A Survey of Clustering Explainability Robustness Transparency Trustworthy AI Unsupervised Learning and Interpretability

www.surveyx.cn

Abstract

This survey paper provides a comprehensive examination of the interrelated concepts of clustering, explainability, robustness, transparency, trustworthy AI, unsupervised learning, and interpretability within artificial intelligence (AI) and machine learning. Emphasizing the importance of these concepts in creating reliable and ethical AI systems, the paper explores the foundational roles of clustering in unsupervised learning, highlighting various techniques and their applications across domains. Explainability and interpretability are discussed as critical for making AI decisions transparent, particularly in high-stakes areas like healthcare. Robustness is examined in the context of maintaining AI performance under varying conditions, while transparency is linked to the development of trustworthy AI systems. The survey also addresses the integration of these key concepts, identifying potential synergies and conflicts. Challenges such as algorithmic biases, computational complexity, and the need for innovative clustering methods are discussed, alongside advancements that enhance AI system performance and ethical alignment. The paper concludes by identifying future research directions, including refining clustering techniques, integrating additional data types, and exploring privacy-preserving methods. These efforts aim to advance the robustness, interpretability, and applicability of AI systems across diverse domains.

1 Introduction

1.1 Structure of the Survey

This survey is designed to elucidate the interplay among clustering, explainability, robustness, transparency, trustworthy AI, unsupervised learning, and interpretability within artificial intelligence. It commences with an **Introduction** that emphasizes the importance of these concepts in fostering reliable, ethical AI systems aligned with human values. The subsequent **Background and Definitions** section provides in-depth explanations of each core concept, detailing their roles and significance in AI and machine learning.

Following this, the survey explores **Clustering in Unsupervised Learning**, examining various clustering techniques, their applications, and the challenges encountered in this domain, alongside recent advancements. The discussion then transitions to **Explainability and Interpretability**, highlighting the necessity of making AI decisions comprehensible to humans and reviewing diverse methods and frameworks for achieving this objective.

In the section on **Robustness in AI Systems**, the focus is on the capacity of AI systems to sustain performance under diverse conditions, discussing strategies and methodologies for enhancing robustness, including model adaptation techniques. The importance of **Transparency and Trustworthy AI** is analyzed, outlining how transparency fosters the development of trustworthy AI systems and exploring relevant frameworks and guidelines.

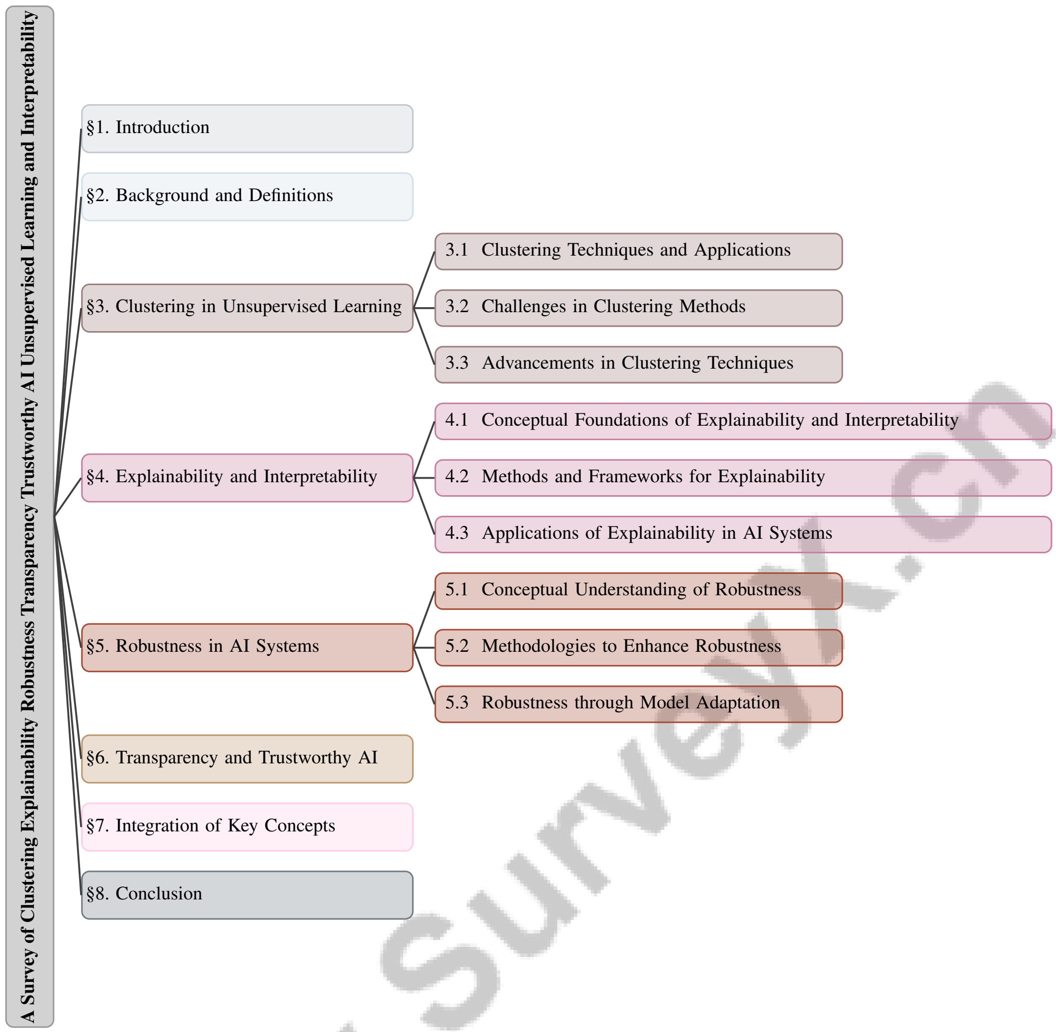


Figure 1: chapter structure

The penultimate section, **Integration of Key Concepts**, examines how clustering, explainability, robustness, transparency, and interpretability can be synergistically integrated to cultivate trustworthy AI systems, while also addressing potential conflicts among these concepts. The survey concludes with a **Conclusion** that synthesizes the key points discussed and reflects on future directions and research opportunities in trustworthy AI, particularly concerning the integration of the surveyed concepts.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Significance of Key Concepts in AI

The advancement of AI systems, both technically and ethically, hinges on key concepts such as clustering, explainability, robustness, transparency, trustworthy AI, unsupervised learning, and interpretability. Clustering, a fundamental aspect of unsupervised learning, is essential for identifying patterns in unlabeled datasets, crucial in domains with scarce labeled data, thus requiring innovative prediction strategies [1]. Traditional methods like k-means lack the adaptability of Bayesian models, necessitating more flexible algorithms [2]. Additionally, challenges such as long-tailed distributions and intra-class variability, particularly in fields like dermatology, demand clustering methods that can generalize effectively [3].

AI-generated, for reference only.

Explainability and interpretability are critical for transparent AI decision-making, fostering trust in high-stakes areas such as healthcare and criminal justice, where comprehending model outputs is vital for accountability [4]. However, existing frameworks often fail to adapt to dynamic problems, leading to suboptimal solutions [5].

Robustness ensures AI systems maintain performance across varying conditions. Inefficient data processing currently hampers the analysis of large datasets, affecting machine learning models and highlighting the need for robust methodologies [6]. Noise and subspace intersections further necessitate robust benchmarks for clustering algorithms [7]. The underperformance of deep learning models on certain data subpopulations, due to biased training datasets, underscores the importance of robust and equitable AI systems [8].

Transparency is foundational for developing trustworthy AI systems. Achieving fairness in clustering, particularly in deep learning, is complex due to potential biases from representation learning, necessitating transparent and fair techniques [9]. Existing computer vision models, often task-specific, lack adaptability across various applications, underscoring the need for flexible and transparent AI models [10].

Integrating these concepts is crucial for addressing the multifaceted challenges of modern datasets. Together, they enhance AI systems' technical capabilities while ensuring alignment with human values, fostering trust and ethical considerations. Developing robust methodologies, such as those improving anomaly detection through complementary data, highlights the importance of these concepts in enhancing AI performance under diverse conditions [11]. Furthermore, addressing limitations in handling non-metric distances and computational complexity in clustering, as seen in the sets-k-means method, underscores the ongoing need for innovation [12].

In recent years, the exploration of clustering techniques in unsupervised learning has gained considerable attention due to its applications across various domains. As illustrated in Figure **??**, this figure depicts the hierarchical structure of these techniques, effectively categorizing both traditional and advanced methods. It not only highlights the challenges faced when dealing with complex datasets but also outlines significant advancements that have been made to enhance clustering performance and adaptability. Such a comprehensive overview is crucial for understanding the evolution of clustering methodologies and their practical implications in contemporary research.

Figure 2: This figure illustrates the hierarchical structure of clustering techniques in unsupervised learning, highlighting traditional and advanced methods, challenges faced, and recent advancements. It categorizes techniques based on their approaches and applications, identifies challenges in handling complex datasets, and outlines significant advancements that improve clustering performance and adaptability.

3 Clustering in Unsupervised Learning

3.1 Clustering Techniques and Applications

Method Name	Methodological Variants	Application Domains	Advanced Considerations
DP-means[2]	Dp-means Algorithm	Dna Sequence Analysis	Fairness, Robustness
CAA[13]	Hierarchical Clustering Methods	Financial Market Predictions	Reduce Noise
nHDP[14]	Nested Hierarchy	Hormone Clustering Analysis	Spatial Dependencies
DFDC[9]	Fair Clustering Algorithms	Visual And Tabular	Fairness Constraints
JBGNN[15]	Spectral Clustering	Citation Datasets	Computational Complexity
CF[10]	Cross-attention Clustering	Vision Tasks	Transparent Decision-making
PCN[3]	Subspace Clustering	Healthcare	Semantic Enhancement
K-modes[16]	K-modes Algorithm	Handwritten Digits	Robustness Against Outliers
SIC[17]	Hierarchical Methods	Image Classification	Semantic Enhancement
FL+HC[18]	Hierarchical Clustering	Image Classification	Privacy-preserving Techniques

Table 1: This table provides a comprehensive overview of various clustering methods, highlighting their methodological variants, application domains, and advanced considerations. The table serves as a resource for understanding the diverse approaches to clustering, their specific applications, and the advanced features they incorporate to address challenges such as fairness, robustness, and computational complexity.

Clustering is a fundamental technique in unsupervised learning, organizing data based on inherent similarities without predefined labels. Traditional methods like k-means are valued for simplicity and computational efficiency, employing Euclidean metrics to define clusters. However, these may not capture the complexities of high-dimensional data, prompting alternative approaches. The DP-means algorithm enhances k-means by integrating Dirichlet process mixture models, allowing dynamic cluster formation and offering greater flexibility [2].

Table 1 presents a detailed comparison of different clustering techniques, illustrating their methodological variants, application domains, and the advanced considerations that enhance their applicability and effectiveness in various contexts.

Subspace clustering is crucial for datasets with inherent subspace structures, identifying features in lower-dimensional spaces, with applications in DNA sequence analysis and image classification [19]. Hierarchical methods, such as Statistically Validated Networks (SVN), enhance financial market trend predictions by leveraging hierarchical structures [13]. The Nested Hierarchical Dirichlet Process (nHDP) models global and local clusters through nested hierarchies, providing a Bayesian nonparametric approach for capturing complex data hierarchies [14].

Advanced clustering methods incorporate fairness and robustness considerations. The Deep Fair Discriminative Clustering method integrates fairness constraints into deep models, ensuring equitable clusters [9]. Similarly, the Just Balance GNN (JBGNN) minimizes imbalance in graph-structured data clustering through a focused loss function [15].

In visual data contexts, methods like ClusterFormer utilize a recurrent cross-attention mechanism to update cluster centers, demonstrating clustering's utility in universal visual representations [10]. The UDIS method identifies subpopulations where deep learning models underperform, highlighting clustering's role in revealing biases [8].

Clustering applications are diverse, including healthcare, where prototypical clustering networks address rare skin condition variability [3]. The K-modes algorithm combines k-means and mean-shift principles for categorical data, enhancing clustering outcomes [16]. Frameworks like C LUST S EG integrate multiple segmentation tasks into a unified neural clustering scheme, improving efficiency in complex scenarios [20]. The Semantic-Enhanced Image Clustering (SIC) method enhances image clustering by mapping images to a semantic space, generating pseudo-labels based on image-semantic relationships [17].

The evolution of clustering methodologies is vital for advancing AI applications across domains. These techniques enhance AI systems' technical capabilities and align them with human values, fostering trust and ethical AI deployment. Hierarchical clustering in federated learning, such as the FL+HC approach, demonstrates clustering's potential to facilitate independent training on specialized models by separating clients based on local updates [18].

3.2 Challenges in Clustering Methods

Clustering methods face challenges, particularly with complex, high-dimensional datasets. Traditional algorithms, like k-means, rely on Euclidean distance metrics, which may not capture intricate structures, leading to suboptimal outcomes [2]. Computational complexity, often requiring $O(n^2)$ time for n data points, poses challenges for real-time clustering, making them impractical for large-scale datasets [21].

The need to specify cluster numbers limits performance, especially with noisy features [16]. Manual interventions, such as the elbow method, introduce error and limit automation [16]. Inadequate cluster separation in feature space can lead to poor performance, highlighting the need for effective differentiation techniques [22].

Deep learning-based clustering struggles with interpretability, often failing to leverage detailed image features, limiting insights [9]. These methods may not account for semantic differences among visually similar images, reducing effectiveness [17]. Hierarchical clustering often lacks the ability to distinguish high-quality from low-quality clusterings, particularly in metric spaces [13].

Non-iid data in federated learning complicates clustering, as varying client data distributions produce unreliable joint model approximations [18]. Estimating global clustering patterns amidst local heterogeneity and absent functional identity information presents further challenges [14].

Figure 3 illustrates these challenges faced in clustering methods, categorizing them into traditional algorithms, deep learning-based approaches, and federated learning. Each category highlights specific issues such as limitations of Euclidean metrics, interpretability concerns, and challenges with non-iid data.

Addressing these challenges requires innovative algorithms capable of managing high-dimensional and noisy data, incorporating nonlinear transformations, and adapting to dynamic data streams. Enhancements in computational efficiency and integration of inter-instance relationships are crucial for advancing clustering methodologies. Future research should expand clustering methods beyond binary classification to accommodate multiclass problems, improving scalability and flexibility. Overcoming limitations of fixed priors and lack of effective inference is essential for enhancing performance [20].

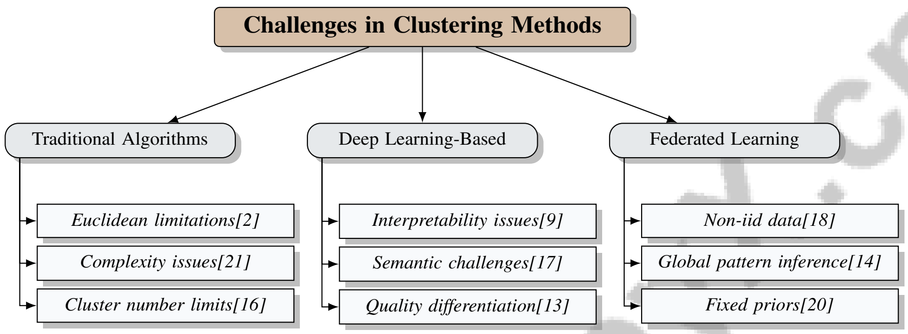


Figure 3: This figure illustrates the challenges faced in clustering methods, categorized into traditional algorithms, deep learning-based approaches, and federated learning. Each category highlights specific issues such as limitations of Euclidean metrics, interpretability concerns, and challenges with non-iid data.

3.3 Advancements in Clustering Techniques

Method Name	Computational Efficiency	Data Type Applicability	Robustness and Adaptability
MNC[23]	Computationally Intensive	Noisy Data	Improved Robustness
K-modes[16]	Computationally Efficient	Nonconvex Data	Robust Against Outliers
JBGNN[15]	Reduced Computational Complexity	Attributed Graphs	Maintain Effective Clustering
LSPP[24]	Fast Local-search	Real-world Datasets	Noisy Environments
AE-l2[22]	-	Mnist And Usps	Noisy Environments
FL+HC[18]	Quicker Convergence	Non-iid Data	Improves Performance
SIC[17]	-	Image Data	Semantic Information

Table 2: Comparative Analysis of Clustering Methods: This table presents a detailed comparison of various clustering methodologies, highlighting their computational efficiency, data type applicability, and robustness. The methods analyzed include MNC, K-modes, JBGNN, LSPP, AE-l2, FL+HC, and SIC, each offering unique advantages in handling complex datasets.

Recent advancements in clustering methodologies have improved complex dataset analysis and interpretation, offering innovative solutions to longstanding challenges. The max-norm introduction as a tighter convex relaxation for clustering provides superior recovery guarantees compared to the trace-norm [23], enhancing clustering accuracy in complex scenarios.

The K-modes algorithm exemplifies computational efficiency by identifying K modes corresponding to meaningful clusters, improving outcomes for categorical data [16]. This approach is advantageous in applications involving categorical data, providing a robust alternative to traditional methods.

In graph-structured data, the Just Balance GNN (JBGNN) achieves competitive performance, excelling in efficiency and convergence speed compared to existing Graph Neural Network methods [15]. This advancement underscores graph-based models' potential to enhance clustering processes, particularly within complex network data.

A fast local-search algorithm operating in $O(nk^2)$ time represents a significant leap, achieving a bicriteria approximation and improving upon previous methods [24]. This algorithm addresses computational challenges associated with large-scale datasets, making clustering more feasible and scalable.

Innovations in subspace clustering introduce a less computationally demanding method that maintains robust performance in noisy environments, providing a benchmark for evaluating clustering algorithms under challenging conditions [7]. This approach is instrumental in applications where data experiences significant noise and variability.

Integrating l_2 normalization during auto-encoder training improves clustering accuracy and anomaly detection performance [22]. This technique enhances robustness in unsupervised learning scenarios where data labeling is infeasible.

In federated learning, introducing a clustering step allows more effective training on non-iid data, enhancing adaptability and accuracy [18]. This innovation facilitates handling heterogeneous data distributions.

The Semantic-Enhanced Image Clustering (SIC) method performs clustering in image and semantic spaces, using pseudo-labels generated from image-semantic relationships [17]. This dual-space approach enriches image clustering depth and granularity, yielding more meaningful insights into visual data.

These advancements contribute to clustering techniques' evolution, setting new standards for future research and applications. The authors address challenges of managing intricate and high-dimensional datasets by introducing innovative techniques that enhance robustness, efficiency, and adaptability. Their approach includes leveraging auxiliary information within documents to improve outcomes, using advanced models for microclustering tailored to applications like entity resolution, and developing a core-set strategy for the sets-k-means problem to ensure effective clustering across various metrics. Table 2 provides a comprehensive comparison of recent advancements in clustering techniques, evaluating their computational efficiency, data type applicability, and robustness, thereby illustrating their contributions to the field. These innovations pave the way for more effective analysis and classification of complex data structures [25, 26, 12, 27].

4 Explainability and Interpretability

The progress of artificial intelligence (AI) is closely linked to explainability and interpretability, essential for ensuring AI systems are effective and user-friendly. Grasping these concepts is crucial for making AI models transparent and understandable. This section delves into these foundational elements, setting the stage for a detailed exploration of methodologies and frameworks that enhance AI system explainability and interpretability.

4.1 Conceptual Foundations of Explainability and Interpretability

Explainability and interpretability are vital for connecting complex algorithmic functions with human understanding, promoting accountability and trust in AI applications [28]. Explainability clarifies AI models' decision-making processes for non-experts, often through visualization tools and simplified representations. For example, LCS-DIVE employs feature-tracking scores and IF:THEN rules to illustrate feature importance, aiding comprehension of underlying data patterns [28].

In Figure 4, the conceptual foundations of explainability and interpretability in AI are visually represented, highlighting key components such as explainability, interpretability, and recent advancements, including the AI-Interpret framework and Semantic-Enhanced Clustering. This figure serves to reinforce the discussion by providing a visual summary that encapsulates the intricate relationships between these concepts.

Interpretability, on the other hand, focuses on model transparency and the ability to show how inputs influence outputs, balancing accuracy with generalizable explanations across the feature space [4]. Techniques like mutual information-based clustering reveal significant patterns in complex datasets, providing insights in genetics and medicine [29].

Recent advancements have introduced frameworks to enhance explainability and interpretability. The AI-Interpret framework clusters complex policy demonstrations to extract high-performing, interpretable decision rules, improving user comprehension [30]. Cluster stability across samples indicates validity, suggesting that consistent clusters are more interpretable [31]. This underscores the need for robust methodologies to achieve reliable clustering outcomes.

Additionally, the Semantic-Enhanced Image Clustering (SIC) method integrates semantic information to boost neighborhood consistency and prediction confidence, enhancing clustering performance [17]. Such enriched data representations are crucial for deriving meaningful insights from clustering.

The drive for explainability and interpretability stems from the need for AI systems to align with human values and ethics. As AI progresses, developing methods that enhance transparency and understanding will be crucial for integrating AI into various domains while maintaining user trust and engagement. Algorithm and hyperparameter choices significantly impact clustering outcomes, highlighting the need for careful design in interpretable AI systems [32].

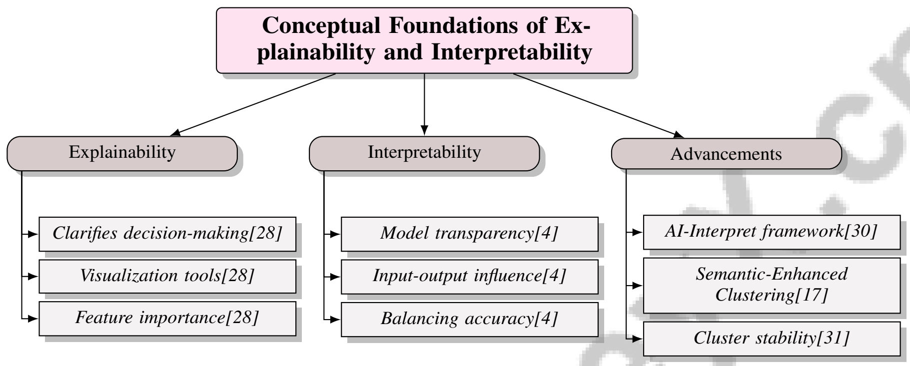


Figure 4: This figure illustrates the conceptual foundations of explainability and interpretability in AI, highlighting key components such as explainability, interpretability, and recent advancements, including AI-Interpret framework and Semantic-Enhanced Clustering.

4.2 Methods and Frameworks for Explainability

Achieving explainability in AI involves methodologies and frameworks that enhance AI systems' comprehensibility for users. Techniques such as reinforcement learning for interpretable decision rules, unsupervised methods for visual concept extraction in deep learning models, and interactive machine learning with human input contribute to making AI decisions more accessible [33, 34, 35, 30, 26]. These approaches demystify complex models' decision-making, fostering trust and informed decision-making.

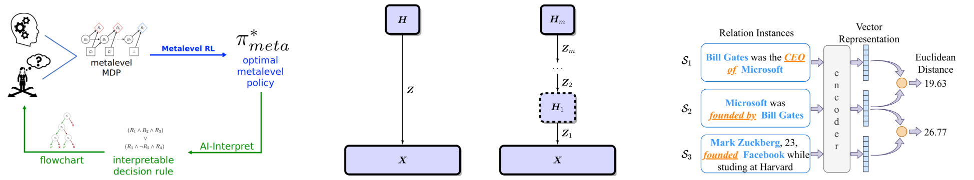
The AI-Interpret framework simplifies intricate decision-making policies into understandable flowcharts, aiding stakeholders in navigating complex decisions through clear visual representations [30]. LCS-DIVE uses scikit-ExSTraCS to model data via rules and feature-tracking scores, elucidating feature-outcome relationships to enhance transparency [28].

In deep learning, CNN-INTE interprets convolutional neural networks (CNNs) by using a two-level clustering algorithm to generate meta-level training data, enhancing interpretability through random forests as base learners [4]. This highlights the importance of understanding neural networks' internal representations for transparency.

Integrating differential privacy into clustering algorithms like K-means allows stakeholders to query data while preserving privacy, balancing transparency with privacy concerns [36]. These methods are essential in sensitive data contexts, ensuring AI systems remain interpretable and secure.

Moreover, the Maximum Activation Groups Extraction (MAGE) and Multiscale Interpretable Visualization (Ms-IV) methods provide a comprehensive framework for understanding model decisions. MAGE identifies feature combinations forming meaningful concepts, while Ms-IV visualizes these concepts to highlight their significance in decision-making [33].

These methodologies represent significant advancements toward explainability in AI by leveraging innovative approaches such as AI-Interpret for transforming opaque decision policies into interpretable rules, enhancing text mining through auxiliary information, utilizing the CLARITY method for comparing heterogeneous datasets, and developing unsupervised techniques for generating representative interpretations in CNNs [37, 26, 35, 30]. These developments address the critical need for transparency and understanding in AI systems, ensuring effective integration into various domains while maintaining user trust and engagement.



(a) Optimal Metalevel Policy for Decision Making in Complex Systems[30]

(b) Sequential Learning with Hierarchical Representation Learning[38]

(c) Relation Instances and Vector Representation[39]

Figure 5: Examples of Methods and Frameworks for Explainability

As illustrated in Figure 5, explainability and interpretability are increasingly significant in AI and machine learning, particularly as models grow more complex. The figure showcases methods and frameworks that enhance the explainability of decision-making processes within complex systems. The first example, "Optimal Metalevel Policy for Decision Making in Complex Systems," demonstrates the transformation of a decision-making flowchart into an interpretable decision rule, generating an optimal metalevel policy symbolized by "meta." This underscores the importance of deriving clear policies from intricate decision processes. The second example, "Sequential Learning with Hierarchical Representation Learning," illustrates a structured flowchart emphasizing the hierarchical nature of learning, depicting connections between different learning stages. Finally, the "Relation Instances and Vector Representation" example shows a neural network model encoding relation instances into vector representations, facilitating the calculation of Euclidean distances between these instances. This model aids in understanding relational dynamics between entities, contributing to the interpretability of complex data relationships. Collectively, these examples underscore ongoing efforts to develop methodologies that enhance model performance while ensuring transparency and interpretability for human users [30, 38, 39].

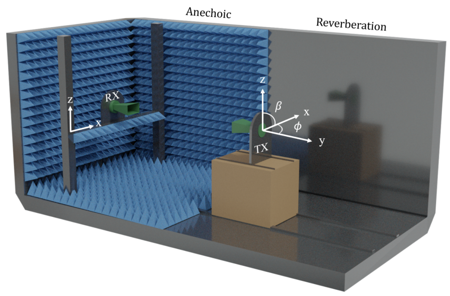
4.3 Applications of Explainability in AI Systems

Explainability in AI systems is crucial for enhancing human understanding and fostering trust, particularly in complex models like deep neural networks. Its application spans various domains, aiding in demystifying AI decision-making processes. In web log analysis, for instance, explainability techniques group similar user navigation patterns, providing insights into user behavior and preferences [40]. This application enhances understanding of user interactions and tailors web experiences to better meet user needs.

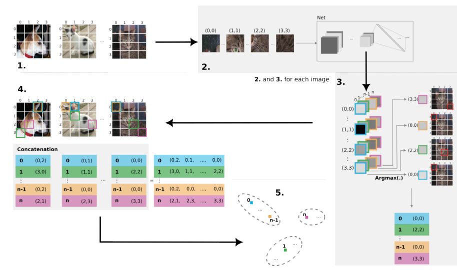
In deep learning, CNN-INTE has proven instrumental in providing global interpretations of CNNs, extracting meaningful insights without sacrificing accuracy, thus serving as a reliable tool for understanding model behavior in applications ranging from image recognition to natural language processing [4]. By elucidating CNNs' internal workings, CNN-INTE enhances the transparency and accountability of AI systems.

Moreover, MAGE and Ms-IV exemplify the application of explainability in AI, enhancing the interpretability of CNNs through clear visualizations of model knowledge [33]. MAGE identifies feature combinations forming meaningful concepts, while Ms-IV visualizes these concepts, highlighting their significance in the model's decision-making process. This dual approach improves interpretability and aids in identifying potential biases and areas for improvement.

Integrating explainability into AI systems is essential for ensuring these technologies operate transparently and ethically. By equipping stakeholders with clear insights into AI models' decision-making processes, explainability enhances trust and promotes informed decision-making, particularly in fields like healthcare. Decision aids such as flowcharts and decision trees can mitigate biases and improve outcomes. Recent advancements like AI-Interpret, which transforms opaque policies into interpretable decision rules, further support this goal by enabling clearer communication of complex AI strategies. Additionally, methods like MAGE and Ms-IV enhance the interpretability of deep learning models, making it easier for non-experts to grasp the model's decision logic. These developments signify a vital step toward fostering accountability and transparency in AI-driven decision-making processes [37, 35, 33, 30]. As AI continues to permeate various sectors, the application of explainability remains a critical factor in the successful deployment and acceptance of AI technologies.



(a) Anechoic Reverberation Chamber[41]



(b) DeepLabv3+ with Spatial Attention for Semantic Segmentation[33]

Figure 6: Examples of Applications of Explainability in AI Systems

As shown in Figure 6, explainability and interpretability are vital for understanding and trusting AI systems, particularly in applications requiring transparent decision-making processes. Two notable examples are illustrated, highlighting diverse applications. The first example, an anechoic reverberation chamber, demonstrates the use of explainability by providing a controlled environment to observe and interpret acoustic phenomena. The second example involves using DeepLabv3+ with spatial attention for semantic segmentation, a process that divides an image into meaningful parts for easier analysis. This method enhances explainability by visually representing how AI models segment images and identify objects, allowing for more intuitive interpretation of results. Both examples underscore the importance of making AI systems transparent and interpretable, ultimately fostering greater trust and usability across various applications [41, 33].

5 Robustness in AI Systems

Robustness is a critical attribute of AI systems, pivotal for their reliability and efficacy in real-world applications. As AI systems are deployed across diverse environments, understanding the principles underpinning their robustness is essential for developing resilient models capable of performing well under varying data conditions and potential adversarial challenges. The following subsection delves into the conceptual foundations of robustness, emphasizing its significance and the theoretical frameworks guiding its application in AI.

5.1 Conceptual Understanding of Robustness

Robustness ensures AI systems maintain stability and reliability across a spectrum of challenging and unpredictable conditions. This feature is vital for performance under uncertainty, exemplified by mean robust optimization, which balances computational efficiency with conservative decision-making. Advancements in interpretable decision-making and clustering-based anomaly detection highlight the necessity of designing AI systems that adapt to diverse contexts while providing transparent outputs [42, 30, 43, 26, 44]. A robust AI system is characterized by consistent performance across various scenarios, enhancing trustworthiness in critical domains such as healthcare, finance, and autonomous systems.

To further illustrate this concept, Figure 7 depicts the hierarchical structure of robustness in AI systems, highlighting key aspects such as performance under uncertainty, generalization capabilities, and defense against attacks. Robustness also extends to the generalization capabilities of AI models across different datasets and conditions. For instance, the AID-C method improves document clustering and classification accuracy by integrating auxiliary information, thereby enhancing robustness in data processing [26]. Similarly, the DAD method balances positive and negative relationships, ensuring stable alliances against external threats, further exemplifying robustness [45].

AI-generated, for reference only.

Addressing clustering algorithms' vulnerability to adversarial attacks is another critical aspect of robustness. Research shows attackers can exploit weaknesses in these algorithms, necessitating robust defenses in unsupervised learning contexts [46]. The DBHT method in hierarchical clustering enhances robustness by utilizing the topological properties of Planar Maximally Filtered Graphs (PMFGs), improving clustering and hierarchy detection [47]. Additionally, l2 normalization bolsters clustering performance by ensuring features are suitable for Euclidean distance metrics [22]. The LoR method underscores the importance of robust parameter choices in clustered regression frameworks [48].

Theoretical perspectives on robustness consider factors like fixed-parameter tractability and the implications of the Gap-Exponential Time Hypothesis (Gap-ETH) on approximation limits, enriching the understanding of robustness in AI systems [49]. These theoretical frameworks underscore robustness's foundational role in trustworthy AI systems, ensuring adaptability to real-world complexities. Advancing robust methodologies is vital for the ethical implementation of AI technologies, facilitating innovative models that enhance decision-making, optimize clustering, and ensure transparency, ultimately fostering greater user trust [34, 50, 30, 26, 44].

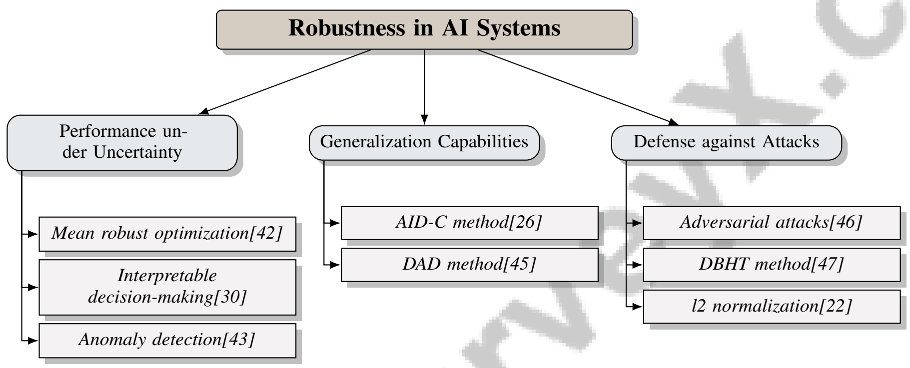


Figure 7: This figure illustrates the hierarchical structure of robustness in AI systems, highlighting key aspects such as performance under uncertainty, generalization capabilities, and defense against attacks.

5.2 Methodologies to Enhance Robustness

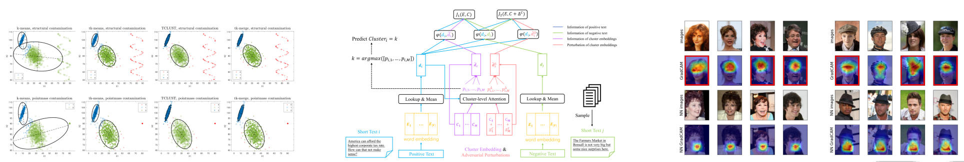
Enhancing robustness in AI systems is crucial for reliable performance across diverse and potentially adverse conditions. Various methodologies have emerged, each offering unique strategies to bolster the stability and resilience of AI models. Notably, cluster-level attention and robust adversarial training facilitate a more integrated learning process, enhancing clustering algorithms by focusing on informative data points and mitigating adversarial perturbations [44].

Integrating clinical and digital data in patient clustering significantly improves clustering coherence and recommendation accuracy, demonstrating the benefits of leveraging diverse data sources to enhance robustness [51]. Entropy-based regularization in Supervised Fuzzy Partitioning (SFP) enhances model flexibility and performance by managing memberships and feature weights, thereby improving robustness in complex environments [52].

In semi-supervised learning, robust estimation techniques that minimize the impact of misassigned data points lead to improved parameter estimation and prediction accuracy, emphasizing the importance of addressing data noise [1]. Mean Robust Optimization (MRO) constructs uncertainty sets around clustered data to minimize worst-case costs while ensuring probabilistic guarantees of constraint satisfaction, thereby enhancing decision-making robustness under uncertainty [42].

The Just Balance Graph Neural Network (JBGNN) achieves strong clustering performance with reduced computational complexity, showcasing the potential of graph-based models to enhance AI robustness by efficiently managing complex network data [15]. The Adversarial Deep Anomaly Unit (ADAU) employs adversarial deep learning to align data distributions across different units, facilitating improved anomaly detection performance [11]. The LoR method enhances clustered regression frameworks by selecting subsets of similar experiments, underscoring the significance of robust parameter selection [48].

Federated Learning with Hierarchical Clustering (FL+HC) allows for model specialization based on local data distributions, improving accuracy and communication efficiency compared to traditional federated learning methods [18]. Collectively, these methodologies contribute to enhancing robustness in AI systems, ensuring models maintain performance across diverse and challenging conditions. The continuous development of robust methodologies is critical for deploying reliable and trustworthy AI systems, encompassing improved text mining techniques for data classification and advanced reinforcement learning methods for creating interpretable decision aids. The integration of these methodologies aims to enhance the effectiveness of AI applications, ultimately leading to better decision-making outcomes across various domains [26, 30].



(a) Comparison of clustering methods under structural and pointmass contamination[53]

(b) Adversarial Perturbation for Cluster Embedding in Text Classification[44]

(c) GradCAM and NN GradCAM: Visualizing the Importance of Different Parts of the Face in Image Classification[8]

Figure 8: Examples of Methodologies to Enhance Robustness

As illustrated in Figure 8, robustness is a critical attribute in AI, ensuring systems maintain performance despite adversities such as noise, adversarial attacks, or unexpected data variations. The methodologies to enhance robustness are diverse and innovative. For instance, the comparison of clustering methods under structural and pointmass contamination evaluates techniques like k-means, tk-means, TCLUST, and tk-merge for their ability to handle noise in multi-cluster datasets, emphasizing the importance of appropriate clustering strategies. Adversarial perturbation for cluster embedding in text classification enhances resilience against adversarial inputs, demonstrating the need for strategic perturbation techniques to safeguard model integrity. Lastly, GradCAM and NN GradCAM visualize crucial facial regions affecting classification outcomes, enhancing model interpretability and robustness. These examples collectively illustrate the multifaceted approaches required to strengthen AI robustness across various domains [53, 44, 8].

5.3 Robustness through Model Adaptation

Model adaptation is vital for enhancing AI robustness, allowing systems to adjust to varying data distributions and environmental conditions. This adaptability is particularly significant in dynamic environments, where maintaining performance amid changes in input data characteristics is essential. One effective approach is the integration of adversarial training techniques, which bolster model resilience against adversarial attacks by facilitating the learning of robust feature representations [44].

In anomaly detection, the Adversarial Deep Anomaly Unit (ADAU) exemplifies model adaptation by utilizing adversarial deep learning to align data distributions across different units, improving anomaly detection performance in diverse settings [11]. This method highlights the importance of adapting models to leverage varied data sources and conditions.

The Mean Robust Optimization (MRO) framework further exemplifies model adaptation by constructing uncertainty sets around clustered data, minimizing worst-case costs while ensuring probabilistic guarantees of constraint satisfaction, thus enhancing decision-making robustness under uncertainty [42]. In federated learning scenarios, hierarchical clustering techniques enable model specialization based on local data distributions, improving accuracy and communication efficiency compared to traditional federated learning methods [18].

Model adaptation is crucial for maintaining high performance across challenging conditions, including effectively handling sparse data representations in tasks like short text clustering and integrating semi-supervised learning techniques that leverage both labeled and unlabeled data. By employing adversarial training and innovative frameworks like the mixture of experts model, AI systems can be optimized for complex and heterogeneous environments, significantly enhancing functionality and reliability [44, 1]. Incorporating techniques such as adversarial training, robust optimization, and hierarchical clustering enables AI models to adapt to varying data landscapes, ensuring reliability and trustworthiness in real-world applications.

6 Transparency and Trustworthy AI

6.1 Role of Transparency in Clustering Algorithms

Transparency in clustering algorithms is essential for ensuring the understandability and accountability of AI systems, particularly in large-scale applications where the complexity of clustering can obscure the rationale behind data groupings. Traditional clustering methods often rely on assumptions that may not hold true across diverse datasets, necessitating a clear understanding of cluster formation criteria to evaluate result validity and quality. As these techniques increasingly influence significant decision-making processes, clarity in data organization and categorization is imperative to uphold fairness and accuracy [54, 55, 56, 57, 26].

Enhancing transparency in clustering algorithms can improve their interpretability and trustworthiness. Scalable algorithms that operate in constant rounds and require sublinear local space, as demonstrated by Altieri et al., ensure scalability for big data while maintaining processing clarity [58]. In financial contexts, methods analyzing data streams to identify regime changes and classify market behaviors provide clear insights into market dynamics [59]. Additionally, integrating differential privacy mechanisms allows stakeholders to query data while preserving privacy, balancing transparency and data protection [26, 55]. These methods foster trust and accountability by elucidating privacy-preserving techniques.

Enhancing transparency through scalable, interpretable, and privacy-preserving methods is crucial for fostering trust in AI systems. Techniques like cryptographic protocols for secure multi-party computation and local differential privacy mechanisms contribute to effective knowledge discovery while ensuring data privacy [57, 26, 55, 60]. As AI evolves, transparency remains a critical focus, enabling the deployment of trustworthy and ethical technologies.

6.2 Differential Privacy and Transparency

Integrating differential privacy into clustering algorithms significantly enhances transparency, ensuring processes and outcomes are interpretable and secure. This approach employs structured noise mechanisms to protect individual data points from privacy breaches while maintaining the utility of insights derived from the data. In distributed settings, privacy-preserving techniques like local differential privacy and cryptographic protocols enhance methods such as k-means and DBSCAN, enabling effective data analysis without compromising sensitive information [55, 60]. Balancing privacy and interpretability is crucial for achieving transparency in AI systems.

Advancements such as differentially private mechanisms that add structured noise to clustering results improve privacy protections without significantly degrading utility, ensuring data remains useful for analysis and decision-making [36]. The kCluster method exemplifies this by achieving (,)-local differential privacy, providing robust privacy guarantees that enhance clustering process interpretability [60]. These innovations foster trust by allowing users to engage confidently with AI systems while respecting individual privacy.

Differential privacy's role in achieving transparency is underscored by providing clear and accountable mechanisms for data protection. By incorporating advanced privacy-preserving techniques into clustering algorithms, AI systems can securely analyze distributed datasets without compromising sensitive information. This integration enhances clustering results' interpretability, builds user trust, and promotes ethical AI implementation, ensuring compliance with privacy regulations and safeguarding individual data rights [55, 60]. As AI evolves, developing and implementing differential privacy mechanisms remain critical for transparent and trustworthy AI systems.

6.3 Frameworks for Fair and Transparent Clustering

Developing frameworks that promote fairness and transparency in clustering is essential for ensuring ethical AI systems accountable for their decision-making processes. Fairness in clustering algorithms is particularly crucial in scenarios where data-driven decisions significantly impact society, such as healthcare, finance, and criminal justice. Achieving fairness requires a comprehensive approach addressing biases from data imbalances and algorithm design, including identifying and applying appropriate fairness metrics derived from expert demonstrations and considering multiple sensitive attributes for equitable representation within clusters [57, 61, 9, 62].

Fairness-aware clustering algorithms incorporate fairness constraints into the clustering process, ensuring resulting clusters do not disproportionately disadvantage any group. For instance, the Deep Fair Discriminative Clustering method integrates fairness constraints into deep clustering models, ensuring equitable clusters concerning protected status variables [9]. This highlights the importance of embedding fairness considerations directly into algorithmic frameworks to mitigate bias and promote equitable outcomes.

Transparency is achieved through methods rendering the clustering process understandable and accountable. Integrating differential privacy mechanisms enhances transparency by allowing stakeholders to query data while preserving privacy [36]. This approach balances interpretability with privacy concerns, ensuring clustering processes remain transparent and secure. Frameworks utilizing hierarchical clustering techniques, such as Federated Learning with Hierarchical Clustering (FL+HC), enhance transparency by accommodating heterogeneous data and allowing specialization in model training based on local distributions [18].

Establishing comprehensive frameworks to enhance fairness and transparency in clustering methodologies is essential for fostering ethical and accountable AI systems, especially as they increasingly influence critical decisions affecting individuals' lives. Recent advancements in fair clustering research emphasize expanding normative principles, incorporating multiple sensitive attributes, and refining fairness metrics based on expert demonstrations. These efforts aim to address existing method limitations and ensure clustering processes effectively group similar entities while representing diverse sensitive attribute groups equitably, mitigating biases and promoting inclusivity in AI applications [57, 61, 9, 62]. By integrating fairness constraints and transparency-enhancing techniques into clustering algorithms, AI systems can provide equitable and understandable outcomes, fostering trust and facilitating their integration across various domains. The pursuit of fairness and transparency remains a critical focus as AI evolves, enabling the deployment of trustworthy and ethical technologies.

7 Integration of Key Concepts

7.1 Synergies Between Clustering and Explainability

The integration of clustering and explainability in AI enhances model interpretability and transparency. Clustering organizes datasets into groups based on intrinsic similarities, aiding data interpretation and supporting applications like statistics, image segmentation, and community discovery. The effectiveness of clustering depends on the algorithm chosen, each with unique strengths and weaknesses [50, 63, 64]. When combined with explainability methods, clustering deepens the understanding of data patterns, thus enhancing AI model interpretability.

The LCS-DIVE framework exemplifies this by automating model interpretation and clarifying complex data associations [28]. By grouping similar data points, LCS-DIVE simplifies visualizing intricate data structures, improving feature relationship accessibility. This enhances model interpretability and helps identify significant patterns not evident through traditional analysis.

The synergy between clustering and explainability also enhances AI system transparency. By organizing data into coherent clusters, these methods generate simplified representations that can be communicated to non-expert stakeholders. In domains like healthcare and finance, understanding AI-driven decision-making rationales is crucial for accountability and trust. Algorithms like AI-Interpret transform complex decision-making policies into interpretable rules, enhancing human decision-making and addressing transparency issues in traditional black-box models [34, 35, 30, 26, 28].

Clustering and explainability frameworks also enhance bias detection in AI models. These methods facilitate identifying underlying data distributions, improving decision-making interpretability and promoting equitable outcomes in applications where AI influences critical human decisions [57, 65, 30]. Analyzing cluster compositions can uncover disparities in data subgroup treatment, fostering fairer AI systems. This integration underscores the necessity of aligning AI technologies with human values and ethical considerations.

The combination of clustering techniques and explainability frameworks significantly enhances AI systems' interpretability, transparency, and fairness. This synergy improves clustering processes' reliability and supports developing interpretable decision-making tools, such as decision trees and flowcharts. By leveraging advanced clustering algorithms and methods like normalized crowd agreement estimation, AI systems can provide clearer insights into their decision-making processes, fostering greater user trust and facilitating responsible AI deployment across various fields, including healthcare and data analysis [37, 66, 30]. As AI evolves, integrating these concepts will remain a focal point for developing trustworthy and ethical technologies.

7.2 Enhancing Robustness Through Clustering

Clustering techniques enhance AI system robustness by providing structured methodologies for data organization and interpretation, essential for maintaining performance across diverse conditions. In the financial sector, clustering groups traders based on behavior, facilitating expert predictions' effective aggregation and improving prediction system robustness [13].

The Just Balance Graph Neural Network (JBGNN) exemplifies clustering's role in enhancing robustness by delivering efficient clustering solutions that sustain performance under varying conditions. JBGNN streamlines the clustering process, ensuring AI systems can adapt to dynamic environments while maintaining computational efficiency [15]. This adaptability is crucial for AI models operating in complex network data, where robustness is imperative.

The Nested Hierarchical Dirichlet Process (nHDP) method enhances robustness by inferring global clusters from locally distributed data, ensuring reliable and consistent clustering outcomes even in distributed settings [14]. This capability is vital for AI systems navigating heterogeneous data landscapes, ensuring reliability and trustworthiness.

Clustering serves as a foundational tool for enhancing AI systems' robustness, enabling them to sustain performance across diverse conditions. The continuous evolution of clustering methodologies is essential for the reliable and ethical application of AI technologies, ensuring that clustering techniques improve decision-making processes that significantly impact human lives while aligning with broader human values and ethical standards. Recent research highlights the importance of fair clustering, addressing fairness in data grouping and emphasizing diverse normative principles. Additionally, integrating deep learning and semi-supervised methods optimizes clustering performance, facilitating more effective data representation and improved outcomes across various applications [57, 26, 50, 67].

7.3 Integrating Interpretability with Clustering and Robustness

Integrating interpretability with clustering and robustness in AI systems enhances transparency, reliability, and trustworthiness, enabling better decision-making through clear model behavior explanations, addressing data sparsity and noise challenges, and facilitating user-friendly decision aids for complex scenarios [26, 44, 65, 30]. Interpretability allows stakeholders to comprehend AI decision-making processes, ensuring outcomes align with human values and ethical considerations. By merging interpretability with clustering and robustness, AI systems yield nuanced insights into complex data, fostering comprehensible and resilient models.

Clustering techniques organize data into coherent groups, significantly enhancing AI model interpretability. By revealing intrinsic data structures, clustering helps identify patterns and relationships not immediately apparent through traditional analysis. This capability is particularly valuable in domains like healthcare and finance, where understanding AI-driven decision rationales is crucial for accountability and trust. Integrating clustering with interpretability methods, such as those in the LCS-DIVE framework, automates model interpretation and clarifies complex data associations [28].

Robustness ensures AI systems maintain consistent performance across diverse conditions. Integrating interpretability with robustness allows AI models to provide transparent behavior explanations under varying scenarios, enhancing stakeholder confidence in reliability. Adversarial training techniques bolster model resilience against adversarial attacks, enabling the learning of robust feature representations that are both interpretable and reliable [44].

Moreover, this integration aids in identifying and mitigating biases within AI models. By analyzing cluster compositions and their impact on model outcomes, practitioners can detect disparities in data subgroup treatment, promoting the development of fairer AI systems. This capability highlights the critical need for an integrated approach combining advanced AI methodologies with ethical frameworks and societal values, ensuring AI technologies enhance decision-making and data analysis while operating transparently and equitably [37, 26, 44, 30].

Integrating interpretability with clustering and robustness significantly enhances AI systems' transparency, reliability, and fairness. This synthesis fosters user trust and facilitates AI technology adoption across domains, as demonstrated by methods like CLARITY, which quantifies consistency across heterogeneous datasets, and AI-Interpret, which simplifies complex decision-making policies into interpretable rules. These advancements help mitigate biases in human decision-making and ensure AI systems are effectively understood and utilized in fields ranging from healthcare to economics [30, 37]. As AI evolves, developing methods that effectively combine these concepts will remain critical for deploying trustworthy and ethical technologies.

7.4 Challenges and Conflicts in Concept Integration

Integrating clustering, explainability, robustness, transparency, and interpretability in AI systems presents challenges and conflicts that must be addressed to ensure trustworthy and ethical technologies. A significant challenge is the reliance on clustering indices, complicating merging clustering methods with models not utilizing clustering, potentially limiting clustering-based approaches' applicability in broader AI contexts [68].

The quality of demonstrations in learning processes impacts the accuracy of inferred fairness constraints. Biased or incomplete demonstrations may lead to unfair AI systems, underscoring the importance of high-quality, representative data during fairness integration with clustering and other AI concepts [61]. This challenge highlights the need for robust mechanisms to assess and mitigate biases in training data.

Another challenge arises from potential overestimation of cluster numbers in noisy data scenarios, as seen in the LCS-DIVE framework. This limitation can hinder effective explainability integration with clustering, particularly in high-variability environments [28]. Addressing this issue requires developing methods to accurately determine optimal cluster numbers in diverse data landscapes.

The selection of algorithm-hyperparameter combinations poses challenges, as current methods may not account for all configurations, potentially missing optimal setups that could enhance concept integration [32]. This limitation emphasizes the need for comprehensive exploration of algorithmic parameters to identify effective configurations for integrating clustering, robustness, and interpretability.

Determining the optimal number of clusters for k-means algorithms used in interpretability processes remains a challenge, impacting the clarity and usefulness of interpretations, especially in complex datasets where intrinsic structures are not readily apparent [4]. Overcoming this challenge necessitates adaptive methods that dynamically adjust clustering parameters based on data characteristics.

Integrating clustering, explainability, robustness, transparency, and interpretability presents a complex landscape of challenges, particularly as researchers strive to develop methods that transform opaque decision-making policies into interpretable formats. The AI-Interpret algorithm exemplifies this, utilizing techniques like imitation learning and clustering to enhance human decision-making across domains. Additionally, the CLARITY method addresses integrating heterogeneous datasets by quantifying consistency and identifying inconsistencies, showcasing the need for robust approaches that accommodate qualitative data differences while maintaining interpretability and transparency [37, 30]. Addressing these issues is essential for ensuring AI technologies are reliable and aligned with ethical principles, fostering trust and facilitating adoption across various domains. As AI evolves, ongoing research and innovation will be crucial for overcoming these challenges and achieving seamless integration of these key concepts.

8 Conclusion

8.1 Future Directions and Research Areas

The development of trustworthy AI systems hinges on sustained research efforts aimed at enhancing their robustness, interpretability, and adaptability across various domains. A significant focus for future research involves refining clustering techniques, especially for handling large-scale datasets with applications in fields such as image processing, bioinformatics, and geology. This includes optimizing algorithms to improve performance with small sample sizes and extending their applicability to sectors like healthcare and environmental studies. The integration of embedding learning techniques holds promise for boosting clustering performance in real-world scenarios.

Further investigation into clustering algorithms should consider the development of incremental versions of Supervised Fuzzy Partitioning (SFP) and kernelized adaptations to improve computational efficiency and streamline hyperparameter tuning. Additionally, validating clustering methods across diverse data types and public datasets could enhance their generalizability, particularly in patient profiling and healthcare contexts.

Exploring lightweight architectures and knowledge distillation techniques offers a promising pathway, particularly for unsupervised re-identification tasks. Advancements in data preprocessing and the exploration of methods such as Dynamic Level Analysis (DLA) across various domains remain critical areas for future research.

In the realm of interactive machine learning, applying the iML approach to NP-hard problems like protein folding, and integrating sophisticated methods for incorporating human feedback into machine learning processes, will be crucial. Enhancing clustering processes by integrating additional relevant features and validating methods across varying conditions is essential for bolstering AI reliability.

Research into complex datasets should explore the effects of different kernel functions on clustering performance and assess the robustness of frameworks like Mean Robust Optimization (MRO) in challenging scenarios. Investigating non-Gaussian noise models and unbalanced mixture models presents opportunities to address the statistical-to-computational gap. Enhancements to algorithms based on k-means principles, spectral relaxations, and extensions to Bayesian nonparametric processes could further advance the field.

Future research should also focus on improving clustering algorithms, potentially replacing k-means with DBSCAN, and exploring decision sets for enhanced interpretability. The application of JBGNN in tasks such as graph pooling for deep GNN architectures and adapting methods for unbalanced clusters are promising areas of exploration.

Additional development areas include exploring local bandwidths for each data point and applying K-modes to nonconvex low-dimensional manifolds. Optimizing running times and developing scalable algorithms for various clustering objectives are potential research directions. Furthermore, refining methods to manage larger, more complex networks and examining dynamic relationship changes over time are critical considerations.

Efforts to refine clustering mechanisms and explore additional segmentation tasks can enhance capabilities. The development of methods for generating higher-quality pseudo-labels and extending theoretical results to improve self-supervised learning applications are promising directions. Finally, examining the impact of privacy-preserving techniques on clustering effectiveness, scalability to larger datasets, and resilience against malicious clients is crucial for advancing federated learning frameworks.

These research directions highlight the ongoing need for innovation and refinement in developing trustworthy AI systems, ensuring they remain robust, interpretable, and ethically aligned with societal values.

AI-generated, for reference only.

References

[1] Oh-Ran Kwon, Gourab Mukherjee, and Jacob Bien. Semi-supervised learning of noisy mixture of experts models, 2024.

[2] Brian Kulis and Michael I. Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics, 2012.

[3] Viraj Prabhu, Anitha Kannan, Murali Ravuri, Manish Chablani, David Sontag, and Xavier Amatriain. Prototypical clustering networks for dermatological disease diagnosis, 2018.

[4] Xuan Liu, Xiaoguang Wang, and Stan Matwin. Interpretable deep convolutional neural networks via meta-learning, 2018.

[5] Huiyan Sang, Alan E. Gelfand, Chris Lennard, Gabriele Hegerl, and Bruce Hewitson. Interpreting self-organizing maps through space–time data models, 2009.

[6] Markus Maier, Matthias Hein, and Ulrike von Luxburg. Optimal construction of k-nearest neighbor graphs for identifying noisy clusters, 2009.

[7] Reinhard Heckel and Helmut Bölcskei. Noisy subspace clustering via thresholding, 2013.

[8] Arvindkumar Krishnakumar, Viraj Prabhu, Sruthi Sudhakar, and Judy Hoffman. Udis: Unsupervised discovery of bias in deep visual recognition models, 2021.

[9] Hongjing Zhang and Ian Davidson. Deep fair discriminative clustering, 2021.

[10] James C. Liang, Yiming Cui, Qifan Wang, Tong Geng, Wenguan Wang, and Dongfang Liu. Clusterformer: Clustering as a universal visual learner, 2023.

[11] Gabriel Michau and Olga Fink. Unsupervised transfer learning for anomaly detection: Application to complementary operating condition transfer, 2020.

[12] Ibrahim Jubran, Murad Tukan, Alaa Maalouf, and Dan Feldman. Sets clustering, 2020.

[13] Wojciech Wisniewski, Yuri Kalnishkan, David Lindsay, and Siân Lindsay. Temporal distribution of clusters of investors and their application in prediction with expert advice, 2024.

[14] XuanLong Nguyen. Inference of global clusters from locally distributed data, 2011.

[15] Filippo Maria Bianchi. Simplifying clustering with graph neural networks, 2022.

[16] Miguel Á. Carreira-Perpiñán and Weiran Wang. The k-modes algorithm for clustering, 2013.

[17] Shaotian Cai, Liping Qiu, Xiaojun Chen, Qin Zhang, and Longteng Chen. Semantic-enhanced image clustering, 2023.

[18] Christopher Briggs, Zhong Fan, and Peter Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data, 2020.

[19] Subspace clustering in high-dimensions: Phase transitions statistical-to-computational gap.

[20] James Liang, Tianfei Zhou, Dongfang Liu, and Wenguan Wang. Clustseg: Clustering for universal segmentation. *arXiv preprint arXiv:2305.02187*, 2023.

[21] Omur Arslan and Daniel E. Koditschek. Anytime hierarchical clustering, 2014.

[22] Caglar Aytekin, Xingyang Ni, Francesco Cricri, and Emre Aksu. Clustering and unsupervised anomaly detection with l2 normalized deep auto-encoder representations, 2018.

[23] Ali Jalali and Nathan Srebro. Clustering using max-norm constrained optimization, 2012.

[24] MohammadHossein Bateni, Vincent Cohen-Addad, Alessandro Epasto, and Silvio Lattanzi. A scalable algorithm for individually fair k-means clustering, 2024.

[25] Wilhelmiina Hämäläinen, Mike Joy, Florian Berger, and Sami Huttunen. Clustering students' open-ended questionnaire answers, 2018.

[26] Jinju Joby and Jyothi Korra. Accessing accurate documents by mining auxiliary document information, 2016.

[27] Giacomo Zanella, Brenda Betancourt, Hanna Wallach, Jeffrey Miller, Abbas Zaidi, and Rebecca C. Steorts. Flexible models for microclustering with application to entity resolution, 2016.

[28] Robert Zhang, Rachael Stolzenberg-Solomon, Shannon M. Lynch, and Ryan J. Urbanowicz. Lcs-dive: An automated rule-based machine learning visualization pipeline for characterizing complex associations in classification, 2021.

[29] Alexander Kraskov and Peter Grassberger. Mic: Mutual information based hierarchical clustering, 2008.

[30] Julian Skirzyński, Frederic Becker, and Falk Lieder. Automatic discovery of interpretable planning strategies, 2021.

[31] Erel Levine and Eytan Domany. Resampling method for unsupervised estimation of cluster validity, 2000.

[32] Antoine Zambelli. Ensemble method for cluster number determination and algorithm selection in unsupervised learning, 2021.

[33] Caroline Mazini Rodrigues, Nicolas Boutry, and Laurent Najman. Unsupervised discovery of interpretable visual concepts, 2023.

[34] Andreas Holzinger, Markus Plass, Katharina Holzinger, Gloria Cerasela Crisan, Camelia-M. Pintea, and Vasile Palade. A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop, 2017.

[35] Peter Cho-Ho Lam, Lingyang Chu, Maxim Torgonskiy, Jian Pei, Yong Zhang, and Lanjun Wang. Finding representative interpretations on convolutional neural networks, 2021.

[36] Nikhil Ravi, Anna Scaglione, Sachin Kadam, Reinhard Gentz, Sean Peisert, Brent Lunghino, Emmanuel Levijarvi, and Aram Shumavon. Differentially private k-means clustering applied to meter data analysis and synthesis, 2022.

[37] Daniel J. Lawson, Vinesh Solanki, Igor Yanovich, Johannes Dellert, Damian Ruck, and Phillip Endicott. Clarity – comparing heterogeneous data using dissimilarity, 2021.

[38] George Trigeorgis, Konstantinos Bousmalis, Stefanos Zafeiriou, and Bjoern W. Schuller. A deep matrix factorization method for learning attribute representations, 2015.

[39] Jun Zhao, Tao Gui, Qi Zhang, and Yaqian Zhou. A relation-oriented clustering method for open relation extraction, 2021.

[40] V. Chitraa and Antony Selvadoss Thanamani. Web log data analysis by enhanced fuzzy c means clustering, 2014.

[41] Alejandro Ramírez-Arroyo, Luz García, Antonio Alex-Amor, and Juan F. Valenzuela-Valdés. Artificial intelligence and dimensionality reduction: Tools for approaching future communications, 2022.

[42] Irina Wang, Cole Becker, Bart Van Parys, and Bartolomeo Stellato. Mean robust optimization, 2023.

[43] Zeyu Fang, Ming Gu, Sheng Zhou, Jiawei Chen, Qiaoyu Tan, Haishuai Wang, and Jiajun Bu. Towards a unified framework of clustering-based anomaly detection, 2024.

[44] Wei Zhang, Chao Dong, Jianhua Yin, and Jianyong Wang. Attentive representation learning with adversarial training for short text clustering, 2021.

[45] Emmanuel Arrighi, Zhidan Feng, Henning Fernau, Kevin Mann, Xingqin Qi, and Petra Wolf. Defensive alliances in signed networks, 2024.

[46] Anshuman Chhabra, Abhishek Roy, and Prasant Mohapatra. Suspicion-free adversarial attacks on clustering algorithms, 2019.

[47] Won-Min Song, T. Di Matteo, and Tomaso Aste. Hierarchical information clustering by means of topologically embedded graphs, 2011.

[48] Kishor Barman and Onkar Dabeer. Clustered regression with unknown clusters, 2011.

[49] Vincent Cohen-Addad, Anupam Gupta, Amit Kumar, Euiwoong Lee, and Jason Li. Tight fpt approximations for k-median and k-means, 2019.

[50] Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. Semi-supervised clustering for short text via deep representation learning, 2017.

[51] Dongjin Choi, Andy Xiang, Ozgur Ozturk, Deep Shrestha, Barry Drake, Hamid Haidarian, Faizan Javed, and Haesun Park. Patient clustering via integrated profiling of clinical and digital data, 2023.

[52] Pooya Ashtari, Fateme Nateghi Haredasht, and Hamid Beigy. Supervised fuzzy partitioning, 2019.

[53] Luca Insolia and Domenico Perrotta. Tk-merge: Computationally efficient robust clustering under general assumptions, 2022.

[54] Pokkuluri Kiran Sree and Inampudi Ramesh Babu. Improving quality of clustering using cellular automata for information retrieval, 2014.

[55] Daniel Hurtado Ramírez and J. M. Auñón. Privacy preserving k-means clustering: A secure multi-party computation approach, 2020.

[56] Jana Gauss, Fabian Scheipl, and Moritz Herrmann. Dcsi – an improved measure of cluster separability based on separation and connectedness, 2025.

[57] Deepak P. Whither fair clustering?, 2020.

[58] Federico Altieri, Andrea Pietracaprina, Geppino Pucci, and Fabio Vandin. Scalable distributed approximation of internal measures for clustering evaluation, 2021.

[59] Zacharia Issa and Blanka Horvath. Non-parametric online market regime detection and regime clustering for multidimensional and path-dependent data structures, 2023.

[60] Lin Sun, Jun Zhao, and Xiaojun Ye. Distributed clustering in the anonymized space with local differential privacy, 2019.

[61] Sainyam Galhotra, Sandhya Saisubramanian, and Shlomo Zilberstein. Learning to generate fair clusters from demonstrations, 2021.

[62] Savitha Sam Abraham, Deepak P, and Sowmya S Sundaram. Fairness in clustering with multiple sensitive attributes, 2020.

[63] Matt Barnes and Artur Dubrawski. Clustering on the edge: Learning structure in graphs, 2016.

[64] Manoj Kr Gupta and Pravin Chandra. A comparative study of clustering algorithms. In *2019 6th international conference on computing for sustainable global development (INDIACom)*, pages 801–805. IEEE, 2019.

[65] Maximilian Fleissner, Maedeh Zarvandi, and Debarghya Ghoshdastidar. Decision trees for interpretable clusters in mixture models and deep representations, 2024.

[66] Dong Huang, Jian-Huang Lai, and Chang-Dong Wang. Combining multiple clusterings via crowd agreement estimation and multi-granularity link analysis, 2016.

[67] Elie Aljalbout, Vladimir Golkov, Yawar Siddiqui, Maximilian Strobel, and Daniel Cremers. Clustering with deep learning: Taxonomy and new methods. *arXiv preprint arXiv:1801.07648*, 2018.

[68] Sudarsun Santhiappan, Nitin Shravan, and Balaraman Ravindran. Clustering indices based automatic classification model selection, 2023.

Disclaimer: