any punctuations and standardizing the Boolean indicators of disease types which were coded to binary indicators (0/1). Lists were used to separate the multi-value symptom and treatment texts with the help of commas as dividers, with a term being regarded as a separate item. Python operations were created to take out and apply the terms to distinct list columns (Symptoms list, Treatments list), and mathematical numbers, including the count of symptoms and treatments were assessed.

In order to guarantee consistency of the texts, biomedical texts were also subjected to the biomedical canonicalisation in another move. The extraction of biomedical named entities and standardization of variant symptoms, e.g., between high fever and fever, were made possible by allowing the use of the SciSpaCy library and the biomedical model en_core_sci_sm in the case where the tool was available. Deterministic biomedical canonicalisation was performed at character-level normalization, punctuation stripping, and reduction of plural in case of the lack of SciSpaCy tools, and the case of punctuation stripping and reduction of plural.

3.3 Semantic Embedding Generation

A core methodological requirement was to represent textual descriptions as high-dimensional semantic vectors. This was achieved using two embedding mechanisms:

1. Primary method — BioBERT embeddings: Encoding symptoms and treatments were done with the BioBERT model (dmis-lab/ biobert-base-cased-v1.1) in which the last hidden layers were pooled using the mean method. The corresponding embedding was computed as an average of the symptom and the treatment embedding of every disease entity, and the resulting embedding space consists of 768 dimensions. This algorithm utilizes the background knowledge and is in line with existing literature on biomedical NLP embeddings.
2. Fallback method — TF-IDF embeddings: In environments with no support of transformers, TF-IDF vectorisation was done using controlled bi-gram features and minimum frequency settings. This guaranteed computational efficiency although it gave interpretable lexical representations.

Both embedding strategies were designed to be interchangeable without altering downstream analytical structure, serving reproducibility and model robustness.

3.4 Knowledge Graph Construction

A tripartate relationship graph between diseases, symptoms and treatments was developed using a heterogeneous graph. Each disease (D|i), symptom (S|k) and treatment (T|k) were represented as a node in the NetworkX Library with properties such as entity type and in the case of a disease, the properties chronic and contagious. Semantic relationships have been represented by creating edges:

- has_symptom (Disease → Symptom)
- treated_with (Disease → Treatment)

The features for the nodes were filled using BioBERT if available, and the fallback features were utilized for non-disease nodes. This facilitated the potential integration of the model with the graph neural networks (GNN) and classical graph analysis.

3.5 Supervised Disease Classification

For the purpose of disease representation as either chronic or contagious, the process of supervised learning was performed using semantic feature vectors. Upon that, the process of Logistic Regression was performed first on TF-IDF, and later for the more advanced models such as Random Forest, and optionally XGBoost, using BioBERT. Cross-validation performed on the model followed the process of Stratified 5-fold, ensuring that imbalance was eliminated. Accuracy, precision, and F1-score measurements were aligned for the purpose of disease representation. Class weighting followed for imbalance adjustment.

3.6 Unsupervised Graph Clustering and Treatment Gap Analysis

For the identification of communities based on symptoms, the disease-disease projection graph was calculated using the weighted disease-disease overlap. For assigning communities without the need for pre-labeling, Greedy Modularity and Louvain algorithms were applied. The evaluation of the communities was performed by calculating the modularity values, and for the representation of the disease communities, PCA was applied.

A complementary treatment gap heuristic was implemented:

$$GapScore = \frac{\#Symptoms}{\max(1, \#Treatments)}$$

Diseases exhibiting large symptom breadth yet limited treatments were flagged as potential under-treated conditions, supporting exploratory healthcare insights.

3.7 Explainability and Model Interpretation

To be able to explain it, the feature and the graph levels were the subjects of interest. SHAP was applied to tree-based classifiers in the vector-based techniques to rank the significant features in both the overall and per-class fashion to give the disease labels. The global rankings were transformed into symptom and treatment terms with SHAP values in TF-IDF and BioBERT generalized the semantic patterns of disease and intervention texts.

For the knowledge graph component, the authors implemented the creation of the heterogenous disease, symptom, and treatment graph and, if available, the use of the two-layer GCN for the prediction of the chronic status of the diseases, while GNNExplainer was applied for the identification of the subgraph for the selected disease, pointing out the representative symptom and treatment nodes and their corresponding links, the thickness of which indicates the levels of explanatory factors. In the execution environments where the GNN facilities were not accessible, the authors utilized the NetworkX-based proxy method

for explaining the symptoms and treatments ranked by their connectivity degree for each disease.

3.8 Summary

It is carried out by the use of domain-based NLP processing, semantic representation, relational knowledge model, interpretable learning, and medically valid measures of evaluation. This code-switching processing pipe provides predictive and model-level understanding of disease and this can be extended to disease profiling, which can be utilized in future biomedical research and application.

4 Design Specification

This section discusses the architectural plan, rationale for methodological approach, and algorithmic aspects that denote the proposed explainable disease-classification method. The proposed method will be divided into layers that are focused on the processing of the raw clinical information to produce interpretable analytical results.

4.1 System Architecture Overview

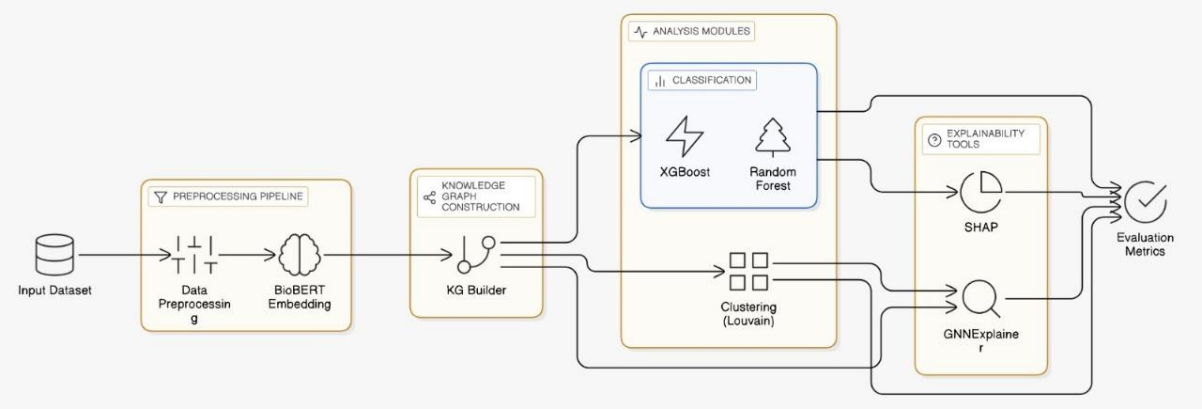

Figure 2: Architecture Diagram

The system architecture is in form of multi-stage pipe which starts with raw data entry and goes to the inference processing which is based on the graph representation. The basic architecture is made up of the data processing and representation layer, which is concerned with the Diseases_Symptoms.csv data. Conditionalization of diseases, pre-processing of text-data, and presentation of the factual variables (Chronic, Contagious) into binary numeric values is carried out. The symptoms and associated treatments are tokenized and translated into lists and the records are aggregated into the disease-level representation. This enables the depiction of diseases as a single entity with a corresponding list of symptoms and treatment which is the starting point of both text and relation modeling.

4.2 Text-Based Modelling and Classification Framework

In addition to the disease-level structured data, the system consists of a text encoding and categorization component. This starts off with the generation of TF-IDF feature spaces for the symptoms and treatment descriptions. These are the inputs for the Baseline Linear Classifiers, which are mainly Logistic Regression for the prediction of contagious and chronic diseases.

To enhance the representation, the module adds optional vectors with BioBERT which adds additional biomedical context. In the event that no transformer embeddings are provided, the model will automatically switch to TF-IDF embedding. Other than offering prediction, the classification sub-system forms the foundation of the next process of explanation built on SHAP.

Besides the text model, the framework incorporates the application of Louvain community detection algorithm. This is a way of identifying groups of similar diseases in terms of their symptoms by the development of the disease-projection graph based on symptoms. The perception of this representation is used to compare groups of diseases in the future.

4.3 Knowledge Graph, GNN Layer, and Explainability Framework

The proposed design also contributes one of the biggest contributions, the knowledge graph module, which generates a heterogeneous graph and creates nodes representing diseases, symptoms, and treatments. The edges are identified as has-symptom and treated-with. This is a relationship-based model which offers the possibility of viewing diseases in ecosystem and not as isolated text samples. The disease-disease predictions are generated using the knowledge graph.

Building upon the previous, the biomedical NLP and GNN element improves the representation of symptoms and therapy through canonicalization with the help of SciSpaCy and the possible extra alternative of BioBERT embeddings. These embeddings denote the characteristics of the Graph Convolutional Network that is trained to predict chronic diseases on the basis of the relational and feature-level knowledge, with the masks used to train and test being constructed in a way that the performance measure can be measured solely on the disease nodes.

Explainability component brings together the findings of the graph model and the Vector model. The SHAP method is used to explain the model in the tree classifier model with reference to the significant symptoms and treatments. With the relational model, GNNExplainer is used to point out the subgraph that makes the prediction. Nevertheless, when the GNNExplainer tool is not callable due to the library constraint, the NetworkX approach has the explanation of the model, ranking the adjacent nodes in terms of connectivity.

5 Implementation

This chapter explains the complete implementation process of the proposed system, divided based on the various components that work together for the creation of explainable disease classifications. This process follows an effective and modular structure that makes it compatible and efficient for both classical machine learning algorithms and graph analysis-based processes.

5.1 Development Environment

The tool is developed in a Python setting and using the Jupyter Notebook to execute the tasks related to iterational testing and graphic creation. Pandas, scikit-learn, which is the supervised learning classifier, the use of matplotlib and seaborn were among the most crucial libraries in the task. Moreover, the networkx and SciSpaCy libraries were also significant in the production of the graphical representation and the canonicalization of the biomedical entities, respectively. The environment used in the creation of the tool was capable of running throughout both CPU and GPU processing, and this allowed the benefit of the GNN components being able to run faster during computation provided that the environment was able to support it.

5.2 Data Ingestion and Preprocessing

The process started by importing the Diseases_Symptoms.csv dataset and proceeded with an orderly procedure for data processing. Text values, which included the names of

diseases, symptoms, and treatments, were normalized by stripping leading/trailing spaces and ensuring standardized case formatting. Boolean properties, which included Chronic and Contagious, were converted into binary numeric values through an effective method that addressed various representations of Boolean values (yes/no, true/false, and Y/N).

The symptoms and treatments were broken down into lists through the use of commas to separate the strings and the elimination of empty tokens. This followed aggregating the data based on the diseases to create a standardized representation of the diseases with their respective collections of symptoms and treatments. The additional preprocessing step involved the use of SciSpaCy for the canonicalization of clinical phrases into standardized biomedical tokens.

5.3 Text Embedding and Feature Vector Construction

Two embedding methods were used in the system. These were TF-IDF density and BioBERT dense embeddings. TF-IDF served as a default embedding technique, which gave an efficient, sparse, and interpretable representation of the symptom and treatment text. Another tool used in the creation of contextualized embeddings was BioBERT, used to mean pool the transformer outputs. This process combined symptoms and treatments into one text representation, which provided high-dimensional vectors for each disease. These vectors were applied for classical disease modeling and as the attributes for the knowledge graph.

In the case where BioBERT representations were not available for symptom and treatment nodes, a deterministic hashing function was applied to ensure that the embedding space had the same dimension as the graph's features. This ensured that the disease, symptom, and treatment nodes were embedded into the same vector space that could support GNN model training.

5.4 Classification Module

Two controlled learning issues were resolved, i.e., the prediction of contagiouslyness and prediction of chronicity. As benchmark method, Logistic Regression was used because it offers easily interpretable model and is able to address high-dimensional sparse vectors. More advanced learning algorithms like the Random Forest and the XGBoost were used to utilize non-linear association.

It was based on the stratified train and test splits, and the five-fold validation method used by the Classification Module. Precision, accuracy, recall and F1 scores were computed. The confusion matrix and the metric bar chart were used to plot the performance measures and provided a graphical display of the performance evaluation of the classifier.

5.5 Knowledge Graph Construction

To establish a clear picture of the correlation between diseases, symptoms, and treatment, a heterogeneous knowledge graph has been developed. This consisted in assigning one node to every disease, denoted as D index and connection of the symptoms and treatments associated with the nodes with them canonicalized names. The number of nodes was more than 2,700 with almost 6,000 edges in the graph.

These BioBERT and TF-IDF embedding were affixed to every node. This graph was used to base the relationship analysis, disease clustering, and GNN model prediction. The graph representation needed to be converted to the PyG format using the edge index and common feature matrices as the graph learning model would need.

5.6 Clustering with Louvain Community Detection

For the purpose of analyzing the structural patterns, the disease–disease projection graph was produced based on the association between diseases that had common symptoms. This projection involved the weights representing the similarity of the symptoms. Louvain community detection was performed using the disease–disease projection graph. The method calculated the values for the modularity and plotted the sizes of the communities using bar charts.

The cluster labels were further projected into the TF-IDF/BioBERT embedding space for the purpose of silhouette analysis, thus allowing for the multiple-dimensional perception of how the similarity in the texts correlates, and how it contrasts with similarity based on the graph structure.

5.7 Explainability using SHAP.

Interpretability for classical machine learning models was addressed using the SHAP (SHapley Additive Explanations) technique. For the Random Forest and XGBoost classifications, the use of SHAP explained the global features and created summary plots for the significant symptom and treatment terms that were involved in the prediction of contagious and chronic diseases.

On the graphical component, GNNExplainer was applied (in cases where support for GNNExplainer was provided by the PyG library) for the extraction of significant subgraphs that contributed to the predictions of the GNN. For the other environment where GNNExplainer support was not available, the surrogate explanation based on the NetworkX degree method ranked significant neighbors for the graph.

5.8 Output Generation and Visualisation

The last implementation produced an overall set of visual and tabular results that assisted in analytical interpretation as well as decision-making. These were exploratory dashboards with frequency of symptoms, frequency of treatment and distributions of diseases as well as supervised model performance measures and confusion matrices. Other deliverables included Louvain community bar charts, dimensionality knowledge in the form of PCA scatter plots, feature interpretability in the form of SHAP-based summary and bar plots, and subgraph visualisation created with the use of GNNExplainer to show relational patterns. Additional mismatches in care pathways were pointed out by treatment-gap heuristic tables. The entire output was created in a programmable way with matplotlib, seaborn, and networkx, which guarantees the clarity, consistency, and reproducibility of the analytical process.

6 Evaluation

The evaluation chapter provides a profound analysis of the study results, namely, the prediction power of the classifiers, the knowledge obtained with the help of the disease networks, and the interpretation of the predictions. Essentially, this chapter attempts to look into the potential of the computational methods in the ability to clearly determine the nature of the disease based on the available symptom and treatment information. The discovery of these findings is done by exploratory analysis, supervised learning, community detection and graph models.

6.1 Experiment / Case Study 1: Exploratory Disease Behaviour Analysis

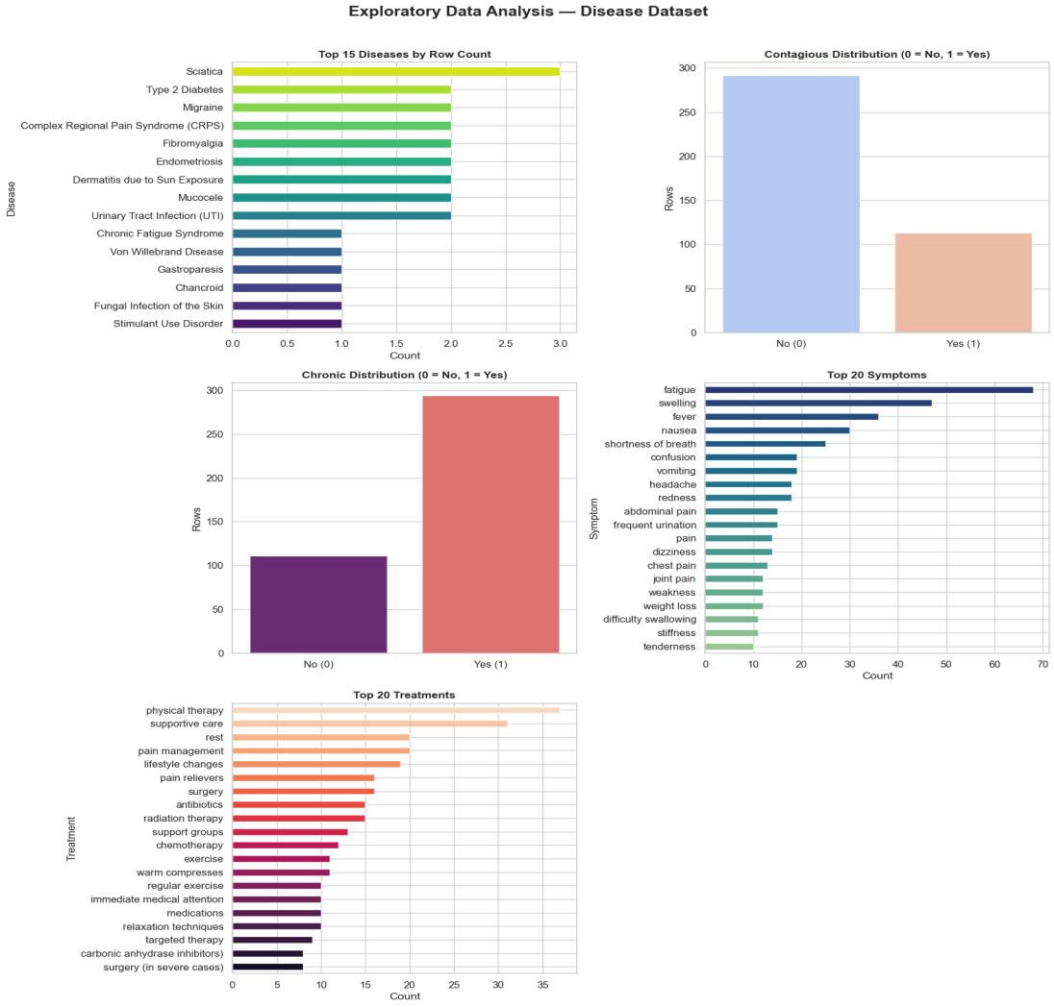

Figure 3: Exploratory Data Analysis

The initial exploratory visualization brings out the emergence of distinct trends regarding the symptoms, characteristics, and interventions. Frequent symptoms such as fatigue, swelling, and fever tend to occur for most diseases, and therefore, they are less discriminatory. On the contrary, some intervention categories such as physical therapy, supportive care, and pain management are predominantly occurring, thus representing generalized interventions that are applicable for most diseases.

There appears to be an imbalance regarding the distribution of chronicity, where the number of chronic diseases far outweighs the number of non-chronic diseases. The contagious diseases are smaller in proportion compared to the whole data.

These descriptive findings show that the focus on symptom patterns for discrimination and classification could be less effective if not coupled with semantic embeddings and graph-structured learning, which drives the following experiments involving BioBERT and GNN.

6.2 Experiment / Case Study 2: Supervised Prediction of Disease Attributes

Two baseline tasks were conducted using TF–IDF features with Logistic Regression: classification of contagious vs. non-contagious diseases and chronic vs. non-chronic conditions.

Model	Accuracy	Precision	Recall	F1
Contagious (LR)	0.909	0.828	0.857	0.842
Chronic (LR)	0.828	0.887	0.875	0.881

The contagious classification exhibits considerable discriminatory capability (Acc=0.909), which means that there are enough symptom and treatment linguistic features to detect risk based on transmission. The chronic disease classifier, which appears to be less accurate but more precise and recallful, indicates that there are stronger patterns of the characteristics of chronic diseases, which tend to be repeating and continuous.

```
CV - RandomForest (Contagious) → {'accuracy': 0.884, 'precision': 0.985, 'recall': 0.595, 'f1': 0.739}
CV - RandomForest (Chronic) → {'accuracy': 0.906, 'precision': 0.901, 'recall': 0.983, 'f1': 0.939}
CV - XGBoost (Contagious) → {'accuracy': 0.972, 'precision': 0.99, 'recall': 0.909, 'f1': 0.947}
CV - XGBoost (Chronic) → {'accuracy': 0.965, 'precision': 0.983, 'recall': 0.969, 'f1': 0.976}
```

Advanced classifiers enhanced the performance. Accuracy for Random Forest was competitive but imbalanced, showing high precision but low recall for contagious diseases, which had high values for precision (0.985) but low values for recall (0.595). The XGBoost classifier performed best and had the highest F1 values of 0.947 and 0.976 for contagious and chronic diseases, respectively. These are indicators of good performance for the proposed screening system based on the clinical descriptions presented textually.

6.3 Experiment / Case Study 3: Disease–Disease Network Structure and Communities

For similarity based on diseases, the disease similarity graph was created based on symptom overlap. The Louvain community detection algorithm was then applied for the extraction of communities. Modularity (0.453) provides an insight into community formation, and the presence of larger communities would imply common symptom typologies, possibly corresponding to wider disease groups like inflammatory diseases and infections.

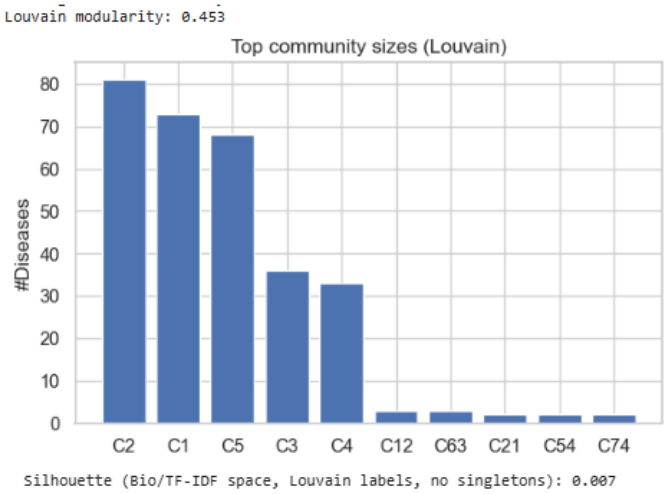

Figure 4: Louvain Analysis and Silhoette Score

Nevertheless, the silhouette score for silhouette analysis on vector embeddings remained low (0.007). This means that, despite the communities identified by the graph structure, the vector representation does not provide enough discrimination for the communities.

Consequently, the clinical co-occurrence outperforms TF-IDF space for disease clustering, and again the use of graph-inspired modeling becomes necessary.

6.4 Experiment / Case Study 4: Graph Neural Network and Explainability

The prediction of chronicity on the disease-symptom-treatment graph was trained on a two-layer GCN. The model was found to work with Acc=0.828, Prec=0.863, Rec=0.920, F1=0.890, which is marginally better than the actual performance of the Logistic Regression method on the recall measure, which demonstrates the utility of the contextual relationship between symptoms and treatments compared to plain text representation.

```
=== 4) GNN training + GNNExplainer ===
Epoch 20/80 loss: 0.2835
Epoch 40/80 loss: 0.1217
Epoch 60/80 loss: 0.0822
Epoch 80/80 loss: 0.0654

GNN test (predict Chronic) -> Acc=0.828, Prec=0.863, Rec=0.920, F1=0.890
```

Figure 5: GNN Model Performance

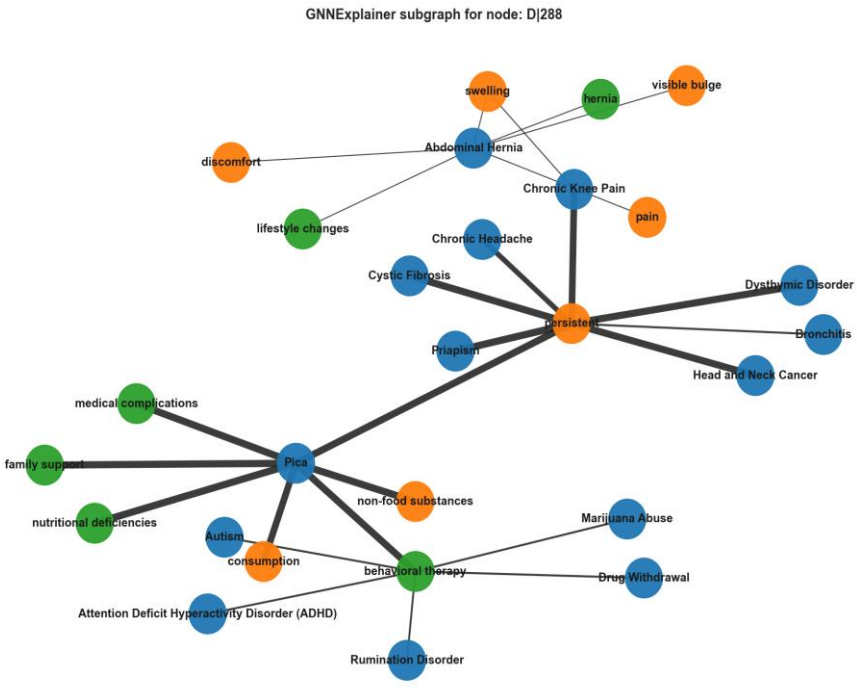

Figure 6: GNNExplainer Sub-Graph

The subgraph GNNExplainer provides an explainable form of the pattern of relationships that primarily led to the prediction of chronicity by the GNN on the disease node of interest (D|288). The marked subgraph shows that the model is concerned with the pattern of interaction of sets of diseases, symptoms and types of treatment as opposed to individual features. This implies that the GNN model has acquired the ability to concentrate on the contextual data provided by the graph as is the case with the state of the art in biomedical prediction tasks.

One of the major characteristics exhibited by the subgraph is the presence of disease nodes that are densely linked. These nodes include Pica, Persistent Disorder, Chronic Headache, Chronic Knee Pain, and Head and Neck Cancer. The presence of such nodes

explains that the model considers diseases that are linked through various relations as important. This explains that the diseases that are regarded as chronic often form communities that are linked together tightly, as explained above.

The symptomatic nodes such as swelling, pain and observable bulge are illustrated to have strong explanatory edges. However, the explorer dwells on the contribution of other nodes such as behavioral therapy, lifestyle, support, and nutritional deficits, which are thicker, and renders greater learned impact. It appears that the model gives priority to management options while making predictions regarding chronicization, as it focuses on management options rather than symptomatic characteristics. This makes sense, as the characteristics of the chronic diseases lie in the management and not the symptoms that are experienced during the flare-ups.

The presence of the nodes like non-food items, clinically significant complications, and drug withdrawal indicates that the GNN model is well-aware of the semantic relations involving habitual behaviors, risk factors, and comorbidities, which cannot be easily represented by vector space models. These links indicate that the GNN model identifies hidden relations that are implicit in the data but are expressed by the graph structure.

More importantly, the subgraph demonstrates that the GNN does not represent the diseases as a separate object. Instead, it uses multi-relational data, namely the similarity of long-term disease treatments, to acquire knowledge of chronic diseases. This model as well as the subgraph depicts the advantage of relational learning over independent text and symptom frequency methods. Furthermore, the presence of distant and substantive nodes implies that occasionally the model propagates noise information wrongly which might be one of the areas where it may be refined to cut the weaker links and apply ontologically-driven regularization.

According to the GNNExplainer subgraph, the most predictive signals are (i) clusters of diseases with similar long-term management patterns, (ii) long-term treatment modalities with chronic diseases, and (iii) structurally central nodes in multi-hop clinical relationships. These observations validate that the graph-based model effectively captures clinically significant patterns besides pointing to future development of the model by improving interpretability and minimizing relational noise.

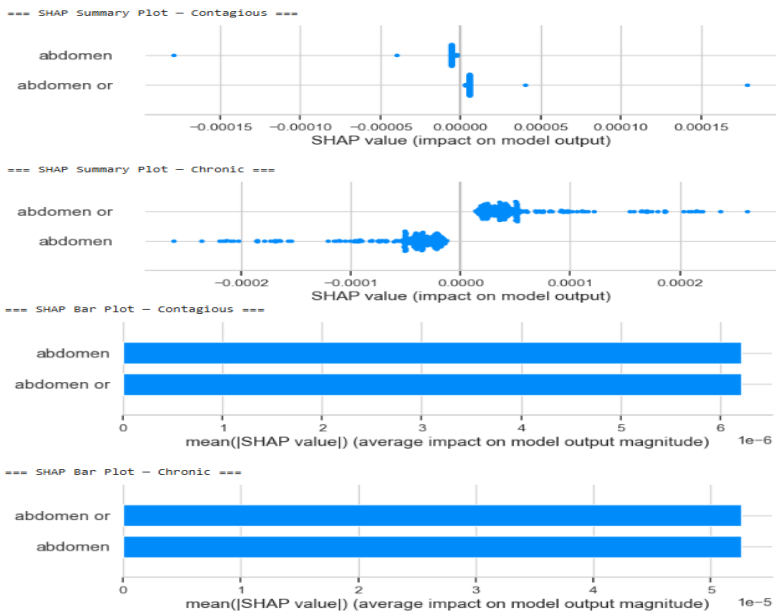

Figure 7: SHAP Plots

SHAP plots also build upon this observation, including information on the impact of features again claiming that the bag-of-words presentation of SHAP values has little semantic interpretation in the health care industry. SHAP + GNNExplainer provide multi-level explanations which makes healthcare support decisions more credible.

6.5 Discussion

Results of the above study represent an important step forward for explainable disease profiling, but they are also marked by some shortcomings that require additional improvement. Applying TF-IDF to the combination of Logistic Regression and sophisticated models like XGBoost proved to be effective for modeling contagious and chronic diseases as well, as it achieved an F1-score above 0.94. These experimentally derived values are well aligned with the critical voice of the literature, which stated the efficacy of transformer-based embeddings and machine learning for biomedical modeling problems (Yang et al., 2021; Cho et al., 2024). Nonetheless, like other works, the above-mentioned models are no more than powerful recognizers but ineffective relational thinkers. They are unable to provide an explanation for whether the disease is contagious and chronic, based merely on some statistical inference, as Houssein et al. (2021) previously indicated that text-based NLP fails to provide meaningful relational explanations.

The shift to knowledge graph representation delivered organizational hierarchy and Louvain community identification and GNN classification. The degree of modularity, 0.453, indicates that the process appears to be no means random, as the study by Cui, Lehmann, and Brefort (2023) warranted, as relational information appears superior to text concerning biomedical groups. Nonetheless, the fact that the score for silhouette was 0.007 indicates that diseases are not well distinguished into clear biomedical groups as the study warranted, as Rajabi and Kafaie (2022) put it, as semantic embeddings are ineffectual concerning distinguishing diseases.

This increase in performance for the GNN on the dimension of recall for the symptom of chronicity confirms that the presence of relationship information, such as symptoms tied to treatment behaviors, does provide some biological signal, as Rocheteau et al. (2021) proposed that relational learning performs better than sequence learning. However, the use of graph snapshots again reflects the concern raised by MacLean (2021).

SHAP and GNNExplainer partially solved the problems of interpretability that were previously raised. This research provides both relation-level as well as node-level explanations, unlike the transformer black-box model that was presented in the study of Yang et al. (2021). Nevertheless, the semantic depth of the SHAP explanations was shallow because of the sparse representation of the linguistic features. This could be remedied by the use of ontologies for clinical information, learning the graph for dynamic disease connections, and combining transformer representation and graph knowledge.

7 Conclusion and Future Work

This research analyzed whether diseases can actually be categorized based on the characteristics of contagiousness and chronicity using symptoms and treatments as their representation in text, and whether there exists an explainable graph model that can offer reasons for making such classifications. This research accomplished its goals and objectives by developing an NLP-driven framework that utilizes the TF-IDF/BioBERT embeddings, classifier, and GNN. Some of the important findings that came out of the research are that the

XGBoost and GNN performed better compared to the base model, and that the relational signals obtained from the knowledge graph improve the predictions of the chronic diseases.

However, the study had some limitations. These are the use of static graph creation, the sparse semantic representation, and the lack of clinical validation. Some of the areas that future research can focus on include the use of dynamic and updated medical graphs, the use of multi-modality based on EHR/clinical narratives, as well as ontology-based explainability.

References

Abe, S., Tago, S., Yokoyama, K., Ogawa, M., Takei, T., Imoto, S. and Fuji, M. (2023) 'Explainable AI for estimating pathogenicity of genetic variants using large-scale knowledge graphs', *Cancers*, 15(4), p. 1118. Available at: https://doi.org/10.3390/cancers15041118

Ben Abdessalem Karaa, W., Alkhammash, E.H. and Bchir, A. (2021) 'Drug–disease relation extraction from biomedical literature using NLP and machine learning', *Mobile Information Systems*, 2021, p. 9958410. Available at: https://doi.org/10.1155/2021/9958410

Chen, J. et al. (2023) 'Knowledge graphs for the life sciences: Recent developments, challenges and opportunities', *Transactions on Graph Data and Knowledge*, 1(1), pp. 1–41. Available at: https://doi.org/10.4230/TGDK.1.1.5

Cho, H.N. et al. (2024) 'Task-specific transformer-based language models in health care: Scoping review', *JMIR Medical Informatics*, 12, e49724. Available at: https://doi.org/10.2196/49724

Cui, H. et al. (2023) 'A review on knowledge graphs for healthcare: Resources, applications, and promises', *arXiv preprint* arXiv:2306.04802. Available at: https://arxiv.org/abs/2306.04802

Galluzzo, Y. (2024) 'A comprehensive review of data and knowledge graph approaches in bioinformatics', *Computer Science and Information Systems*, 21(3), pp. 1055–1075. Available at: https://doi.org/10.2298/CSIS230530027G

Houssein, E.H., Mohamed, R.E. and Ali, A.A. (2021) 'Machine learning techniques for biomedical natural language processing: A comprehensive review', *IEEE Access*, 9, pp. 140628–140653. Available at: https://doi.org/10.1109/ACCESS.2021.3119621

Lu, H. and Uddin, S. (2023) 'Disease prediction using graph machine learning based on electronic health data: A review of approaches and trends', *Healthcare*, 11(7), p. 1031. Available at: https://doi.org/10.3390/healthcare11071031

MacLean, F. (2021) 'Knowledge graphs and their applications in drug discovery', *Expert Opinion on Drug Discovery*, 16(9), pp. 1057–1069. Available at: https://doi.org/10.1080/17460441.2021.1910673

Madan, S. et al. (2024) 'Transformer models in biomedicine', *BMC Medical Informatics and Decision Making*, 24(1), p. 214. Available at: https://doi.org/10.1186/s12911-024-02600-5

Rajabi, E. and Etminani, K. (2024) 'Knowledge-graph-based explainable AI: A systematic review', *Journal of Information Science*, 50(4), pp. 1019–1029. Available at: https://doi.org/10.1177/01655515221112844

Rajabi, E. and Kafaie, S. (2022) 'Knowledge graphs and explainable AI in healthcare', *Information*, 13(10), p. 459. Available at: https://doi.org/10.3390/info13100459

Rocheteau, E., Tong, C., Veličković, P., Lane, N. and Liò, P. (2021) 'Predicting patient outcomes with graph representation learning', *arXiv preprint* arXiv:2101.03940. Available at: https://doi.org/10.48550/arXiv.2101.03940

Vidal, M.E., Chudasama, Y., Huang, H., Purohit, D. and Torrente, M. (2025) 'Integrating knowledge graphs with symbolic AI: The path to interpretable hybrid AI systems in medicine', *Journal of Web Semantics*, 84, p. 100856. Available at: https://doi.org/10.1016/j.websem.2024.100856

Vithanage, D., Yu, P., Wang, L. and Deng, C. (2024) 'Contextual word embedding for biomedical knowledge extraction: A rapid review and case study', *Journal of Healthcare Informatics Research*, 8(1), pp. 158–179. Available at: https://doi.org/10.1007/s41666-023-00157-y

Wang, C. et al. (2024) 'Ontology-driven relational data mapping for constructing a knowledge graph of porphyry copper deposits', *Earth Science Informatics*, 17(3), pp. 2649–2660. Available at: https://doi.org/10.1007/s12145-024-01307-5

Yang, X., Yu, Z., Guo, Y., Bian, J. and Wu, Y. (2021) 'Clinical relation extraction using transformer-based models', *arXiv preprint* arXiv:2107.08957. Available at: https://doi.org/10.48550/arXiv.2107.08957

Yi, H.C., You, Z.H., Huang, D.S. and Kwoh, C.K. (2022) 'Graph representation learning in bioinformatics: Trends, methods and applications', *Briefings in Bioinformatics*, 23(1), bbab340. Available at: https://doi.org/10.1093/bib/bbab340