Analyzing impacts of SNPs on proteins related to Type 2 Diabetes Mellitus

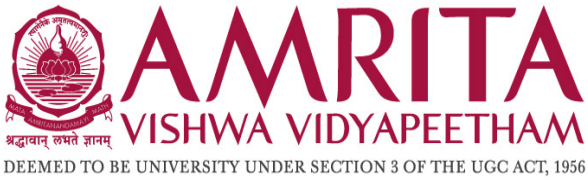
PROJECT REPORT

Submitted by

A P Devanampriya
(CB.SC.U4AIE23001)
K Prerana
(CB.SC.U4AIE23038)
Niharika Sharma
(CB.SC.U4AIE23048)
Patel Srikari Shasi
(CB.SC.U4AIE23053)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE ENGINEERING - ARTIFICIAL INTELLIGENCE



COMPUTER SCIENCE ENGINEERING - ARTIFICIAL INTELLIGENCE

AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE

AMRITA VISHWA VIDYAPEETHAM

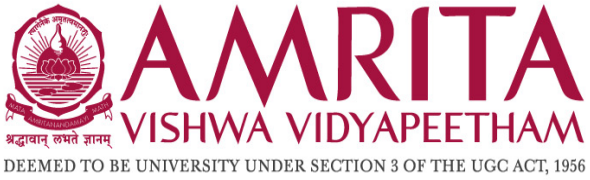
COIMBATORE—641 112 (INDIA)

APRIL - 2025

COMPUTER SCIENCE ENGINEERING - ARTIFICIAL INTELLIGENCE

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112



BONAFIDE CERTIFICATE

This is to certify that the thesis entitled **" Analyzing Impacts of SNPs on Proteins Related to Type 2 Diabetes Mellitus,"** submitted by **A. P. Devanampriya (CB.SC.U4AIE23001), K. Prerana (CB.SC.U4AIE23038), Niharika Sharma (CB.SC.U4AIE23048), and Patel Srikari Shasi (CB.SC.U4AIE23053)** for the award of the **Degree of Bachelor of Technology** in the **"COMPUTER SCIENCE ENGINEERING - ARTIFICIAL INTELLIGENCE"** is a bonafide record of the work carried out by them under our guidance and supervision at Amrita School of Artificial Intelligence, Coimbatore.

Dr. Harishchander A
Assistant Professor (Sr.Gd.),
School of Artificial Intelligence,
Amrita Vishwa Vidyapeetham
Coimbatore

Ms. Rema M
Assistant Professor
School of Artificial Intelligence
Amrita Vishwa Vidyapeetham
Coimbatore

Submitted for the university examination held on 11th April 2025

INTERNAL EXAMINER **EXTERNAL EXAMINER**

AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112

DECLARATION

We, **A. P. Devanampriya (CB.SC.U4AIE23001), K. Prerana (CB.SC.U4AIE23038), Niharika Sharma (CB.SC.U4AIE23048), and Patel Srikari Shasi (CB.SC.U4AIE23053),** hereby declare that this thesis, entitled **"Analyzing Impacts of SNPs on Proteins Related to Type 2 Diabetes Mellitus"**, is the record of the original work done by me under the guidance of **Dr.Harishchander A**, Assistant Professor (Sr. Gd.), and **Ms. Rema M.**, Assistant Professor, Amrita School of Artificial Intelligence, Coimbatore. To the best of my knowledge, this work has not formed the basis for the award of any degree/diploma/associateship/fellowship or a similar award to any candidate in any university.

Place: Ettimadai **Signature of the Students**

Date: 11/04/2025

COUNTERSIGNED

Dr. K.P.Soman
Professor and Dean
Amrita School of Artificial Intelligence
Amrita Vishwa Vidyapeetham

Contents

Acknowledgement

We would like to express our sincere gratitude to all those who have supported and guided us throughout the course of this project. First and foremost, we would like to thank our project supervisors, **Dr. Harishchander A. and Ms. Rema M.**, for their invaluable guidance, constant encouragement, and insightful feedback that helped us stay focused and refine our ideas. Their mentorship was essential in shaping the direction of this work.

We are also grateful to the faculty members and staff of the **Amrita School of Artificial Intelligence**, **Amrita Vishwa Vidyapeetham**, for providing the necessary infrastructure and a research-friendly environment.

We would like to thank our friends and peers for their moral support, collaborative discussions, and continuous motivation during the project development.

This project, titled *"Analyzing impacts of SNPs on proteins related to Type 2 Diabetes Mellitus,"* has been a rewarding experience, and we are thankful to everyone who contributed to its successful completion.

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Full Form
AA	Amino Acid
ADIPOQ	Adiponectin, C1Q and Collagen Domain Containing
BLOSUM	Blocks Substitution Matrix
ClinVar	Clinical Variant Database
dbSNP	Database of Single Nucleotide Polymorphisms
DPP-IV	Dipeptidyl Peptidase IV
GCK	Glucokinase
GWAS	Genome-Wide Association Study
INSR	Insulin Receptor
KCNQ1	Potassium Voltage-Gated Channel Subfamily Q Member 1
LLM	Large Language Model
SLC22A3	Solute Carrier Family 22 Member 3
SLC30A8	Solute Carrier Family 30 Member 8
SMOTE	Synthetic Minority Over-sampling Technique
SNP	Single Nucleotide Polymorphism
T2DM	Type 2 Diabetes Mellitus
TCF7L2	Transcription Factor 7 Like 2
VUS	Variant of Uncertain Significance

Abstract

In recent times, the rapid expansion of genomic data has posed both a challenge and an opportunity in the realms of biomedical research and clinical diagnostics. Determining the specific genetic mutations that lead to diseases is a complicated and time-consuming task when performed manually. In order to tackle this issue, the current project suggests the creation of an AI-powered system capable of accurately categorizing and ranking genetic mutations based on their potential to cause diseases. The suggested system utilizes a mix of data preprocessing methods—including dealing with missing values, converting categorical variables into numerical representations, and employing the synthetic minority oversampling technique (SMOTE)—to generate high-quality input data. The classifier is trained using historical mutation datasets to distinguish between pathogenic and non-pathogenic mutations. The model assigns probability-based critical scores to each mutation, indicating their potential clinical significance. These scores are utilized to evaluate mutations, allowing researchers and healthcare professionals to prioritize the most significant and hazardous variants. The performance of the model is evaluated using commonly used metrics like accuracy, ROC-AUC, and goodness-of-fit, guaranteeing both reliability and robustness. By automating the process of identifying and prioritizing disease-causing mutations, this system seeks to greatly decrease diagnostic delays and enhance the effectiveness of genetic research and precision medicine. This project successfully connects extensive genomic data with practical clinical knowledge, enabling quicker, more precise, and AI-driven genetic diagnostics.

Chapter 1

Introduction

1.1 Literature Survey

Key research examining single nucleotide polymorphisms (SNPs) linked to Type 2 Diabetes Mellitus (T2DM) is compiled in this section. Using a variety of molecular and statistical methods, each study emphasizes how certain gene variants affect susceptibility, and insulin secretion.

1. TCF7L2 Polymorphisms in Asian Indians (2017)

This study explored the association of rs12255372 (G/T) and rs7903146 (C/T) SNPs in the *TCF7L2* gene with T2DM. Genotyping was performed using TaqMan assays on case-control samples. Both variants were linked to impaired insulin secretion and glucose metabolism. These findings suggest a strong genetic predisposition to T2DM in the studied population and support early risk screening. [1]

2. KCNQ1 Variants and Diabetes Risk in China (2021)

Using PCR-RFLP and qPCR, this study analyzed SNPs rs163177, rs163184, rs2237895, and rs2283228 in the *KCNQ1* gene. The first three SNPs were associated with increased

T2DM risk, while rs2283228 had a protective effect. Lower KCNQ1 expression in T2DM patients supports its role in β-cell function and glucose regulation.[2]

3. ADIPOQ Gene and Glycemic Control (2022)

This study focused on rs2241766 in the *ADIPOQ* gene, using PCR-RFLP and ELISA to assess adiponectin levels. The GG genotype was associated with reduced adiponectin and poorer glycemic control. These results suggest a functional role for ADIPOQ variation in insulin sensitivity.[3]

4. DPP-IV Gene and T2DM Susceptibility (2022)

SNPs rs3788979 and rs7608798 in the *DPP-IV* gene were studied using PCR and Sanger sequencing. The G allele of rs3788979 and T allele of rs7608798 were significantly associated with increased T2DM risk, likely through altered incretin degradation affecting insulin regulation.[4]

5. SLC22A3 Polymorphisms in the Chinese Population (2023)

This GWAS-based study used high-throughput genotyping to analyze rs555754, rs3123636, and rs3088442 in *SLC22A3*. While rs555754 and rs3123636 were associated with higher T2DM risk, rs3088442 was not. A haplotype involving rs3088442 and rs3123636, however, showed significant risk association.[5]

1.2 Problem statement

The rapid growth of genetic data has posed a significant challenge in identifying and prioritizing mutations that contribute to complex diseases such as Type 2 Diabetes Mellitus (T2DM). Traditional manual approaches are time-consuming and prone to human er-

ror, making them inadequate for timely diagnosis and effective clinical decision-making. Therefore, there is a pressing need for intelligent systems that can automatically classify mutations as pathogenic or non-pathogenic and prioritize them based on their potential clinical relevance. In this project, we aim to develop a machine learning model that utilizes historical genetic data related to T2DM to predict the pathogenicity of single nucleotide polymorphisms (SNPs). The model incorporates data preprocessing techniques, including handling missing values and encoding categorical variables using class-compensated methods. By ranking SNPs based on their likelihood of contributing to T2DM, this system is designed to support clinical research by enabling more efficient and accurate identification of high-impact variants, thereby aiding in early diagnosis and targeted therapeutic strategies.

1.3 Objectives

The primary objectives of this project are as follows:

1. **Analyze** the impact of single nucleotide polymorphisms (SNPs) on protein structures and interactions, particularly in relation to insulin regulation and glucose metabolism in Type 2 Diabetes Mellitus (T2DM).

2. **Investigate** genetic variations associated with T2DM to identify key SNPs contributing to disease susceptibility and progression.

3. **Model** gene–protein interaction networks using graph algorithms to uncover complex relationships and patterns between SNPs and their functional contexts.

4. **Develop** a predictive framework by integrating machine learning techniques with graph-based models to identify significant co-occurrence patterns and predict the functional impact of SNPs.

5. **Discover** potential biomarkers through graph-based analysis, with the aim of enabling early diagnosis and supporting the development of personalized treatment strategies for T2DM.

Chapter 2

Background

Type 2 Diabetes Mellitus (T2DM) is a complex, multifactorial metabolic disorder characterized by insulin resistance and impaired insulin secretion. It accounts for the majority of diabetes cases globally and is influenced by both environmental and genetic factors. In recent years, advances in genomic technologies have revealed that single nucleotide polymorphisms (SNPs) play a crucial role in the genetic predisposition to T2DM. These SNPs can influence gene expression, protein structure, and protein–protein interactions, ultimately affecting pathways involved in glucose metabolism, insulin signaling, and beta-cell function.

Several genome-wide association studies (GWAS) and clinical databases such as ClinVar have identified numerous SNPs associated with increased or decreased susceptibility to T2DM. These variants are often located in genes that encode key regulatory proteins such as transcription factors, ion channels, hormone receptors, and transporters. Understanding how these genetic variations affect protein structure and function is essential for uncovering the molecular mechanisms underlying the disease.

This project aims to integrate SNP data with protein-level insights to investigate

how genetic mutations contribute to T2DM progression. By combining evolutionary conservation analysis, functional domain mapping, and graph-based modeling of SNP–protein interactions, the study seeks to prioritize pathogenic variants and identify potential biomarkers. Such integrative approaches can enhance our understanding of T2DM pathogenesis and support the development of precision medicine strategies.

Chapter 3

Proposed Work

The primary objective of this work is to analyze the functional and biochemical impacts of Single Nucleotide Polymorphisms (SNPs) on proteins implicated in Type 2 Diabetes Mellitus (T2DM), utilizing a feature-rich, graph-based clustering framework. The approach emphasizes the integration of domain knowledge with computational modeling to uncover complex interactions that influence disease phenotypes.

3.1 Feature Engineering and Normalization

Key biological and biochemical attributes are extracted for each SNP to quantify their potential impact on protein function:

- **Biochemical Impact:** Missense mutations are analyzed for their influence on amino acid properties such as hydrophobicity, charge, and molecular size. For instance, a charge-altering mutation in the *INSR* gene may disrupt insulin binding, compromising receptor functionality.

- **Functional Domain Mapping:** SNPs are mapped to protein domains such as

catalytic sites or ligand-binding regions. A mutation within the active site of *Glucokinase (GCK)*, for example, can significantly impair glucose metabolism, a critical process in T2DM regulation.

To ensure fair contribution of each feature, **Min-Max Normalization** is applied to scale values to a uniform range of 0–1. The following normalized features are included:

- `norm_pvalue`: Reflects the statistical significance of SNP-disease associations.
- `norm_effect`: Captures the strength of association using odds ratios or beta coefficients.
- `norm_blosum`: Assesses evolutionary conservation of amino acid substitutions.
- `norm_hydro`: Quantifies disruption due to hydropathy changes in protein structure.

3.2 Graph-Based Modeling of SNP-Protein Interactions

Traditional SNP analyses often treat mutations as independent events. In contrast, the proposed method constructs a **graph representation** to model interdependencies and co-functional relationships:

- **Nodes:** Represent both SNPs and associated proteins.
- **Edges:** Indicate functional interactions, weighted by a composite similarity score derived from the normalized features.

This structure facilitates:

- **Pattern Recognition:** Clustering of pathogenic variants based on shared features.
- **Interaction Mapping:** Revealing networks of SNPs converging on the same biological pathways.
- **Prioritization:** Highlighting high-impact SNPs central to T2DM mechanisms.

3.3 Clustering via the Leiden Algorithm

To uncover meaningful biological communities within the SNP-protein graph, the **Leiden algorithm** is employed, offering superior modularity and connectivity compared to conventional methods:

1. **Initialization:** Each node begins as its own cluster.
2. **Local Moving:** Nodes are reassigned to communities that maximize modularity.
3. **Refinement:** Poorly connected communities are split to enhance internal cohesion.
4. **Aggregation:** Communities are merged into supernodes, forming a reduced graph.
5. **Iteration:** Steps are repeated until no further improvements in modularity are observed.

3.4 Centrality-Based Hotspot Detection

To prioritize high-impact nodes in the SNP-protein interaction network, multiple centrality measures were computed, reflecting the importance of each node within the graph topology. These metrics help identify key mutation hotspots that may play crucial roles in T2DM pathology.

Centrality Measures Computed

- **Degree Centrality:** Represents the number of direct connections a node has, normalized by total nodes.
- **Strength (Weighted Degree):** Sum of edge weights connected to the node, reflecting cumulative interaction strength.
- **Closeness Centrality (Weighted):** Inversely proportional to the total shortest path distance from a node to all others. A manually implemented Dijkstra-based algorithm was used for accurate weighted closeness computation.
- **Betweenness Centrality:** Measures how often a node lies on shortest paths between other nodes, computed using NetworkX's built-in algorithm with weight consideration.
- **Eigenvector Centrality:** Captures influence of a node by considering both its direct and indirect neighbors. Nodes connected to other influential nodes score higher.

Each centrality score was **Min-Max normalized** to bring values into the 0–1 range:

$$\text{Normalized Value} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Hotspot Score Computation

To rank nodes by potential functional importance, a composite **Hotspot Score** was derived:

$$\text{Hotspot Score} = \text{Norm\_Degree} + \text{Norm\_Strength} + \text{Norm\_Betweenness} + \text{Norm\_Eigenvector}$$

Nodes with the highest Hotspot Score are prioritized for downstream biological analysis. This integrated approach captures local connectivity, global importance, and evolutionary relevance.

3.5 Result Interpretation and Visualization

The resulting centrality matrix was saved as `mutation_hotspots.csv`, with the top-ranked nodes further annotated:

- **dbSNP ID Matching:** Nodes starting with "rs" are flagged as SNPs using a prefix match rule.
- **Cluster Assignment:** Each node is tagged with its Leiden cluster ID for community-level tracking.
- **Top Proteins:** The top 10 non-SNP nodes (proteins) with the highest Hotspot Scores are listed along with their top three SNP neighbors.

This prioritization scheme enables high-confidence detection of mutational hotspots, which may represent critical nodes in the pathogenesis of Type 2 Diabetes Mellitus. Clustering and centrality complement each other to improve interpretability and guide validation efforts.

The outcome is a set of well-connected SNP-protein communities with shared functional roles, suitable for downstream biological interpretation.

3.6 Results and Discussion

3.6.1 Graph Plotting

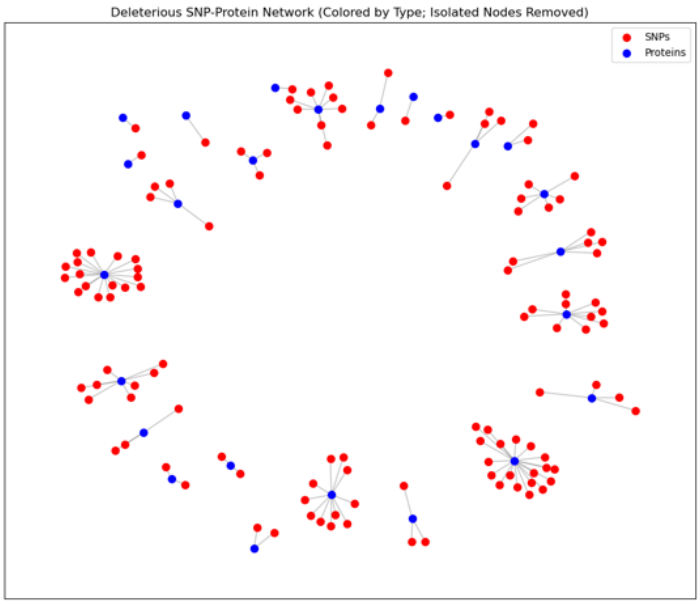


Figure 3.1: Plot of SNPs and Proteins

Figure 3.1 presents the result of the graph obtained after plotting unique SNPs and proteins and removing isolated nodes and assigning respective colors for SNPs and proteins.

3.6.2 Leiden Clustering of SNP-Protein Interaction Network

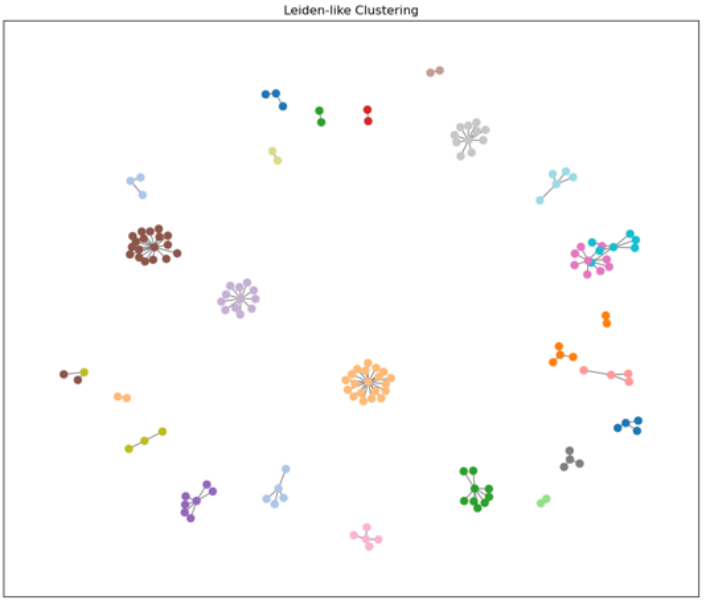


Figure 3.2: Leiden Clustering

Figure 3.2 presents the result of applying the Leiden community detection algorithm to the SNP-protein interaction network, revealing a total of 25 distinct clusters. Each node in the graph represents a single nucleotide polymorphism (SNP) or a protein, while the edges reflect meaningful biological associations, such as mapped pathogenic effects, co-occurrence, or gene-protein links. The Leiden algorithm, which optimizes modularity in a multilevel fashion, effectively partitions the graph into densely connected subgroups or communities, each visualized using a different color.

3.6.3 Pathogenicity Distribution Across Clusters

Figure 3.3 depicts a heatmap showing the distribution of pathogenicity classes—derived from ClinVar annotations—across the 25 Leiden clusters. The y-axis enumerates the

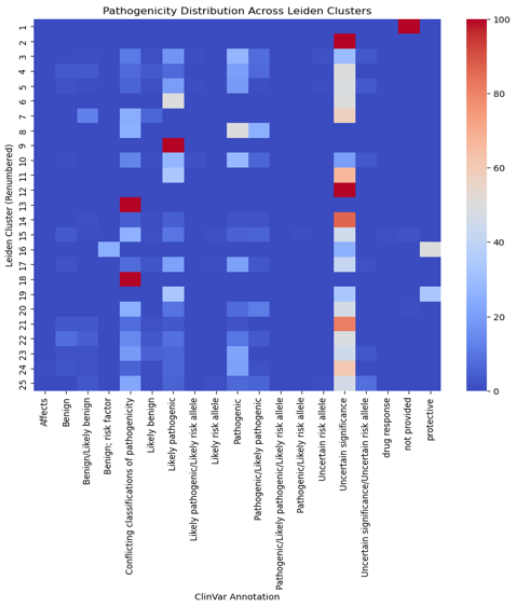


Figure 3.3: Pathogenicity of SNPs

cluster identifiers, while the x-axis represents various ClinVar clinical significance categories, including "Benign," "Likely benign," "Uncertain significance," "Likely pathogenic," and "Pathogenic," among others.

The color intensity denotes the number of SNPs in each category per cluster. This visualization offers two key insights: (1) some clusters exhibit high concentrations of pathogenic or likely pathogenic SNPs, indicating strong associations with disease phenotypes, while (2) other clusters are dominated by benign or uncertain variants, suggesting a more neutral or ambiguous role in disease etiology.

3.6.4 Scalability and Performance of the Leiden Algorithm

Table 3.1: Leiden Algorithm Performance Across Graph Sizes

Nodes	Edges	Initial Modularity	Final Modularity	Time (s)
50	100	-0.1050	0.4964	0.1738
100	200	-0.0473	0.5335	1.4313
200	400	-0.0249	0.5645	5.1813
400	800	-0.0124	0.5732	24.4045
800	1600	-0.0061	0.5745	44.6931

Table 3.1 presents the modularity optimization performance of the Leiden community detection algorithm applied to SNP–protein interaction graphs of increasing size. For each graph configuration, the algorithm begins with a low or negative initial modularity and steadily improves it through iterative refinement. As the number of nodes and edges increases, the final modularity shows consistent improvement, indicating well-formed community structures. The runtime also scales with graph size, from under a second for 50 nodes to over 44 seconds for 800 nodes. These results highlight the Leiden algorithm's robustness and scalability in detecting biologically meaningful clusters in large, complex genomic networks.

Figure 3.4 provides a performance evaluation of the Leiden algorithm by examining how both execution time vary with the size of the input graph.

The plot illustrates a clear, nonlinear increase in execution time as the number of nodes increases. This is expected, given that larger graphs involve more iterations and deeper hierarchical refinement, especially when optimizing modularity across multiple resolution levels.

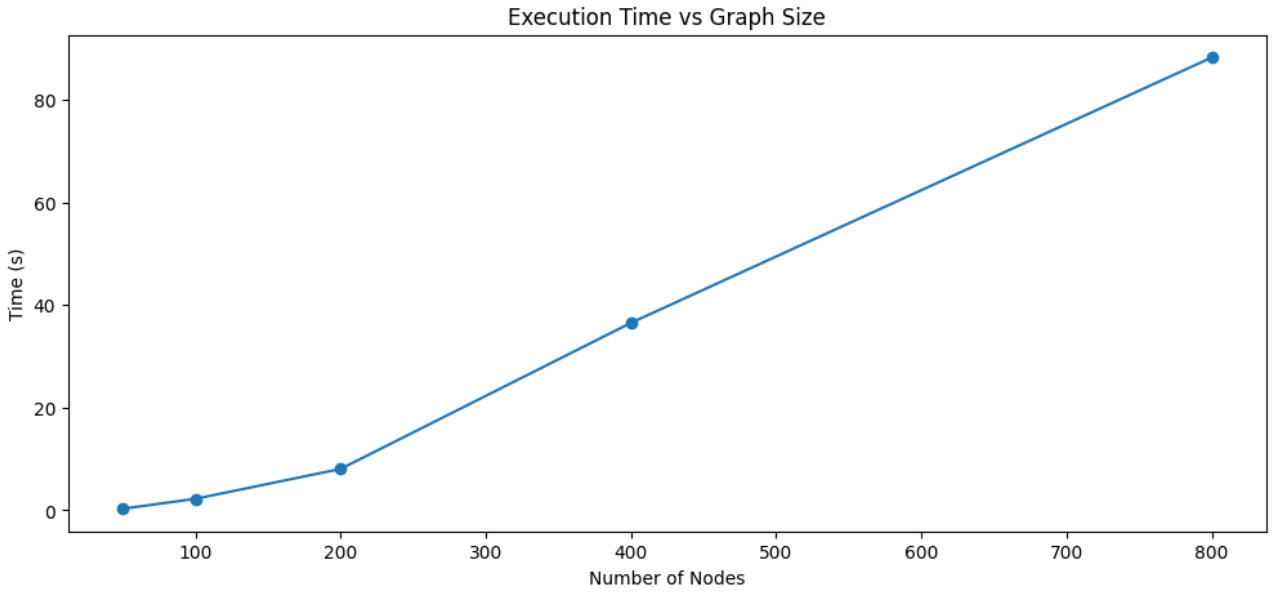


Figure 3.4: Leiden Comparison

Table 3.2: Top 10 Mutation Hotspots and Associated SNPs

Gene (Node)	**Hotspot Score**	**Top Connected SNPs (Edge Weights)**
KCNQ1	4.000	rs2237896 (1.043), rs2237897 (1.003), rs163182 (0.996)
CDKAL1	2.594	rs10440833 (0.999), rs35612982 (0.986), rs7766070 (0.969)
HNF1B	1.379	rs11651052 (0.925), rs10908278 (0.914), rs4430796 (0.902)
GLIS3	1.335	rs7041847 (1.054), rs79103584 (1.054), rs10814916 (1.054)
PPARG	0.934	rs11709077 (0.954), rs71304101 (0.951), rs3963364 (0.936)
WFS1	0.909	rs9998835 (0.924), rs4996963 (0.917), rs4458523 (0.908)
SLC30A8	0.651	rs531347476 (1.089), rs6997279 (0.984), rs13266634 (0.949)
GCK	0.604	rs2971669 (0.790), rs2908286 (0.784), rs730497 (0.781)
INSR	0.384	rs8101064 (1.045), rs17175860 (0.892), rs75253922 (0.891)
ABCC8	0.370	rs67254669 (1.030), rs4148646 (0.919), rs757110 (0.782)

Table 3.2 lists the top 10 mutation hotspot proteins identified from the SNP–protein interaction network using centrality analysis. Each entry includes the gene node, computed Hotspot Score, and the top three connected SNPs ranked by interaction strength (edge weights). Higher Hotspot Scores indicate central nodes that may play key roles

in the pathogenesis of Type 2 Diabetes Mellitus (T2DM). Notably, genes like KCNQ1, CDKAL1, and HNF1B exhibit strong interactions with high-weight SNPs, highlighting them as potential targets for deeper functional and clinical investigation. These results have been used to train the LLM.

3.6.5 LLM Integration Results

Figures 3.5 and 3.6 showcase the integration of an LLM-powered explanation system with an SNP lookup interface. This tool allows users to enter a dbSNP ID and receive both structured genomic information and an AI-generated interpretation of the variant's biomedical significance.

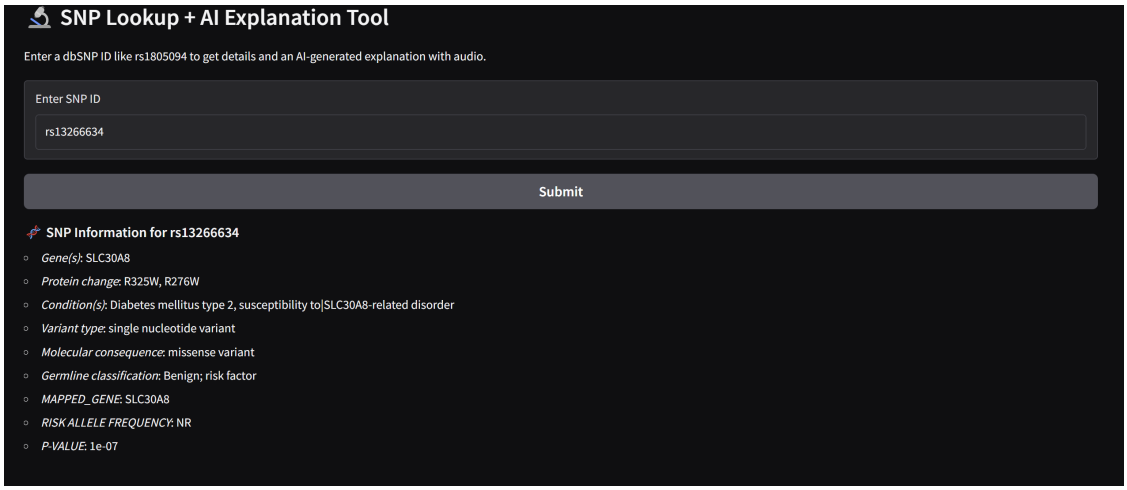


Figure 3.5: Data Collection: The user interface displays structured SNP information for the input rsID (e.g., `rs13266634`). Information retrieved includes gene name, protein changes, variant type, molecular consequence, germline classification, mapped gene, risk allele frequency, and p-value.

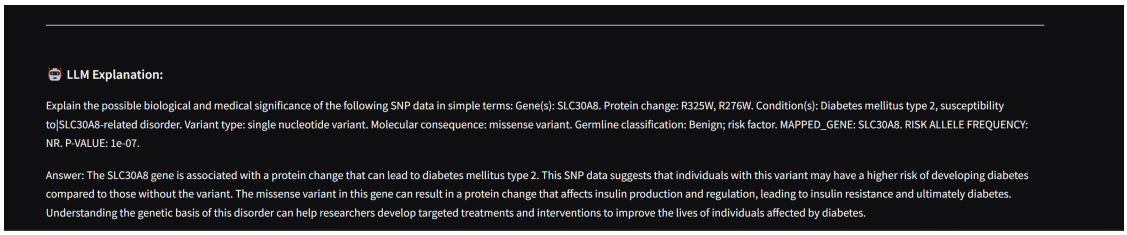


Figure 3.6: LLM Explanation: An AI-generated textual explanation summarizes the clinical and biological significance of the SNP. It interprets the missense variant in the SLC30A8 gene and its association with increased risk for type 2 diabetes.

The LLM provides a simplified narrative, explaining that the SNP `rs13266634` in the *SLC30A8* gene results in a missense mutation (R325W, R276W) that may affect insulin production. This genetic variation has been linked to increased susceptibility to type 2 diabetes. The explanation aids in understanding the potential disease mechanism, which could support future therapeutic development or risk stratification.

The inclusion of large language models in this workflow enhances interpretability, offering domain-aware summaries that bridge the gap between raw genomic data and actionable insights for researchers and clinicians.

Chapter 4

Conclusion

This study presents a comprehensive framework for analyzing the impact of Single Nucleotide Polymorphisms (SNPs) on proteins involved in Type 2 Diabetes Mellitus (T2DM) using biochemical feature extraction, graph-based modeling, and advanced clustering via the Leiden algorithm. By incorporating normalized biochemical and statistical features—such as hydropathy changes, evolutionary conservation (BLOSUM62), and effect size—each SNP is contextualized in terms of its functional significance and disease relevance. The graph-based representation, with SNPs and proteins as nodes and weighted functional interactions as edges, enables the detection of biologically meaningful communities. The Leiden algorithm effectively clusters these nodes, uncovering mutation patterns and interaction networks that are otherwise overlooked in traditional single-point SNP analysis. Comparative analysis with ClinVar annotations further enhances the interpretability of the clustering results by highlighting pathogenicity trends across identified communities. Clusters enriched in pathogenic or likely pathogenic variants provide a powerful tool for prioritizing high-risk mutations for further investigation. Moreover, the identification of mutation hotspots within con-

served and functionally critical protein regions reinforces their potential biological and clinical relevance.

To support this interpretation, a heatmap matrix is constructed with the following structure:

- **Rows:** Renumbered Leiden clusters (sequentially from 1 to N).
- **Columns:** ClinVar classification categories – *Pathogenic*, *Likely Pathogenic*, *Variants of Uncertain Significance (VUS)*, and *Benign*.
- **Cells:** Represent the percentage of SNPs in each cluster that fall into the corresponding ClinVar category.

The color intensity in the heatmap provides quick visual cues:

- **Dark red cells** indicate clusters heavily enriched in pathogenic SNPs, which are potential high-risk mutation hubs.
- **Mixed color cells** suggest heterogeneity in the cluster, potentially including benign or uncertain variants.

This visualization aids in identifying mutation hotspots and selecting clusters that warrant further biological investigation and validation.

Overall, this integrative approach not only improves the biological interpretation of SNP datasets but also offers a scalable strategy for identifying candidate variants with probable involvement in T2DM pathogenesis.

References

1. Dhanasekaran Bodhini, Venkatesan Radha, Monalisa Dhar, Nagarajan Narayani, Viswanathan Mohan, *The rs12255372(G/T) and rs7903146(C/T) polymorphisms of the TCF7L2 gene are associated with type 2 diabetes mellitus in Asian Indians*, 2017.

2. Xu J, Zhang W, Song W, Cui J, Tian Y, Chen H, Huang P, Yang S, Wang L, He X, Wang L, Shi B, Cui W, *Relationship Between KCNQ1 Polymorphism and Type 2 Diabetes Risk in Northwestern China*, 2021.

3. Alfaqih MA, Al-Hawamdeh A, Amarin ZO, Khader YS, Mhedat K, Allouh MZ, *Single Nucleotide Polymorphism in the ADIPOQ Gene Modifies Adiponectin Levels and Glycemic Control in Type Two Diabetes Mellitus Patients*, 2022.

4. Archna Bhargave, Kiran Devi, Imteyaz Ahmad, Anita Yadav, Ranjan Gupta, *Genetic Variation in DPP-IV Gene Linked to Predisposition of T2DM: A Case Control Study*, 2022.

5. Zhongyu Li, Xiangmin Yuan, Xin Liu, Yuping Yang, Li Huang, Qiuhong Tan, Cuilin Li, *The Influence of SLC22A3 Genetic Polymorphisms on Susceptibility to Type 2 Diabetes Mellitus in Chinese Population*,2023.

List of Publications based on this Research Work

1. M. V. Prashanth, N. T. Jose, K. H. Verma, B. Sreevidya, M. Rajesh, *Prognostic Modeling of Diabetes Mellitus in Women: A Comparative Analysis of Machine Learning Algorithms Integrating Clinical and Genetic Data*, 2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS), Bengaluru, India, pp. 1–6, IEEE. (Published)
2. P. Dharmarajan, B. Rajathilagam, *Performance Analysis of Various Membership Functions using ANFIS for the Detection of Type-II Diabetes*, 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), Kuala Lumpur, Malaysia, pp. 1–5, IEEE. (Published)
3. C. P. Prathibhamol, J. Chandrakiran, S. Santhosh, M. Meenakshi, S. A. Menon, M. Nair, *A Pangenome Graph-Based Approach for Predicting Alzheimer's Disease*, Chapter in Book: Data Science & Exploration in Artificial Intelligence*, 1st Edition, CRC Press, 2025, Pages: 6.
4. K. Sreekar, V. Aswin, V. Narayanan, S. K. Thangavel, K. Somasundaram, S. K. Shanmugam, *MedBot – An NLP based ChatBot for Diabetes Prediction*, Proceedings of the 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-11.