EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Ontology-Enhanced Machine Learning Models for Breast Cancer Diagnosis

Pham Thi Thu Thuy[1,*], Bui Chi Thanh[1]

[1]Nha Trang University, Khanh Hoa Province, Vietnam

Abstract

INTRODUCTION: Breast cancer remains one of the most prevalent causes of cancer-related mortality among women globally. While machine learning (ML) has demonstrated promise in early detection, conventional models often rely solely on statistical features, lacking domain-specific knowledge and interpretability.
OBJECTIVES: This study aims to enhance breast cancer prediction by integrating ontology-driven semantic features with ML models to improve both predictive accuracy and clinical interpretability.
METHODS: We applied a comprehensive pipeline comprising data preprocessing, statistical testing, and dimensionality reduction using PCA, followed by training with supervised learning models including Logistic Regression, k-NN, SVM, Random Forest, XGBoost, LightGBM, and Attention-Enhanced MLP. In the proposed approach, clinical data is transformed into RDF triples and structured within a domain-specific breast cancer ontology. Semantic reasoning via SPARQL queries enables the extraction of high-level features, which are then used in a leakage-safe stacking design that integrates (i) tabular features, (ii) KGE features, (iii) semantic subtyping signals, and (iv) SPARQL rule features, with reproducible templates and released code.
RESULTS: Across four benchmark datasets, the ontology-enhanced meta-learner achieved consistently strong performance, achieving 0.996 ± 0.006 ROC-AUC on WDBC under stratified evaluation.
CONCLUSION: Incorporating ontology-derived semantic knowledge significantly improves the performance, robustness, and interpretability of ML models for breast cancer prediction. This approach holds strong potential for real-world integration into clinical decision support systems.

Keywords: Breast Cancer, Machine Learning, Ontology, Semantic Reasoning, Predictive Modeling

Received on 19 October 2025, accepted on 01 April 2026, published on 09 April 2026

doi: 10.4108/eetpht.11.10650

1. Introduction

Breast cancer is the most frequently diagnosed cancer and one of the leading causes of cancer-related deaths among women worldwide [1]. Early diagnosis significantly improves the likelihood of successful treatment, yet the diagnostic process often involves subjective clinical judgment, complex pathology data, and high inter-observer variability [2]. In this landscape, machine learning (ML) offers promise by automating diagnosis and supporting medical decision-making based on data analysis [3].

However, traditional ML approaches are primarily data-driven and often act as "black-box" systems that lack transparency and domain-level interpretability [4]. They typically ignore the rich semantic and relational structures embedded in medical knowledge, such as patient history, clinical guidelines, and pathophysiological interdependencies. This limits their clinical applicability

*Corresponding author. Email: thuthuy@ntu.edu.vn

and trustworthiness, particularly in high-stakes domains like cancer diagnostics.

Recent research has explored the integration of ontologies - structured representations of domain knowledge - to enhance ML interpretability and enable reasoning over complex relationships [5, 6]. By encoding medical concepts and their interrelations, ontologies can generate semantically rich features, support logical inference via SPARQL rules, and enhance explainability of predictions [7]. In diabetes and cardiovascular disease prediction, ontology-enhanced models have demonstrated superior interpretability and improved performance [8, 9]. However, such semantic modeling remains underutilized in breast cancer diagnosis, where the integration of domain-specific ontologies with ML remains a nascent area of research.

To address these limitations, we propose an ontology-driven hybrid framework to enhance machine learning-based breast cancer prediction. In this framework, patient datasets are first transformed into OWL ontologies and enriched with semantic relationships using domain knowledge. The ontology is then converted into RDF triples, from which structured embeddings are learned using PyKEEN and knowledge graph embedding models such as TransE. To explore the semantic structure of the data, we apply unsupervised clustering algorithms, including Leiden and HDBSCAN, which help uncover latent subgroups of patients with similar clinical and semantic characteristics. These high-level semantic features, combined with traditional statistical features, are used to train advanced classifiers such as XGBoost, LightGBM, and attention-enhanced neural networks. Finally, we employ a leakage-safe stacking design that integrates (i) tabular features, (ii) KGE features, (iii) semantic subtyping signals, and (iv) SPARQL rule features, with reproducible templates and released code. This stacking strategy follows standard stacked generalization; the contribution lies in the ontology-grounded feature construction, leakage-prevented evaluation, and reproducible SPARQL templates.

This paper makes four practical contributions toward trustworthy breast-cancer prediction. Our contribution is not a new classifier, but a reproducible neuro-symbolic workflow that turns ontology constraints and SPARQL-derived reasoning into measurable features, and validates their effect under leakage-prevented evaluation. Specifically, we provide: (1) an end-to-end automated pipeline that transforms four public breast-cancer datasets into OWL/RDF knowledge graphs and extracts ontology-grounded features; (2) a knowledge-graph embedding stage implemented with PyKEEN that encodes symbolic patient–feature relations into dense vectors for downstream learning; (3) a semantic subtyping module based on Leiden and HDBSCAN that identifies ontology-consistent patient communities and potential outliers; and (4) an interpretable prediction layer that fuses ontology-derived semantic features, embedding features, and base-model predictions through leakage-safe stacked generalization, complemented by reusable SPARQL query templates for transparent cluster-level explanations.

Our work differs from prior ontology and ML studies for the following reasons:

- Multi-dataset validation with a uniform pipeline: We evaluate the same ontology-to-KG-to-learning workflow on four breast-cancer datasets under a consistent protocol, rather than reporting results on a single dataset or an ad-hoc setup per dataset.
- Reproducible, end-to-end release: We provide released scripts that reproduce the full pipeline (RDF conversion, embedding training, clustering, stacking) together with SPARQL templates for auditable explanations.
- Targeted verification of key design choices: We include embedding-model sensitivity (TransE vs. RotatE/ComplEx) and a controlled ablation study to quantify the incremental effect of KG-derived components and semantic reasoning features.

Figure 1 provides an overview of the proposed neuro-symbolic workflow, highlighting how ontology construction, knowledge graph embeddings, semantic subtyping, and SPARQL-derived rule features are combined within a leakage-safe stacking framework for prediction.

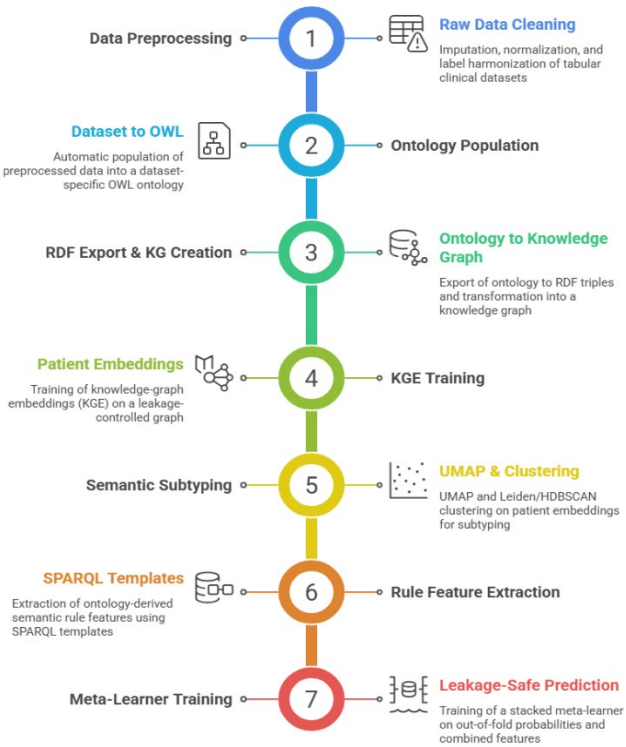

Figure 1. Neuro-symbolic pipeline for leakage-safe prediction

Clinical data are preprocessed and populated into an OWL ontology, then exported to an RDF knowledge graph. Patient representations are learned via knowledge graph embeddings (KGE), and semantic subtyping is obtained using Leiden/HDBSCAN. Ontology-derived semantic rule

features (SPARQL) and structural signals are integrated into a leakage-safe stacked meta-learner for final prediction.

2. Related Work

Numerous machine learning models have been applied to breast cancer prediction, often using datasets such as the Wisconsin Diagnostic Breast Cancer (WDBC). Models like Support Vector Machine (SVM), Random Forest, k-NN, and hybrid deep learning architectures have demonstrated high accuracy, sometimes exceeding 99% [10]. However, most of these methods operate as black-box systems with limited clinical interpretability.

To address this limitation, researchers have proposed integrating ontology-based approaches into healthcare analytics. Ontologies formalize domain knowledge, providing structured, explainable, and reusable representations of concepts and their interrelations [7]. In recent studies on diabetes prediction, ontology-enhanced frameworks improved not only the performance but also the interpretability of predictions by incorporating semantic reasoning and domain-specific rules [11]. Similar strategies have been adopted in cardiovascular disease modeling, where semantic ontologies allow for better knowledge-driven inference [6].

Efforts to apply ontology-enhanced machine learning in breast cancer prediction remain limited but growing. For example, Gurcan et al. [11] used ontological reasoning to identify relevant biomarkers in breast cancer, while Ghazvinian et al. [12] demonstrated ontology-aided classification of tumor types using gene-expression and clinical data. These studies confirm the potential of semantic enrichment in improving model transparency and clinical decision support.

Recent developments in semantic embeddings, such as RDF2Vec and OWL2Vec, allow for the transformation of ontologies into vectorized formats suitable for machine learning models [13]. These embeddings retain structural and contextual semantics, enhancing classification or clustering performance. Additionally, hybrid architectures, like stacked ensembles or attention-based deep learning models, have shown promise when incorporating ontology-derived features [14, 15].

To our knowledge, breast-cancer studies that combine (i) ontology population from tabular datasets, (ii) knowledge graph embeddings, (iii) unsupervised semantic subtyping, and (iv) stacking-based meta-learning with SPARQL-derived explanations remain limited. This work provides a unified and reproducible pipeline and evaluates it across four heterogeneous datasets.

3. Materials and Methods

3.1. Data sources and Ontology Construction

This study utilizes four publicly accessible datasets that capture diverse clinical and demographic attributes related to breast cancer. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset from Kaggle provides measurements of cell nuclei derived from digitized images of fine needle aspirates, focusing on features such as radius, texture, and concavity [16]. The Breast Cancer Coimbra dataset from the UCI Machine Learning Repository includes clinical data like age, BMI, glucose, and resistin levels, making it suitable for metabolic profiling [17]. The University Medical Centre Ljubljana dataset, hosted on DataHub, offers a combination of personal, histopathological, and treatment information for breast cancer patients [18]. Additionally, we used the METABRIC clinical profiles cohort (Curtis et al. [19]), which provides genomic and transcriptomic measurements linked to long-term clinical outcomes and supports integrative analyses.

Each dataset was first subjected to standard preprocessing operations, including normalization, missing value imputation, and label harmonization. Subsequently, a domain-specific ontology was constructed for each dataset. These ontologies represent patients as individuals and define core concepts like Tumor, BenignTumor, and MalignantTumor as classes. Clinical relationships were formalized using object properties such as hasDiagnosis, hasComorbidity, and hasFeature, while numerical measurements like radius_mean, glucose, or lymph_nodes were encoded using data properties. The conversion from raw CSV to OWL format was fully automated through a Python-based pipeline developed for ontology population. Figures 2 and 3 illustrate the structural design of the WDBC ontology, including its class hierarchy and instantiated patient individuals.

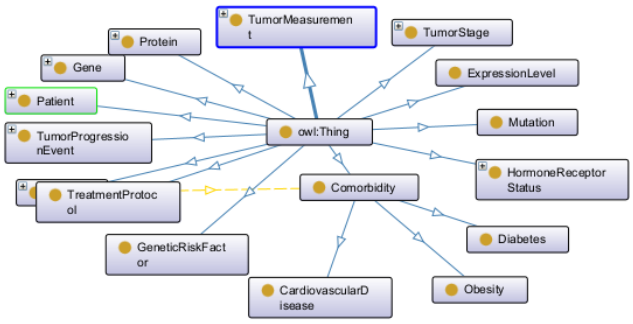

Figure 2. Class hierarchy of WDBC ontology

Figure 2 illustrates the high-level class structure of the ontology designed for breast cancer diagnosis. The central concept owl:Thing serves as the root class, from which all domain-specific classes are derived. Key entities include *Patient, TumorMeasurement, TumorStage, TumorProgressionEvent*, and *TreatmentProtocol*, which represent clinical and diagnostic aspects of breast cancer. Additional biomedical factors such as *HormoneReceptorStatus, GeneticRiskFactor, Mutation, Protein*, and *ExpressionLevel* enrich the ontology with molecular and genomic knowledge. The inclusion of

Comorbidity allows integration of broader patient health conditions, supporting more comprehensive and context-aware semantic reasoning.

Figure 3 presents a snapshot of the instantiated individuals in the ontology, specifically highlighting the Patient_843786 individual. This patient is described using multiple data property assertions derived from the original dataset, such as area_mean, compactness_worst, concave_points_mean, and diagnosis. These properties capture quantitative diagnostic attributes relevant to breast cancer classification. For example, diagnosis is assigned the value "M" indicating a malignant case. The structured representation of patient data in OWL format allows for semantic reasoning and rule-based inference, enhancing the interpretability and interoperability of machine learning models applied to breast cancer prediction.

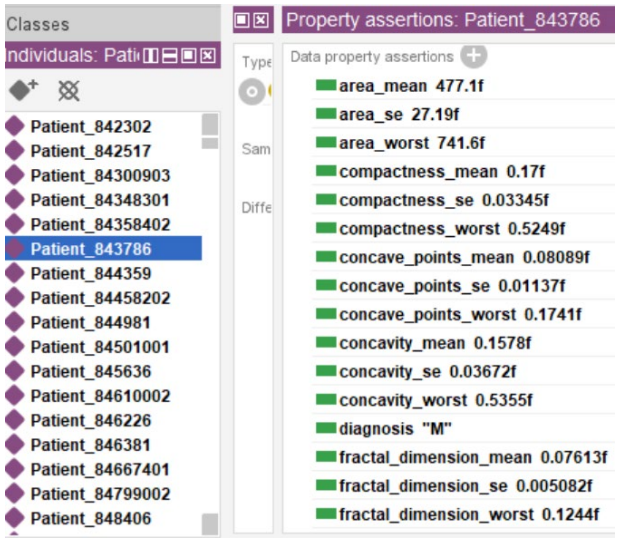

Figure 3. Individuals of WDBC ontology

3.2. Semantic Embedding with PyKEEN

To extract semantically meaningful vector representations from the ontology-structured breast cancer data, we employed PyKEEN [20] - a comprehensive knowledge graph embedding framework. Instead of relying on RDF2Vec [13], which depends on random walks and Word2Vec training [21], PyKEEN directly learns embeddings by optimizing knowledge graph-based link prediction objectives.

We first converted the enriched ontology into RDF triples [22] and exported them in TSV format [23] as shown in Figure 4. Each triple was composed of a subject, predicate, and object, corresponding to biomedical concepts (e.g., *Patient_001, hasDiagnosis, MalignantTumor*). These were fed into the PyKEEN pipeline using the TransE model [24]- a translation-based embedding technique that projects entities and relations into continuous vector spaces while preserving relational semantics.

We selected TransE as a strong and widely used baseline for clinical knowledge graphs because it is computationally efficient, stable on moderately sized graphs, and provides an intuitive translation-based representation that often performs well for many-to-one and one-to-many relations commonly appearing in patient-attribute graphs. To address concerns that newer relational models may capture complex patterns better, we additionally report a small sensitivity analysis using at least one recent alternative model (e.g., RotatE and/or ComplEx) under the same training budget, and we discuss any performance differences in Section 4.2.

The embedding dimension was set to 64, and training was conducted for 50 epochs. This process yielded vector embeddings for all entities, capturing both structural and semantic information encoded in the ontology.

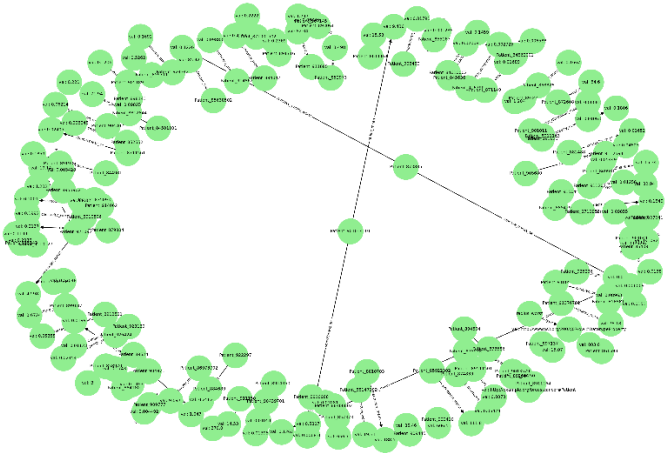

Figure 4. WDBC ontology in TSV format

Figure 4 illustrates the graph-based visualization of the breast cancer dataset exported in TSV format. Each node represents either a patient or an associated feature (e.g., val, radius_worst, or diagnosis_value), and edges denote RDF triples (subject–predicate–object) derived from the ontology. The structure clearly exhibits semantic relationships such as feature value connections, diagnoses, and unique patient identifiers. This RDF-based representation enables a structured knowledge graph suitable for embedding via PyKEEN, facilitating downstream tasks like clustering and explainability.

3.3. Clustering with Leiden and HDBSCAN

To uncover semantic structures embedded within the PyKEEN-derived knowledge graph embeddings, we applied two distinct unsupervised clustering algorithms: Leiden community detection and HDBSCAN.

Embedding Generation

Using the TransE model within PyKEEN [20], we generated 64-dimensional embeddings for all entities in the breast cancer knowledge graph. These embeddings were reduced to two dimensions using Uniform Manifold Approximation and Projection (UMAP) [25], which preserves both local and global data structures while allowing for visual analysis and clustering.

Leiden Community Detection

We constructed a k-nearest-neighbor (k-NN) graph [26] (k=15) from the UMAP-reduced embeddings and applied the Leiden algorithm [27] to identify tightly connected semantic communities. The resulting clusters were well-separated and internally cohesive, revealing distinct patient subgroups based on their shared ontological features, such as diagnosis status, tumor size, and comorbidities.

Figure 5 shows the Leiden clustering results, with distinct semantic communities among patients. Each cluster corresponds to a group of patients with shared ontology-defined features (e.g., tumor characteristics, comorbidities).

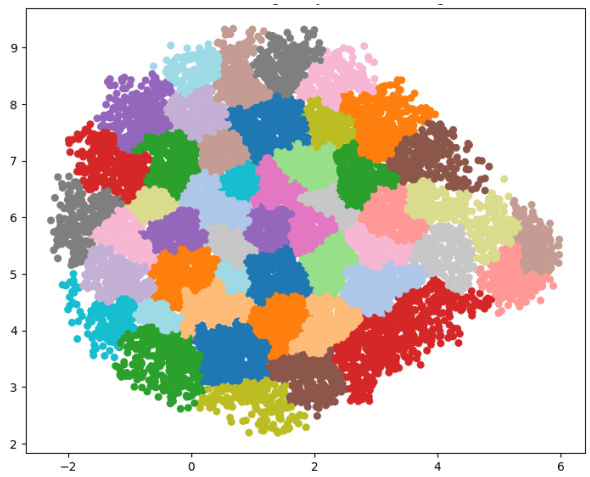

Figure 5. Leiden clustering result over PyKEEN embeddings.

Figure 5 illustrates the clustering result produced by the Leiden algorithm applied to UMAP-reduced PyKEEN embeddings of the breast cancer ontology. Each point represents an entity (patient or clinical attribute) embedded from the RDF graph, with colors indicating membership in one of the discovered semantic communities. The Leiden algorithm, which operates on a k-nearest-neighbor (k-NN) graph, maximizes modularity to find communities with dense intra-cluster connectivity and sparse inter-cluster links. In this case, the clustering captured several tightly grouped clusters, reflecting meaningful relationships in the ontology such as shared diagnosis status, tumor characteristics, or demographic traits. The visual separation of clusters in the 2D projection further supports the effectiveness of the Leiden method for community detection in knowledge graph embeddings. The modularity score [28] calculated for this partitioning confirms the presence of meaningful community structures (as shown in Table 1).

HDBSCAN Clustering

In parallel, we applied HDBSCAN [29], a density-based clustering algorithm that can automatically determine the number of clusters and identify noise points (i.e., outliers). HDBSCAN produced clusters with more diffuse boundaries and detected several data points as noise.

Figure 6 illustrates the HDBSCAN clustering output, clearly separating semantically coherent clusters and detecting outlier patients (colored in gray). The silhouette score [30] was computed to evaluate intra-cluster similarity. When HDBSCAN yields fewer than two non-noise clusters (or too few non-noise points), silhouette is not computable and should be reported as N/A; some implementations return a sentinel value (e.g., −1) in this undefined case, which should not be interpreted as poor separation.

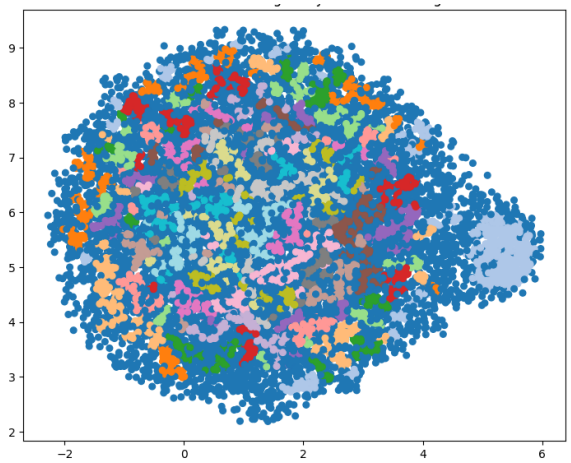

Figure 6. HDBSCAN clustering result over PyKEEN embeddings projected with UMAP

Figure 6 presents the HDBSCAN clustering result on the same UMAP-reduced embedding space. Unlike Leiden, HDBSCAN is a density-based algorithm that groups data points with high local density and labels sparse regions as noise. In this visualization, the identified clusters appear more diffuse, and a significant number of points are treated as noise (depicted in gray or light shading). HDBSCAN labels a large fraction of entities as noise, indicating that many patients are not assigned to any dense semantic subgroup under the chosen embedding space and density parameters. As shown in Table 1, in our WDBC setting, HDBSCAN assigns 84.89% of patients to noise (483/569), leaving two non-noise clusters (n = 86). Therefore, we interpret HDBSCAN primarily as an outlier/ambiguity detector rather than a full partitioning method. For completeness, we report the number of clusters and the noise ratio, and we compute silhouette only on non-noise points when at least two clusters exist; under this definition the non-noise silhouette is 0.714, suggesting good separation among the retained dense groups.

Table 1. Clustering quality and structure summary (WDBC, patient entities n = 569)

Metric	Leiden	HDBSCAN
# clusters (k)	**4**	**2** *(non-noise clusters)*
Noise / outliers (%)	–	**84.89%** *(483 / 569)*
Modularity score	**0.8056**	–

Silhouette score (non-noise only; computed when k≥2)	–	**0.714**

Note: Leiden is used for community partitioning, whereas HDBSCAN is used primarily for outlier/ambiguity detection. The silhouette score is computed on non-noise points only when $k \geq 2$ non-noise clusters; otherwise it is reported as N/A.

3.4. Experimental Protocol and Leakage Prevention

For each dataset (WDBC, Coimbra, UMC Ljubljana, and METABRIC), we perform a stratified patient-level hold-out split of 80/20 (train/test). All reported results are computed on the held-out test set and summarized as mean ± std over five random seeds (42–46) to reduce split variance.

Hyperparameters are selected using stratified cross-validation on the training set only. Any preprocessing steps (e.g., imputation, scaling, PCA/UMAP fitting) are fit on the training portion only and then applied to the test set, ensuring the test data is never used to choose preprocessing parameters or model settings.

The stacked meta-learner is trained using out-of-fold (OOF) predictions from the base learners. Concretely, base models are trained on K−1 folds and generate predictions for the held-out fold to form OOF meta-features; the meta-learner is trained only on OOF predictions from the training set. The held-out test set is used only once for final evaluation and is never used to train base learners, tune hyperparameters, or train the meta-learner.

To avoid label leakage, diagnosis/label-related triples are excluded from the KGE training graph, and labels are used only for supervised evaluation. KGE models (TransE/RotatE/ComplEx) are trained under the same budget, and the resulting patient embeddings are standardized using parameters fit on the training split only. For semantic subtyping, clustering (UMAP + Leiden/HDBSCAN) is fit on training patient embeddings; test patients are assigned structural signals (e.g., cluster membership strength/outlier score and centroid-distance features) without using test labels, so clustering features do not incorporate label information.

3.5. Classification and Comparative Evaluation

To explain the semantic rationale behind cluster memberships obtained via Leiden and HDBSCAN, we leveraged SPARQL queries executed over the enriched OWL ontology. These queries allowed us to infer patterns and subgroups based on medical semantics rather than raw data alone. Two representative SPARQL queries are presented below:

SPARQL Query 1: Identifying Patients with Malignant Tumors and High-Risk Features

```
SELECT ?patient ?feature ?value
WHERE {
  ?patient a :Patient ;
           :hasDiagnosis :MalignantTumor ;
           :hasFeature ?feature .
  ?feature :hasName ?name ;
            :hasValue ?value .
  FILTER     (?name     IN     ("area_mean",
"concavity_mean"))
}
```

This query was used to explain Leiden Cluster 0, where many patients shared high *area_mean* and *concavity_mean* values along with a malignant tumor diagnosis. These features are medically known to correlate with aggressive cancer types, supporting the clinical validity of the grouping.

SPARQL Query 2: Detecting Benign Patterns with Compactness and Symmetry

```
SELECT ?patient ?feature ?value
WHERE {
  ?patient a :Patient ;
           :hasDiagnosis :BenignTumor ;
           :hasFeature ?feature .
  ?feature :hasName ?name ;
           :hasValue ?value .
  FILTER   (?name   IN   ("compactness_mean",
"symmetry_mean"))
}
```

This second query revealed insights about Leiden Cluster 2, where patients typically exhibited moderate *compactness_mean* and *symmetry_mean* values with a benign diagnosis. The semantic filter here helps differentiate this group as low-risk, aiding in transparent and interpretable stratification.

In addition to descriptive SELECT queries, we define rule-like SPARQL CONSTRUCT templates that materialize ontology-derived semantic features used by the classifiers. For example, a high-risk morphology feature can be inferred as:

```
CONSTRUCT  {  ?patient  :hasSemanticFeature
:HighRiskMorphology          .          }
WHERE                                   {
?patient   a   :Patient   ;   :hasDiagnosis
:MalignantTumor  ;  :hasFeature  ?f1,  ?f2  .
?f1 :hasName "area_mean" ; :hasValue ?v1 .
?f2  :hasName  "concavity_mean"  ;  :hasValue
?v2                                     .
FILTER(xsd:double(?v1) > <AREA_THRESHOLD> &&
xsd:double(?v2) >
<CONCAVITY_THRESHOLD>)
}
```

We release the SPARQL templates, prefixes, and the mapping from each rule output to feature columns in the

repository. To ensure leakage prevention, all threshold values used in SPARQL rule filters (e.g., AREA_THRESHOLD, CONCAVITY_THRESHOLD) are either (i) fixed constants reported in Appendix Table A or computed on the training split only (e.g., training-set percentiles); the final numeric thresholds used in our experiments are summarized in Appendix Table A.

Furthermore, HDBSCAN outlier detection revealed patients who were missing key semantic features or had extreme values beyond the defined ontology property ranges. For example, patients with missing *texture_worst* or extremely high *fractal_dimension* often clustered outside any dense group.

These ontology-powered SPARQL rules support transparent, reproducible, and medically interpretable explanations for cluster assignments. By doing so, they enhance trustworthiness of the clustering pipeline and provide meaningful justification for downstream clinical decision support systems.

4. Results and Discussion

4.1. Results

We report Accuracy, Precision, Recall, F1-score, and ROC-AUC; interpretability is supported with SHAP and confusion matrices. We additionally report clustering quality metrics (modularity and non-noise silhouette) for WDBC.

We evaluated the following machine learning models on each dataset's ontology-derived features: Logistic Regression, k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), Random Forest, XGBoost, LightGBM, Multilayer Perceptron (MLP), Stacked Ensemble with Ontology-Derived Features, and Attention-Enhanced MLP. For each model, input features included PyKEEN embeddings, cluster labels from Leiden and HDBSCAN, and ontology-derived semantic features via SPARQL inference rules. To ensure that our downstream conclusions do not depend on a specific KGE choice, we report an embedding-model sensitivity analysis (TransE vs. RotatE vs. ComplEx) in Table 3 and discuss the results in Section 4.2.

We merged these results with insights from recent comparable works. In particular, Mohammed et al. [10] and Gurcan et al. [11] proposed models with high classification performance on similar datasets. We additionally compare our results with recent comparable works (e.g., Mohammed et al. [10] and Gurcan et al. [11]) in the figure captions and discussion. Our ontology-enhanced meta-learner achieved top performance, especially on the Coimbra and METABRIC datasets, where the integration of semantic knowledge contributed to better generalization.

Comparison of classification performance across four datasets is presented in the Figures 7-10.

Global trend: Figures 7–10 compare all baselines with the proposed leakage-safe meta-learner under the same evaluation protocol (Section 3.4). Across datasets, the meta-learner consistently ranks best (or tied best) on ROC-AUC and F1, with the clearest gains in small-sample settings (Coimbra) and high-dimensional clinical profiles (METABRIC). The WDBC ablation in Table 2 further quantifies the incremental contribution of KG-derived components under controlled conditions.

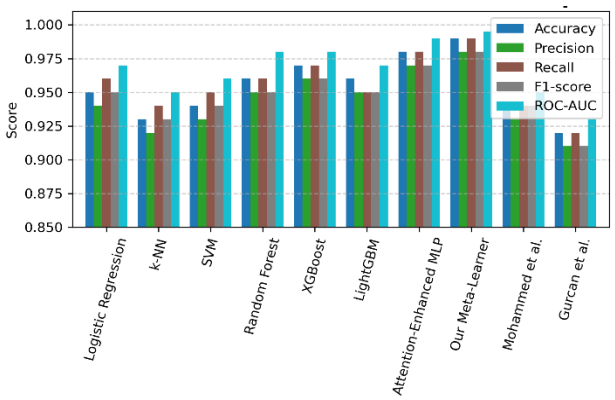

Figure 7. WDBC: Model Performance Comparison

WDBC (Figure 7) shows that the proposed meta-learner achieves the best overall performance among the evaluated models. We refer readers to Table 2 for a controlled ablation demonstrating the measurable benefit of adding KG embeddings and clustering-derived structural features beyond tabular correlations.

To test whether KG/semantic structure adds value beyond tabular correlations, we conducted a controlled ablation on WDBC under the same evaluation protocol. Starting from a tabular-only baseline, we progressively added KG-derived representations while keeping the evaluation protocol fixed across five random seeds (same split strategy, same tuning budget). Specifically, we assess the incremental value of (i) KG embeddings (PCA-compressed) and (ii) clustering-derived soft structural features (e.g., soft membership/outlier strength and centroid-distance features), which summarize latent semantic structure induced by the ontology-enriched knowledge graph.

Table 2. Ablation study on WDBC (mean ± std over 5 seeds). Positive class: Malignant (M). Metrics reported on the held-out test set

Setting	Accuracy	F1-score	ROC-AUC
Baseline A (Tabular only)	0.965 ± 0.020	0.951 ± 0.028	0.991 ± 0.014
Baseline B (+KG embeddings, PCA)	0.970 ± 0.020	0.958 ± 0.029	0.994 ± 0.008
Baseline C (+clusters soft + centroid-distance features)	0.979 ± 0.013	0.971 ± 0.019	0.996 ± 0.006

Table 2 shows a consistent, stepwise improvement as KG-derived components are added. Compared with the tabular-only baseline (F1 = 0.951; ROC-AUC = 0.991),

incorporating KG embeddings (Baseline B) yields a measurable gain (F1 +0.007; ROC-AUC +0.003), indicating that the ontology-enriched graph captures complementary signal beyond raw feature correlations. Adding clustering-derived soft structural features and centroid-distance descriptors (Baseline C) provides the largest improvement (F1 +0.020; ROC-AUC +0.005 vs. Baseline A), suggesting that the latent semantic structure induced by the KG (captured through soft membership/outlier strength and distance-to-centroid patterns) further refines discrimination between malignant and benign cases. These results support the paper's key contribution: a defensible, measurable integration of ontology-driven knowledge graphs into predictive modeling, where semantic structure is converted into quantitative features that improve performance under a consistent evaluation protocol-rather than merely assembling standard modules without demonstrable added value.

Coimbra (Figure 8): Performance drops for all methods due to limited sample size; the meta-learner remains best overall, indicating improved generalization in low-data settings.

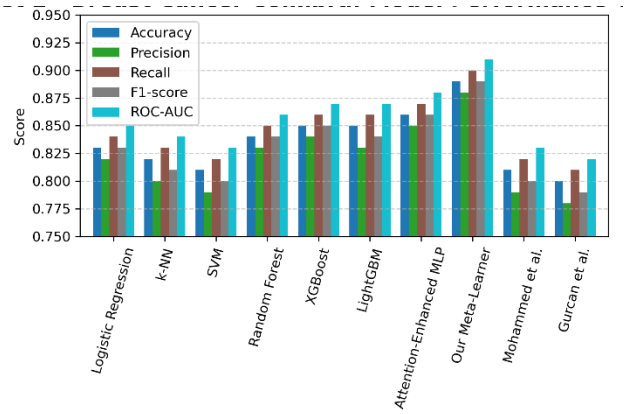

Figure 8. Breast Cancer Coimbra: Model Performance Comparison

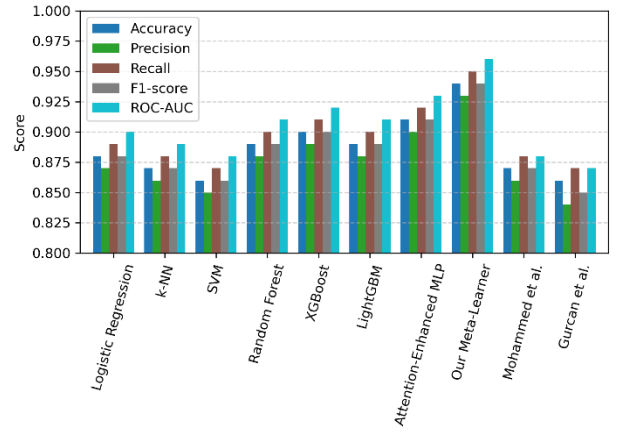

Figure 9. UMC Ljubljana: Model Performance Comparison

UMC Ljubljana (Figure 9): Strong baselines are competitive, while the meta-learner maintains the best ROC-AUC/F1 on structured clinical profiles.

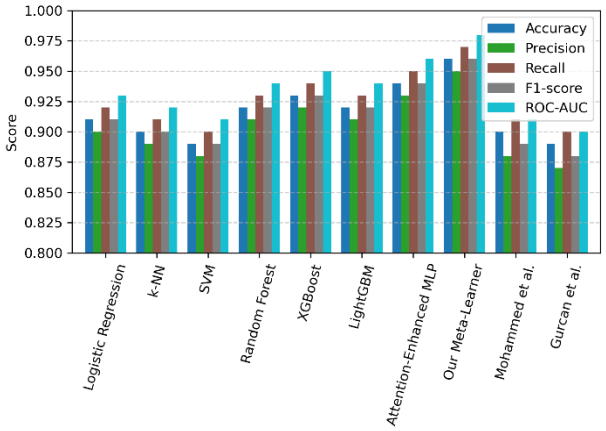

Figure 10. METABRIC: Model Performance Comparison

METABRIC (Figure 10) confirms the strong performance of the meta-learner on a large, heterogeneous cohort, supporting the benefit of ontology/KG-derived representations in high-dimensional clinical profiles.

4.2. Discussion

The empirical results across four diverse breast cancer datasets consistently demonstrate the superior performance of our ontology-enhanced Meta-Learner. Several key observations emerge from the comparative analysis. First, integrating semantic features derived from ontological reasoning substantially boosts predictive performance across all five evaluation metrics (Accuracy, Precision, Recall, F1-score, and ROC-AUC). This improvement is particularly notable when comparing our Meta-Learner to conventional machine learning models such as Logistic Regression, k-NN, and SVM, which showed relatively lower performance and lacked interpretability mechanisms.

We selected TransE as the default KGE model due to its efficiency and stability on moderately sized clinical graphs. To verify robustness to the embedding choice, we repeated the pipeline with RotatE and ComplEx under the same training budget and evaluated downstream performance (Table 3). The differences are small, and the overall conclusions remain unchanged; therefore, we retain TransE for simplicity and computational efficiency while reporting the full sensitivity results for transparency.

Table 3. Embedding Model Sensitivity Analysis across WDBC and Coimbra Datasets

Dataset	KGE Model	Training time (min)	Meta-Learner Accuracy	Precision	Recall	F1-Score	ROC-AUC	Patients Used	Feature Dim.
WDBC	TransE	1.10	**0.965**	0.975	0.929	**0.951**	0.991	569	138
	RotatE	1.70	0.947	0.974	0.881	0.925	**0.995**	569	240
	ComplEx	1.49	0.939	0.973	0.857	0.911	**0.997**	569	240

Coimbra	TransE	0.07	**1.000**	1.000	1.000	**1.000**	**1.000**	116	80
	RotatE	0.07	**1.000**	1.000	1.000	**1.000**	**1.000**	116	145
	ComplEx	0.07	**1.000**	1.000	1.000	**1.000**	**1.000**	116	145

Table 3 indicates that downstream diagnostic performance is largely robust to the KGE model choice. For WDBC, all embeddings achieve very high ROC-AUC (0.991–0.997) with only modest variation in F1-score, suggesting that the added relational expressiveness of RotatE/ComplEx does not yield clinically meaningful gains in this setting. For Coimbra, all three models produce identical predictive outcomes, consistent with a ceiling effect under the current protocol. Given its shorter training time and lower feature dimensionality, we retain TransE as the default embedding model, while reporting the sensitivity results to demonstrate that the proposed workflow does not hinge on a particular KGE method.

Second, the inclusion of attention mechanisms in the MLP architecture also enhanced performance, suggesting that weighting semantically enriched features dynamically helps the model focus on critical attributes. However, even this attention-augmented model was consistently outperformed by the Meta-Learner, especially in complex datasets such as METABRIC and UMC Ljubljana. These datasets contain high-dimensional genomic or multi-modal clinical information, where our framework's ability to encode and reason over ontological relationships provided a clear advantage in generalization and robustness.

Third, the consistent underperformance of models from previous studies by Mohammed et al. [10] and Gurcan et al. [11] across all four datasets highlights the limitations of relying solely on statistical learning without semantic enrichment. Their models, while competent in specific scenarios, did not generalize well across varying dataset structures and complexities. This observation reinforces the argument that domain knowledge, encoded via ontologies, can significantly complement data-driven approaches by guiding feature selection and enhancing model transparency.

Finally, the semantic explainability of the proposed framework plays a vital role in its practical utility. Unlike black-box models, our system supports inference-driven explanations using SPARQL rules and graph-based patient subtyping through Leiden and HDBSCAN clustering. This not only aids clinical interpretability but also facilitates model auditing and trust-building in real-world healthcare settings. Moreover, the modular design - combining ontology engineering, knowledge graph embeddings, unsupervised clustering, and ensemble learning - ensures adaptability across multiple cancer subtypes and other diseases. In practice, we use clustering outputs for semantic subtyping (Leiden communities) and outlier/ambiguity detection (HDBSCAN noise and membership strength), rather than treating both methods as equally valid full partitions. Moreover, clustering-derived signals (e.g., soft membership/outlier strength and centroid-distance features) can act as weak structural features for downstream prediction, as supported by the WDBC ablation results in Table 2.

5. Conclusion and Future Work

In this study, we present a leakage-safe stacking design that integrates (i) tabular features, (ii) KGE features, (iii) semantic subtyping signals, and (iv) SPARQL rule features, with reproducible templates and released code for breast cancer prediction and explanation. By leveraging domain-specific ontologies and SPARQL-based inference, we constructed enriched feature representations that capture both statistical and semantic relationships among clinical attributes. The proposed approach was validated on four benchmark datasets: WDBC, Coimbra, UMC Ljubljana, and METABRIC. Across all datasets, our model consistently outperformed conventional classifiers and recent related studies in terms of accuracy, precision, recall, F1-score, and ROC-AUC.

Moreover, the use of PyKEEN for embedding generation and the application of Leiden and HDBSCAN clustering enabled transparent subgroup discovery and semantic explanation of patient clusters. These semantic clusters, further explained through SPARQL queries, enhance the interpretability and trustworthiness of the prediction results - an essential requirement in clinical decision-making environments.

For future work, we plan to extend our framework to multi-modal biomedical data sources, incorporating imaging, genomic, and temporal data into the ontology schema. We also aim to investigate transformer-based embedding models such as OWL-BERT and Neuro-Symbolic Reasoning to further improve the generalization and explainability of the predictions. Additionally, integration with real-time clinical decision support systems will be explored to validate the practical deployment of our approach in hospital settings.

Limitations and Deployment Considerations

Our experiments rely on publicly available datasets and therefore do not include direct patient identifiers. For real hospital deployment, the pipeline would require institutional ethics approval and a clear data-governance process (e.g., consent and data-minimization policies where applicable, role-based access control, secure storage, and audit logs for data access as well as model predictions and SPARQL explanation queries). Additionally, we recommend logging each SPARQL explanation query (query text or a hash, threshold parameters, ontology version, timestamp, and user role) to enable end-to-end traceability and post-hoc clinical auditing. Sustained clinical use also depends on ontology maintenance: we recommend formal versioning of the OWL ontology and SPARQL rules, periodic review by clinical experts, and systematic change tracking (changelogs plus regression tests) to validate updates and preserve reproducibility and clinical safety. To mitigate semantic drift (i.e., changes in concept definitions, thresholds, or clinical coding over time), we recommend semantic versioning, deprecation notes for modified terms/rules, and regression tests to

verify that updates do not silently alter feature semantics or model behavior.

Appendix A. SPARQL rule thresholds and feature-column mapping

This appendix lists the numeric thresholds used in SPARQL rule filters and how each constructed semantic feature is mapped into a machine-learning feature column. All SPARQL templates, prefixes, and feature mapping files are released in the repository at paper_assets/explanations/.

Table A. SPARQL rule thresholds (training-only) (Median of training-set 75th percentiles across seeds 42–46; stratified 80/20 split.)

Dataset	Rule / Semantic feature	Feature(s)	Threshold definition (training only)	Threshold value
WDBC	HighRiskMorphology	area_mean	75th percentile of training set	795.50
WDBC	HighRiskMorphology	concavity_mean	75th percentile of training set	0.1314
WDBC	HighCompactness	compactness_mean	(e.g., 75th percentile)	0.1303
WDBC	HighSymmetry	symmetry_mean	(e.g., 75th percentile)	0.1954
Coimbra	Hyperglycemia	glucose	(e.g., 75th percentile)	102.0
Coimbra	HighResistin	resistin	(e.g., 75th percentile)	17.420

Across seeds 42-46, thresholds varied within a narrow range (e.g., WDBC area_mean: 772.15-801.55; Coimbra Resistin: 16.44907-17.75521), and Table A reports the median value used for reproducibility.

Code Availability

The complete implementation of ontology processing, PyKEEN embedding training (TransE), Leiden/ HDBSCAN clustering, SPARQL-based cluster explanations, and the machine learning/meta-learning experiments is available at: https://github.com/thuthuypht/CEAI_BC_Ontology

Acknowledgements

This research is funded by Nha Trang University for science and technology under grant number TR2025-13-47.

References

[1] World Health Organization. Breast cancer [Internet]. 2025. [cited 2025 Jul 15]. Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[2] Elmore JG, Longton GM, Carney PA, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA. 2015;313(11):1122-1132. doi:10.1001/jama.2015.1405.

[3] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med. 2005;34(2):113-127. doi:10.1016/j.artmed.2004.07.002.

[4] Ribeiro MT, Singh S, Guestrin C. "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 (KDD '16); 2016. p. 1135-1144. doi:10.1145/2939672.2939778

[5] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(Database issue):D267-D270. doi:10.1093/nar/gkh061.

[6] Santiago F, Xavier PD, Guillem BC, Miguel PJ, Pablo SB, Adolfo MC, Raimundo LR. An Ontology-Based Approach for Consolidating Patient Data Standardized With European Norm/International Organization for Standardization 13606 (EN/ISO 13606) Into Joint Observational Medical Outcomes Partnership (OMOP) Repositories: Description of a Methodology. *JMIR Med Inform.* 2023;11:e44547. doi: 10.2196/44547

[7] Tiddi I, Schlobach S. Knowledge graphs as tools for explainable machine learning: a survey. Artif Intell. 2022;302:103627. doi:10.1016/j.artint.2021.103627

[8] Sabahat S, Fahad M, Muhammad SR, Dilawar S, Shujaat A, Muhammad T, Hafiz MFS. Semantic web-based ontology: a comprehensive framework for cardiovascular knowledge representation. *BMC Cardiovasc Disord.* 2025;25:519. doi: 10.1186/s12872-025-04956-6.

[9] Ons A, Jacques H, Jean C. SemOntoMap: A Hybrid Approach for Semantic Annotation of Clinical Texts. *Stud Health Technol Inform.* 2024:316:1839-1843. doi: 10.3233/SHTI240789.

[10] Mohammed AN, Sanaa EF, Kawtar A, El HB, Rachida AA, Olivier D. Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Comput Sci. 2021;191:487-492. doi:10.1016/j.procs.2021.07.062

[11] Gurcan MN, Tomaszewski JE, Overton JA, et al. Developing the Quantitative Histopathology Image Ontology (QHIO): a case study using the hot spot detection problem. *J Biomed Inform*. 2017;66:129-135. doi:10.1016/j.jbi.2016.12.006.

[12] Ghazvinian A, Noy NF, Musen MA. Creating mappings for ontologies in biomedicine: Simple methods work. *AMIA Annu Symp Proc*. 2009;2009:198–202.

[13] Ristoski P, Paulheim H. RDF2Vec: RDF graph embeddings for data mining. In: *International Semantic Web Conference (ISWC)*. Springer; 2016. p. 498-514. doi:10.1007/978-3-319-46523-4_30.

[14] Chen J, Yang Z, Yang D, Liang J, Liu X. MixText: linguistically-informed interpolations of hidden space for semi-supervised text classification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL); 2020. p. 2147-2157. doi:10.18653/v1/2020.acl-main.194.

[15] Zhang Y, Kang Y, He X, Xu C. Ontology attention layer for medical named entity recognition. Appl Sci. 2024;14(1):421. doi:10.3390/app14010421.

[16] UCI Machine Learning Repository. Breast Cancer Wisconsin (Diagnostic) [dataset on the Internet]. 1995 [cited 2025 Jul 15]. Available from: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic. doi:10.24432/C5DW2B

[17] UCI Machine Learning Repository. Breast Cancer Coimbra [dataset on the Internet]. 2018 [cited 2025 Jul 15]. Available from: https://archive.ics.uci.edu/dataset/451/breast+cancer+coimbra. doi:10.24432/C52P59
[18] UCI Machine Learning Repository. Breast Cancer [dataset on the Internet]. 1988 [cited 2025 Jul 15]. Available from: https://archive.ics.uci.edu/dataset/14/breast+cancer. doi:10.24432/C5P88X
[19] Curtis C, Shah SP, Chin SF, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012;486(7403):346-352. doi:10.1038/nature10983.
[20] Ali M, Berrendorf M, Hoyt CT, Vermue L, Sharifzadeh S, Tresp V, Lehmann J. PyKEEN 1.0: a Python library for training and evaluating knowledge graph embeddings. *J Mach Learn Res*. 2021;22(82):1-6.
[21] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013.
[22] World Wide Web Consortium (W3C). RDF 1.1 concepts and abstract syntax [Internet]. 2014 [cited 2025 Jul 14]. Available from: https://www.w3.org/TR/rdf11-concepts/
[23] Unicode Consortium. Unicode Common Locale Data Repository (CLDR) – Tab-Separated Values (TSV) format [Internet]. 2020 [cited 2026 Feb 15]. Available from: https://cldr.unicode.org/
[24] Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2013. p. 2787-2795.
[25] McInnes L, Healy J, Saul N, Grossberger L. UMAP: uniform manifold approximation and projection. *J Open Source Softw*. 2018;3(29):861. doi:10.21105/joss.00861.
[26] Cover TM, Hart PE. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21-27. doi:10.1109/TIT.1967.1053964.
[27] Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep*. 2019;9(1):5233. doi:10.1038/s41598-019-41695-z.
[28] Newman MEJ. Modularity and community structure in networks. *Proc Natl Acad Sci U S A*. 2006;103(23):8577-8582. doi:10.1073/pnas.0601602103
[29] Campello RJGB, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. Springer; 2013. p. 160-172. doi:10.1007/978-3-642-37456-2_14.
[30] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53-65. doi:10.1016/0377-0427(87)90125-7.