# **PROJECT REPORT ON**

**Aid Amyotrophic Lateral Sclerosis via Knowledge Graph**

**Table of contents**

[INTRODUCTION 3](#_Toc149644610)

[RELATED WORK 4](#_Toc149644611)

[APPROACH 5](#_Toc149644612)

[Data/Problem Analysis 5](#_Toc149644613)

[Survey of Tools and Resources Available 6](#_Toc149644614)

[Software Design 7](#_Toc149644615)

[Source Code Description 8](#_Toc149644616)

[EVALUATION AND RESULTS 9](#_Toc149644617)

[Metrics used 9](#_Toc149644618)

[CONCLUSION 13](#_Toc149644619)

[REFERENCES 15](#_Toc149644620)

[APPENDIX 16](#_Toc149644621)

# **Introduction:**

# Neurodegenerative diseases, which include a variety of conditions that progressively harm the nervous system, are a big challenge in healthcare today. Amyotrophic Lateral Sclerosis (ALS), popularly referred to as Lou Gehrig's disease, is one of these illnesses. Genetics, environment, and lifestyle all have a role in the complicated disease known as ALS. Loss of muscle control results from an attack on the brain's and spinal cord's motor neurons. It is challenging for physicians and scientists to treat this dangerous illness, which has a significant negative impact on people's quality of life. Finding new approaches to research and comprehend ALS is critical because the disease worsens swiftly and there are now no effective treatments. The complicated anatomy of diseases such as ALS presents challenges for traditional medical research approaches. This project introduces the concept of an ALS Knowledge Graph (KG), a tool for organizing and visualizing information that explains the relationships between various elements such as genes, chemicals, and diseases. The ALS-KG is particularly significant as it simplifies the understanding of complex connections associated with ALS, offering a comprehensive map of relationships that can assist medical professionals and researchers.

# **Develop an ALS Knowledge Graph using Advanced Language Models:**

In order to accomplish this, we make use of advanced Large Language Models (LLMs) that have proven efficient in comprehending and placing complex biomedical data in context. We seek to extract entities and relationships relevant to ALS from the last ten years (2013–2023) of peer-reviewed literature. We are able to go through enormous volumes of textual data and discover important elements like genes, illnesses, chemicals, and their complex interactions by using sophisticated language models like ERNIE-Health and Stanford OpenIE.

# **Validate and Refine the ALS-KG for Accuracy and Reliability:**

# Recognizing the inherent challenges in automated knowledge extraction, we implement a robust validation process to ensure the accuracy and reliability of the ALS-KG. Manual validation, involving domain experts, offers a meticulous review of the constructed graph, correcting any inaccuracies or misinterpretations. Simultaneously, we explore innovative techniques for automated validation, incorporating confidence measures and combination strategies. This dual-validation approach aims to create a comprehensive and trustworthy ALS-KG, bridging the gap between automated extraction and expert-reviewed precision.

# **Related work:**

# Amyotrophic Lateral Sclerosis (ALS) has been extensively studied in the field of diseases that harm the nerves in the brain and spinal cord. In our project, we will use information from previous studies to create a detailed ALS Knowledge Graph (ALS-KG). This will help us better understand ALS. This review of previous studies looks at different important research areas that have led to our project and helped us understand ALS better [2].

# Over time, scientists have realized that knowledge graphs are useful for organizing and showing complex relationships in biomedical fields. People have been trying to make knowledge graphs specifically for ALS by using different methods for understanding and processing language. These knowledge graphs have been very important in connecting ALS, its genetic causes, and possible treatments. By studying and getting ideas from previous research, our project aims to make a better and broader knowledge system for ALS. This system will include the latest tools and methods to help us understand disease better. Bean et al. (2020) [1].

# A lot of ALS research has been studying the genes that are involved in causing the disease. Researchers have found that certain genes, including SOD1, C9orf72, FUS, and TARDBP, are linked to cases of ALS that run-in families. These findings about genes and disease have given us important information about how disease works. This information could help us come up with better ways to treat it. We want to use genetic information to better understand ALS and how it is related to genes [3].

# Neurodegenerative disorders represent a major worldwide health concern, with Amyotrophic Lateral Sclerosis (ALS) being one of the most complex and mysterious conditions. This review of the literature seeks to give a thorough overview of current studies on neurodegenerative diseases, knowledge graph embedding techniques, and applications, along with how artificial intelligence tools are being used to better understand and treat these affecting illnesses.

# *Neurodegenerative Diseases and Metabolic Deficiency*

# In a seminal work by Muddapu et al. (2020), the authors delve into the potential root causes of neurodegenerative diseases, exploring the intriguing link between these conditions and metabolic deficiency. The review highlights the intricate interplay between metabolic dysregulation and the onset of neurodegenerative diseases, shedding light on potential therapeutic avenues (Muddapu et al., 2020). This foundational work sets the stage for understanding the multifaceted nature of ALS and its underlying metabolic intricacies[20].

# *Knowledge Graph Embedding Approaches*

# Wang, Q., Mao, Z., Wang, B., & Guo, L (2017) provide a comprehensive survey of knowledge graph embedding approaches in their work. The review spans various techniques and applications, offering insights into the evolving landscape of knowledge representation. This knowledge is crucial for constructing a robust ALS Knowledge Graph (ALS-KG), emphasizing the need for embedding models to capture intricate relationships between entities in the biomedical domain (Wang et al., 2017)[18].

# *Applications of Knowledge Graphs in Tracing Outbreaks*

# Shen et al. (2023) present a compelling case for the application of knowledge graphs in tracing the source of large-scale outbreaks. The study, focused on the Beijing municipality outbreak in 2020-2021, demonstrates the versatility of knowledge graphs in epidemiological research (Wang et al., 2017). This real-world application highlights the potential of knowledge graphs not only in biomedical domains but also in addressing broader public health challenges[22].

# *Biological Applications of Knowledge Graph Embedding Models*

# The work by Mohamed et al. (2021) delves into the biological applications of knowledge graph embedding models. The review explores how these models contribute to unraveling complex biological relationships, a perspective valuable for constructing a nuanced ALS-KG (Mohamed et al., 2021). By examining the intersection of knowledge graph embeddings and biology, the review provides insights into potential methodologies and challenges in the context of neurodegenerative diseases[19].

# *Artificial Intelligence Tools in Neurodegenerative Diseases*

# Tauţan et al. (2021) offer a comprehensive review of artificial intelligence tools in neurodegenerative diseases, with a specific focus on machine learning techniques [5]. The integration of artificial intelligence tools into the study of ALS holds promise for diagnostic assistance, personalized treatment, and deeper insights into the disease's complexities (Tauţan et al., 2021). This review serves as a guide for leveraging advanced technologies in constructing the ALS-KG [4].

# *Integrative, Interactive Machine Learning in Biomedical Informatics*

# Holzinger and Jurisica (2014) advocate for integrative, interactive machine learning solutions in biomedical informatics. Their work emphasizes the importance of collaborative and interactive approaches in handling complex biomedical data. This perspective is crucial for designing the ALS-KG construction process, ensuring a dynamic and user-friendly interface for researchers and clinicians [6].

# *Pretrained Contextualized Embeddings on Structured Electronic Health Records*

Rasmy et al. (2021) presents an impactful contribution with their work on Med-BERT, a pretrained contextualized embedding model trained on large-scale structured electronic health records. This work holds relevance in the context of constructing the ALS-KG, as it showcases the effectiveness of leveraging pre-existing structured data for disease prediction. Med-BERT's approach, grounded in the rich information present in electronic health records, can serve as a guiding principle for the extraction of relevant entities and relationships from the vast landscape of biomedical literature (Rasmy et al., 2021). The Med-BERT model excels in capturing contextualized information from structured data, which aligns with the objectives of the ALS-KG project. The structured nature of electronic health records mirrors the organized format sought in constructing a knowledge graph.By adopting similar principles of contextualized embeddings, the ALS-KG can benefit from a more nuanced representation of relationships between various entities associated with ALS,

# facilitating a deeper understanding of the disease and its underlying mechanisms [15].

# *Building Knowledge Graphs from Scientific Literature*

# Delmas et al. (2021) contribute valuable insights into the systematic construction of knowledge graphs from both public databases and scientific literature. Their work provides a roadmap that is particularly pertinent to the ALS-KG project, emphasizing the importance of a systematic approach in handling diverse sources of information (Delmas et al., 2021). The methodologies proposed by Delmas et al., especially regarding the extraction of associations between chemicals and diseases, offer a robust foundation for the construction phase of the ALS-KG (Delmas et al., 2021). In the context of neurodegenerative diseases, where the available information is vast and diverse, the structured approach advocated by Delmas et al. becomes crucial. By systematically extracting and organizing information from scientific literature, the ALS-KG can ensure a comprehensive representation of the relationships between various biomedical entities associated with ALS. This structured knowledge base is fundamental for researchers and clinicians seeking to navigate the intricate landscape of neurodegenerative diseases [16].

# *Generating Knowledge Graphs from Scientific Literature of Degenerative Diseases*

# Rossanez and dos Reis (2019) focus specifically on generating knowledge graphs from scientific literature related to degenerative diseases. While their emphasis is not on ALS, their approach and methodologies offer valuable guidance for constructing a knowledge graph dedicated to neurodegenerative diseases. The practical techniques they present for systematically extracting structured information from scientific literature align with the goals of the ALS-KG project. In neurodegenerative diseases, where literature is rich with information but dispersed across various sources, the methodology proposed by Rossanez and dos Reis becomes particularly relevant. By employing their systematic techniques for information extraction, the ALS-KG can ensure that insights from diverse scientific literature are captured accurately. This inclusive approach contributes to a more holistic representation of the knowledge landscape surrounding ALS, aiding researchers in uncovering hidden connections and patterns [5].

# *Mining Alzheimer's Disease-Related Knowledge Graph for Drug Repurposing*

# Nian et al. (2022) present a mining approach focused on Alzheimer's disease-related knowledge graphs, with the aim of identifying potential semantic triples for drug repurposing. While their primary focus is on Alzheimer's disease, the methodology they employ is transferable to the ALS-KG project, providing valuable insights into leveraging knowledge graphs for identifying potential therapeutic strategies. In neurodegenerative diseases, the challenge of finding effective treatments is paramount. The approach by Nian et al. offers a strategic methodology for repurposing existing knowledge to identify potential therapeutic avenues. By adapting their mining techniques to the context of ALS, the ALS-KG can become a valuable resource for researchers seeking novel insights into potential drug candidates or repurposing opportunities. This cross-disciplinary application of methodologies reinforces the versatility of knowledge graphs in advancing biomedical research [21].

# *Techniques for Translating Embeddings in Multi-Relational Data*

# Bordes et al. (2013) delve into techniques for translating embeddings in multi-relational data [12]. Their study provides a foundational understanding of the intricacies of embedding models, which is crucial for the construction of the ALS-KG. The techniques discussed in their work are directly relevant to capturing nuanced relationships within the biomedical domain. In neurodegenerative diseases, the relationships between various entities are often complex and multifaceted. Bordes et al.'s techniques for translating embeddings offer a nuanced approach to represent these intricate relationships within the ALS-KG accurately. By incorporating these techniques, the ALS-KG can provide a more sophisticated understanding of the associations between different entities, contributing to a richer and more comprehensive knowledge representation [7].

# *Embedding Entities and Relations for Learning and Inference in Knowledge Bases*

# Yang et al. (2014) makes a significant contribution by discussing the embedding of entities and relations for learning and inference in knowledge bases [13]. Their work provides fundamental insights into embedding techniques, which are essential for accurately representing entities and relationships within the ALS-KG. The ALS-KG, as a knowledge base, relies heavily on the accurate representation of entities and their relationships. Yang et al.'s work offers a theoretical foundation for embedding techniques, guiding the ALS-KG project in constructing a robust knowledge representation. By incorporating these embedding techniques, the ALS-KG can capture the nuanced associations between different entities, facilitating more accurate inferences and insights for researchers and clinicians [8].

# **Methodology:**

1. **Corpus selection:**

We used a Python script to execute web scraping in order to automatically collect data from PubMed, a well-known database for biomedical literature, during the paper extraction stage of our study. Our script was specifically designed to look for publications published between 2013 and 2023 that explored Amyotrophic Lateral Sclerosis (ALS).

By constantly reviewing the search results pages, the script taken the manual search procedure by going through the pages of PubMed. It was designed to traverse up to five pages to ensure a comprehensive collection of data. Three essential pieces of data were extracted by the script for each article it found: the abstract, which offers a succinct synopsis of the article's contents, the paper's title, and its distinct PubMed identification number (PMID).

The pandas library, which excels at handling data, was then used to carefully arrange this acquired data into a structured format known as a DataFrame. The resulting DataFrame served as the foundation for our dataset, ready for the subsequent steps of entity extraction and analysis.

1. **Entity Extraction Using PubTator:**

In the entity extraction phase, we programmed our Python script to interact with the PubTator API for each collected paper's abstract. An recognized list of biomedical items, including genes and disorders, together with their classifications was obtained by PubTator after a batch of PMIDs was systematically submitted there by the script. These annotations were recorded and added to a text file so that they could be combined.

The resulting program then parsed the text file's accumulated annotations to extract entities and their corresponding categories. Next, these were added to the already-existing DataFrame such that each paper's title and abstract were accompanied by comprehensive biomedical keywords. Through this methodical process, we transformed our preliminary collection of ALS research papers into a dataset intricately annotated with biomedical entities, setting the stage for advanced analysis and insight generation.

After collecting and interpreting the annotations, we improved the data in our DataFrame. A list of biomedical things, such as particular genes, diseases, and chemical compounds, each labelled with its associated kind, was now included to the record of each paper. The careful classification made it easier to comprehend each abstract's content and providing a rich context for the papers in our corpus.

**Consolidating Data for Analysis**

Our study relied heavily on the enhanced Data Frame, which provided a list of entities and their categories for every publication. This made it easier to comprehend the studies in more detail than just their titles and abstracts and enabled a thorough examination of the state of ALS research. The successful extraction and categorization of entities marked a pivotal step towards synthesizing a comprehensive knowledge base from the amassed literature.

1. **Entity Merging:**

Phase three involves entity merging in the creation of an ALS knowledge graph. Refinement of the data was a crucial step after extracting entities from the scientific literature on ALS and making sure that different phrases relating to the same notion were correctly identified. As a necessary step for a consistent comparison, we first normalized the entities in this phase by changing all of the terms to lowercase and eliminating all punctuation. For Lemmatization, we used spacy model which condensed words to their dictionary or basic form, was also used. One example of this would be the simplification of "running," "ran," and "runs" into just "run."

Next, in order to assess the similarity of phrases, we employed a method that included the Levenshtein distance and the Jaccard similarity. The Levenshtein distance calculates the number of changes required to turn one term into another, while the Jaccard similarity gauges the resemblance between sample sets. We were able to determine with accuracy which two terms were variations of one another by combining these two measurements.

Duplicate terms were detected for each entity type because they were judged to be extremely similar based on a predetermined threshold. We made sure that each distinct notion was only ever represented once in our dataset by using a standard phrase to describe each group of related things. This merging was not random.

Now that redundant terms had been removed from the information and it was standardized, a framework for mapping and identifying the relationships between these entities was established. This mapping serves as the foundation for the next stage of the ALS knowledge graph construction, which will clarify the relationships between various elements, such genes linked to ALS or the effects of various treatments on the disease. With this knowledge graph, we hope to present a thorough overview of the state of ALS research, possibly providing novel viewpoints and directing future investigations.

## **Relation extraction:**

During the relationship extraction stage, we focused on defining the connections among the biomedical objects we had located. Our goal was to understand the complex connections found in the scholarly literature on ALS (amyotrophic lateral sclerosis).

To do this, we used Stanza, a Natural Language Processing (NLP) tool, which enabled us to methodically look at the context of each entity's appearance in the abstracts. The Stanford University-developed Stanza toolbox is proficient in processing and comprehending natural language. We could parse phrases and determine significant correlations between terms by utilizing its skills.

Each abstract was processed through the CoreNLP package of Stanza to identify the triplets of 'subject', 'relation', and 'object'. These triplets represent the fundamental structure of a relationship within a sentence. For instance, if an abstract included the following sentence: "Riluzole delays the progression of ALS." In this case, the subject would be identified as "Riluzole," the relation would be "delays," and the object would be "progression of ALS."

This process was automated for the entire dataset using a Python script, which iterated through each abstract, calling the CoreNLP server with the appropriate parameters. We added the structured relationships to our DataFrame after receiving the server's output. The intricate interactions between many things, including genes, diseases, therapies, and outcomes, as they were addressed in the literature, were thus able to be briefly captured.

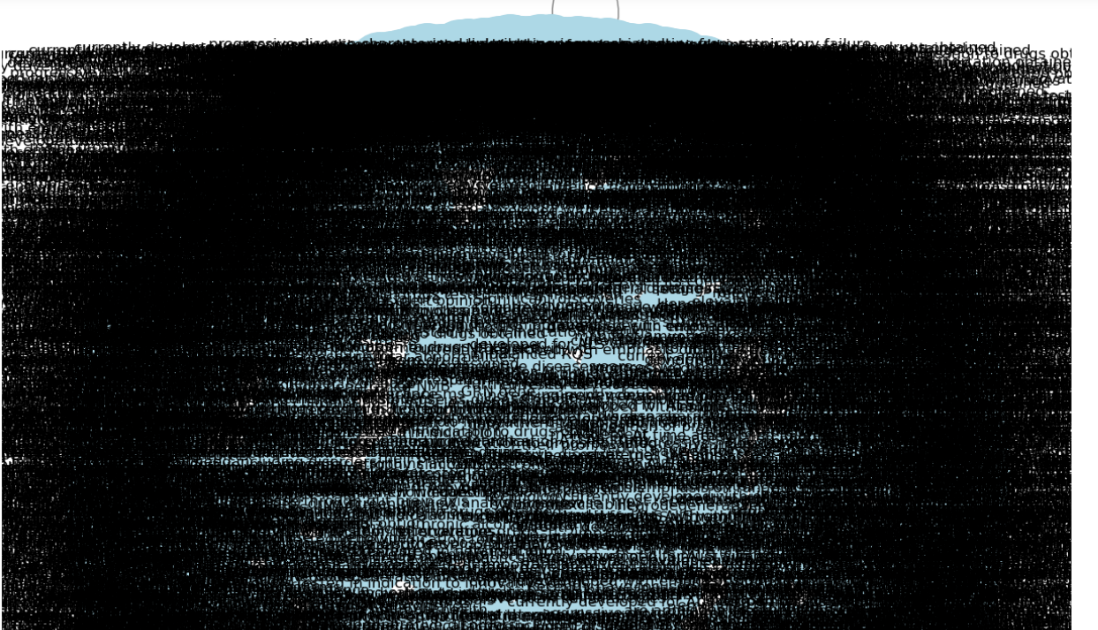
With this detailed of extracted relationships, we are now prepared to advance to the next phase of our project: building the knowledge graph itself. In the next phase, we will combine the data from our relationship extraction to build a network that illustrates the complex relationships found in the field of ALS research. By doing so, we will be able to find previously unknown patterns and insights that may result in advances in our knowledge of and ability to treat this disease.

1. **Constructing Initial Knowledge Graph:**

Using Python code, we produced a graphic representation of the associations we had extracted in the first step of building the knowledge graph. This required creating a graph structure with the NetworkX library, in which each node stands for an entity (such a gene or disease) and each edge for the relationships between the entities.

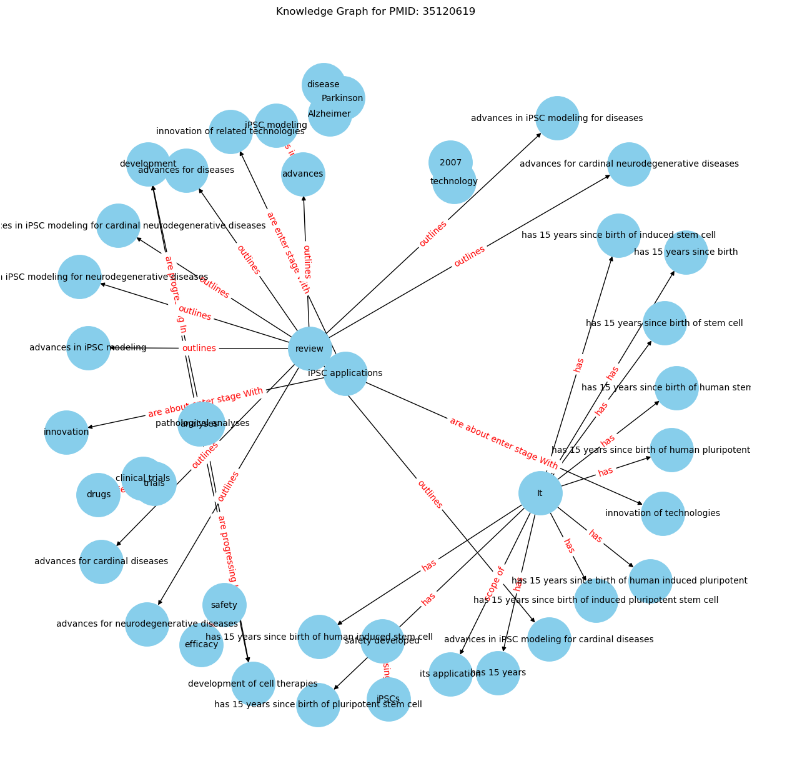
Based on the correlations in our dataset, we added edges to the graph. Each edge had a term, such as "causes" or "treats," indicating the kind of relationship it represented. For clarity, we set the size of the nodes and the colour of the edges while drawing and displaying this graph using Matplotlib. The spring layout algorithm, which places nodes apart for improved visibility, was used to design the graph's layout.

The resulting graph, which was represented as a network of connected nodes, provided a crude visual summary of the intricate interactions in ALS research. This graph acts as a foundation for a more complex knowledge graph, which we want to develop and enhance in later project stages.



**Knowledge Graph for One PMID:**

Following the creation of knowledge graphs for all PMIDs, we decided to focus on just one PMID in order to provide a more understandable analysis because the large amount of data presented an image that was a little hazy. We focused on a single PMID in order to completely understand and present the complex web of relationships in a more readable format. This strategic move was designed to fine-tune our approach, ensuring each connection between entities was mapped out with precision, therefore avoiding the complications that arise from working with larger, more complex datasets.



**Future Work:**

For the Amyotrophic Lateral Sclerosis Knowledge Graph (ALS-KG) to be more accurate and reliable, this project needs to focus on its validation and improvement. In order to better extract and evaluate complicated data from an expanding corpus of biomedical literature, the next steps will require the use of Large Language Models (LLMs). These models are anticipated to provide deeper insights into ALS research, potentially uncovering subtle connections that have not been previously identified.

Future research will also focus on refining our relation extraction and entity merging techniques, which make use of modern algorithms and tools for natural language processing. By doing this, the study hopes to provide a more thorough knowledge of ALS by encompassing a wider range of biological concepts and relationships.

In order to maintain the ALS-KG's lead in neurodegenerative illness research, we intend to incorporate data from several peer-reviewed sources and maybe conduct comparative analyses with other disorders that are similar. It is anticipated that this all-encompassing strategy would strengthen the knowledge graph and make it a vital resource for both scholars and medical professionals.

The goal of the project is to create a dynamic, self-updating ALS-KG that can serve as a trustworthy knowledge base. This database will play a critical role in shaping health policies related to ALS and other neurodegenerative illnesses, as well as in guiding future research and therapeutic trials. This knowledge graph needs to be developed and expanded further in order to help us fight these deadly diseases.

## **Source Code Description**

Our source code is done in a Python Jupyter notebook file with different sections for easy management.

|  |  |  |
| --- | --- | --- |
| Filename | Section/Cell | Description |
| ALS-KG.ipynb | Topic header | Tells what we are doing in the project |
|  | Import | This part/parts in our code are helping us get access to external libraries we need for our project. |
|  | Corpus selection | We used pubmed for corpus selection by doing webscraping using Beautiful soup successfully parsing articles from 2013 to 2023 |
|  | Entity extraction | We use spaCy or other NLP libraries to extract entities |
|  | Relation extraction | We extract relations between entities |
|  | Knowledge Graph construction | We build a knowledge graph using the merged entities and extracted relations |

As our approach summary, we analyzed data, looked at available tools, and made a software solution. This method helped us easily make the ALS Knowledge Graph. We used different resources and technologies to create a useful tool for people studying and treating ALS.

# **Conclusion:**

# Building an Amyotrophic Lateral Sclerosis Knowledge Graph (ALS-KG) is a novel effort that this project offers. Our goal is to create a dynamic and integrated knowledge base by utilizing modern methods for entity and relation extraction. Using methods that don't require much human effort not only makes building the knowledge graph faster but also sets the stage for making it bigger and useful in more ways. Our work is important for reasons other than only ALS; it advances our knowledge and fight against neurodegenerative illnesses in general. Through the fusion of advanced language models, systematic methodologies, and the commitment to unveiling hidden relationships, this project stands as a beacon in biomedical research.

# **References:**

# [1] Bean, D. M., Al-Chalabi, A., Dobson, R. J., & Iacoangeli, A. (2020). A knowledge-based machine learning approach to gene prioritisation in amyotrophic lateral sclerosis. *Genes*, *11*(6), 668.

# [2] Das, T., Kaur, H., Gour, P., Prasad, K., Lynn, A. M., Prakash, A., & Kumar, V. (2022). Intersection of network medicine and machine learning towards investigating the key biomarkers and pathways underlying amyotrophic lateral sclerosis: a systematic review. *Briefings in bioinformatics*, *23*(6), bbac442.

# [3] Hu, J., Lepore, R., Dobson, R. J., Al-Chalabi, A., M. Bean, D., & Iacoangeli, A. (2021). DGLinker: flexible knowledge-graph prediction of disease–gene associations. *Nucleic acids research*, *49*(W1), W153-W161.

# [4] A.-M. Tăuţan, B. Ionescu, and E. Santarnecchi, "Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques," Artificial Intelligence in Medicine, vol. 117, p. 102081, 2021.

# [5] A. Rossanez and J. C. dos Reis, "Generating knowledge graphs from scientific literature of degenerative diseases," in SEPDA@ ISWC, 2019, pp. 12–23.

# [6] A. Holzinger and I. Jurisica, "Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions," in Interactive knowledge discovery and data mining in biomedical informatics, Springer, 2014, pp. 1–18.

# [7] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in Advances in neural information processing systems, 2013.

# [8] B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," arXiv preprint arXiv:1412.6575, 2014.

# [9] C.-H. Wei, H.-Y. Kao, and Z. Lu, "Pubtator: a web-based text mining tool for assisting biocuration," Nucleic acids research, vol. 41, no. W1, pp. W518–W522, 2013.

# [10] J. Collins, J. K. Parker, M. McKean, G. P. Lovell, and L. Hogarth, "Examining the Relationship Between Countermovement and Squat Jump Measures Amongst Elite Conditioning," International Journal of Strength and Development Female Football and Rugby Players, vol. 3, no. 1, 2023.

# [11] G. Angeli, M. J. Johnson Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 344–354.

# [12] B. I. Jansen, "Vernieuwd emissiemodel houtkachels," TNO, 2016.

# [13] J. K. Parker, M. McKean, L. Hogarth, and G. P. Lovell, "Joe Collins1," International Journal of Strength and Conditioning, vol. 3, no. 1, 2021.

# [14] K. Canese and S. Weis, "Pubmed: the bibliographic database," The NCBI handbook, vol. 2, no. 1, 2013.

# [15] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," NPJ digital medicine, vol. 4, no. 1, pp. 1–13, 2021.

# [16] M. Delmas et al., "Building a knowledge graph from public databases and scientific literature to extract associations between chemicals and diseases," Bioinformatics, vol. 37, no. 21, pp. 3896–3904, 2021.

# [17] Q. Wang, S. Dai, B. Xu, Y. Lyu, Y. Zhu, H. Wu, and H. Wang, "Building Chinese biomedical language models via multi-level text discrimination," arXiv preprint arXiv:2110.07244, 2021.

# [18] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 12, pp. 2724–2743, 2017.

# [19] S. K. Mohamed, A. Nounu, and V. Nováček, "Biological applications of knowledge graph embedding models," Briefings in bioinformatics, vol. 22, no. 2, pp. 1679–1693, 2021.

# [20] V. Muddapu, A. Ranjith, S. Chakravarthy, and M. Gromiha, "Neurodegenerative diseases – is metabolic deficiency the root cause?" Frontiers in Neuroscience, vol. 14, 2020.

# [21] Y. Nian, X. Hu, R. Zhang, J. Feng, J. Du, F. Li, Y. Chen, and C. Tao, "Mining on Alzheimer’s diseases related knowledge graph to identity potential ad-related semantic triples for drug repurposing," arXiv preprint arXiv:2202.08712, 2022.

# Y. Shen, Y. Liu, X. Jiao, Y. Cai, X. Xu, H. Yao, and X. Wang, "Knowledge graph: Applications in tracing the source of large-scale outbreak—beijing municipality, china, 2020–2021," China CDC Weekly, vol. 5, no. 4, p. 90, 2023.

**Contributions:**

**Tulasi Makenini:**

I contributed to the detailed study of Amyotrophic Lateral Sclerosis (ALS) by examining the research conducted up to now. This included identifying the major advantages and disadvantages of previous studies. Additionally, I was involved in selecting corpora from PubMed through web scraping using Beautiful Soup, successfully parsing articles from 2013 to 2023. I also assisted other team members with various coding tasks.

**Pravallika Naraharasetty:**

I have contributed to a detailed study of prior research, focusing mainly on the use of machine learning models in Amyotrophic Lateral Sclerosis (ALS) by referencing various research papers. My major contribution involves extracting entities from PubTator. I have successfully extracted six bio-concepts, such as drugs, chemicals, and diseases, from PubTator, and this information has been saved in a text file. Additionally, I have also assisted my teammates with other parts of our project.

**Manasa Ullam:**

I have contributed to a detailed study on the use of large language models in building knowledge graphs. My major contribution involves processing entities of different types extracted from PubTator. I added these entities to a dataframe, replacing empty values with NaN, and then implemented entity merging. This merging process involved a three-step method: normalizing, lemmatizing, and removing special characters. I combined techniques such as Jaccard similarity and Levenshtein distance to detect duplicates. I have successfully implemented these steps and am now focusing on the remaining work in relation extraction, exclusively using Stanford's Open IE for this purpose.