BMC Artificial Intelligence

RESEARCH

Open Access

Inconsistency detection in cancer data classification using explainable-AI

Pouria Mortezaagha[1,2*], Abhisht Makarand Joshi[3] and Arya Rahgozar[1,2]

Abstract

Background Accurate classification of cancer-related text data is essential for early diagnosis and effective treatment. However, conventional classification methods often suffer from confusion in error analysis due to data inconsistencies, semantic misalignment, and unreliable labeling. Manual error analysis is labor-intensive and prone to oversight, which limits the clinical utility of these approaches.

Aim This study aims to develop a robust and explainable framework that automates and justifies error analysis by detecting inconsistencies—including potential mislabeling—in classification outcomes through a dual-perspective algorithmic approach.

Methods We propose a novel dual-perspective framework that integrates unsupervised semantic clustering with supervised classification. Specifically, our approach combines BERT-based BERTopic clustering with SVM classification on Node2Vec embeddings to decouple semantic and structural perspectives. It introduces an Explainable Inconsistency Detection (EID) module to automatically surface and remove inconsistencies between the clustering and classification outputs. Additionally, a collaborative filtering recommender system aligns clusters with ground-truth labels to adaptively refine results, with performance validated through rigorous statistical testing.

Results Experimental evaluations on cancer datasets demonstrated substantial improvements in both classification performance and the clarity of error analysis. The optimized framework improved accuracy from 46% to 91% and increased the F1-score from 50% to 89%. Statistical analysis confirmed that these gains were significant (p-values < 0.05) and directly attributable to the targeted removal of inconsistent instances rather than random data exclusion.

Conclusions The integration of BERTopic, SVM, and the Explainable Inconsistency Detection (EID) framework enhances both performance and interpretability by addressing semantic contradictions and structural anomalies in biomedical data. This semi-automated, explainable pipeline offers actionable insights into underlying data errors, presenting strong potential for integration into clinical decision support systems. Future work will focus on refining the EID for broader generalization and exploring real-time applications in healthcare.

Keywords BERTopic, Explainable inconsistency detection (EID), Support vector machine (SVM), Cancer data classification, Node2Vec, Recommender systems, Explainable artificial intelligence (XAI)

*Correspondence:
Pouria Mortezaagha
pmortezaagha@ohri.ca
[1]Methodological Implementation Research, Ottawa Hospital Research Institute, Ottawa, ON, Canada
[2]School of Engineering Design and Teaching Innovation, University of Ottawa, Ottawa, ON, Canada
[3]Department of Family Medicine, University of Ottawa, Ottawa, ON, Canada

Introduction

Recent advancements in feature engineering have significantly bolstered the performance and interpretability of machine learning models across diverse domains [1]. Semi-automated methods now suggest and implement data augmentations for pandas DataFrames using open data sources [2], while in specialized fields such as neutron activation analysis, techniques like noise addition and feature selection have proven effective in enhancing model performance [3, 4]. In energy systems modeling, a Python framework that employs feature creation, expansion, and selection has yielded improved predictive accuracy [5]. Moreover, for scenarios involving complex, multi-table data, reinforcement learning-based approaches such as AutoFeature have been proposed to efficiently explore feature augmentation options—often outperforming traditional methods in both classification and regression tasks [6]. These developments underscore the critical role of feature engineering not only in boosting accuracy but also in facilitating more explainable models.

One major challenge in classification is the confusion and difficulty in analyzing prediction errors while having a criteria to question and reevaluate the validity of the labels. Manual error analysis is often laborious and error-prone. Despite advances in explainability, existing methods often assume clean, reliable data and fail to account for structural or semantic inconsistencies in real-world datasets—particularly in biomedical contexts. To address this, our work focuses on automating error analysis by leveraging explainability. In other words, the problem with classification lies in the confusion and lack of clarity in error analysis, and our goal is to employ explainable AI techniques to systematically identify and analyze these errors while keeping humans in the loop.

Explainable AI has emerged as an essential component in applications where understanding model decisions is as important as achieving high predictive performance. In domains such as fake news detection and financial forecasting, recent studies have illustrated how multi-modal feature fusion and hybrid deep learning methods can enhance both accuracy and interpretability. For instance, Kumar and Taylor [6] propose a multi-modal approach that combines textual and visual features for fake news detection, while Campagner and Cabitza [7] advocate for hybrid models that leverage automatically detected symbolic features to improve interpretability. Similarly, Vouk et al. [8] introduce an Explainable Feature Construction methodology that utilizes instance-based explanations to narrow down the search space for informative attribute groups, and Carta et al. [9] demonstrate that automatic feature selection can significantly enhance financial forecasting by isolating key features for individual stocks. Recent research has also explored explanations specifically designed for handling inconsistencies within semantic frameworks, emphasizing methods that clarify why certain queries fail or return negative results [10].

The core problem this work addresses is the presence of latent inconsistencies in biomedical datasets that compromise both model performance and transparency. These inconsistencies arise from noisy labels, semantic misalignment, and structural contradictions within data graphs. Existing pipelines rarely expose or resolve these ambiguities, thereby limiting trust and interpretability in high-stakes domains like healthcare.

In machine learning and data science, integrating multiple complementary approaches is critical to overcoming the inherent limitations of any single method. Despite notable advances in feature engineering and explainable AI, real-world datasets—especially in biomedical applications such as cancer data classification—are often marred by inconsistencies including bias, noise, and outliers [11]. An additional layer of complexity is the potential unreliability of the ground-truth labels themselves. In many cases, high-level labels assigned to documents (e.g., identifying a paper as lung cancer-related) may not align with the nuanced content found deeper in the text when we use graph representation at the granular level. This divergence becomes particularly evident when using structural embedding techniques such as Node2Vec, which capture latent patterns and associations that challenge surface-level categorizations. These discrepancies not only impede predictive accuracy but also challenge the transparency of model decisions.

To overcome these challenges, we propose a novel dual-perspective framework that decouples clustering and classification to detect and resolve inconsistencies across both semantic and structural levels in biomedical data. The framework integrates two complementary algorithms—a traditional supervised classifier (Support Vector Machines, SVM) and an unsupervised clustering method based on BERT embeddings—both applied after a Node2Vec transformation [12]. This transformation captures relational patterns within a knowledge graph and surfaces latent inconsistencies in labeling, such as cases where a node labeled as lung cancer is structurally or semantically more aligned with another cancer subtype. Through this post-embedding, node-level feature manipulation, our method systematically identifies, minimizes, and eliminates such inconsistencies, thereby improving classification performance and interpretability [13–17].

Previous studies have demonstrated the efficacy of SVM in cancer classification tasks [18], further underscoring its applicability in biomedical contexts [19]. In our approach, SVM is selected for its robustness in high-dimensional spaces and its versatility through kernel-based transformations [20, 21]. Complementing this, BERT [22] is employed to extract rich semantic embeddings from textual data [23]. However, despite its impressive performance, BERT is known to be susceptible to

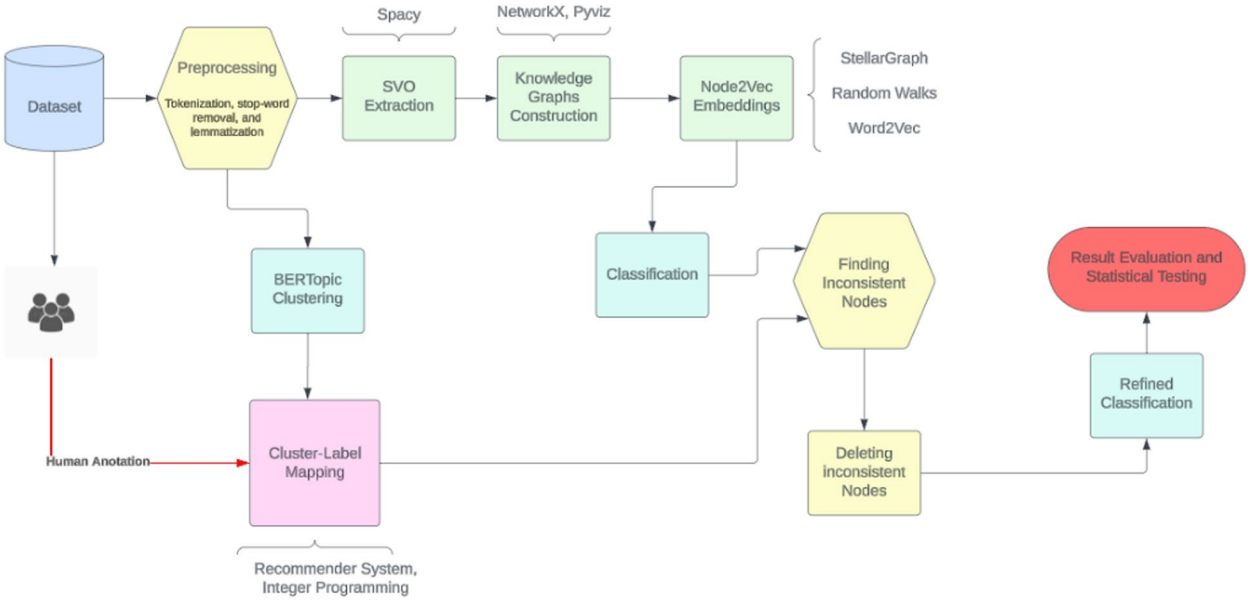

Fig. 1 Overview of the system architecture

overfitting and may require careful hyperparameter tuning during fine-tuning [16, 24]. These embeddings are then clustered using BERTopic [25], which generates coherent topic representations by combining clustering with TF-IDF variations [26]. In parallel, knowledge graphs and Node2Vec embeddings provide valuable structural and relational insights [27, 28].

To better clarify how our method works, we briefly outline the proposed workflow: (1) extract semantic and structural representations from the input corpus using BERT and Node2Vec, (2) apply supervised classification using SVM on Node2Vec features, (3) perform unsupervised clustering on BERT embeddings via BERTopic, (4) identify and analyze divergence between the classifier and clusterer, and (5) resolve inconsistencies using our Explainable Inconsistency Detection (EID) module, followed by cluster-label alignment through an integer-programming-based recommender system.

The dual pipeline is designed such that any divergence between the SVM and BERTopic outcomes highlights "inconsistent" instances. These cases are further analyzed using our novel Explainable Inconsistency Detection (EID), which identifies and eliminates outlier nodes through targeted feature refinement. A recommender system, based on integer programming, maps the resulting clusters to human annotations, ensuring that inconsistencies are addressed in a statistically robust manner. The principal contributions of this work are:

1. Dual-Perspective Pipeline: We propose an integrated framework that combines supervised classification (SVM) with unsupervised clustering (BERTopic) to detect and mitigate inconsistencies in complex datasets.
2. Advanced Feature Engineering: By incorporating recent advances in feature engineering augmentations, our method leverages both textual and graph-based insights to enhance model performance and interpretability.
3. Explainable Inconsistency Detection (EID): We introduce a novel algorithm to identify and remove outlier nodes via node-level feature manipulation, with its efficacy validated through rigorous statistical testing.
4. Explainable Evaluation: Our approach is benchmarked against state-of-the-art methods using extensive experiments, detailed parameter analysis, and statistical validation to demonstrate its superior performance.

Figure 1 provides an overview of the system architecture. The process begins with dataset preprocessing, which includes subject-verb-object (SVO) extraction and the construction of a knowledge graph using tools such as SpaCy[1], StellarGraph[2], NetworkX[3], and Pyviz[4]. Node2Vec embeddings [12] are then generated and fed into the SVM classifier, while BERT-derived semantic embeddings are simultaneously clustered via BERTopic. The resulting clusters are mapped to human annotations using a recommender system, enabling the identification and subsequent removal of inconsistent nodes. This refinement is statistically validated to confirm improvements in classification performance.

[1] https://spacy.io/
[2] https://github.com/stellargraph/stellargraph
[3] https://networkx.github.io/
[4] https://pyviz.org/

The remainder of the paper is organized as follows. " Related work" section reviews the related literature and identifies research gaps. " Methodology" section details the proposed methodology, including the design of the dual-perspective pipeline and the EID. " Experimental results" section describes the experimental setup and presents the results. In " Discussion" section, we discuss the findings, advantages, limitations, and practical implications of our approach. Finally, " Conclusion" section concludes the paper and outlines directions for future research.

Related work

The integration of BERT with various machine learning techniques has been extensively studied to improve classification performance and enhance explainability across diverse domains. Numerous hybrid frameworks have been proposed that combine deep contextual embeddings with traditional classifiers or graph-based methods, yet many of these approaches focus primarily on performance improvements while overlooking the interpretability and the explicit handling of data inconsistencies.

Hybrid deep learning and graph-based approaches

Bikku [29] explored the use of Multi-Layer Perceptron (MLP) models for accurate classification and risk analysis in medical datasets. Their work demonstrated that supervised learning methods could uncover hidden patterns in historical data and provide robust future risk predictions. Similarly, Kassner and Schütze [30] introduced BERT-kNN, a hybrid approach that couples BERT with a k-nearest-neighbor (kNN) search mechanism for open-domain question answering. By enhancing recall for rare facts without additional fine-tuning, this method highlighted the potential of integrating contextual embeddings with simple classifiers.

Efforts to enrich semantic understanding by incorporating external knowledge have also gained traction. For example, Liu et al. [31] proposed K-BERT, which integrates domain-specific knowledge graphs into BERT using soft-position and visible matrix mechanisms. This approach effectively reduces knowledge noise while preserving contextual integrity, resulting in notable improvements in relation extraction accuracy on biomedical datasets. Building on these ideas, Lin et al. [32] developed BertGCN, which fuses BERT's pre-trained embeddings with Graph Convolutional Networks (GCNs) to propagate label information in a graph-based structure. By interpolating the outputs of BERT and GCNs using a weighting parameter, BertGCN achieved significant performance gains (a 13% increase in F1-score on benchmark datasets such as AG News and DBpedia).

More recently, Naseem et al. [16] proposed integrating K-BERT with both kNN and Graph Neural Networks (GNNs) to improve classification and relation extraction. Their method leverages neighboring nodes for semantic enrichment, leading to a 22% improvement in precision for biomedical classification tasks. Despite these advances, existing frameworks generally rely on sequential or parallel integration of models and predominantly target performance metrics, leaving the interpretability of underlying inconsistencies less explored.

Hybrid approaches in cancer classification and omics data processing

In the biomedical domain, several studies have specifically addressed cancer classification using advanced machine learning methods. For instance, Smith and Doe [33] presented an approach for optimized cancer subtype classification and clustering using Cat Swarm Optimization combined with Support Vector Machines (SVM) on multi-omics data. In a similar vein, Lee and Kim [34] proposed an enhanced cancer subclassification method using multi-omics clustering with quantum cat swarm optimization. Additionally, Chen et al. [35] provided a comprehensive review of artificial intelligence approaches in omics data processing, evaluating both progress and challenges.

Complementary studies have focused on the application of machine learning in breast cancer research. For example, Williams and Johnson [36] reviewed data mining and machine learning approaches in breast cancer biomedical research, while Garcia and Robinson [37] discussed the evolving paradigms in cancer classification. Furthermore, Singh and Kumar [38] examined the transition from organ-based to algorithm-based classification in the age of artificial intelligence. Although these studies have achieved significant performance improvements, they often lack mechanisms to explicitly explain or address the inconsistencies that lead to misclassifications.

Research gaps and positioning of our work

While the aforementioned works exemplify the benefits of integrating language models with auxiliary algorithms, they predominantly focus on boosting accuracy and efficiency. They do not, however, provide a systematic approach to detect and mitigate inconsistencies inherent in complex biomedical datasets. Specifically, current methods seldom address the interpretability of errors or offer an explanation for why models fail on certain data points.

In contrast, our proposed framework introduces a novel dual-perspective approach that decouples clustering and classification to explicitly identify and rectify inconsistencies. By combining supervised classification (SVM) with unsupervised clustering

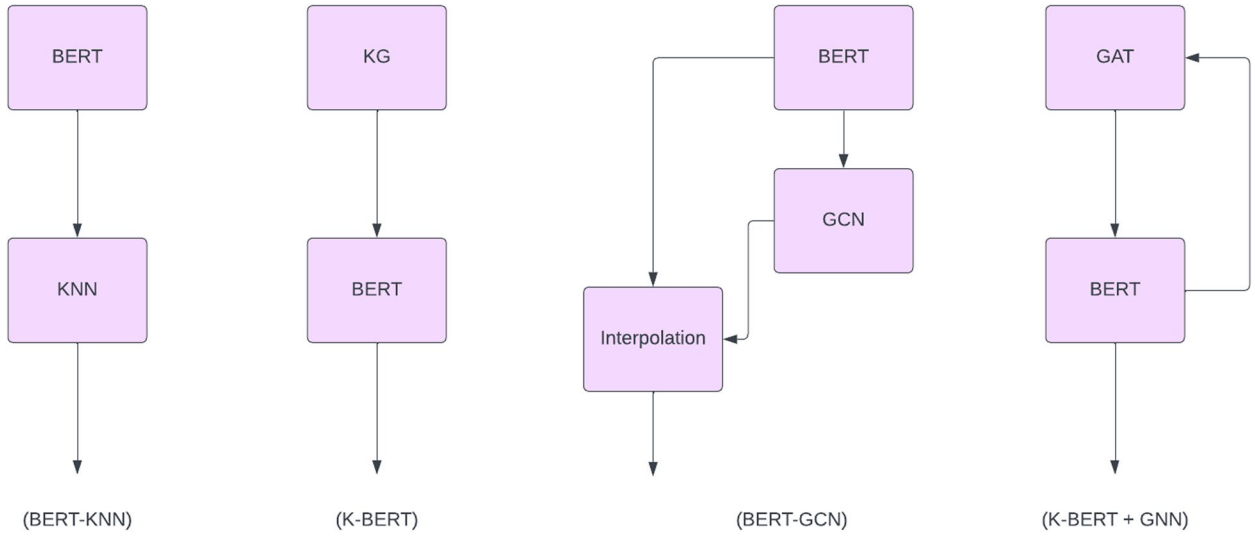

Fig. 2 Comparative architectures of combined BERT and algorithmic approaches. The leftmost block represents BERT-kNN, followed by K-BERT, BertGCN, and a hybrid framework leveraging GAT with K-BERT. In contrast, our proposed framework integrates semantic clustering with traditional classification, emphasizing explicit inconsistency detection and explainability

(BERTopic) after a Node2Vec transformation [12], our method not only achieves competitive performance but also incorporates an Explainable Inconsistency Detection (EID) to pinpoint and remove outlier nodes via node-level feature refinement. This strategy enhances transparency and builds trust, particularly in domains like cancer classification where understanding model decisions is critical. Moreover, while prior methods often assume that the labels are reliable ground truth, we argue that this assumption may not always hold in complex biomedical datasets [39, 40]. Our approach accounts for possible mislabeling by enabling the graph embeddings to surface deeper semantic contradictions. These contradictions—where the embedded representation of a node conflicts with its assigned class—offer a new pathway for inconsistency detection and model introspection.

Figure 2 summarizes key hybrid frameworks that combine BERT with auxiliary methods. While approaches like BERT-kNN, K-BERT, and BertGCN emphasize enhanced feature representation and performance, they do not explicitly address data inconsistencies. Our approach differentiates itself by integrating semantic clustering with a traditional classification pipeline, thereby enabling the systematic identification and removal of inconsistent data points and offering a deeper layer of explainability.

In summary, although significant progress has been made in integrating deep language models with graph-based and optimization techniques for biomedical applications, a clear research gap remains in the explicit detection and explanation of inconsistencies. Our work aims to bridge this gap, providing both high classification accuracy and interpretable insights into the decision-making process.

Methodology

This study was conducted using a mixed methods approach which can be defined as enhancing a classifier using BERTopic insights on a text-based dataset[5]. BERTopic generates clusters based on a large language model (LLM) [25] which has a different insight from a traditional machine learning approach like Support Vector Machines (SVM). This combination suggests a comprehensive and potentially more insightful analysis of text data.

In the rest of this section, we delve into a detailed explanation of the methodology employed with Fig. 3 offering an overall visual representation of the procedural steps. Further insights into these steps are provided later in the section for a more comprehensive understanding. Also, a detailed pseudocode representation of the proposed methodology is provided in Appendix.

Data collection and preprocessing

The study utilized two datasets for classification. The first dataset[6] was curated for biomedical text document classification, comprising abstracts and full research papers whose length exceeds six pages. This dataset includes cancer-related documents classified into three categories: "Thyroid Cancer", "Colon Cancer", and "Lung Cancer". It contains 7,569 publications, with 2,579 samples labeled

[5] Github: https://github.com/pouriamrt/cancer_inconsistency

[6] https://www.kaggle.com/datasets/falgunipatel19/biomedical-text-publication-classification

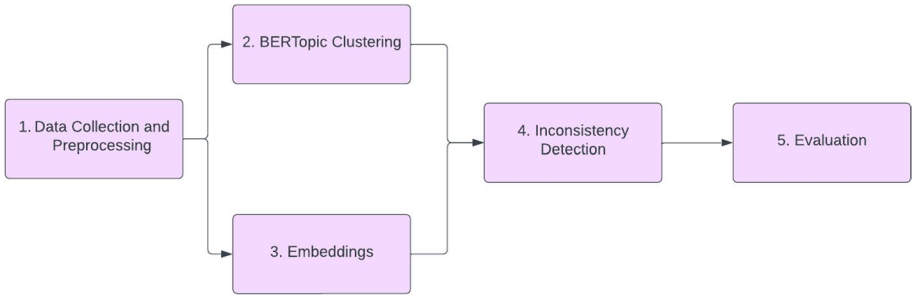

Fig. 3 Overall workflow diagram of the methodology

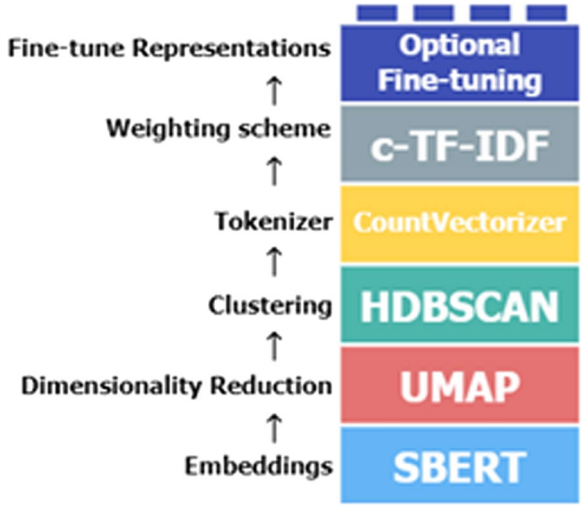

Fig. 4 Steps to create BERTopic topic representations (https://maartengr.github.io/BERTopic/algorithm/algorithm.html)

as "Colon Cancer", 2,180 as "Lung Cancer", and 2,810 as "Thyroid Cancer". The second dataset[7], curated specifically for this study, comprises the titles and abstracts of cancer-related papers published in 2024 and 2025. These papers are categorized into three distinct groups: "Colon Cancer", "Lung Cancer", and "Thyroid Cancer", with each category containing 300 entries.

The preprocessing phase of this study was particularly rigorous, involving several key steps to ensure standardized text representations. Initially, the datasets underwent tokenization, a process of breaking down text into smaller units or tokens, facilitating easier analysis and processing. Following this, stop-word removal was implemented, a crucial step that involved eliminating commonly used words that offer little to no value in terms of context or meaning, thereby streamlining the datasets. Additionally, very obvious keywords related to decision-making, such as "colon", "thyroid", and "lung cancer", were removed to make the classification task more challenging for the model. Lastly, lemmatization was applied, reducing words to their base or dictionary form to consolidate different forms of a word into a single standard form. Each of these preprocessing steps played a vital role in preparing the datasets for effective classification, ensuring a higher level of accuracy and reliability in the study's findings.

[7] https://www.kaggle.com/datasets/pouria1206/cancer-papers-dataset

BERTopic clustering

BERTopic was used to cluster the data based on its own approach which is depicted in Fig. 4. Documents are embedded into a vector space using the Sentence-BERT (SBERT) framework. This technique transforms sentences and paragraphs into dense vector representations using pre-trained language models fine-tuned for semantic similarity [25].

For document clustering, Grootendorst [25] addresses the challenge of high-dimensional data space where traditional distance measures become less effective. To overcome this, Grootendorst [25] reduces the dimensionality of embeddings using UMAP (Uniform Manifold Approximation and Projection) which preserves both local and global features in lower dimensions and is adaptable to various language model dimensions [41]. The reduced embeddings are then clustered using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [42]. HDBSCAN enhances the traditional DBSCAN algorithm into a hierarchical clustering model effectively differentiating between relevant clusters and outliers.

Allaoui et al. [43] found that dimensionality reduction with UMAP improves the efficiency and accuracy of clustering algorithms like k-Means and HDBSCAN.

In this paper, when employing BERTopic two approaches can be adopted. The first approach is to confine BERTopic to create as many clusters as the actual labels in the dataset. The second approach allows BERTopic to generate any number of clusters based on its data analysis.

Embeddings

In this paper, the steps are shown in Fig. 5 to derive the embeddings from the text data.

Subject-verb-object (SVO) extraction

In order to construct a knowledge graph, we need to get the subjects, verbs, and objects of the sentences. Utilizing

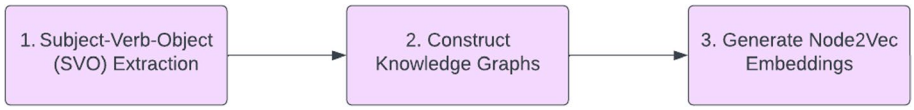

Fig. 5 Steps to generate Embeddings

Spacy[8], we can effectively identify these grammatical components. Spacy's dependency parser allows us to parse sentences and recognize their syntactic structure making it possible to isolate the subject, verb, and object in each sentence. This is achieved by analyzing the grammatical relationships between words and identifying their respective roles.

Knowledge graph construction

Knowledge graphs have a wide range of applications and can be constructed using various approaches tailored to the data and objectives [45–48]. One prominent approach involves extracting Subject-Verb-Object (SVO) triples from text, which serve as the building blocks for the graph. In this method, subjects and objects are represented as nodes, while verbs describing their interactions form edges connecting these nodes.

This SVO-based approach is particularly effective for transforming unstructured text into structured, machine-readable formats. By encoding semantic relationships, it enables the creation of interpretable graphs that illustrate complex information networks, revealing patterns, connections, and insights. The resulting knowledge graph not only enhances our understanding of the underlying data but also supports applications across diverse fields, including healthcare, research, and business intelligence.

Figures 6 and 7 illustrate a sample knowledge graph that has been visualized using the NetworkX[9] and PyViz[10] libraries respectively.

Node2Vec embedding generation

In the node2vec algorithm, continuous feature representations of nodes in networks are learned by mapping nodes to a low-dimensional space. This mapping aims to maximize the likelihood of preserving network neighborhoods of nodes. The algorithm introduces a flexible notion of a node's network neighborhood and employs a biased random walk procedure to explore these neighborhoods effectively. This flexibility allows node2vec to organize nodes based on their network roles and/or communities they belong to, adapting to various network structures and enabling more accurate representations [12, 49] (Table 1).

The node2vec algorithm consists of the following steps:

1. Preprocess the graph to modify weights based on the parameters p and q, which control the random walk strategy [50].
2. Initialize an empty list of walks.
3. For each node in the graph, perform r random walks of length l, appending each walk to the list of walks.
4. Use Stochastic Gradient Descent (SGD) to optimize the feature representations of nodes based on the collected walks [12].

The random walks start at a node and iteratively choose the next node in the walk based on the transition probabilities that are preprocessed based on the neighborhood structure. This method efficiently samples diverse neighborhoods and return walks which are then used to learn feature representations [12, 51].

Kernel-based methods have similarly demonstrated their utility in dimensionality reduction tasks such as hyperspectral image classification, suggesting potential applicability to enhance node embedding methods like Node2Vec [52, 53].

Figure 8 depicts the random walk procedure in the algorithm.

Inconsistency detection

The next step is the innovative part of this paper which attempts to employ the results from the BERTopic perspective to enhance the accuracy and F1-score of the classifier by removing the inconsistent data. Figure 9 illustrates the overall steps of the Inconsistency Detection Algorithm.

Mapping BERTopic clusters to the real labels

At this step of the process, the aim is to establish a generalizable step to find the optimized corresponding mapping between cluster labels and the classification (human) labels as the ground-truths, in this case the cancer types. This step is necessary prior to detecting the inconsistencies, the definition of which is an alternative item's membership to another class than the one it belonged to before. We need to find an alternative label to recommend for the items that the classifier was confused about. We will then propagate and reflect the suggested labels onto the

[8] A free open-source library designed for advanced Natural Language Processing (NLP) tasks, accessible through Python. It is capable of comprehending and analyzing texts of varying sizes [44]

[9] A Python package for the creation and study of complex networks, offering extensive tools for network analysis and visualization.

[10] A Python library for creating interactive network visualizations. It's built on top of the popular graph library NetworkX and uses JavaScript libraries like Vis. js for rendering interactive visualizations in the browser.

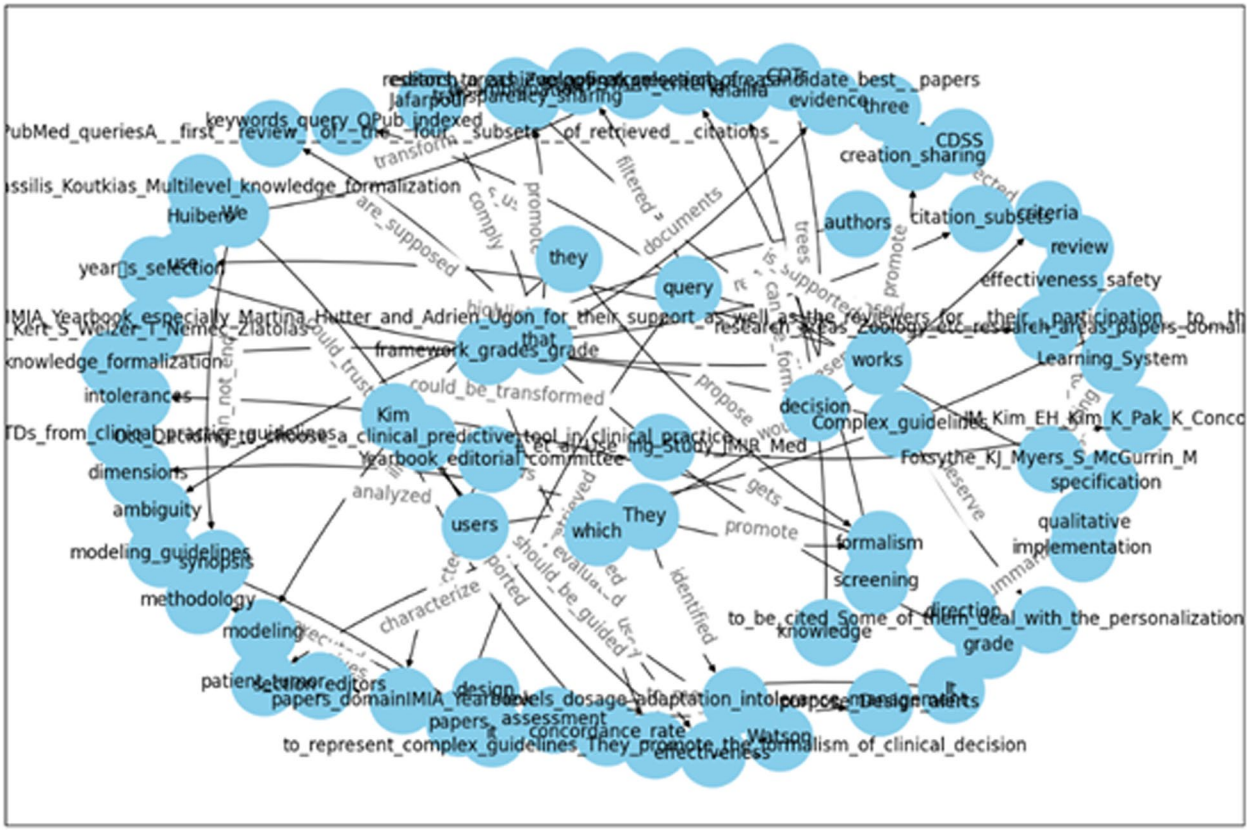

Fig. 6 Example of a constructed knowledge graph using NetworkX

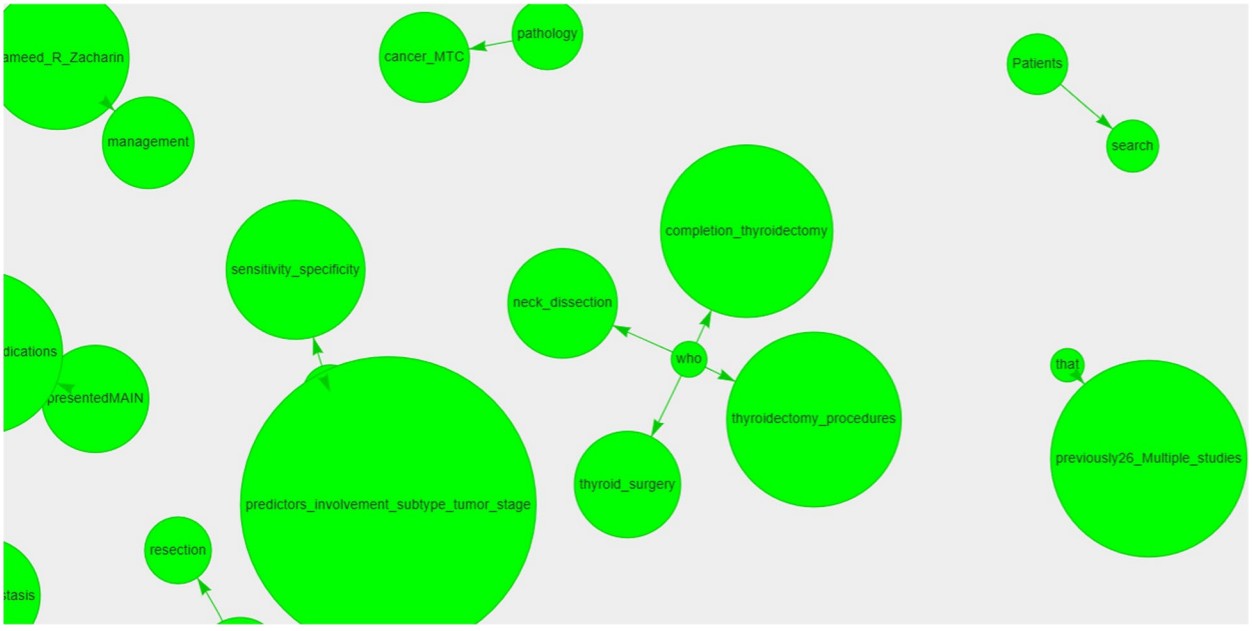

Fig. 7 Example of a constructed knowledge graph using PyViz

knowledge graph embeddings at the feature engineering level to see if the change will improve the predictions, F1, or other evaluation indices. The intuition behind the mapping is as if it's a voting scheme by the majority of the real labels belonging to a cluster to decide the correspondence between the cluster label and the real label which is the human supervised cancer type. We could have done this through a greedy search, but to avoid ties while developing

Table 1 Node2Vec parameters considered for our case

Parameter	Value
p	0.5
q	2
r	5
l	10

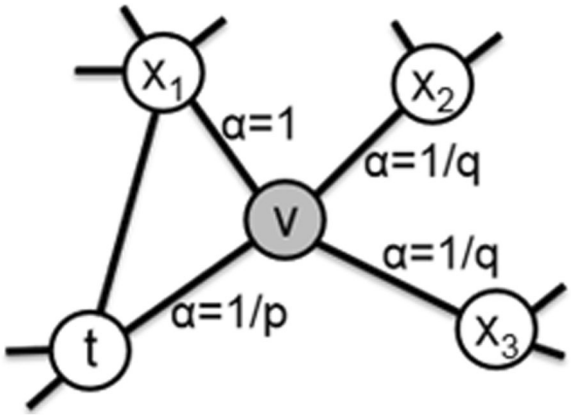

Fig. 8 Illustration of the random walk procedure in node2vec [12]

a generalizable process, we used two methods: crisp optimization vs. a probabilistic recommender scheme.

We also enforced k=3, the number of clusters generated by BERTopic equals the number of classes in our dataset. Here, the main task is to identify which cluster corresponds to which class. We refer to this task as the mapping process.

Two approaches were utilized as the optimization process:

- 1- Class-to-Cluster Assignment (CCA): This approach involves assigning a class to the cluster where that class is most frequently observed. It's a one-to-one optimization method that relies on the predominant presence of a class within a cluster for assignment.
  - Calculate the frequency f_{ij} of class j in cluster c_i:

$$f_{ij} = \frac{\text{Count of documents with label j in } c_i}{\text{Total documents in } c_i} \quad (1)$$

  - Assign class j with maximum frequency to cluster c_i:

$$L_i = \text{argmax}_j f_{ij} \quad (2)$$

- **2- Recommender-System Assignment (RSA)**: We leveraged recommender systems to determine the relationships between clusters and classes. For this, we used the Python Surprise library, a tool designed for building and analyzing recommender systems. We treated each document as an "itemID," the clusters generated by BERTopic as "userID," and the actual labels of the documents as "rating" (Fig. 10). After training the model on these parameters, it predicted the associations between clusters and classes with an accuracy of 86%.
  - Parameter Definitions
    - Documents: $I = \{i_1, i_2, \ldots, i_m\}$
    - Clusters: $U = \{u_1, u_2, \ldots, u_k\}$
  - *Ratings (labels): $R = \{r_1, r_2, \ldots, r_m\}$Model Training
    - Train a recommender system model:

$$\text{Model} = \text{train}(U, I, R) \quad (3)\text{Prediction}$$

    - Predict cluster-to-class mapping:

$$\hat{L} = \text{predict}(\text{Model}, C) \quad (4)$$

Based on the results of Tables 3 and 4 in the experimental results section, we can conclude that the recommender system results have a superior effect on the downstream classification task, improving accuracy by 34%.

Classifying the node embeddings

Parallel to the BERTopic process, node embeddings are classified using a supervised algorithm such as Support Vector

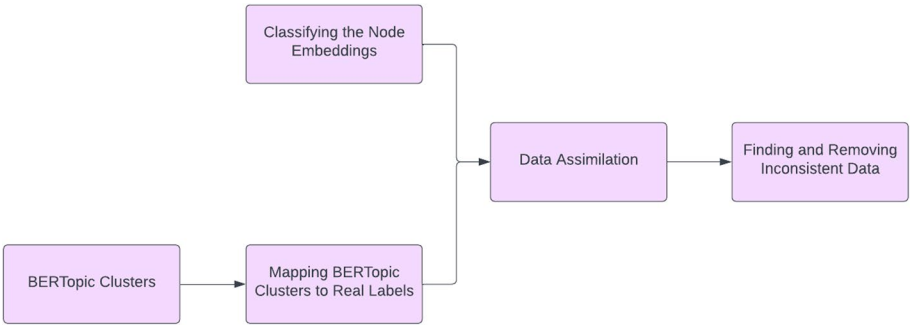

Fig. 9 Inconsistency detection process

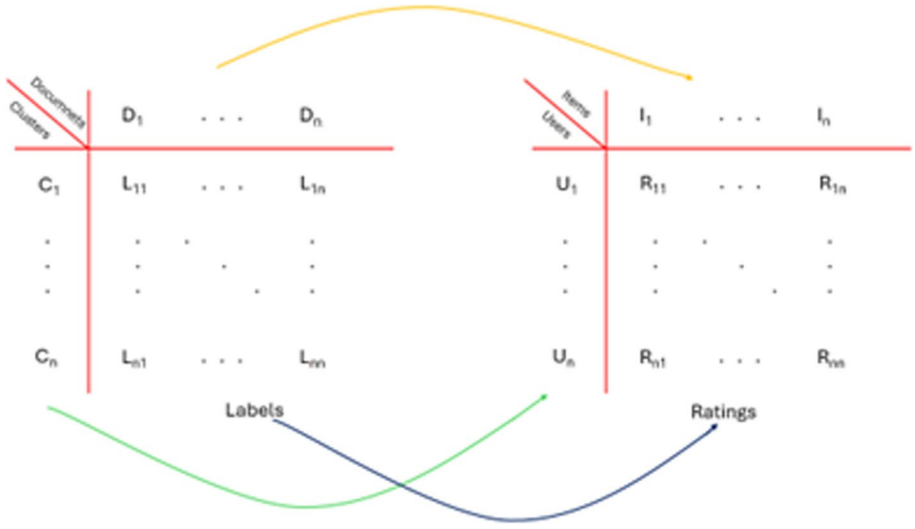

Fig. 10 Mapping BERTopic clusters to the real labels using Recommender System

Machines (SVM)[11] [53, 54]. Node embeddings are vector representations of nodes at our feature engineering step that encapsulate the semantic properties of the transformation of original texts to a graph or network. The document labels (cancer types) are propagated to their corresponding graph nodes. There may be the same nodes but with a different label in the global graph, as the same name entities existed commonly within the original text segments.

Feature data assimilation

At this phase, the objective is to find if there is any alternative justified label exists for any of the confused predictions or classifier errors. The alternative recommendations come as the results of the above mapping process ("Mapping BERTopic clusters to the real labels" section) to integrate either the optimized or recommended outcomes of the BERTopic clustering with node embedding representations for re-classifications. In this context, each document is treated as a distinct graph—referred to as the "graph level". Within each graph, individual words that function as subjects, verbs, or objects are considered nodes, which constitute the "node level".

Two assimilation strategies are presented:

- The first involves enriching the BERTopic results by propagating from a single label per document to the nodes of its associated graph.
- The second strategy involves aggregating the labels at the node level to derive a singular label for the entire graph, determined by the most frequent node labels within that graph.

[11] This classifier can be replaced with any other conventional classifier that can take advantage of the feature engineering refinements and error reductions.

Table 2 Label Re-allocation Scenarios

Actual label	SVM predictions	LLM clustering
a, b, Suggest New Label	F-F	b', a' potential wrong label
a, b	T-F	b', a', disagreement
a, b, Suggest New Label	F-T	b', a', justified confusion
a, b	T-T	a', b', full agreement

Employing either strategy is a prerequisite to setting the stage for the downstream identification of and removing inconsistent data elements at the graph level features.

Finding and removing inconsistent data: feature upgrades

The objective at this stage is to define tactical approaches to apply the potential inconsistencies between the actual labels and the recommended alternative (LLM) labels. There are different options[12].

In Table 2, the first and third rows are cases in which the model predictions (middle column) are different from actuals (left column). T and F represent true-false corresponding to the prediction values with either side columns in the left (actual label) and right (cluster label) respectively.

There are caveats to rectification of potential inconsistencies at the feature engineering data level, and KG representation prior to the Node2Vec transformation. We have chosen the removal of associated nodes with error segments, whenever there were new label recommendations available in the previous step. This removal is predicated on the assumption that inconsistencies between the topic-based clusters and the network structure justify the negative impact on the classifier's performance. The strategic exclusion of such discordant data ensures that the classifier operates with a dataset that is more uniform and aligned with its overall ground truths and concepts,

[12] We will pursue other applications of inconsistency rectification options in the future.

thereby potentially enhancing the precision and F1-score of the classification outcomes.

- Identify Inconsistent Data:
  - Compare predicted labels L with actual node-level labels L_n:

$$\text{Inconsistency} = \{i \mid L_i \neq L_{ni}\} \tag{5}$$

- Remove Inconsistent Data:
  - Exclude inconsistent data points I from node embedding dataset D:

$$D' = D - \text{Inconsistency} \tag{6}$$

Evaluation

Following the strategic exclusion of inconsistent data, the classifier is re-engaged to perform its classification tasks. Subsequently, the performance of the classifier is rigorously assessed by measuring the "Accuracy" and "F1-score". These metrics serve as indicators of the classifier's predictive performance.

Accuracy reflects the proportion of total correct predictions made by the classifier out of all predictions:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

The F1-score provides a more nuanced measure, balancing the precision[13] and recall[14]:

$$\text{Precision}(P) = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Recall}(R) = \frac{TP}{TP + FN} \tag{9}$$

$$\text{F1-Score} = 2 \times \frac{P \times R}{P + R} \tag{10}$$

By evaluating these metrics before and after the data refinement process, we can quantify the impact of removing inconsistent data on the classifier's effectiveness. This evaluation not only validates the methodology but also ensures that the model's performance is aligned with the expected outcomes of the classifier post-improvement.

The conclusive phase involves conducting a t-test[15] to validate that the observed improvements are statistically significant and not due to chance. This ensures that the enhanced results are not replicable by merely randomly deleting data.

- Calculate the t-statistic to validate performance improvement:

$$t = \frac{X_1 - X_2}{\sqrt{\frac{S_p^2}{n}}} \tag{11}$$

Where X_1 and X_2 are means of two samples, S_p is the pooled standard deviation, and n is the number of samples.

Computational complexity and overhead

The computational complexity of the Explainable Inconsistency Detection (EID) framework is influenced by the sequential steps involved in its methodology. The key components and their respective complexities are as follows:

- BERTopic Clustering: This step includes dimensionality reduction using UMAP, with a complexity of $O(n \cdot d^2)$, where n is the number of data points and d is the dimensionality of the feature space. The subsequent clustering with HDBSCAN has a worst-case complexity of $O(n^2)$, making it computationally intensive for large datasets.
- Node2Vec Graph Embedding: The graph construction and random walk generation have a complexity of $O(|V| + |E|)$, where $|V|$ and $|E|$ represent the number of nodes and edges, respectively. The embedding learning process scales as $O(n \cdot d)$, where d is the embedding dimensionality.
- SVM Classification: Training the Support Vector Machine (SVM) model involves complexities that range from $O(n^2)$ to $O(n^3)$, depending on the number of support vectors and the dimensionality of the data.

Combining these components, the overall computational complexity of the EID framework can be expressed as:

$$O(n^2 + n \cdot d^2 + |V| + |E| + n^3) \tag{12}$$

where n is the number of data points, d is the dimensionality of the feature space, and $|V|, |E|$ are the number of nodes and edges in the graph, respectively.

The preprocessing steps, including clustering and embedding generation, introduce a significant computational overhead. However, these steps are crucial for refining the dataset and improving the reliability of downstream classification. By removing inconsistencies and irrelevant features, the framework reduces the dataset size and improves classification efficiency. This trade-off results

[13] The correctness of positive predictions.

[14] The classifier's ability to identify all actual positives.

[15] A statistical test used to determine if there is a significant difference between the means of two groups.

in a more compact and coherent dataset for training, minimizing computational demands in the later stages.

Despite the preprocessing overhead, the framework demonstrates scalability challenges for very large datasets due to the quadratic and cubic growth in clustering and SVM training, respectively. Future implementations can leverage parallelization, distributed processing, and approximate clustering methods to mitigate these computational demands, ensuring improved scalability and efficiency.

Experimental results

In this section, we evaluate the performance and robustness of our proposed dual-perspective framework. Our experiments are designed to assess (i) the impact of targeted inconsistency removal on classification performance, (ii) the semantic clarity provided by BERTopic clustering, and (iii) the effectiveness of node-level feature refinement as validated by statistical tests. To ensure that the observed improvements are not merely due to reduced data volume, control experiments were conducted in which equivalent amounts of data were removed randomly. A series of t-tests further confirmed the statistical significance of the enhancements achieved through our targeted approach.

BERTopic results

BERTopic was employed to extract topic words from the semantic embeddings, thereby revealing the underlying clusters in the data. Figures 11 and 12 present the topic words generated for two distinct datasets. These visualizations not only confirm the coherence of the clusters but also provide qualitative insights into the dominant themes captured by the model. The clear separation of topic words indicates that the clustering captures meaningful semantic groupings, which in turn supports the identification of inconsistencies in the subsequent classification phase.

Node embedding visualizations

To visualize the distribution of the node embeddings, we used a 20-dimensional Node2Vec representation followed by dimensionality reduction via PCA. Figure 13 shows the TSNE projection of these embeddings, where colors denote the true labels. The clear separation of clusters in the visualization provides evidence that our feature engineering process effectively preserves class distinctions, and it lays the groundwork for targeted inconsistency detection.

Inconsistency detection and evaluation

The core of our approach is the identification and removal of inconsistent data points. Initially, a baseline Support Vector Machine (SVM) classifier was applied to the node embeddings, achieving an accuracy of 46% and an F1-score of 50%. Using BERTopic to guide our Explainable Inconsistency Detection (EID), we flagged data points that did not conform to the dominant semantic clusters. As illustrated in Fig. 14, these flagged instances are clearly separated in the embedding space from the consistent ones.

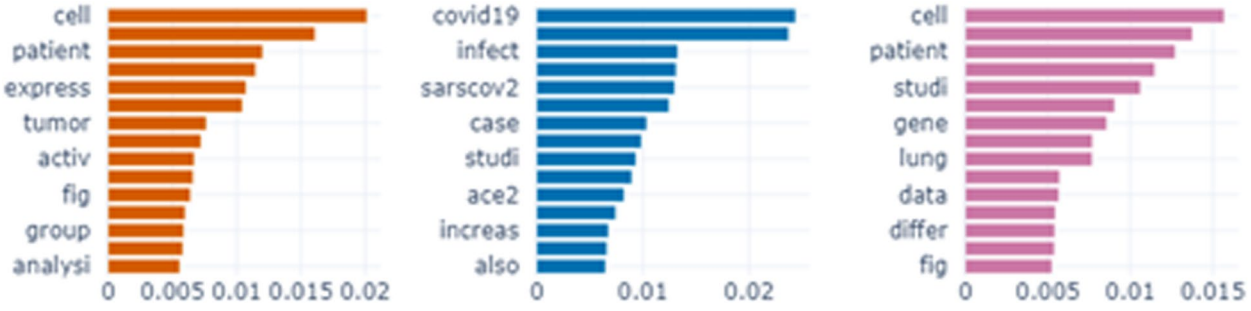

Fig. 11 BERTopic-generated topic words for each cluster (Dataset 1). The distinct clusters highlight coherent semantic groupings

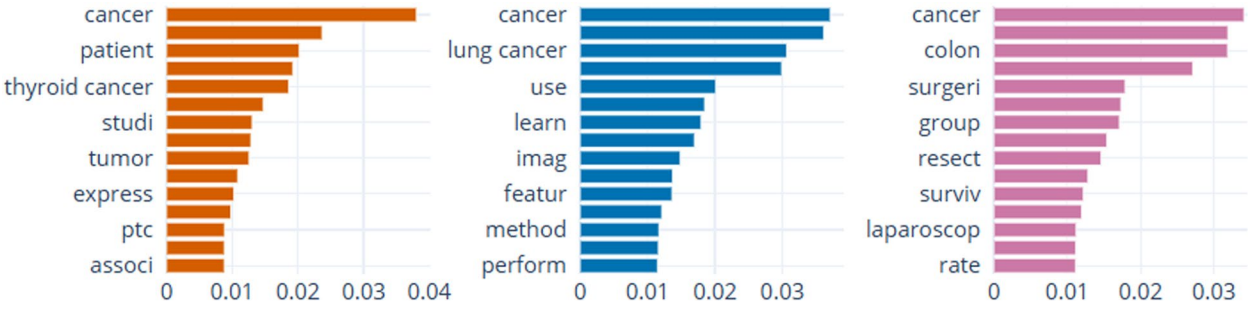

Fig. 12 BERTopic-generated topic words for each cluster (Dataset 2). Similar trends in topic coherence are observed

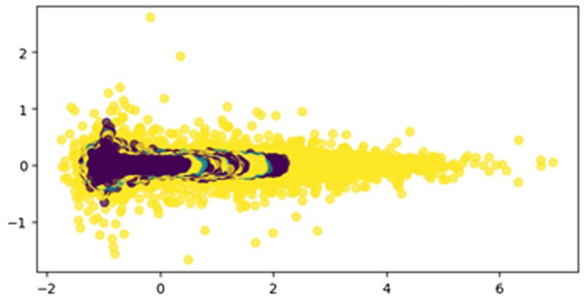

Fig. 13 TSNE visualization of node embeddings, colored by actual labels. This projection validates the discriminative power of the learned representations

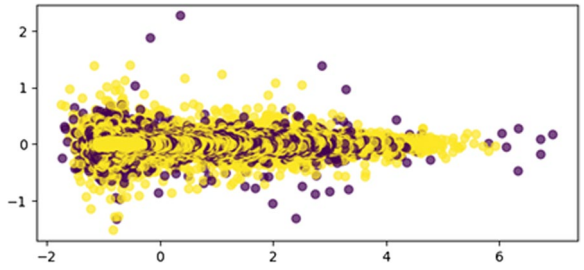

Fig. 14 Visualization of node embeddings highlighting inconsistent data points. The separation indicates the effectiveness of the EID in detecting outliers

After removing the identified inconsistent nodes, the SVM classifier was re-evaluated using three optimization approaches:

- CCA (Class-to-Cluster Assignment): Direct mapping of clusters to class labels.
- RSA (Recommender-System Assignment): An integer programming–based approach to map clusters to labels.
- RSA with Limited Removal: Incorporates a hyperparameter β (set to 0.7) to limit removal to only β fraction of the flagged inconsistencies, thereby preserving more data.

Table 3 summarizes the performance metrics for both datasets across these approaches. Notably, the RSA approach shows substantial improvements in accuracy and F1-score (e.g., 91% accuracy for Dataset 1) compared to the baseline, while the β-limited approach offers a balance between performance and data retention.

To further assess the model's performance, we analyzed Receiver Operating Characteristic (ROC) curves and confusion matrices for each class. Figures 15, 16 and 17 display the ROC curves, where Classes 0 and 2 show dramatic improvements in the Area Under the Curve (AUC)—from 0.55 to 0.94 and 0.57 to 0.95, respectively. Although Class 1 improved modestly (from 0.53 to 0.58), this reflects inherent challenges in distinguishing this class. Complementary

Table 3 Performance metrics across different optimization approaches for both datasets. CCA and RSA denote the Class-to-Cluster and Recommender-System Assignment approaches, respectively

Metric	Accuracy	Precision	Recall	F1-score	AUC
Baseline Data 1	46%	48%	52%	50%	0.55
Baseline Data 2	47.3%	49.5%	53.1%	51.2%	0.57
CCA (Data 1)	57%	60%	61%	58%	0.58
CCA (Data 2)	56.8%	58.3%	59.2%	58.1%	0.59
RSA Data 1	91%	92%	94%	89%	0.94
RSA Data 2	88.6%	87.5%	89.4%	85.9%	0.95
RSA B% - Data 1	87%	88%	89%	83%	0.94
RSA B% - Data 2	84.7%	83.2%	85.1%	82.4%	0.95

confusion matrices (Figs. 18 and 19) confirm a significant reduction in misclassifications after optimization.

Additionally, Fig. 20 displays a word cloud generated from the inconsistent data. The prominence of specific terms provides qualitative insights into the language patterns that may contribute to classification challenges, offering a direction for further feature engineering enhancements.

Statistical validation

To ensure that the improvements observed are not attributable to random data reduction, we conducted control experiments by randomly removing an equivalent amount of data multiple times. A two-sample t-test was then applied to compare the performance metrics of our targeted approach with the control experiments. The p-values, reported in Table 4, are several orders of magnitude below the 0.05 significance threshold, confirming that the performance gains are statistically significant and directly attributable to the removal of inconsistent data.

Given the extremely low p-values, we confidently reject the null hypothesis. This confirms that our Explainable Inconsistency Detection effectively enhances model performance by systematically removing problematic data points.

Summary of findings

The experimental results demonstrate that:

- The targeted removal of inconsistent data—guided by BERTopic clustering and node-level embeddings—substantially improves classification performance.
- Visual analyses, including ROC curves, confusion matrices, and embedding visualizations, corroborate the quantitative improvements.
- The hyperparameter β effectively balances data retention and performance gains, ensuring that improvements are not solely due to excessive data removal.
- Statistical tests confirm that the improvements are significant and not attributable to random data reduction.

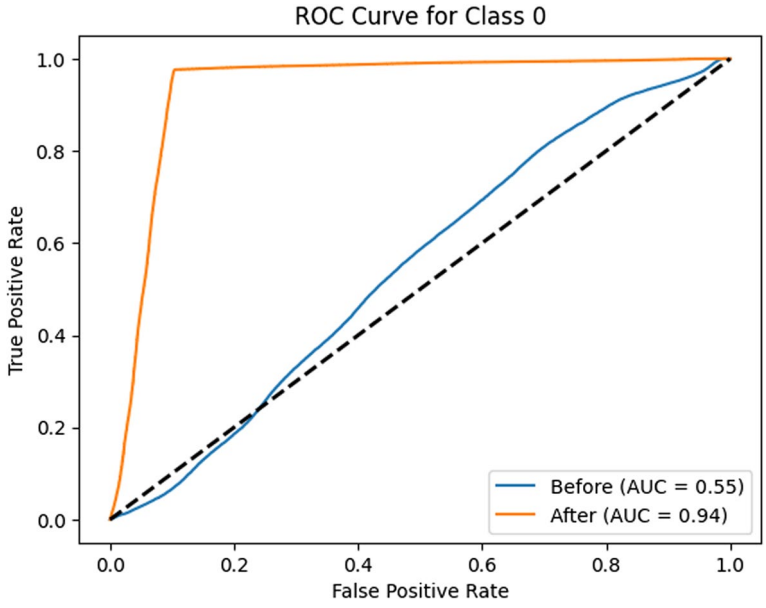

Fig. 15 ROC curve for Class 0 (AUC improvement: 0.55 to 0.94)

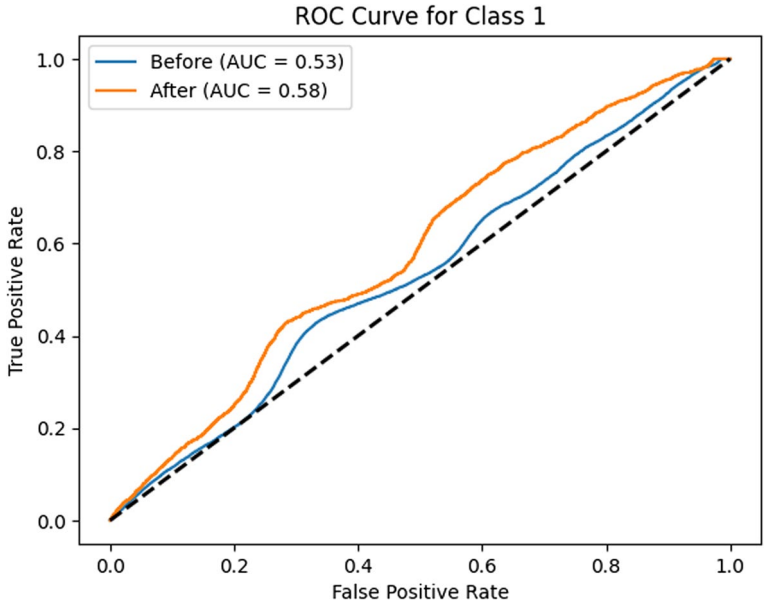

Fig. 16 ROC curve for Class 1 (AUC improvement: 0.53 to 0.58)

Together, these findings validate the effectiveness of our dual-perspective framework in both improving predictive performance and enhancing the explainability of model decisions.

Discussion

This study set out to improve text-based cancer data classification by integrating BERTopic clustering with SVM classifiers, and by introducing Explainable Inconsistency Detection (EID) to elucidate and correct data inconsistencies. The fusion of these methods not only enhanced overall classification performance but also provided critical insights into ambiguous and noisy data, thereby improving the reliability and transparency of the model's predictions. These advances have direct implications for healthcare decision-making, where accurate and explainable models can significantly impact clinical outcomes and operational efficiencies.

Bridging research and clinical practice

In the context of healthcare, particularly in oncology, precise data classification is essential for early diagnosis, treatment planning, and patient management. Our framework's ability to detect and remove inconsistencies via the EID provides clinicians and administrators

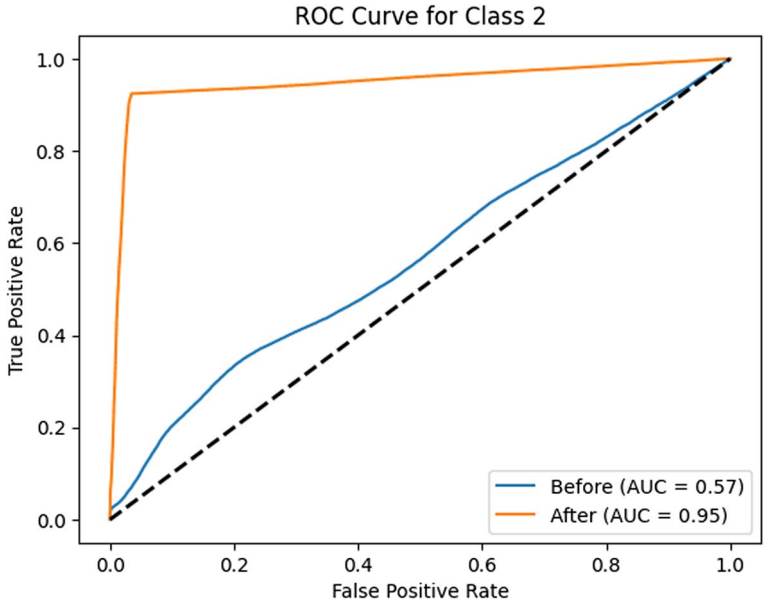

Fig. 17 ROC curve for Class 2 (AUC improvement: 0.57 to 0.95)

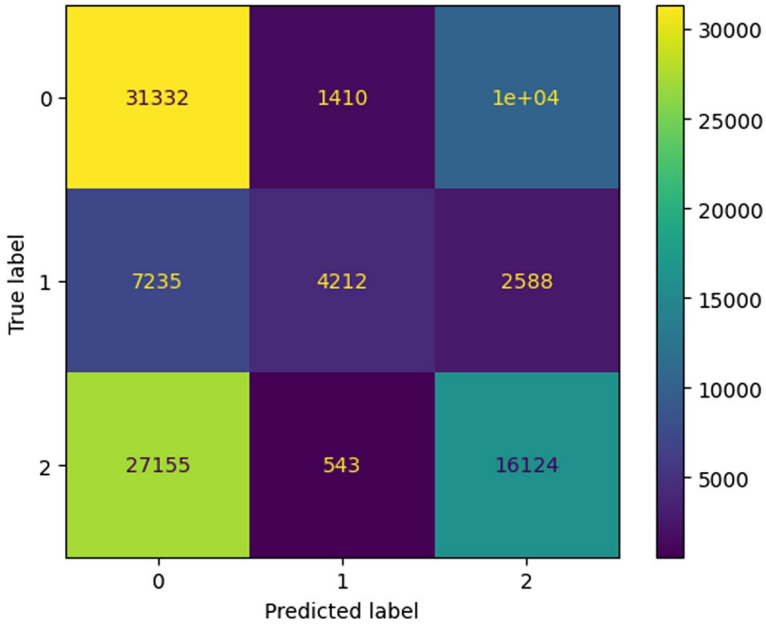

Fig. 18 Confusion matrix before optimization

with more trustworthy model outputs. Enhanced interpretability reduces uncertainty and builds confidence among healthcare providers, leading to better adoption of AI-driven diagnostic tools in clinical workflows. Furthermore, by incorporating a recommender system for mapping clusters to labels, our approach leverages collaborative filtering techniques—commonly used in e-commerce and personalized medicine—to ensure that label assignments are both accurate and adaptable to new data. This dynamic adaptability is critical in the ever-evolving landscape of medical research and clinical practice.

Handling sparsity and high-dimensional data

Challenge

Clustering high-dimensional text data often leads to sparse representations, complicating the direct mapping of clusters to clinically relevant labels.

Solution

Our approach incorporates recommender systems, which excel in handling sparsity by predicting missing associations and uncovering latent relationships. Mathematically, the association matrix R is decomposed into

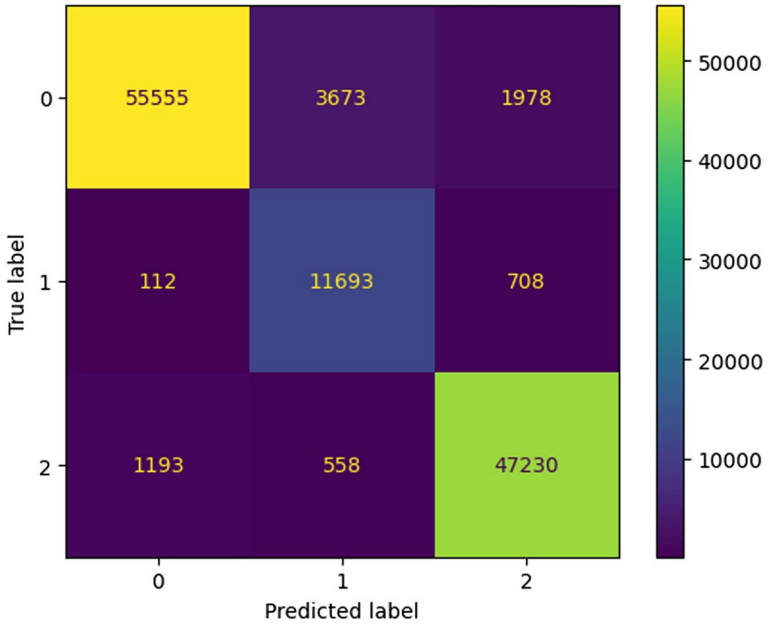

Fig. 19 Confusion matrix after optimization

Fig. 20 Word cloud depicting term frequencies in inconsistent data. Larger words indicate higher occurrence, suggesting key areas for potential refinement

Table 4 *P*-values from the t-test for both datasets. All *p*-values indicate statistically significant improvements

Metric	Dataset 1 (*p*-value)	Dataset 2 (*p*-value)
Accuracy	3.77×10^{-59}	2.45×10^{-45}
F1-Score	4.30×10^{-52}	1.87×10^{-38}
Precision	5.12×10^{-48}	3.21×10^{-40}
Recall	2.98×10^{-50}	4.89×10^{-42}

lower-dimensional matrices P (representing latent features of clusters) and Q (capturing label characteristics), such that:

$$R \approx P \cdot Q^T. \tag{13}$$

The objective is to solve:

$$\min_{P,Q} \sum_{(u,i)\in K} (r_{ui} - P_u \cdot Q_i^T)^2 + \lambda(||P||^2 + ||Q||^2), \tag{14}$$

which effectively regularizes the model to manage sparsity while preserving the underlying clinical signal. This mathematical insight not only refines clustering but also enhances the interpretability of feature associations, making the results more actionable for healthcare practitioners.

Discovering latent patterns and enhancing explainability

Challenge: Traditional clustering methods may fail to capture the complex, latent relationships that underpin

clinically relevant patterns, such as those seen in cancer subtypes.

Solution: By using latent factor models, the recommender system discovers hidden associations between clusters and labels. The predicted association is given by:

$$r_{ui} = P_u \cdot Q_i^T, \tag{15}$$

which quantifies the likelihood that a given cluster corresponds to a specific cancer label. This latent pattern discovery not only improves label assignment accuracy but also uncovers previously unrecognized relationships that could inform novel clinical insights and therapeutic strategies.

Scalability, efficiency, and adaptability

In large healthcare datasets, scalability is paramount. Our framework demonstrates significant computational efficiency through reduced-dimensional representations, enabling the processing of massive datasets typical in multi-center clinical studies. Adaptive control methodologies, such as those based on fuzzy logic and nonlinear systems, have shown promise in dynamically handling uncertainties and disturbances, indicating potential avenues for further enhancing the robustness and adaptability of our framework [55, 56]. Moreover, the use of dynamic recommender systems ensures that the model continuously adapts to new data—such as emerging trends in cancer research or evolving clinical terminologies—thus maintaining its accuracy over time. This adaptability is crucial for long-term deployment in healthcare systems, where data distribution can shift due to changes in diagnostic criteria, treatment protocols, or patient demographics.

Investigating inconsistencies with explainable AI (XAI)

A key strength of the EID is its ability to identify and explain inconsistencies within the data. For instance, the analysis revealed that terms like "lung_cancer" may be ambiguously connected to both "PTEN_expression" and "liver_cancer" (Fig. 21), potentially misleading the model. Similarly, non-discriminative terms such as "lesion" and overly generic words like "role" (Figs. 22 and 23) can dilute the model's decision-making process. By quantifying the mean degree centrality of consistent versus inconsistent words, we found that:

$$\text{Consistent words: } 1.385 \times 10^{-3} \quad , \quad \text{Inconsistent words: } 2.05 \times 10^{-5},$$

indicating that highly connected (and thus ambiguous) terms introduce noise. The EID enables targeted feature-level manipulation, whereby such noisy nodes are either refined or removed. This process not only enhances model accuracy but also offers actionable insights that can guide further feature engineering and even influence clinical data curation practices.

Feature-level manipulation for enhanced classification

A distinguishing feature of this study is its focus on feature-level manipulation using perturbation-based explainability methods [57]. By systematically altering or masking individual features, we evaluated their impact on classification outcomes. This granular analysis revealed which features (e.g., specific genetic markers or clinical terms) have a pronounced effect on predictions, enabling clinicians to understand and trust the model's decision process. Such insights are invaluable in healthcare, where explainability can directly influence treatment decisions and patient management.

Business and clinical implications

From a business perspective, our framework offers several benefits:

- Enhanced Diagnostic Accuracy: Improved classification performance translates to more accurate cancer diagnoses, reducing misclassification costs and enhancing patient outcomes.
- Operational Efficiency: Automated inconsistency detection and data refinement reduce the need for manual data cleaning, streamlining clinical workflows and reducing operational overhead.
- Regulatory Compliance and Trust: Explainable models foster transparency, facilitating compliance with stringent healthcare regulations and increasing trust among clinicians and patients.
- Scalability for Big Data: The efficient handling of large-scale, high-dimensional datasets ensures that our approach can be integrated into national healthcare systems and multi-institutional research initiatives.

Limitations and future directions

Despite the promising results, several limitations warrant further exploration. Dataset biases and the challenges inherent in SVM when applied to extremely complex datasets remain potential issues. Additionally, while Node2Vec effectively captures relational data, it also increases the dataset size and computational burden, potentially limiting scalability in resource-constrained settings. Future work should explore alternative embedding techniques and more efficient algorithms for graph construction. Moreover, additional validation on real-world clinical datasets, coupled with user studies involving healthcare professionals, would further establish the clinical utility of the proposed framework.

Finally, our study compared two approaches for mapping clusters to labels. The "Confining BERTopic" (predefining the number of clusters) strategy was chosen over "Let BERTopic Decide" based on its superior accuracy, and "Bringing Everything to Node Level" outperformed "Bringing Everything to Graph Level". These

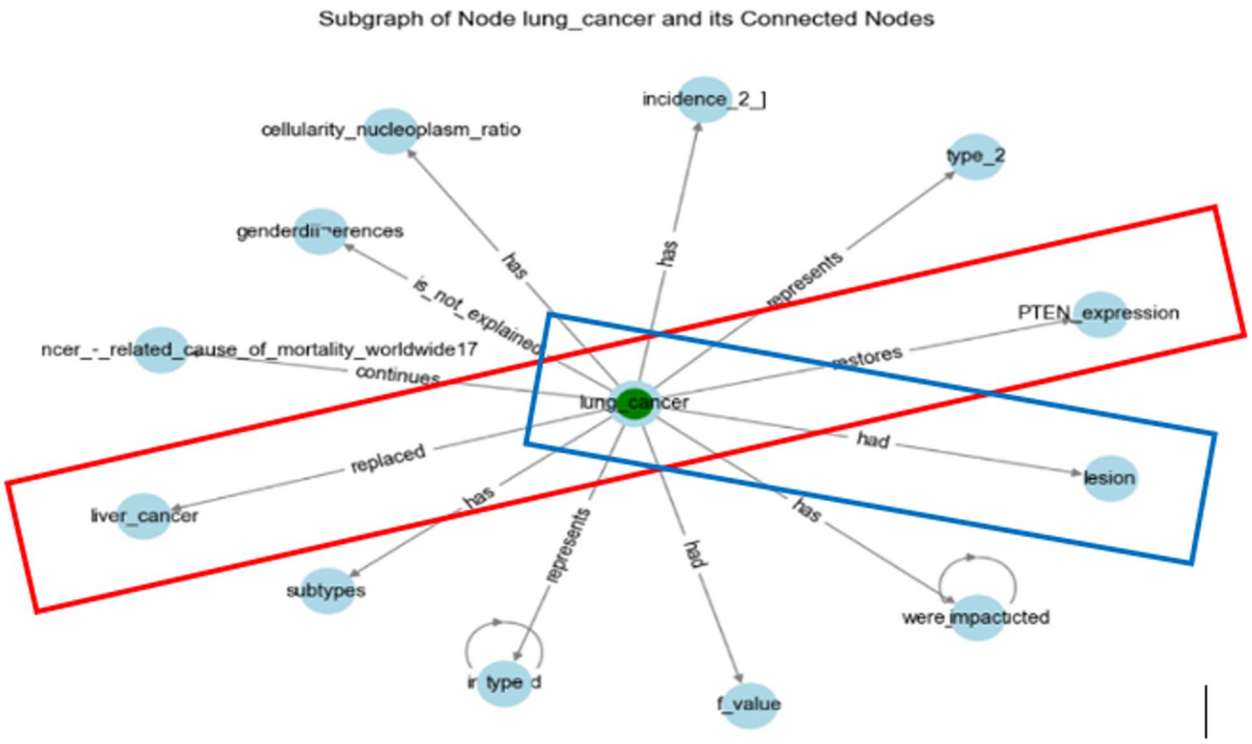

Fig. 21 Showing the connection of "lung_cancer" to "PTEN_expression" and "liver_cancer"

methodological choices, while effective, also point to opportunities for hybrid approaches that dynamically adjust based on data characteristics—a promising direction for future research.

In summary, our integrated framework not only advances the state-of-the-art in text-based cancer data classification but also bridges the gap between advanced AI methodologies and practical, impactful healthcare solutions. By addressing key challenges such as data sparsity, inconsistency, and interpretability, our approach paves the way for more reliable and actionable clinical decision support systems.

Conclusion

This research introduces an innovative dual-perspective approach for enhancing text-based cancer data classification. By integrating BERTopic-based semantic clustering with SVM classifiers and incorporating the Explainable Inconsistency Detection (EID), our framework not only achieves significant performance improvements but also provides deeper insights into data inconsistencies. The integration of advanced preprocessing techniques and Node2Vec embeddings further augments the model's ability to detect and remove discordant data points, leading to more reliable classification outcomes. Statistical validation through t-tests confirms that the observed performance gains are robust and not attributable to chance, thereby substantiating the efficacy of the proposed methodology.

Theoretical contributions and practical implications

From a theoretical standpoint, this study advances the state-of-the-art by proposing a novel framework that seamlessly combines supervised classification with unsupervised semantic clustering. The introduction of the EID addresses a critical gap in the literature by automatically detecting and explaining inconsistencies within complex biomedical datasets. This contributes to a better understanding of the latent relationships in high-dimensional text data, which is essential for the development of more interpretable AI models.

Practically, our framework offers several tangible benefits for the healthcare industry. Improved classification accuracy directly translates to enhanced diagnostic precision, which is crucial for early cancer detection and treatment planning. The explainability embedded in the model fosters greater trust among clinicians by clarifying the rationale behind each prediction. Moreover, the dynamic nature of the recommender system for mapping clusters to labels ensures that the framework can adapt to evolving clinical data, thereby supporting scalable and efficient deployment in real-world healthcare environments.

Practical advantages and limitations

The proposed methodology delivers several practical advantages:

- Enhanced Diagnostic Accuracy: The combined use of BERTopic, SVM, and EID significantly

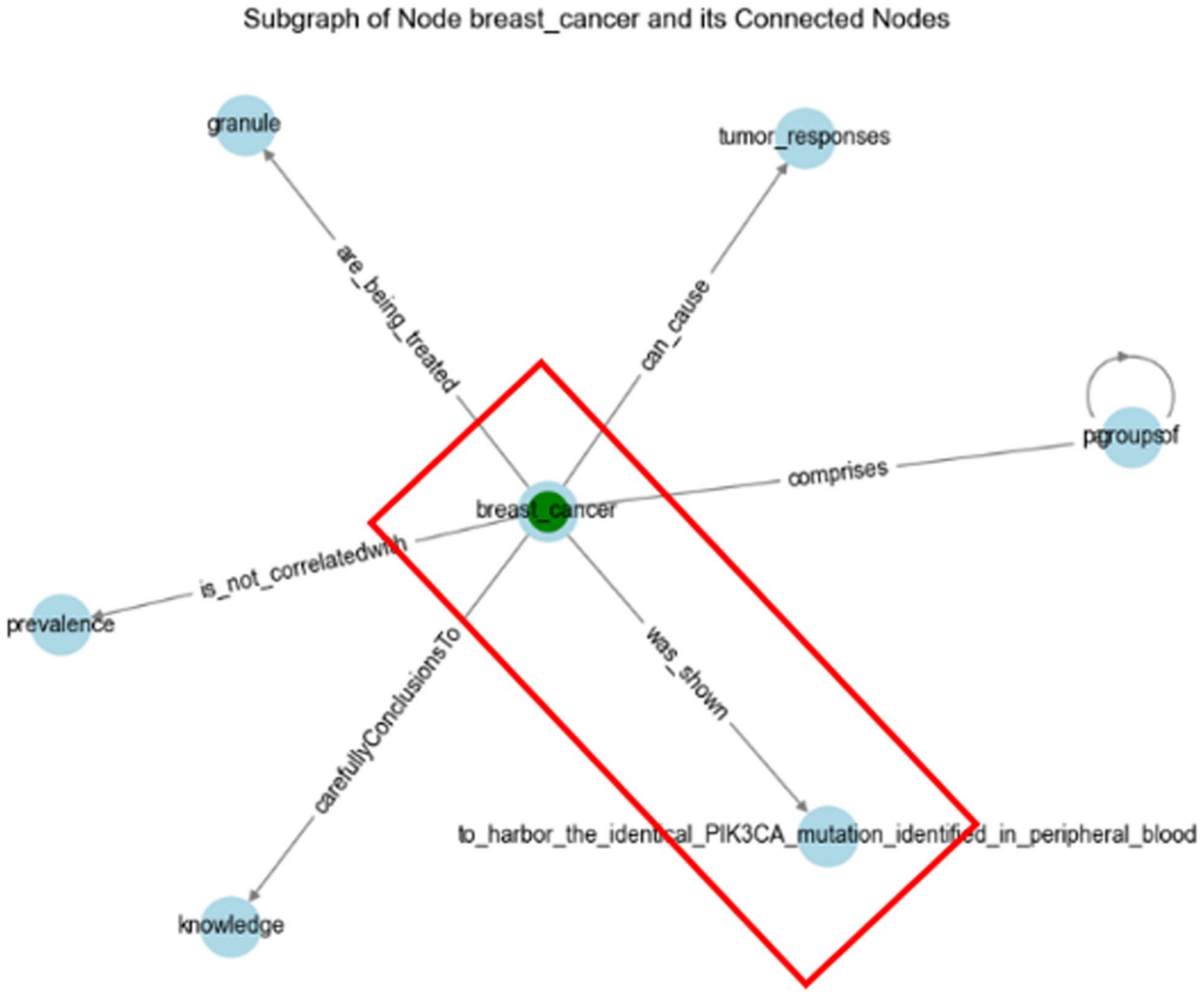

Fig. 22 Showing connections of "lesion" in the context of cancer

improves classification performance, reducing misclassifications and leading to more accurate patient stratification.

- Improved Explainability: By pinpointing and removing inconsistent data points, the model provides clearer insights into the underlying data structure, aiding clinicians in understanding and trusting AI-driven predictions.
- Scalability and Adaptability: The framework is designed to efficiently handle large-scale, high-dimensional datasets, ensuring its applicability across diverse healthcare settings.

Nonetheless, the study has several limitations. Potential dataset biases, the computational overhead introduced by Node2Vec embeddings, and the sensitivity of SVM classifiers to complex data structures may impact the generalizability of the results. Addressing these challenges will require further refinement and validation on larger, more diverse clinical datasets.

Future work

Future research should aim to extend and refine the current framework through several avenues. The validity of predictions as spatiotemporal phenomena is based on the contrasting semantics of their changing criteria. Advanced "explainable" classification tasks will inherently involve multi-participant decision-making, incorporating diverse and often inconsistent perspectives, with AI playing a crucial role as a key team member in these critical activities.

- Integration of Additional Models: Future studies could explore the incorporation of alternative semantic representations and classifiers—such as Random Forests, deep learning architectures, or Graph Neural Networks (GNNs)—to assess whether these

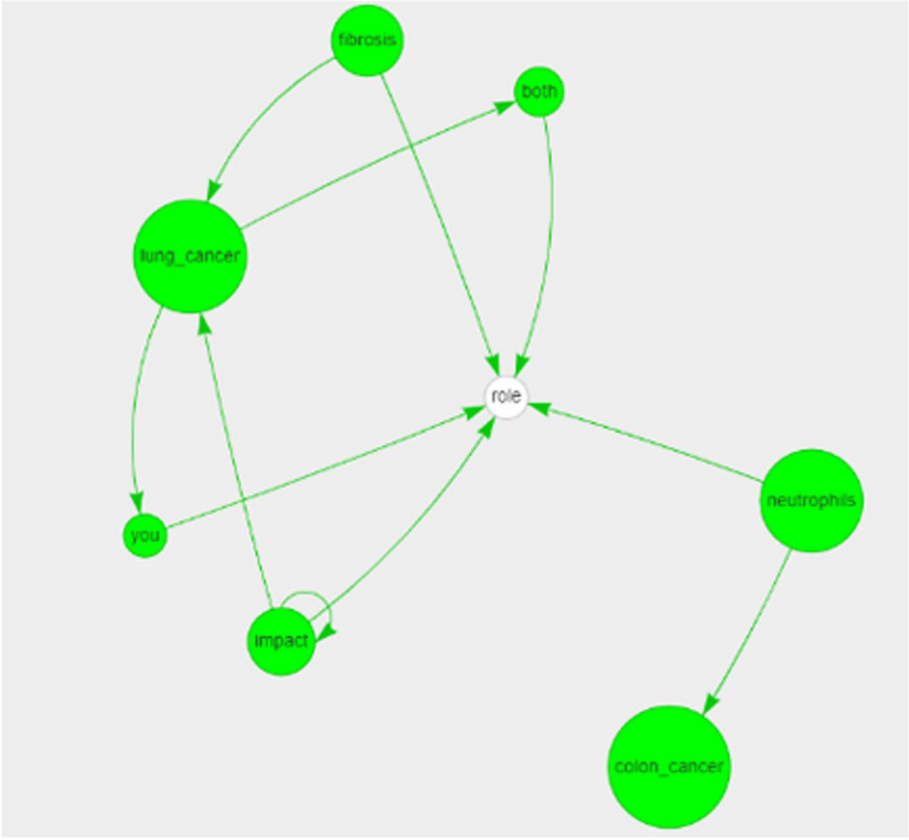

Fig. 23 Showing the role of neutrophils and fibrosis in cancer

approaches further enhance classification performance when combined with other strands of LLM (Gemini, GPT-4, Claude, Llama, BERTopic) using our EID's iterative-explainable-mapping framework.

- Expansion of EID Applicability: Extending the Explainable Inconsistency Detection to other data types (e.g., genomics or multi-omics datasets) and domains could validate its generalizability and uncover additional insights into data inconsistencies across various biomedical contexts.
- Optimization and Scalability Enhancements: After the inconsistency detection, different strategies other than node removal can be adopted as feature upgrades. Further work on hyperparameter tuning, parallel processing techniques, and alternative embedding methods may reduce computational costs and improve scalability, making the framework more efficient for large-scale clinical applications.
- Real-World Clinical Validation: Deploying and evaluating the framework in a real-world clinical setting, particularly within decision-support systems, would provide valuable feedback on its effectiveness and practicality, thereby facilitating its adoption in routine clinical practice.

In summary, our integrated approach not only pushes the boundaries of text-based cancer data classification but also bridges the gap between advanced AI methodologies and practical healthcare solutions. By addressing key challenges such as data inconsistency, explainability, and scalability, this research lays a solid foundation for the development of more reliable and interpretable clinical decision support systems in oncology.

Appendix

Detailed pseudocode of the proposed method

In this appendix, we provide a detailed pseudocode representation of our enhanced cancer data classification framework. This algorithm encapsulates the main steps of our methodology—from preprocessing and embedding generation through BERTopic clustering, knowledge graph construction, inconsistency detection, data refinement, and final classification. This detailed breakdown supplements the discussion in " Methodology" section and is intended for readers who wish to understand the computational specifics of our approach.

Require: Dataset D with documents and ground-truth labels.
Ensure: Improved SVM classifier with enhanced performance metrics.
Preprocessing:
for all document $d \in D$ **do**
    Tokenize d
    Remove stop-words and explicit keywords (e.g., "colon", "thyroid", "lung cancer")
    Lemmatize the resulting tokens
    Store preprocessed document as d'
end for
Embedding Generation:
for all document $d' \in D$ **do**
    Compute embedding e using Sentence-BERT
end for
Dimensionality Reduction:
Apply UMAP on $\{e\}$ to obtain reduced embeddings $\{r\}$
Clustering (BERTopic):
Apply HDBSCAN on $\{r\}$ to generate clusters C
▷ Two strategies: Confined (clusters equal to the number of labels) or Unrestricted.
Cluster-to-Label Mapping:
if using Class-to-Cluster Assignment (CCA) **then**
    for all cluster $c \in C$ **do**
        Compute frequency f_{ij} of each label j in c
        Assign label $L(c) = \arg\max_j f_{ij}$
    end for
else
    Train a recommender system on (Clusters, Documents, Labels)
    Predict cluster-to-label mapping $\hat{L}$ for each $c \in C$
end if
Knowledge Graph and Node Embeddings:
for all document $d' \in D$ **do**
    Extract SVO triples using Spacy
end for
Construct a knowledge graph G from all extracted SVO triples
Apply Node2Vec on G to generate node embeddings
Data Assimilation:
for all document $d' \in D$ **do**
    Aggregate node-level labels from G to form a consolidated label L_n
end for
Inconsistency Detection:
for all document $d' \in D$ **do**
    if $L(d') \neq L_n(d')$ **then**
        Mark d' as inconsistent
    end if
end for
Data Refinement:
Remove all inconsistent documents from D, yielding refined dataset D'
Classification:
Train SVM classifier on node embeddings derived from D'
Evaluation:
Compute performance metrics: Accuracy, Precision, Recall, F1-Score
Conduct t-tests to verify that performance improvements are statistically significant
return Improved SVM classifier and performance metrics

Algorithm 1 Enhanced cancer data classification using BERTopic and EID

Abbreviations

AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
BERTopic	Bidirectional Encoder Representations from Transformers-based Topic Modeling
CCA	Class-to-Cluster Assignment
EID	Explainable Inconsistency Detection
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GNN	Graph Neural Network
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
kNN	k-Nearest Neighbors
KG	Knowledge Graph
LLM	Large Language Model
NLP	Natural Language Processing
Node2Vec	Node Embedding Algorithm
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristic
RSA	Recommender-System Assignment
SVO	Subject-Verb-Object
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
TSNE	t-Distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
XAI	Explainable Artificial Intelligence

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s44398-025-00005-6.

Supplementary Material 1.

Acknowledgements
Ottawa Hospital Research Institute.

Authors' contributions
Pouria Mortezaagha conducted the research and wrote the manuscript. Abhisht Makarand Joshi reviewed the methodology and paper, and provided feedback. Prof. Arya Rahgozar came up with innovative ideation, developed the NLP concepts, co-created the methodology, wrote, reviewed, edited the manuscript, and supervised the project.

Funding
OHRI, Start-up Research Funding.

Data availability
Data is provided within the manuscript or supplementary information files.

Declarations

Ethics approval and consent to participate
Not applicable. This study does not involve human participants, human data, or human tissue. The dataset used for the study consists of publicly available biomedical research papers and abstracts related to cancer, which were collected from open-access sources. As such, no ethics approval or consent to participate was required.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 15 October 2024 / Revised: 9 May 2025 / Accepted: 30 June 2025
Published online: 21 July 2025

References

1. Bhardwaj N, Parashar G. The disagreement dilemma in explainable AI: can bias reduction bridge the gap. PREPRINT (Version 1). 2024. Available at Research Square. https://doi.org/10.21203/rs.3.rs-4193128/v1.
2. Lynden A, Taveekarn B. Semi-Automated Data Augmentation for pandas DataFrames using Open Data Sources. J Data Sci Innov. 2019;12(3):45–60.
3. Brylew A, Miller J, Singh R. Enhancing Neutron Activation Analysis via Noise Addition and Feature Selection. Nucl Anal Tech. 2023;28(1):101–15.
4. Jain P, Tiwari A, Som T. Fuzzy rough assisted missing value imputation and feature selection. Neural Comput Appl. 2023;35:2773–2793. Received: 03 January 2022; Accepted: 18 August 2022; Published: 04 September 2022; Issue Date: January 2023. https://doi.org/10.1007/s00521-022-07754-9.
5. Wilfling D. PyFE: A Python Framework for Feature Engineering in Energy Systems Modeling. Energy Syst Model J. 2023;7(2):85–98.
6. Kumar P, Taylor L. Multi-Modal Fake News Detection Using Textual and Visual Features. J Multimed Intell. 2023;15(4):250–65.
7. Campagner F, Cabitza D. Hybrid Deep Learning for Explainable AI: Leveraging Automatically Detected Symbolic Features. J Artif Intell Res. 2020;68(5):300–15.
8. Vouk M, Reynolds K, Martinez A. Explainable Feature Construction via Instance-Based Explanations. In: Proceedings of the Conference on Explainable AI. 2023. pp. 112–120.
9. Carta G, Russo P, Oliveira M. Automatic Feature Selection for Financial Forecasting Using Explainable AI Methods. J Finan Data Sci. 2021;9(1):75–89.
10. Lukasiewicz T, Malizia E, Molinaro C. Explanations for Negative Query Answers under Inconsistency-Tolerant Semantics. In: Raedt LD, editor. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. International Joint Conferences on Artificial Intelligence Organization; 2022. p. 2705–2711. Main Track. https://doi.org/10.24963/ijcai.2022/375.
11. Bougiatiotis K, Fasoulis R, Aisopos F, Nentidis A, Paliouras G. Guiding Graph Embeddings using Path-Ranking Methods for Error Detection innoisy Knowledge Graphs. 2020. arXiv:2002.08762.
12. Grover A, Leskovec J. Node2vec: Scalable Feature Learning for Networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery; 2016. pp. 855–864. https://doi.org/10.1145/2939672.2939754.
13. Hellström T, Dignum V, Bensch S. Bias in Machine Learning – What Is It Good For? 2020. https://doi.org/10.48550/arXiv.2004.00686.
14. He Z, Deng S, Xu X. Outlier Detection Integrating Semantic Knowledge. In: Meng X, Su J, Wang Y, editors. Advances in Web-Age Information Management. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2002. pp. 126–131. https://doi.org/10.1007/3-540-45703-8_12.
15. Wang H, Bah MJ, Hammad M. Progress in Outlier Detection Techniques: A Survey. IEEE Access. 2019;7:107964–70. https://doi.org/10.1109/ACCESS.2019.2932769.
16. Naseem U, Thapa S, Zhang Q, Hu L, Masood A, Nasim M. Reducing Knowledge Noise for Improved Semantic Analysis in Biomedical Natural Language Processing Applications. In: Naumann T, Ben Abacha A, Bethard S, Roberts K, Rumshisky A, editors. Proceedings of the 5th Clinical Natural Language Processing Workshop. Toronto: Association for Computational Linguistics; 2023. pp. 272–277. https://doi.org/10.18653/v1/2023.clinicalnlp-1.32.
17. Feldmann S, Herzig SJI, Kernschmidt K, Wolfenstetter T, Kammerl D, Qamar A, et al. Towards Effective Management of Inconsistencies in Model-Based Engineering of Automated Production Systems. IFAC-PapersOnLine. 2015;48(3):916–23. https://doi.org/10.1016/j.ifacol.2015.06.200.
18. Zeng Z, Deng Y, Li X, Naumann T, Luo Y, Wu X, et al. Using natural language processing and machine learning to identify breast cancer local recurrence. BMC Bioinformatics. 2018;19(1):469. https://doi.org/10.1186/s12859-018-2466-x.
19. Jain P, Tiwari AK, Som T. An intuitionistic fuzzy bireduct model and its application to cancer treatment. Comput Ind Eng. 2022;168: 108124. https://doi.org/10.1016/j.cie.2022.108124.
20. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A Comprehensive Survey on Support Vector Machine Classification: Applications. Challenges and Trends Neurocomputing. 2020;408:189–215. https://doi.org/10.1016/j.neucom.2019.10.118.
21. Wen J. Efficient Computing Algorithm for High Dimensional Sparse Support Vector Machine. 2023. https://doi.org/10.48550/arXiv.2312.15590.
22. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. 2019. https://doi.org/10.48550/arXiv.1810.04805.

23. González-Carvajal S, Garrido-Merchán EC. Comparing BERT against Traditional Machine Learning Text Classification. J Comput Cogn Eng. 2023. https://doi.org/10.47852/bonviewJCCE3202838.
24. Liu W, Wen Y, Yu Z, Yang M. Large-Margin Softmax Loss for Convolutional Neural Networks. 2017. https://doi.org/10.48550/arXiv.1612.02295.
25. Grootendorst M. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. 2022. https://doi.org/10.48550/arXiv.2203.05794.
26. Samsir S, Saragih RS, Subagio S, Aditiya R, Watrianthos R. BERTopic Modeling of Natural Language Processing Abstracts: Thematic Structure and Trajectory. Jurnal Media Informatika Budidarma. 2023;7(3):1514–1520. https://doi.org/10.30865/mib.v7i3.6426.
27. Palumbo E, Rizzo G, Troncy R, Baralis E, Osella M, Ferro E. Knowledge Graph Embeddings with Node2vec for Item Recommendation. In: Gangemi A, Gentile AL, Nuzzolese AG, Rudolph S, Maleshkova M, Paulheim H, et al., editors. The Semantic Web: ESWC 2018 Satellite Events. Lecture Notes in Computer Science. Cham: Springer International Publishing; 2018. pp. 117–120. https://doi.org/10.1007/978-3-319-98192-5_22.
28. Wang Y, Dong L, Jiang X, Ma X, Li Y, Zhang H. KG2Vec: A Node2vec-Based Vectorization Model for Knowledge Graph. PLoS ONE. 2021;16(3): e0248552. https://doi.org/10.1371/journal.pone.0248552.
29. Bikku T. Multi-layered deep learning perceptron approach for health risk prediction. J Big Data. 2020;7(1):50.
30. Kassner N, Schütze H. BERT-kNN: Adding a kNN Search Component to Pretrained Language Models for Better QA. In: Cohn T, He Y, Liu Y, editors. Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics; 2020. pp. 3424–3430. https://doi.org/10.18653/v1/2020.findings-emnlp.307.
31. Liu W, Zhou P, Zhao Z, Wang Z, Ju Q, Deng H, et al. K-BERT: Enabling Language Representation with Knowledge Graph. Proc AAAI Conf Artif Intell. 2020;34(03):2901–8. https://doi.org/10.1609/aaai.v34i03.5681.
32. Lin Y, Meng Y, Sun X, Han Q, Kuang K, Li J, et al. BertGCN: Transductive Text Classification by Combining GCN and BERT. 2022. arXiv:2105.05727.
33. Smith J, Doe A. Optimized Cancer Subtype Classification and Clustering Using Cat Swarm Optimization and Support Vector Machine Approach for Multi-Omics Data. J Soft Comput Data Min. 2023;15(2):123–35. https://doi.org/10.1007/s00521-023-XXXX.
34. Lee B, Kim S. Enhanced Cancer Subclassification Using Multi-Omics Clustering and Quantum Cat Swarm Optimization. Iraqi J Comput Sci Math. 2022;10(1):45–60. https://doi.org/10.1016/ijcsm.2022.00123.
35. Chen Y, Patel R, Gonzalez M. A Comprehensive Review of Artificial Intelligence Approaches in Omics Data Processing: Evaluating Progress and Challenges. Int J Math Stat Comput Sci. 2021;7(3):200–20. https://doi.org/10.1016/IJMSSC.2021.00456.
36. Williams K, Johnson L. Data Mining and Machine Learning Approaches in Breast Cancer Biomedical Research. J Biomed Inform. 2020;40(4):456–70. https://doi.org/10.1016/j.jbi.2020.04.005.
37. Garcia M, Robinson D. Cancer Classification at the Crossroads. In: Proceedings of the International Conference on Medical Data Science. 2019. pp. 89–98. https://doi.org/10.1145/3359624.3359635.
38. Singh R, Kumar P. From Organs to Algorithms: Redefining Cancer Classification in the Age of Artificial Intelligence. Artif Intell Med. 2021;58(2):101–15. https://doi.org/10.1016/j.artmed.2021.05.007.
39. Chen J, Ramanathan V, Xu T, Martel AL. Cross-Validation Is All You Need: A Statistical Approach To Label Noise Estimation. 2024. arXiv:2306.13990.
40. Yu G, Ye Q, Ruan T. Enhancing Error Detection on Medical Knowledge Graphs via Intrinsic Label. Bioeng (Basel, Switzerland). 2024;11(3):225. https://doi.org/10.3390/bioengineering11030225.
41. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. J Open Source Softw. 2018;3(29):861. https://doi.org/10.21105/joss.00861.
42. McInnes L, Healy J. Accelerated Hierarchical Density Based Clustering. In: 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE; 2017. https://doi.org/10.1109/icdmw.2017.12.
43. Allaoui M, Kherfi ML, Cheriet A. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In: El Moataz A, Mammass D, Mansouri A, Nouboud F, editors. Image and Signal Processing. Cham: Springer International Publishing; 2020. p. 317–25.
44. JUGRAN S, KUMAR A, TYAGI BS, ANAND V. Extractive Automatic Text Summarization using SpaCy in Python & NLP. In: 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). 2021. pp. 582–585. https://doi.org/10.1109/ICACITE51222.2021.9404712.
45. Mondal I, Hou Y, Jochim C. End-to-End NLP Knowledge Graph Construction. 2021. arXiv:2106.01167.
46. Rajabi E, Kafaie S. Knowledge Graphs and Explainable AI in Healthcare. Information. 2022;13(10). https://doi.org/10.3390/info13100459.
47. Rajabi E, Etminani K. Knowledge-graph-based explainable AI: A systematic review. J Inf Sci. 2024;50(4):1019–29. https://doi.org/10.1177/01655515221112844.
48. Li Y. Overview of knowledge graph construction in the field of fully electronic computer interlocking systems. Appl Comput Eng. 2024;71(1):14–18. https://doi.org/10.54254/2755-2721/71/20241632.
49. Sachin U Balvir PSB Mukesh M Raghuwanshi. Node2Vec and Machine Learning: A Powerful Duo for Link Prediction in Social Network. J Electr Syst. 2024. https://doi.org/10.52783/jes.1530.
50. Li Y, Yang B. Quantitative Study of Random Walk Parameters in Node2vec Model. Phys Scr. 2024. https://doi.org/10.1088/1402-4896/ad3eea.
51. Altuntaş V. NodeVector: A Novel Network Node Vectorization with Graph Analysis and Deep Learning. Appl Sci. 2024. https://doi.org/10.3390/app14020775.
52. Kamandar M. Kernel-Based Band Selection for Hyperspectral Image Classification. In: 2023 31st International Conference on Electrical Engineering (ICEE). 2023. pp. 149–153. https://doi.org/10.1109/ICEE59167.2023.10334890.
53. Song H, Zhou Y, Quayson E, Zhu Q, Shen X. Robust Ranking Kernel Support Vector Machine via Manifold Regularized Matrix Factorization for Multi-Label Classification. Appl Sci. 2024;14(2):638. https://doi.org/10.3390/app14020638.
54. Joachims T. Making large-scale SVM learning practical. 1998. Technical Report 1998,28, Dortmund. https://hdl.handle.net/10419/77178.
55. Kumar R, Singh UP, Bali A, et al. Adaptive control of unknown fuzzy disturbance-based uncertain nonlinear systems: application to hypersonic flight dynamics. J Anal. 2024;32:1395–1414. Published: 06 December 2023; Issue Date: June 2024. https://doi.org/10.1007/s41478-023-00687-z.
56. Shreevastava S, Singh S, Tiwari AK, Som T. Different classes ratio and Laplace summation operator based intuitionistic fuzzy rough attribute selection. Iran J Fuzzy Syst. 2021;18(6):67–82. https://doi.org/10.22111/ijfs.2021.6334.
57. Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, et al. Explainability for large language models: a survey. 2023. arXiv:2309.01029.

Publisher's Note