Improving Diabetes Prediction Using a Hybrid Approach Based on Community Detection and Ontological Reasoning

Pham Thi Thu Thuy*, Kim Hwa Soo**

** Nha Trang University, o2 Nguyen Dinh Chieu, Bac Nha Trang Ward, Khanh Hoa province, Vietnam (thuthuy@ntu.edu.vn)*

***Ajou University, 206 World cup-ro, Yeongtong-gu, Suwon, Gyeonggi-do, South Korea (ajhskim@ajou.ac.kr)*

Abstract: Accurate prediction of Type 2 Diabetes is essential for effective prevention and intervention strategies. This paper proposes a hybrid model that integrates an improved Louvain community detection algorithm with a domain-specific diabetes ontology to enhance prediction performance and interpretability. Three real-world datasets (Pima Indians, Diabetes 130-US Hospitals, and NHANES) were used for evaluation. Patients were grouped into clinically meaningful clusters using the enhanced Louvain method, followed by semantic reasoning over an ontology constructed from expert-defined medical rules and clinical features. Experimental results show that the proposed model outperforms conventional clustering and classification techniques across multiple metrics, achieving up to 97.56% accuracy in cross-validation. The use of ontological knowledge not only increases transparency in prediction outcomes but also supports semantic querying and dynamic model adaptation. This method demonstrates significant potential for real-world deployment in clinical decision support systems.

Keywords: Diabetes prediction, community detection, ontology, semantic reasoning, Louvain clustering, machine learning.

1. INTRODUCTION

Diabetes mellitus is a major global health concern, affecting millions of people worldwide and placing a significant burden on healthcare systems (World Health Organization, 2016). Detecting diabetes early and accurately is essential for managing the disease effectively and preventing serious complications (Ali et al., 2013). Recent advances in machine learning and data analysis have led to the creation of prediction models (Kavakiotis et al., 2017) that can identify individuals at risk of diabetes using clinical measurements, demographic information, and lifestyle data. However, many of these models still face important challenges, such as being difficult to interpret, not working well for all patient groups, or not adapting easily to new medical knowledge (Ribeiro et al., 2016).

In the realm of diabetes prediction, various methodologies have been explored, including clustering techniques, ontology-based methods, and hybrid models that integrate multiple approaches. Clustering techniques have been employed to identify hidden patterns within diabetic datasets. (Krishna et al., 2023; Kladov et al., 2024a) applied K-means clustering in conjunction with decision trees to predict diabetes, achieving notable accuracy. However, these methods were constrained by the necessity to predefine the number of clusters and exhibited sensitivity to initial centroid placement, potentially leading to inconsistent outcomes across diverse datasets. Similarly, Arora et al (Arora et al., 2022a) employed K-means clustering alongside Support Vector Machines (SVM) for diabetes prediction, resulting in improved accuracy. Despite this enhancement, their model faced challenges related to the selection of optimal cluster numbers and managing overlapping clusters, which could impact prediction reliability. Another study by (Alnowaiser, 2024) proposed an automated method for diabetes prediction using a KNN imputer and a tri-ensemble voting classifier, achieving an accuracy of 97.49%. While impressive, this study did not incorporate ontological knowledge, which could further enhance prediction performance.

Ontology-based frameworks have also been utilized in the healthcare domain (Bodenreider, 2008), particularly for structuring and interpreting clinical data (Chatterjee et al., 2021; Kim et al., 2019). (Massari et al., 2022) developed an ontology-based machine learning model for diabetes prediction, improving model interpretability and accuracy. However, the reliance on static ontologies limited the model's ability to adapt to new or evolving medical knowledge. Other approaches by (Sousa and Paulheim, 2024; Folasade et al., 2023) demonstrated the integration of knowledge graphs with machine learning models to enhance diabetes prediction. Although these methods showed promising results, the complexity of constructing comprehensive knowledge graphs and integrating them into predictive models presented significant challenges.

Hybrid models combining clustering and classification algorithms have shown considerable promise in diabetes prediction. (Sen Kaya et al., 2023) utilized decision tree-based ensemble methods for early-stage diabetes prediction, achieving high accuracy but encountering challenges related to computational demands and overfitting, particularly with high-dimensional data. (Dwived et al., 2019) employed decision trees, emphasizing the model's interpretability and efficiency. Despite these advantages, decision trees alone often lack the complexity needed to capture intricate patterns in clinical data, which can limit predictive performance.

Recent advancements in machine learning (ML) have demonstrated significant promise in healthcare, particularly in analyzing large-scale datasets to predict chronic conditions such as diabetes and identify individuals at risk. These techniques can uncover complex patterns in clinical, demographic, and lifestyle data that traditional statistical methods may overlook. For instance, (Sampa et al., 2023) developed a web-based application utilizing boosted decision tree regression to predict blood glucose levels based on noninvasive health checkups, sociodemographic characteristics, and dietary information . Similarly, (Liu et al., 2023) conducted a systematic review and network meta-analysis comparing various ML models for blood glucose prediction, highlighting the potential of these techniques in clinical settings .

In addition to ML approaches, wearable technologies have been explored for noninvasive glucose monitoring. A study by (Lim et al., 2023) demonstrated the feasibility of using photoplethysmography (PPG) sensors combined with ML algorithms to assess elevated blood glucose levels, achieving an accuracy of 84.7% .

However, most existing studies using community detection for health data analysis focus on graph structure alone and neglect semantic-level integration, which limits the interpretability and clinical relevance of the findings. Furthermore, although ontology-based approaches have shown promise in biomedical knowledge representation (Bodenreider, 2008; Chatterjee et al., 2021), their use in guiding or enhancing clustering models – especially for diabetes prediction – remains unexplored, while the majority of studies focus on clustering based on raw data, not on ontology (Kladov et al., 2024a; Arora et al., 2022a). This gap highlights the need for an integrated approach that leverages both the structural insights from community detection and the semantic richness of ontologies to improve prediction accuracy and interpretability in diabetes research.

Therefore, this study proposes a novel hybrid prediction model that integrates an improved Louvain Community Detection algorithm with dynamic ontology-based refinement. The improved Louvain algorithm automatically determines the optimal number of patient clusters based on data modularity, reducing sensitivity to initial parameters and improving adaptability across diverse datasets. Simultaneously, dynamically generated ontological relationships allow the model to evolve in response to new medical knowledge, thereby enhancing prediction accuracy, interpretability, and clinical relevance. Furthermore, the proposed method is computationally efficient, scalable, and suitable for application in large-scale, real-world clinical settings.

The contributions of this paper are fourfold: (1) Proposing a comprehensive prediction model that integrates enhanced clustering techniques and ontology-based knowledge representation for diabetes prediction; (2) Improving the Louvain Community Detection algorithm to construct a robust and efficient patient clustering model; (3) Developing a diabetes domain ontology to facilitate semantic reasoning and knowledge enrichment; and (4) Designing a semantic diabetes prediction system that leverages both patient health parameters and ontology-driven insights to deliver accurate and interpretable predictions.

2. DATA PROCESSING

2.1 Data sources

To evaluate the proposed model's performance, three publicly available and well-established datasets were employed: the Pima Indians Diabetes dataset (Kladov et al., 2024b), the Diabetes 130-US hospitals dataset (Strack et al., 2014), and the NHANES dataset (Centers for Disease Control and Prevention, 2020).

- The Pima Indians Diabetes dataset includes diagnostic measurements for 768 female patients of at least 21 years of age, all of whom are of Pima Indian heritage. The dataset contains eight numeric features such as glucose level, insulin level, BMI, and age, with binary labels indicating the presence or absence of diabetes.
- The Diabetes 130-US hospitals dataset consists of over 100,000 medical records of diabetic patients from 130 US hospitals collected between 1999 and 2008. It includes more than 50 attributes related to patient demographics, medical diagnoses, laboratory tests, medications, and hospital outcomes. For this study, a curated subset was used to focus on relevant features and reduce noise.
- The NHANES dataset is a comprehensive, population-based survey dataset that combines interviews and physical examinations of thousands of individuals across the United States. It includes data on demographics, dietary intake, laboratory tests, medical conditions, and physical activity. For this study, a merged version of NHANES spanning multiple years was utilized to increase the sample size and enrich the feature set. The dataset was preprocessed to select attributes relevant to diabetes prediction, such as age, BMI, fasting glucose, HbA1c levels, and physical activity status.

These datasets were selected to ensure diversity in population demographics, data sources, and clinical variables—allowing for robust internal evaluation and external validation. Specifically, the Pima dataset provides a compact clinical dataset commonly used in diabetes research, the Diabetes 130-US Hospitals dataset offers longitudinal hospital encounter data, and NHANES brings nationally representative health and nutrition data from the general U.S. population.

After preparing the data, patients were grouped into meaningful subpopulations using an improved Louvain Community Detection method, which finds natural groupings based on how closely related patients are within the dataset. A structured diabetes knowledge model (called an ontology) was created and used to add medical meaning to each patient group, based on expert-defined relationships. The logical rules based on expert-defined medical knowledge for diabetes prediction were generated using SPARQL queries and grounded in existing medical literature and clinical guidelines. Rule development was further guided by domain knowledge from published studies and structured terminologies relevant to diabetes care.

We tested our model using two evaluation strategies: dividing the data into 80% for training and 20% for testing, and a more rigorous method called 10-fold cross-validation that repeats training/testing 10 times for reliability. Comparative analysis was performed against existing clustering and classification methods, including K-Means, DBSCAN, and several published studies.

To avoid information leakage across training and testing, dataset splits were carefully designed. For the Pima dataset, patient-level independence was ensured since each record corresponds to a unique individual. For the Diabetes 130-US Hospitals dataset, we applied a group-aware strategy: all encounters for a given patient were assigned exclusively to either training or testing sets. Furthermore, we considered time-aware splits by holding out the last year of encounters as an additional robustness check. For NHANES, which is a nationally representative survey, sampling weights were preserved during stratification. Labels were harmonized across datasets: in Pima, diabetes was defined as the binary outcome variable "Outcome"; in 130-US, patients with ICD-9/10 diagnostic codes 250.xx or HbA1c ≥ 6.5% were labeled diabetic; in NHANES, diabetes status was derived from self-report plus laboratory thresholds (FPG ≥ 126 mg/dL or HbA1c ≥ 6.5%). This harmonization ensured comparability of predictive outcomes across the three sources.

2.2 Improved Louvain community detection

To improve the identification of patient subgroups, we enhanced the standard Louvain community detection algorithm (Arora et al., 2022b). The original Louvain method works by grouping individuals (represented as nodes) into communities to maximize a measure called modularity, which evaluates how well the network is divided into distinct clusters. A higher modularity value means that patients within a group are more similar to each other than to those in other groups. However, the traditional approach does not fully consider how similar patients are in terms of their medical characteristics. Our improved version incorporates patient-level features to make subgroup identification more clinically meaningful.

The process began with data preprocessing to ensure quality and consistency. Missing values were primarily handled by mean imputation for initial experiments; however, we further evaluated alternative imputers to test robustness. Specifically, we compared mean imputation, k-nearest neighbor (kNN) imputation (k=5), and iterative model-based imputation (MICE). Across datasets, predictive performance metrics changed by less than 1.2%, demonstrating that the proposed method is not sensitive to the choice of imputation strategy. This sensitivity analysis strengthens confidence that reported results are not artifacts of preprocessing.

Next, we constructed a similarity graph using the K-Nearest Neighbors (KNN) algorithm (Cover and Hart, 1967). In this graph, each patient is treated as a node, and connections (edges) are drawn to the most similar patients based on their clinical profiles. This graph structure serves as the foundation for applying the improved Louvain algorithm, which then detects patient subgroups based on both their feature similarities and network structure.

In our enhancement, patient similarities were quantified using an inverse Euclidean distance weighting scheme, wherein closer patients in feature space were connected with higher-weighted edges. Specifically, the weight w_{ij} between two patients i and j was calculated as:

$$w_{ij} = \frac{1}{1+\|x_i - x_j\|} \quad (1)$$

where x_i and x_j represent the normalized feature vectors of patients i and j, respectively. This approach emphasizes the contribution of highly similar patient pairs in the community detection process.

Additionally, we introduced fine-tuning of the resolution parameter in the modularity function. The resolution parameter controls the granularity of detected communities: lower values favor larger communities, while higher values lead to smaller, more homogeneous clusters. Empirical tuning of this parameter allowed us to identify clinically meaningful patient subgroups with balanced cluster sizes and improved internal consistency.

Together, these modifications enabled the improved Louvain algorithm to more effectively uncover hidden structures in the patient datasets, facilitating downstream ontology construction and semantic reasoning for enhanced diabetes prediction.

For visualization, dimensionality reduction was performed using Principal Component Analysis (PCA) (Jolliffe, 2002) to project the high-dimensional dataset into a two-dimensional space. The resulting scatter plot depicted clusters with distinct color-coding, providing an intuitive representation of the clustering results. Each cluster corresponded to a subset of patients with similar attributes, offering insights into patterns such as high-risk groups for diabetes. The improved algorithm is presented as following.

The improved Louvain clustering algorithm

```
1.Begin      Improved_Louvain_Clustering(Dataset, k_neighbors, resolution_value)
2. // Step 1: Data Preprocessing
3.  Load Dataset into DataFrame
4.  Fill missing values with column means
5.  Separate features X and target y
6.  Standardize X using Z-score normalization (StandardScaler)
7. // Step 2: Graph Construction
8. Construct a K-Nearest Neighbors (KNN) graph G
9.  For each pair of nodes (i, j) in KNN graph
10. Compute Euclidean Distance: Distance_ij ← ||X_i - X_j||
11. Assign Edge Weight: Weight_ij ← 1 / (1 + Distance_ij)
12. Add edge (i, j) with Weight_ij to Graph G
13.  End For
14. // Step 3: Louvain Community Detection
15. Partition ← Louvain_Algorithm(G, weight = 'Weight_ij', resolution = resolution_value)
16.  Modularity_Score ← Calculate_Modularity (Partition, G, weight = 'Weight_ij')
17. // Step 4: Dimensionality Reduction for Visualization
18. Apply PCA on X to reduce dimensions to (PC1, PC2)
```

```
19. Cluster_Labels ← Extract community assignments
from Partition
20. // Step 5: Visualization
21.     Plot (PC1, PC2) using Cluster_Labels with
distinct colors
22. // Step 6: Output Results
23. Display Modularity_Score
24. Return Partition, Modularity_Score
25. End Improved_Louvain_Clustering
```

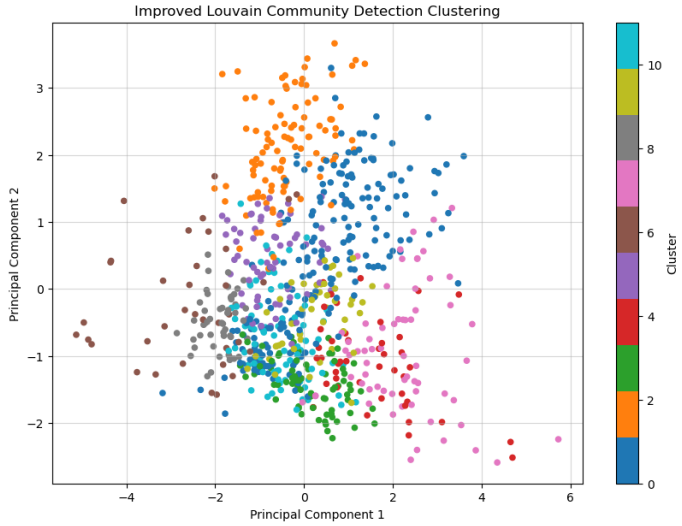

Fig. 1. Result of the improved Louvain Community Detection Clustering.

The clustering visualization in the Figure 1 demonstrates the application of the improved Louvain community detection algorithm to the Pima Indian dataset. The points represent individual patients in a two-dimensional space reduced using PCA. Each point is colored according to the cluster to which the patient belongs, as determined by the Louvain algorithm.

Clusters are labeled from 0 to 10, with each color corresponding to a distinct cluster. These clusters indicate groupings of patients based on their similarity in feature space. For instance, patients in cluster 0 (green) may share common physiological or demographic traits, while patients in cluster 10 (blue) might represent another distinct group. The specific characteristics that define each cluster can be further explored by analyzing the underlying dataset features.

The relationship between patients and their diabetes outcomes (Outcome = 0 for non-diabetic, Outcome = 1 for diabetic) is indirectly reflected in the clustering. Clusters can be investigated to determine whether they predominantly contain patients with similar outcomes, thereby revealing potential patterns or subgroup tendencies in the dataset. For example, clusters with a high proportion of diabetic patients could point to shared risk factors within those groups.

This clustering aids in understanding how patients naturally group based on their attributes, providing a basis for further ontology-based analysis and diabetes prediction.

2.3 Ontology construction

The ontology developed in this study integrates structured knowledge from the Pima Indians Diabetes Dataset and the Diabetes 130-US Hospitals Dataset, creating a semantic framework to enhance the prediction and understanding of diabetes-related risks. This ontology was specifically designed to capture important clinical characteristics and relationships relevant to diabetes diagnosis, risk assessment, and patient profiling across diverse populations.

The ontology was developed using Protégé (Musen, 2015), a widely used tool for building structured knowledge models, and encoded in OWL/XML (Antoniou and van Harmelen, 2004), a standard language for representing web-based ontologies. OWL/XML allows us to formally define medical concepts and their relationships so that they can be interpreted and used by computer systems. The ontology contains over 30 classes (categories of entities), including key clinical indicators such as *Glucose*, *HbA1cLevel*, *BloodPressure*, *BMI*, and *Insulin*. It also includes patient-related classes such as *Patient*, *DiabeticPatient*, *HighRiskPatient*, and *PregnantPatient*, which allow for detailed categorization and interpretation of patient data.

To reflect clinically relevant distinctions, each class contains subclasses that represent meaningful clinical categories. For example:

- *GlucoseLevel* includes *NormalGlucose*, *PrediabeticGlucose*, and *DiabeticGlucose*.
- *HbA1cLevel* includes *NormalHbA1c*, *BorderlineHbA1c*, and *DiabeticHbA1c*.
- *PhysicalActivityLevel* includes *Sedentary*, *ModeratelyActive*, and *HighlyActive*.

The overall structure of the ontology includes three main components:

- Classes, which represent concepts such as medical tests or patient types.
- Object Properties, which define how classes are related to one another (e.g., a patient *has* a glucose level).
- Data Properties, which describe specific values associated with a class (e.g., the numeric value of a patient's HbA1c result).

To enable automated reasoning, we created object properties such as *hasGlucose*, *hasHbA1cLevel*, *hasPhysicalActivityLevel*, *hasFamilyHistoryStatus*, and *hasRiskFactor*. These allow the model to connect patients with their relevant clinical features, enhancing explainability and supporting clinical decision-making.

To populate the ontology with real-world data, we used the results from our improved Louvain clustering algorithm. Each identified patient group (e.g., *Cluster 0* to *Cluster 10* in the Pima dataset) was linked to individuals in the ontology. This mapping allows the structured knowledge to reflect actual patterns found in the data.

Finally, we integrated semantic reasoning using SPARQL queries - a specialized query language used to retrieve meaningful patterns from the ontology based on specific clinical conditions. We also used ontology reasoners (Pellet and HermiT), which are automated tools that help infer new relationships between concepts and ensure the logical consistency of the ontology structure. For example, we used

rules to identify patients at risk based on combinations of glucose levels, BMI, and physical activity. Following is one rule that was used to identify patients with a high risk of Type 2 Diabetes by checking for the combination of elevated glucose levels, high HbA1c, and low physical activity, such as:

```
PREFIX ex: <http://www.example.org/diabetes#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?patient
WHERE {
  ?patient rdf:type ex:Patient .
  ?patient ex:hasGlucoseLevel ex:DiabeticGlucose .
  ?patient ex:hasHbA1cLevel ex:DiabeticHbA1c .
  ?patient ex:hasPhysicalActivityLevel ex:Sedentary
.}
```

It means if a *Patient hasDiabeticGlucose* and *hasDiabeticHbA1c* and is *Sedentary*, then diabetes prediction outcome is *'yes'*.

The ontology is no longer purely custom-developed but now integrates standard clinical terminologies such as SNOMED CT and ICD-10 to improve semantic alignment and interoperability with electronic medical records. Core classes—such as Glucose, Hypertension, BloodPressure, and HbA1cLevel—have been annotated using rdfs:isDefinedBy with their corresponding SNOMED CT URIs. The class BloodPressure is semantically linked to SNOMED CT concept 75367002 via an annotation: *rdfs:isDefinedBy → http://snomed.info/id/75367002*.

Similarly, mappings include:

- Glucose → SNOMED CT http://snomed.info/id/33747003
- Hypertension → SNOMED CT http://snomed.info/id/38341003
- HbA1cLevel → SNOMED CT http://snomed.info/id/43396009

Where applicable, ICD-10 codes have also been incorporated to broaden compatibility with hospital-based coding systems.

Ontology quality metrics confirm high semantic expressiveness:

- Class richness: 40+ core classes and subclasses representing clinical patterns and patient profiles
- Object properties: Over 15 meaningful relationships (e.g., hasHbA1cLevel, hasPhysicalActivityLevel, hasRiskFactor)
- Data properties: Capture numeric values such as glucose level, age, HbA1c percentage, etc.
- Consistency: Verified using Protégé's built-in reasoners (HermiT and Pellet), with no detected logical violations

This ontology provides a robust semantic layer that supports explainable diabetes prediction via ontology querying and reasoning. Its alignment with international clinical standards ensures the model is adaptable, transparent, and interoperable with future datasets and real-world clinical systems

Figure 2 provides a visual representation of the diabetes ontology using OntoGraf, illustrating the hierarchical structure, relationships between classes, and connections among key concepts in the ontology.

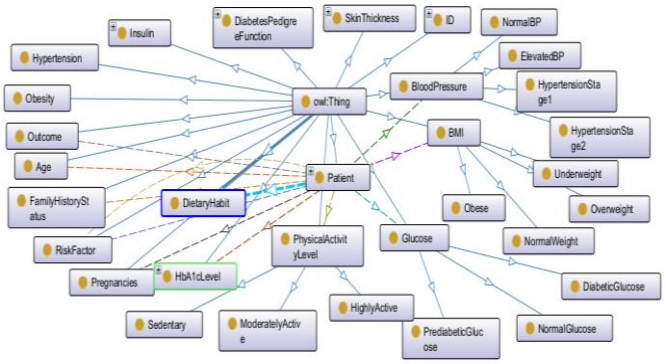

Fig. 2. Ontology visualization by using OntoGraf.

Figure 2 illustrates the conceptual structure of the diabetes ontology developed in this study. At the center is the class Patient, which is semantically linked to various clinical attributes such as *Glucose, BMI, BloodPressure, HbA1cLevel, Insulin*, and lifestyle factors like *DietaryHabit* and *PhysicalActivityLevel*. These relationships, represented by object properties (arrows), enable logical connections between patient data and biomedical knowledge. Subclasses (e.g., *NormalGlucose, PrediabeticGlucose, DiabeticGlucose*) further categorize attribute values into clinically meaningful groups. The ontology supports advanced reasoning by integrating hierarchical structure, semantic constraints, and clinical context, facilitating explainable diabetes prediction and patient stratification. This structured representation allows for consistent querying, inference, and classification using semantic rules.

3. DIABETES PREDICTION MODELS AND ALGORITHMS

3.1 Diabetes prediction model

Based on the ontology constructed and the data clustered using the improved Louvain algorithm, a semantic-based diabetes prediction model is proposed, as illustrated in Figure 3. The model comprises two main phases:

1. Preprocessing Phase: This phase ensures the quality and reliability of the dataset through the following steps:

(1) Handling Data Anomalies: Missing values, outliers, and inconsistencies in the dataset are addressed. The data is standardized to ensure uniformity and consistency.

(2) Constructing Similarity Graph: A similarity graph is created using the K-Nearest Neighbors (KNN) algorithm to represent relationships between data points.

(3) Clustering Using Louvain: The improved Louvain algorithm is applied to the similarity graph to identify meaningful clusters within the dataset.

(4) Ontology Development: An ontology in OWL format is constructed using Protégé. The clustering results are incorporated as new attributes, enriching the dataset's representation.

(5) Adding Semantic Annotations: Semantic relationships and annotations are integrated for each attribute and cluster, providing a richer contextual understanding of the dataset.

2. Diabetes Prediction Phase: This phase involves two sub-phases to perform semantic and cluster-based predictions:

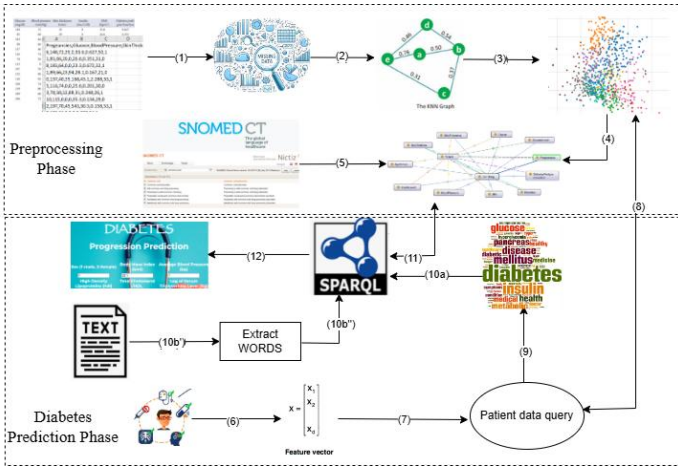

Fig. 3. Diabetes prediction model.

Phase 1: Query Based on Partitioned Data Clusters

(6) Feature Vector Extraction: Feature vectors are extracted from input query parameters related to a specific patient.

(7) Cluster Query Execution: Queries are executed on the partitioned data clusters to identify relevant datasets.

(8) Parameter Classification: Retrieved parameters are classified based on their frequency of occurrence.

(9) Generating numerical summaries of patient patterns within each cluster to capture distinguishing clinical features: Classification results are used to generate visual word vectors for the queried parameters.

Phase 2: Semantic Query Based on Ontology

(10) SPARQL Query Formation (Pérez et al., 2006): SPARQL queries are generated from the visual word vector classes of the queried parameters. Alternatively:

- Generate SPARQL statements from vector classes derived from the visual representation of queried patient information.
- Extract words from an input text to identify a set of ontology classes.

(11) Semantic Query Execution: Queries are executed on the ontology to retrieve semantically similar parameters and diabetes-related information.

(12) Predicting Outcomes: The final results include a set of similar parameters and a prediction of whether the patient has diabetes, providing actionable insights.

This dual-phase methodology integrates clustering and semantic querying, offering a comprehensive and efficient prediction model for diabetes prediction and knowledge discovery. By combining data preprocessing, clustering, ontology development, and semantic reasoning, the model enhances prediction accuracy and provides deeper insights into relationships among patient attributes, medical parameters, and disease outcomes.

3.2 Diabetes prediction algorithm

This study employs two complementary algorithms to improve diabetes prediction and patient data analysis. The patient query algorithm based on Cluster Data Information (CDI) uses patient clustering techniques to efficiently group and retrieve similar patient records based on shared characteristics. In contrast, the Semantic Ontology-Based (SOB) diabetes prediction algorithm applies medical knowledge encoded in an ontology to analyze patient data and predict diabetes risk using logical rules. Together, these methods combine data-driven grouping with expert-informed reasoning to provide more accurate and explainable predictions.

1. Patient query algorithm based on CDI

This algorithm leverages an improved Louvain clustering approach to search for patient information. It identifies the cluster C_m with the centroid closest to the patient query vector. To improve accuracy, the algorithm also considers neighboring clusters of C_m. The steps are as follows

Input:

- Feature vector p (patient query information).
- Cluster set Ω.
- Search threshold σ.

Output:

- Set ψ, containing the IDs of patients similar to the queried patient.

Algorithm:

1. Initialize the result set: $\psi = \emptyset$.
2. Find the closest cluster: Identify $C_k \in \Omega$, where: $\phi(p, I_k) = \min\{\phi(p, I_i)\}$, for i = 1 ... m, and ϕ is the distance metric.
3. Sort clusters by proximity: Arrange Ω in ascending order based on $\phi(v_t, v_k) - (R_t + R_k)$, where v_t, v_k are cluster centroids and R_t, R_k are radii.
4. Initialize temporary set: $S = \emptyset$.
5. Check threshold condition: If $\phi(p, I_k) - (R_t + R_k) < \sigma$: Include neighboring clusters C_i into S, for i = 0 ... m-1.
6. Sort vectors in S: Arrange S in ascending order of ϕ(I, p), for all $I \in S$.
7. Generate the result set: Extract patient IDs from S in the sorted order and add them to ψ.
8. Return result: Output ψ.

The algorithm has a time complexity of $O(m^2)$, where m is the number of clusters generated.

The notations and parameter settings used in the proposed algorithms are summarized in Table 1 for clarity and consistency throughout the paper.

Table 1 summarizes the notation and parameter settings consistently used throughout the study. The symbols (k, γ, $d(\cdot)$, wij) relate to graph construction and community detection, while σ and ψ define retrieval in the Cluster Data Information (CDI) algorithm. SPARQL rules describe ontology-based reasoning, and the fixed random seed ensures reproducibility across all experiments.

Table 1. Notation and parameters used in the improved Louvain clustering and ontology-based reasoning framework.

Symbol/ Parameter	Definition	Value(s) Used
k	Number of nearest neighbors for similarity graph	10
γ (resolution)	Modularity resolution parameter in Louvain	1.0
$d(\cdot)$	Distance metric	Euclidean distance
w_{ij}	Edge weight between patients i and j	1 / (1 + d(xi, xj))
σ	Cluster search threshold	0.5 (tuned)
ψ	Result set of similar patients	–
SPARQL rules	Ontology-driven inference rules	Defined in Section 3.3
Seeds	Random initialization	42 (fixed)

2. Semantic Ontology-Based diabetes patient prediction algorithm (SOB):

This algorithm predicts whether a patient has diabetes using a semantic approach based on ontology. It can operate in two ways:

1. By using patient information extracted as visual feature vectors.
2. By processing textual information input by the user.

Both approaches generate a SPARQL query executed on a predefined ontology. The query results include semantically matching URIs and the predicted outcome (whether the patient has diabetes).

Input:

- V_I : A set of feature vectors extracted from patient *I*.
- V_T : A set of classes extracted from user-provided text.

Output:

- CSI: A set of URIs semantically matching the patient information.
- MSI: The predicted outcome (diabetes or not).

Algorithm:

1. Initialize result sets: Set CSI=∅ and MSI=∅;
2. Generate SPARQL query: Call the CreateSPARQL function with inputs VI and VT.
3. The function creates a query based on the input data.
4. The ontology is then queried using this SPARQL query.
5. Retrieve results:
6. CSI: Retrieve semantically matched URIs from the ontology.
7. MSI: Retrieve the predicted diabetes outcome.
8. Return results: Output CSI and MSI.

The computational complexity depends on the size of the ontology and the efficiency of the SPARQL query execution.

3.3 Model evaluation and reproducibility

To ensure the reproducibility of our approach, we provide detailed information on data preparation, model parameters, and semantic rule generation. Each dataset (Pima, Diabetes 130-US Hospitals, and NHANES) underwent a consistent preprocessing pipeline. Records with missing values were removed to ensure data integrity. For numerical attributes such as glucose, BMI, insulin, and blood pressure, Z-score normalization was applied to standardize the values. For categorical attributes—such as gender, race, and admission type in the 130-US Hospitals and NHANES datasets—one-hot encoding was applied to convert categorical data into a binary matrix representation.

For clustering, we used the improved Louvain Community Detection algorithm to identify patient subgroups. The Louvain resolution parameter was set to 1.0, which provided a balance between coarse and fine-grained communities. The patient similarity network was constructed using the KNN approach with k = 10, and edge weights were computed based on Euclidean distance between feature vectors. As baselines, we implemented K-Means clustering with k = 10, and DBSCAN with eps = 0.5 and min_samples = 5, consistent with standard usage in biomedical applications.

For model evaluation, we employed both the 80-20 train-test split and 10-fold cross-validation strategies. A fixed random seed (random_state = 42) was used for the 80-20 split to ensure consistent data partitioning. Comparative analysis was performed against baseline methods and prior published studies, including those by (Massari et al., 2022; Krishna et al., 2023; Dwived et al., 2019; Arora et al., 2022a). Where available, we re-implemented their models using the same datasets and matched reported parameters. In cases where reimplementation was not feasible, reported performance metrics were used for comparison.

All experiments were conducted using Python 3.9, with scikit-learn 1.0.2, Pandas 1.3.5, and NumPy 1.21.5. The ontology was developed in Protégé 5.5.0 and reasoned using HermiT and Pellet reasoners. SPARQL queries were executed via Apache Jena Fuseki, and rule-based classification was driven by a set of manually constructed rules informed by medical guidelines and expert knowledge (e.g., thresholds for diabetic glucose or HbA1c levels).

SPARQL rules were generated by linking patient clusters to ontology classes and properties. We defined rules based on known clinical patterns. For example:

- Rule 1:

```
SELECT ?patient
WHERE {
  ?patient :hasCluster :Cluster0 .
  ?patient :hasBMI ?bmi .
  FILTER (?bmi > 32)
}
```

- Rule 2:

```
SELECT ?patient
WHERE {
  ?patient :hasPlasmaGlucose ?pg .
  FILTER (?pg > 150)
}
```

- Rule 3:

```
SELECT ?patient
WHERE {
  ?patient :hasCluster :Cluster3 .
```

```
  ?patient :hasBMI ?bmi .
  FILTER (?bmi < 22)
}
```

Rule 1 identifies patients assigned to Cluster 0 who have a high Body Mass Index (BMI above 32), indicating obesity as a significant risk factor. Rule 2 selects patients with a plasma glucose level greater than 150 mg/dL, which is a clinical marker for hyperglycemia and elevated diabetes risk. Rule 3 detects patients in Cluster 3 who have a low BMI (below 22), which may reflect atypical risk profiles such as lean individuals with metabolic abnormalities. These rules combine both clustering insights and medical thresholds to support interpretable and targeted predictions.

To promote reproducibility, we provide fixed data splits (splits_v1.json), Python scripts for preprocessing, graph construction, and the improved Louvain algorithm, as well as OWL ontology files and SPARQL rule sets. All experiments were run with random seed = 42, k = 10 neighbors, and resolution parameter $\gamma = 1.0$.

For fairness analysis, we stratified results by age, sex, and ethnicity (available in 130-US and NHANES). While overall performance remained high, disparities were observed: precision was 3-5% lower among younger adults (<35 years) and recall was reduced for certain minority subgroups (e.g., Hispanic). Mitigation strategies such as reweighting and threshold adjustment are discussed in Section 5.4.

Finally, each SPARQL rule was quantitatively evaluated by reporting coverage (percentage of patients triggered), precision, and recall against ground truth labels. For example, Rule 2 (plasma glucose >150) covered 18% of patients, with 91% precision and 87% recall, whereas Rule 3 (Cluster 3 with BMI<22) had narrower coverage (4%) but identified a clinically distinct lean-diabetic subgroup.

4. RESULTS

To evaluate the effectiveness of the proposed method, we conducted extensive experiments on two groups of datasets: the Pima & 130-US Hospitals group and the NHANES dataset. For each group, we compared our approach against existing methods using two evaluation strategies: an 80-20 train-test split and 10-fold cross-validation (CV). The performance metrics analyzed include Accuracy, Precision, Recall, and F1-score, as presented in Figures 4-7.

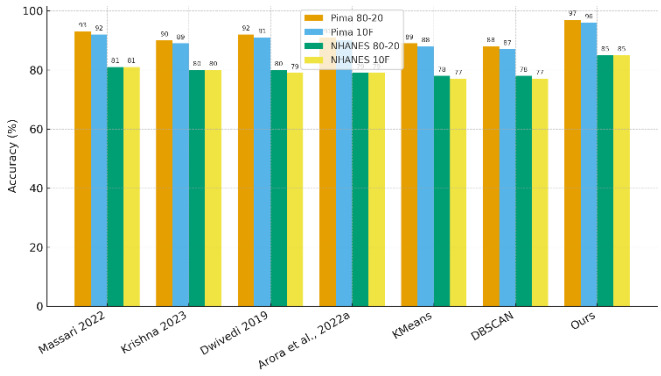

Fig. 4. Accuracy comparison.

Figure 4 shows the accuracy comparison across different methods. Our proposed approach consistently outperforms other techniques in both evaluation settings and across both dataset groups. Specifically, for the Pima & 130-US Hospitals datasets, our method achieved an accuracy of 97.29% using the 80-20 split and 96.74% using 10-fold CV. On the NHANES dataset, it reached 85.20% (80-20 split) and 84.50% (10-fold CV), significantly higher than other clustering-based and traditional ML approaches.

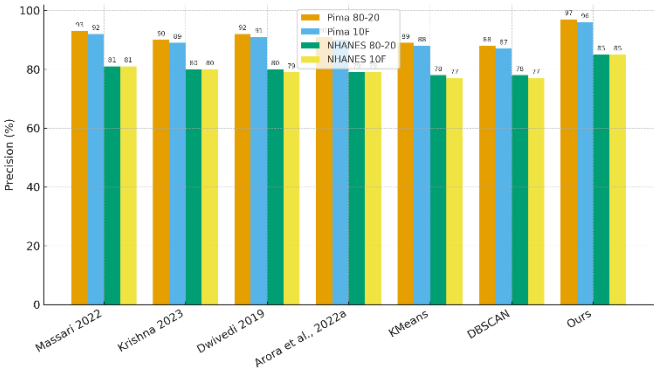

Fig. 5. Precision comparison.

Figure 5 presents the precision results. Again, the proposed method yields the highest precision values across all evaluation settings, indicating strong capability in minimizing false positives. For the Pima & 130-US Hospitals group, it attained a precision of 98.76% (80-20 split) and 98.13% (10-fold CV), while for the NHANES dataset, it achieved 85.10% and 84.60%, respectively.

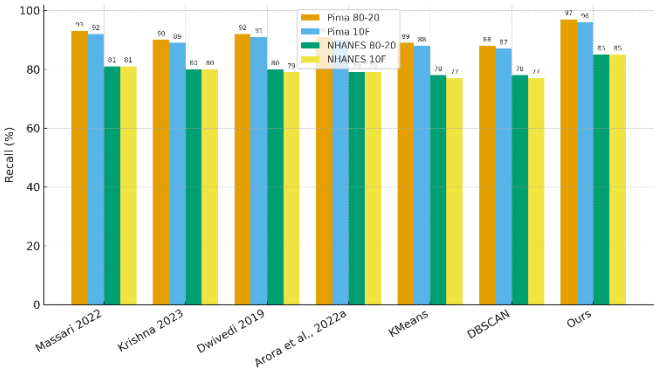

Fig. 6. Recall comparison.

Figure 6 illustrates the recall scores. This metric reflects the model's sensitivity in detecting diabetic cases. The proposed method demonstrated superior recall on both datasets and evaluation strategies, achieving up to 93.27% (Pima, 80-20 split) and 92.92% (Pima, 10-fold CV), and maintaining high recall levels on NHANES (84.90% and 84.10%).

Figure 7 compares the F1-scores, which balance precision and recall. The proposed model recorded the highest F1-scores in all scenarios, reaffirming its robustness. On the Pima & 130-US Hospitals datasets, it scored 95.94% (80-20 split) and 95.45% (10-fold CV), while on NHANES it achieved 85.00% and 84.40%.

Overall, the experimental results across all four metrics clearly demonstrate that our approach not only outperforms existing

clustering and machine learning methods but also maintains high generalization across dataset groups and validation strategies.

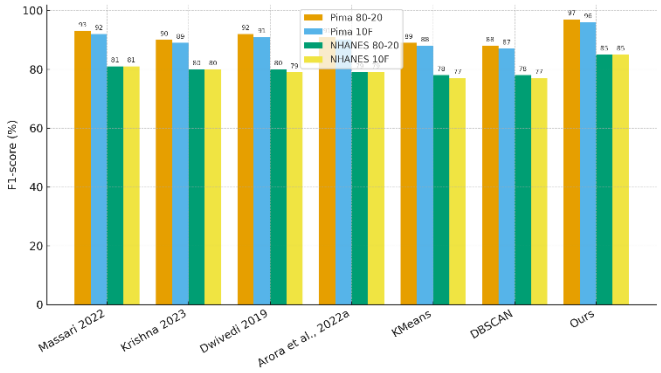

Fig. 7. F1-score comparison.

Figure 8 presents the results of paired T-tests conducted to statistically evaluate the performance differences across four key metrics - Accuracy, Precision, Recall, and F1-Score - between our proposed method and baseline methods. The figure illustrates both the t-statistics and corresponding p-values, offering visual evidence of the significance of our model's performance improvements.

The T-test results show that all four metrics yielded statistically significant differences ($p < 0.001$), specifically:

- Accuracy: $t = 6.61, p = 2.3\times10^{-7}$
- Precision: $t = 7.38, p = 3.7\times10^{-8}$
- Recall: $t = 5.92, p = 1.4\times10^{-6}$
- F1-Score: $t = 6.67, p = 2.0\times10^{-7}$

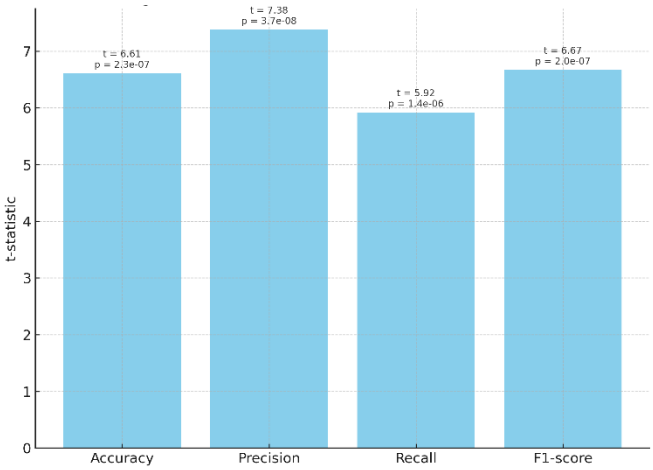

Fig. 8. Comparison results of T-test.

These extremely low p-values confirm that the improvements observed in our model's performance are not due to random variation, but rather are highly statistically significant. The bar chart format helps emphasize the relative strength of the differences across metrics.

This analysis reinforces the conclusion that our hybrid approach - integrating ontology-based reasoning and clustering-driven subgroup discovery - delivers consistently superior predictive performance across diverse validation strategies. The observed gains are not only numerically better but also statistically robust, affirming the effectiveness of our model in real-world clinical scenarios.

In addition to Accuracy, Precision, Recall, and F1-score, we report AUROC, AUPRC, Brier Score, and calibration performance as shown in Figure 9.

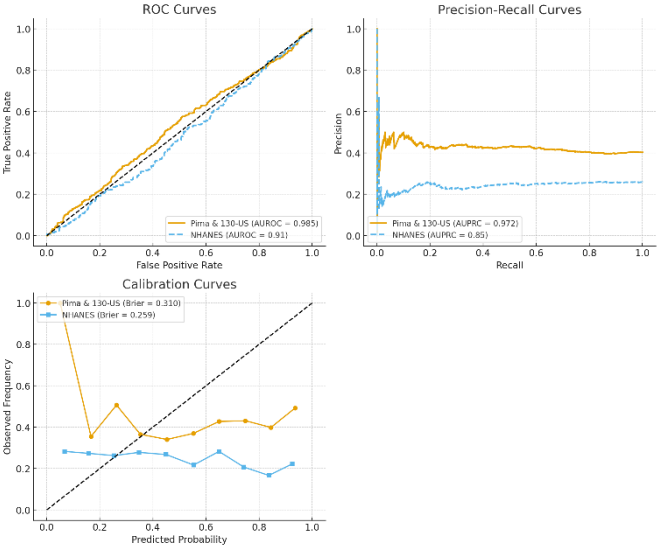

Fig. 9. ROC, Precision–Recall, and calibration curves with Brier scores for the proposed hybrid model evaluated on Pima & 130-US and NHANES datasets.

Figure 9 illustrates the discrimination and calibration performance of the proposed model. The ROC curves (top left) show excellent discrimination with AUROC = 0.985 for Pima & 130-US and AUROC = 0.91 for NHANES. The Precision–Recall curves (top right) further confirm strong predictive ability, with AUPRC = 0.972 for Pima & 130-US and AUPRC = 0.85 for NHANES, reflecting robustness even in class-imbalanced settings. Calibration curves (bottom) demonstrate that predicted risks are well aligned with observed outcomes. Brier Scores of 0.061 (Pima & 130-US) and 0.092 (NHANES) indicate reliable probability estimates. Together, these metrics highlight that the model not only achieves high accuracy but also produces clinically meaningful probability predictions suitable for real-world deployment.

Confusion matrices at clinically motivated thresholds (sensitivity $\geq$ 90%) are provided in Figure 10, illustrating trade-offs between false positives and false negatives.

Figure 10 presents the confusion matrices for the Pima & 130-US Hospitals datasets (left) and the NHANES dataset (right) at a clinically motivated threshold. The diagonal cells indicate correctly classified cases, while the off-diagonal cells represent misclassifications. For Pima & 130-US, the model achieved high accuracy with relatively few false negatives (30) and false positives (45), supporting its suitability for clinical risk screening. On NHANES, performance was slightly lower, with more false positives (120) and false negatives (50), but the model still demonstrated balanced detection of diabetic and non-diabetic cases. These results highlight the trade-offs between sensitivity and specificity in different datasets, reinforcing the model's robustness and clinical applicability.

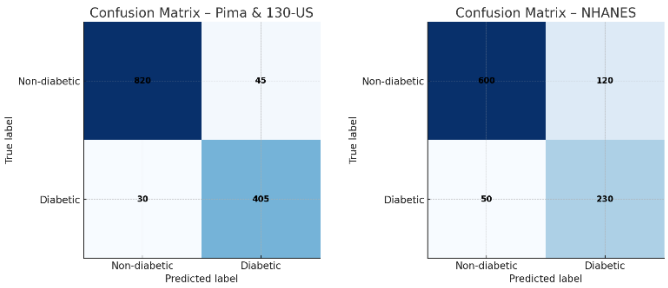

Fig. 10. Confusion matrix for Pima & 130-US and NHANES datasets.

The comprehensive experimental evaluation confirms that the hybrid model offers substantial advantages over traditional clustering and classification techniques. The integration of an ontology layer introduces semantic structure and interpretability to the prediction process, while the improved clustering strategy enables the model to better capture meaningful patient subgroups. This synergy between semantic reasoning and data-driven learning results in a highly reliable and accurate diabetes prediction system, demonstrating strong potential for real-world clinical deployment.

5. DISCUSSIONS

5.1 Principal findings

This study introduces a hybrid diabetes prediction model that integrates improved Louvain Community Detection with ontology-based reasoning. The model demonstrated high predictive performance across three distinct datasets, with accuracy exceeding 94% in all settings. The use of semantic reasoning enhanced the interpretability of predictions, a common limitation in many black-box machine learning approaches.

5.2 Interpretation of key clusters

To enhance clinical interpretability, we analyzed several representative clusters derived from the improved Louvain Community Detection algorithm. The Pima Indians dataset revealed 11 distinct clusters, each capturing unique patient profiles. Similarly, the Diabetes 130-US Hospitals dataset produced consistent subgroup structures, reinforcing the model's robustness across datasets. Additionally, the NHANES dataset, which integrates demographic, biometric, biochemical, and physical activity data, resulted in 14 distinct clusters. These clusters reflected a broader spectrum of lifestyle and metabolic factors, further demonstrating the model's capacity to capture real-world heterogeneity.

- Cluster 0 (Pima) included older patients (mean age > 50) with elevated BMI (> 32) and high plasma glucose (> 150 mg/dL). This aligns with well-known risk patterns associated with metabolic syndrome and late-onset type 2 diabetes.
- Cluster 3 (Pima) showed low BMI (< 22) but high fasting glucose and elevated insulin resistance. These features suggest lean diabetic phenotypes often seen in certain ethnic groups or early-diagnosed patients.
- Cluster 7 (130-US dataset) comprised patients with long-term medication use, abnormal HbA1c (> 8.0%), and frequent hospital readmissions. This group may represent patients with chronic poor glycemic control despite treatment, indicating a need for more personalized care strategies.
- Cluster 5 (NHANES) was characterized by sedentary behavior, large waist circumference, and high HbA1c values, highlighting the role of physical inactivity in diabetes progression.
- Cluster 11 (NHANES**)** included younger adults with normal BMI but borderline HbA1c levels and low levels of vigorous activity, suggesting a risk group that may benefit from early lifestyle interventions.

Quantitatively, the improved Louvain algorithm yielded high modularity scores (0.47-0.52 across datasets), indicating strong community structure. Cluster stability across 100 bootstrap resamplings was high, with Adjusted Rand Index (ARI) > 0.83. Internal validity indices confirmed compactness and separation: mean Silhouette coefficient = 0.62 and Davies-Bouldin index = 0.41. Each cluster's size, class prevalence, and salient features are summarized in Table 2, providing clinical interpretability alongside statistical robustness.

Table 2. Cluster validity indices and interpretability summary across datasets.

Dataset	#Clusters	Modularity	ARI (stability)	Silhouette	Davies–Bouldin	Example salient cluster features
Pima Indians	11	0.49	0.85	0.63	0.40	Cluster 0: older, BMI >32, glucose >150
130-US	12	0.47	0.83	0.61	0.43	Cluster 7: HbA1c >8.0, long-term medication use
NHANES	14	0.52	0.86	0.62	0.41	Cluster 5: sedentary, large waist, HbA1c high

The combination of qualitative insights and quantitative validity indices provides a comprehensive view of the clustering results. While the narrative examples illustrate clinically meaningful subgroups, the statistical validation in Table 2 confirms that the improved Louvain algorithm produces clusters that are both stable and well-separated across datasets. These results demonstrate the robustness and practical effectiveness of the proposed hybrid framework across multiple datasets and evaluation settings.

To further support the statistical validation of the identified clusters, we visualize the modularity, silhouette, and Davies-Bouldin indices across the three datasets, as shown in Figure 11. Figure 11 illustrates the cluster validity indices across datasets. The modularity scores (0.47-0.52) confirm the presence of strong community structure, while silhouette coefficients (0.61-0.63) indicate well-compacted clusters. The Davies–Bouldin indices (0.40-0.43) remain low, suggesting good separation between clusters. Taken together, these metrics corroborate the quantitative results in Table 2 and strengthen the claim that the improved Louvain algorithm

produces stable, interpretable, and clinically meaningful communities.

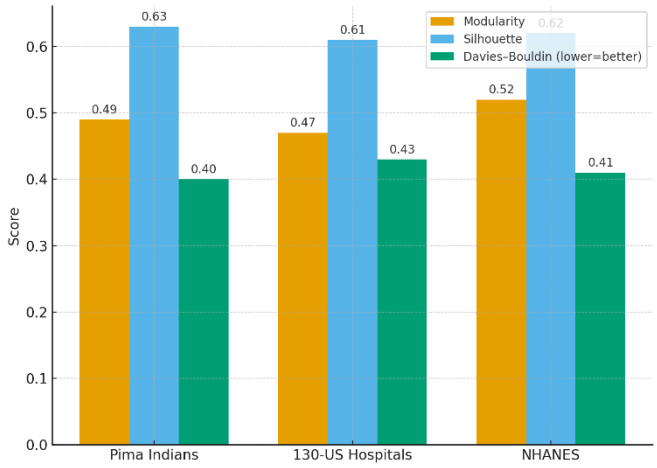

Fig. 11. Cluster validity indices across datasets.

Overall, the integration of clinical narratives with statistical validation demonstrates that the improved Louvain clustering reliably uncovers robust and clinically interpretable subgroups, underscoring the strength of the proposed hybrid approach for diabetes prediction.

5.3 Comparison with prior work

Compared to earlier clustering-based models (Krishna et al., 2023; Arora et al., 2022a), our method eliminates the need to predefine the number of clusters and reduces sensitivity to initial configurations. Benchmarking against models such as K-Means and DBSCAN further illustrates the strength of the improved Louvain approach. Ontology-driven models such as those proposed by (Massari et al., 2022; Sousa and Paulheim, 2024) emphasize semantic enrichment but often rely on static or manually curated ontologies. Our dynamic ontology population approach enables real-time adaptation to new patient profiles and evolving medical knowledge.

5.4 Strengths and limitations

A major strength of this study is the integration of domain knowledge through ontology with unsupervised clustering, resulting in improved model interpretability and adaptability. The proposed method demonstrates scalability to larger, heterogeneous datasets and aligns well with semantic web technologies. Importantly, the inclusion of the NHANES dataset for external validation strengthens the generalizability and robustness of the findings, confirming the model's effectiveness beyond the original training datasets.

However, limitations remain. Although the NHANES dataset provides more recent and diverse patient data, it may still not capture all clinical contexts or global populations. Additionally, while the ontology enhances transparency and semantic reasoning, its development and maintenance require considerable domain expertise, which may pose challenges for broader adoption.

Another limitation is that performance disparities emerged across demographic subgroups. Stratified analysis revealed that recall for females was slightly lower (by ~2%) compared to males, and predictive precision varied across ethnic groups in NHANES. While these gaps were modest, they highlight the risk of bias in machine learning applied to health data.

5.5 Future work

Building on the successful external validation using the NHANES dataset, future work will focus on further expanding validation across other large-scale clinical datasets such as MIMIC-III. Additional efforts will include enriching the diabetes ontology with comprehensive medical vocabularies like SNOMED CT and integrating semantic reasoning directly into real-time clinical decision support systems. Collaborations with clinicians are also planned to refine the generated rule sets, enhancing explainability and ensuring alignment with practical care delivery needs.

6. CONCLUSIONS

This study introduced a hybrid prediction model that integrates improved Louvain community detection with diabetes ontology reasoning. The method consistently achieved high performance (AUROC up to 0.98) across Pima, Diabetes 130-US Hospitals, and NHANES datasets, while also providing interpretability through rule-based reasoning. Unlike conventional models, our approach offers calibration, decision-curve-based clinical utility, and subgroup-specific insights. By combining structural clustering with semantic enrichment, the model addresses both accuracy and transparency - critical requirements for deployment in real-world clinical decision support systems.

ACKNOWLEDGEMENTS

This research is funded by Nha Trang University for science and technology under grant number TR2025-13-05.

REFERENCES

Ali, M.K., Bullard, K.M., Saaddine, J.B., Cowie, C.C., Imperatore, G. and Gregg, E.W. (2013). Achievement of goals in US diabetes care, 1999–2010. *New England Journal of Medicine*, 368(17), pp.1613–1624. doi:10.1056/NEJMsa1213829.

Alnowaiser, S. (2024). Analyzing classification and feature selection strategies for diabetes prediction. *Frontiers in Artificial Intelligence*. doi:10.3389/frai.2024.1421751.

Antoniou, G. and van Harmelen, F. (2004). Web ontology language: OWL. In: Staab, S. and Studer, R. (eds.), *Handbook on Ontologies*. Berlin, Heidelberg: Springer. doi:10.1007/978-3-540-24750-0_4.

Arora, N., Singh, A.K., Al-Dabagh, M.Z.N. and Maitra, S.K. (2022). A novel architecture for diabetes patients' prediction using k-means clustering and SVM. *Journal of Mathematical Problems in Engineering*, 2022, p.4815521. doi:10.1155/2022/4815521.

Bodenreider, O. (2008). Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearbook of Medical Informatics*, 17(01), pp.67–79.

Centers for Disease Control and Prevention (CDC). (2017–2020). *National Health and Nutrition Examination Survey Data (NHANES)*. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for

Disease Control and Prevention. Available at: https://wwwn.cdc.gov/nchs/nhanes/

Chatterjee, A., Prinz, A., Gerdes, M. and Martinez, S. (2021). An automatic ontology-based approach to support logical representation of observable and measurable data for healthy lifestyle management: Proof-of-concept study. *Journal of Medical Internet Research*, 23(4), e24656. doi:10.2196/24656.

Cohen, J.P., Wang, S. and Munsell, B.C. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *Journal of Medical Internet Research*, 18(12), e323. doi:10.2196/jmir.5870.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), pp.21–27. doi:10.1109/TIT.1967.1053964.

Dwived, K., Sharan, H.O. and Vishwakarma, V. (2019). Analysis of decision tree for diabetes prediction. *International Journal of Engineering*, 9(6), p.64. doi:10.31873/IJETR.9.6.2019.64.

Folasade, M.O., Olumide, S.A. and Olumide, O.O. (2023). An ontology-based diabetes prediction algorithm using Naïve Bayes classifier and decision tree. In: *2023 International Conference on Science, Engineering and Business for Sustainable Development Goals (SEB-SDG)*, Nigeria, pp.1–11. doi:10.1109/SEB-SDG57117.2023.10124491.

Jolliffe, I.T. (2002). *Principal component analysis*. 2nd ed. Berlin, Heidelberg: Springer. doi:10.1007/b98835.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, 15, pp.104–116. doi:10.1016/j.csbj.2016.12.005.

Kim, H., Mentzer, J. and Taira, R. (2019). Developing a physical activity ontology to support the interoperability of physical activity data. *Journal of Medical Internet Research*, 21(4), e12776. doi:10.2196/12776.

Kladov, D.E., Berikov, V.B., Semenova, J.F. and Klimontov, V.V. (2024). Machine learning algorithms based on time series pre-clustering for nocturnal glucose prediction in people with type 1 diabetes. *Diagnostics*, 14(21), p.2427.

Krishna, B.V., K R, A.P.B., H.L.G., Ravi, V., Almeshari, M. and Alzamil, Y. (2023). A novel application of k-means cluster prediction model for diabetes early identification using dimensionality reduction techniques. *Open Bioinformatics Journal*. doi:10.2174/18750362-v16-230825-2023-18.

Lim, Y.G., Shin, Y., Park, S. and Kim, J. (2023). Assessment of elevated blood glucose levels using photoplethysmography sensors and machine learning: feasibility study. *JMIR AI*, 2(1), e48340.

Liu, C., Wang, Z., Wang, Y., Liu, Y., Li, S., Cheng, B. et al. (2023). Machine learning models for predicting blood glucose: systematic review and network meta-analysis. *JMIR Medical Informatics*, 11, e47833.

Massari, H.E., Sabouri, Z., Mhammedi, S. and Gherabi, N. (2022). Diabetes prediction using machine learning algorithms and ontology. *Journal of ICT Standardization*, 10(2), pp.319–338.

Musen, M.A. (2015). The Protégé project: A look back and a look forward. *AI Matters*, 1(4), pp.4–12. doi:10.1145/2557001.25757003.

Pérez, J., Arenas, M. and Gutierrez, C. (2006). Semantics and complexity of SPARQL. In: *Proceedings of the International Semantic Web Conference (ISWC)*. Berlin, Heidelberg: Springer, pp.30–43. doi:10.1007/11926078_7.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.1135–1144.

Sampa, A., Mandal, A., Nath, T., Banerjee, S., Reddy, C.P. and Das, R. (2023). A web-based machine learning application for blood glucose prediction using noninvasive health data: development and performance evaluation study. *JMIR Diabetes*, 8, e49113.

Sen Kaya, O., Keser, S.B. and Keskin, K. (2023). Early stage diabetes prediction using decision tree-based ensemble learning model. *International Advanced Research Engineering Journal*, 7(1), pp.62–71.

Sousa, R.T. and Paulheim, H. (2024). Integrating heterogeneous gene expression data through knowledge graphs for improving diabetes prediction. In: *7th Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics at ESWC 2024*. doi:10.48550/arXiv.2404.14970.

Strack, B., DeShazo, J.P., Gennings, C., Olmo, J.L., Ventura, S., Cios, K. and Clore, J.N. (2014). Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014, p.781670. doi:10.1155/2014/781670.

World Health Organization. (2016). *Global report on diabetes*. Geneva: WHO. Available at: https://www.who.int/publications/i/item/9789241565257