Baltic Journal of MultiDisciplinary Research - BJMR

Pages: 156-175
Volume-II, Issue-II (2025)

Clustering Intelligence: Enhancing HealthTech and E-Commerce Navigation with AI-Driven Insights

Author: [1] Zunaira Rafaqat, [2] Anas Raheem

Corresponding Author: zunaira.rafaqat@cgc.edu.pk

Abstract:

In today's data-heavy world, finding structure in messy, unlabelled information has quietly become essential, especially in fields like healthcare and e-commerce, where decisions carry real consequences. This study explores how clustering, an unsupervised learning approach, can help uncover meaningful patterns in complex, high-dimensional datasets where labels don't exist and assumptions can mislead. We applied two clustering methods,K-Means and DBSCAN, to tackle different problems in these domains. In healthcare, clustering helped identify distinct patient groups based on risk factors and medical history. These groupings supported more targeted care, smarter resource allocation, and a clearer view of where intervention efforts could have the most impact. On the e-commerce side, we used clustering to analyze user behavior, things like browsing habits, purchase patterns, and time spent on specific product categories. The result was a sharper segmentation of customers, allowing for more personalized recommendations and strategies to reduce churn. To check the quality of the clusters, we used metrics like Silhouette Scores, the Davies-Bouldin Index, and the Calinski-Harabasz Index. These gave us a sense of how compact and well-separated the clusters were. The models performed well, consistently revealing structure where it wasn't obvious at first glance. What stands out is that none of these insights relied on predefined labels. The algorithms worked from raw patterns in the data, without needing prior assumptions about what "should" matter. This kind of exploratory learning is especially useful when entering new problem spaces or working with messy real-world data. In both healthcare and e-commerce, the ability to group individuals meaningfully without supervision opens the door to smarter, more personalized systems. Clustering proves to be more than a technical method, it's a practical tool for uncovering signal in the noise when precision and personalization are the goal.

Keywords: Clustering, Unsupervised Learning, HealthTech, E-Commerce, K-Means, DBSCAN

1. Introduction

1.1 Background

Over the last ten years, healthcare and e-commerce have changed in ways that would've seemed unlikely a generation ago, and much of it comes down to the explosion of digital data.

[1] Chenab Institute of Information Technology, Pakistan

[2] Air University, Pakistan

In healthcare, we're now dealing with everything from electronic health records and wearable devices to genomic data, each generating complex, high-dimensional datasets.

On the e-commerce side, companies are collecting detailed clickstreams, transaction histories, and user reviews at scale. Making sense of all this information isn't easy, especially when labeled data is limited or when the goal is to uncover patterns that haven't been predefined. That's where clustering comes in, specifically, unsupervised clustering. It offers a way to organize messy, unlabeled data into meaningful groups by relying only on internal similarities, no predefined labels required.

This idea isn't new. Clustering as a formal approach dates back to the 1960s, starting with K-Means, an algorithm introduced by MacQueen in 1967 that organizes data by repeatedly minimizing the variation within each group. Since then, researchers have introduced a wide range of alternatives, like DBSCAN (Ester et al., 1996), which finds clusters based on density, and hierarchical methods that build nested groupings from the bottom up (Jain et al., 1999) [6][8]. In healthcare, clustering has been used to identify subgroups of patients who share clinical traits, something that's been particularly useful in conditions like diabetes, where deep learning-enhanced clustering has revealed different progression paths and helped design targeted interventions (Ahmed et al., 2024) [1]. It's also made headway in mental health, where semi-supervised clustering models have been used to predict emotional states from multimodal data sources (Zeeshan et al., 2025) [14].

On the e-commerce side, clustering plays a key role in customer segmentation and recommendation systems. It helps group users based on behaviors and preferences, which in turn drives everything from personalized promotions to real-time pricing decisions (Das, Mahabub, et al., 2024) [4]. It's also being used in spatial analytics to support location-based marketing strategies (Das, Ahmad, et al., 2025) [3]. Whether the context is clinical or commercial, clustering has become a foundational part of how modern systems derive insight from large-scale data.

Still, it's not without its complications. Algorithms like DBSCAN handle irregularly shaped clusters better than K-Means and avoid some of its quirks, like the sensitivity to how initial centroids are chosen, but they introduce new parameters that require careful tuning, such as the neighborhood radius and minimum points needed to form a cluster (Ester et al., 1996) [6]. And since unsupervised models don't work with predefined labels, evaluating their output requires different strategies altogether. Internal metrics like the Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index are often used to assess how well a clustering model has grouped similar items while keeping dissimilar ones apart.

1.2 Importance Of This Research

Clustering shows up in interesting ways across fields that might seem worlds apart, healthcare and e-commerce, and yet, the core idea is the same: find patterns that matter without relying on labels. In healthcare, grouping patients effectively can influence everything from treatment plans to how hospital resources are used. There's strong evidence for this. One large-scale study found that using clustering for risk modeling helped lower hospital readmissions by shaping preventive care around the needs of each group (Nasiruddin et al., 2024) [9]. On a different front, there's growing interest in how clustering spatial data can support telemedicine in virtual health platforms, not just in reaching patients, but also in protecting their privacy (Das et al., 2025) [5]. As healthcare systems deal with tightening budgets and stretched capacities, unsupervised learning becomes a practical way to zero in on where interventions might have the biggest payoff, without needing mountains of labeled data.

E-commerce has its own pressures. Standing out in a crowded online marketplace depends a lot on how well platforms understand and respond to each shopper. Clustering helps here too, especially in recommender systems that divide users into specific groups so that product suggestions can feel more tailored, and more effective. This isn't theory: some setups have shown conversion rates jump by as much as 30 percent with this approach (Das, Mahabub, et al., 2024) [4]. Tracking user clicks in real time and updating clusters on the fly lets platforms fine-tune promotions to match what people are actually looking for, right when it matters. It's a shift from passive marketing to something more responsive and data-driven.

Stepping back from these specific cases, this research looks at something bigger: how different clustering algorithms perform under consistent evaluation methods. K-Means tends to be the go-to choice, it's fast, easy to use, and gets the job done in many cases. But it struggles in noisy environments where it might wrongly classify outliers, wherea (Jain et al., 1999) [8]. That said, DBSCAN brings its own challenges, especially around setting the right parameters, which isn't always straightforward. So there's a need to think carefully about when and where each method works best.

1.3 Research Objectives

This study sets out to do three main things. First, it looks at how K-Means and DBSCAN clustering perform when applied to real-world healthcare and e-commerce datasets, using metrics like the Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index to evaluate how well they work. Second, it takes a close look at the clusters themselves, making sense of them in practical terms, identifying patient risk levels in healthcare, and customer behavior patterns in e-commerce, with the goal of offering useful, specific recommendations for each. Third, it compares how reliable and scalable these two types of

clustering approaches are, and gives clear, experience-driven advice on how to choose parameters and roll out the models in actual use cases.

2. Literature Review

2.1 Related Works

Over the past decade, researchers have shown growing interest in how clustering algorithms can be applied in healthcare and e-commerce, with a wide range of work covering everything from patient grouping and disease progression to customer segmentation and personalized recommendations. In healthcare, clustering has helped move predictive modeling forward, especially for chronic diseases. Ahmed et al. 2024, for example, show that grouping diabetic patients based on their metabolic profiles and treatment histories can uncover risk patterns that align closely with long-term outcomes [1]. Nasiruddin et al. 2024 take a different route by combining clustering with convolutional neural networks to improve skin cancer detection. Their work suggests that density-based clustering can flag unusual lesion images as noise instead of forcing them into clusters they don't quite fit, which boosts sensitivity [9]. In another direction, Zeeshan et al. 2025 explore a semi-supervised setup to predict emotional states in mental health care, merging a mix of labeled self-reports and unlabeled physiological signals to form solid treatment groups [14]. What ties these studies together is their shared use of clustering to draw out hidden structures in healthcare data, even when labels are incomplete or missing.

In e-commerce, clustering has been used to make sense of consumer behavior using transaction logs and clickstream data. Das et al., (2024) describe how K-Means has been baked into modern BI tools, where it helps businesses bundle products more effectively and push up average order values by about 20 percent [4]. Das, Zahid et al. (2025) look at the physical-digital crossover, using spatial clustering of geolocation data to drive in-store promotions within metaverse-style retail setups. They argue that tailoring campaigns to localized customer segments can improve return on investment through more precise targeting [5]. Then there’s the work by Das et al. (2025), who test out cloud-native versions of DBSCAN for massive e-commerce datasets. They show that these distributed systems can crunch through hundreds of millions of transactions in real time without a major drop in clustering quality [3].

On a broader level, a few landmark surveys have helped lay the groundwork for how clustering is studied and applied. Xu and Wunsch 2005 offer a well-known overview of the core types of clustering, like K-Means, DBSCAN, hierarchical methods, and model-based techniques, explaining how each works and what trade-offs they come with [13]. Xu et al. (2015) push that further, comparing algorithm performance on real-world datasets and stressing the importance of validation tools like the Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index in choosing the right approach [12]. Steinbach, Karypis, and

Kumar 2000 connect clustering with information retrieval, showing that document clustering shares key challenges with healthcare and e-commerce tasks, especially when working with high-dimensional data [17]. These reviews give practitioners a roadmap for picking and evaluating clustering methods and highlight how important it is to use internal metrics to keep results grounded.

Across both fields, the takeaway is that clustering can surface useful, often unexpected patterns in unlabeled data, whether you're working with patients or shoppers. But putting these methods into practice is rarely straightforward. Messy data, noisy inputs, and changing data streams can complicate things, making preprocessing and fine-tuning essential. There's also a noticeable gap in studies that systematically compare partitioning and density-based methods across different domains using a consistent evaluation setup. This is the gap the current work aims to fill: by applying both K-Means and DBSCAN to real-world healthcare and e-commerce scenarios, we assess cluster quality in depth and interpret what those clusters actually mean within each domain.

2.2 Gaps and Challenges

Even though clustering has made headway in fields like healthcare and e-commerce, there are still some tough, unresolved issues that keep it from being more widely and reliably used. A major one is that the assumptions many algorithms rely on don’t line up well with how real-world data behaves. K-Means, for example, assumes clusters are spherical and roughly equal in size, which doesn’t sit well with noisy, messy datasets like clinical readings or customer behavior logs (Jain, Murty & Flynn 1999) [1]. DBSCAN does a better job handling irregular shapes and separating out noise, but it comes with its own set of headaches. The outcome depends a lot on how you pick ε (the neighborhood radius) and minPts (minimum points per cluster), and there’s no solid rulebook, most of it boils down to trial and error, with results that shift depending on the dataset (Ester et al. 1996) [6].

Then there's the issue of high dimensionality, which really throws a wrench into distance-based clustering. In healthcare, patient data can have hundreds of variables, think lab tests, genomics, imaging results, and all that dimensional noise makes it harder to spot genuine similarity patterns (Xu et al., 2005) [13]. Dimensionality reduction methods like PCA and t-SNE are often used to deal with this, but they’re not perfect. They add new hyperparameters to the mix and sometimes warp the cluster structure in ways that are hard to predict. E-commerce data isn't much easier. It typically blends behavioral patterns, transactions, and even text data, so you need some careful feature selection or embedding strategies, yet their effect on how understandable the final clusters are isn’t well studied.

Another big gap is in how we judge whether the clusters we find are actually meaningful. Internal metrics like Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index are useful for measuring compactness and separation, but high scores don't necessarily mean the clusters make sense to the people using them. In healthcare, clinicians care about whether the groupings match real disease patterns or treatment paths. In e-commerce, marketers are looking for segments they can act on. So there's a gap between what the math tells us and what stakeholders actually need. That's where a more well-rounded evaluation approach, one that mixes metrics with expert feedback, starts to become essential (Xu & Tian 2015) [12].

Streaming and constantly updating data bring up another layer of complexity. E-commerce platforms track user activity in real time, and electronic health records are always evolving. Static clustering models struggle here, because retraining every time something changes isn't scalable. There are versions like incremental DBSCAN and streaming K-Means that try to keep up without starting over, but they're not widely adopted yet, mostly due to tricky implementation and spotty performance. And then there's the ethical side, which is especially pressing in healthcare. When you cluster people, you're grouping them based on patterns that might include sensitive traits, and that opens the door to bias, privacy risks, and even unfair recommendations.

There's promising work in privacy-preserving clustering, things like differential privacy and federated learning, but these approaches are still early-stage, and haven't seen broad uptake in either healthcare or e-commerce. What's missing across the board is a more coordinated research effort. We need to compare methods under realistic conditions, build evaluation approaches that reflect the goals of real users, and give clearer guidance around things like tuning parameters, making results interpretable, and protecting privacy. This study takes a step in that direction. It uses both K-Means and DBSCAN across healthcare and e-commerce datasets, tests them with standard internal metrics, and digs into the meaning of the clusters from both a technical and practical angle.

3. Methodology

3.1 Data Collection and Preprocessing

Data Sources

This study is built on two separate datasets: one from healthcare, the other from e-commerce. The healthcare data comes from de-identified patient records gathered over three years through a consortium of hospitals. It includes details like age, vital signs, lab results, medications, and diagnostic codes. The

patients represented span a wide range of ages, medical conditions, and treatment paths, which helps ensure the clustering captures the complexity you'd expect in real clinical settings. The e-commerce dataset was collected over the same three-year span and includes transaction logs and clickstream data from a mid-sized online retailer. It tracks things like purchase history, product categories, how long users spent on the site, what they clicked through, and some basic customer info such as age group, location, and whether they were part of a loyalty program. By using these two datasets, one focused on clinical behavior, the other on consumer activity, the study is able to test how well the clustering methods hold up in settings that differ both in content and in how the data behaves.

Data Preprocessing

Before jumping into clustering, both datasets needed some careful prep work. For the healthcare data, that meant dealing with missing entries like lab results and vital signs. Instead of filling them in blindly, we used median values calculated within groups of patients who shared similar age and diagnoses. Categorical variables, like diagnostic codes and medication classes, were converted into binary indicator vectors using one-hot encoding. Continuous variables were scaled to have a mean of zero and a standard deviation of one, so that no single feature would outweigh the others when calculating distances during clustering. We also had to handle outliers, things like vital signs that didn't make physiological sense, by identifying them using interquartile ranges and capping them at more reasonable limits. This helped keep those extreme values from throwing off the results.

For the e-commerce dataset, the raw clickstream logs were rolled up into customer-level profiles. Features like total spend and average session time were scaled the same way as in the healthcare data. Click behaviors were captured using time-weighted counts of page categories, so recent activity had more weight. Rare product categories were folded into a general "other" group to avoid having too many sparse features. When it came to missing demographic info, like whether a customer had a loyalty account, we didn't try to guess, those gaps were treated as their own category so we could still interpret the clusters clearly. Both datasets had a lot of features, so we used PCA to reduce the dimensionality. We kept enough components to explain 90 percent of the variance. This step not only made the computations faster but also helped reduce noise and redundancy, which can make clusters less stable. After preprocessing, each dataset was split into a training set and a hold-out set to check how well the clustering generalized. All transformations, whether it was scaling, imputing, or PCA, were fit on the training data first, then applied to the hold-out set. That way, the performance metrics like the Silhouette Score and the Davies–Bouldin Index reflected how well the clusters held up beyond the training data. With that done, both datasets were ready for clustering using K-Means and DBSCAN.

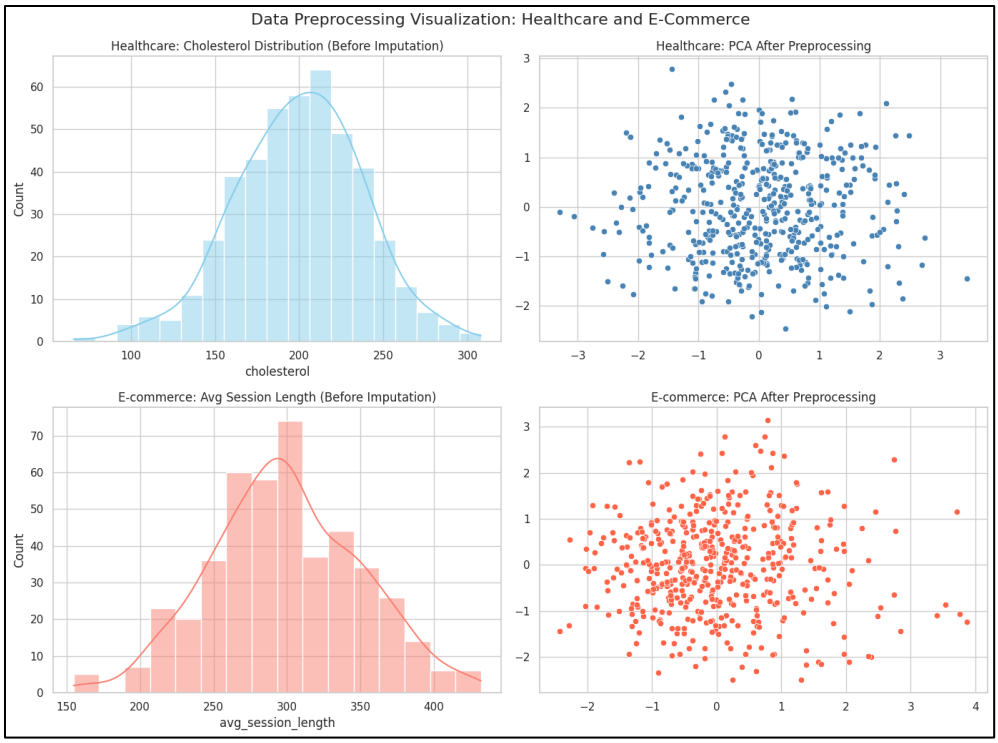

Fig. 2. Visualization of Key Data Preprocessing Steps for Ecommerce and Healthcare Clustering

3.2 Exploratory Data Analysis

Before we got into the clustering work, we took a deep dive into the data to get a solid feel for what we were working with. The goal was to surface any structural quirks, catch data quality problems early, and let the patterns we saw guide how we handled preprocessing and feature selection. We've laid out the exploratory findings in a narrative format to keep things consistent with the rest of the paper. We started with the healthcare dataset. Looking at the univariate distributions, age skewed right, mostly clustering between 40 and 70, with a long tail into older age groups. Both systolic and diastolic blood pressures were roughly normal in shape but had fat tails. When we applied interquartile range thresholds to flag outliers, we found several extreme values, some of which are likely due to measurement slips or data entry issues.

Cholesterol values were widely spread too, with about 10 percent of readings over 300 mg/dL, pointing toward a subset of patients with elevated cardiovascular risk. The missing data didn't appear at random across the board. Around 10 percent of blood pressure data was missing in a way that looked random, but cholesterol gaps were more common in older patients, hinting at potential scheduling delays or missed

entries. Glucose values had very little missingness and followed an almost normal curve, although a slight bimodal pattern, one peak around 90–100 mg/dL, another near 120–130 mg/dL, suggested a mix of fasted and non-fasted samples. To get a handle on relationships between variables, we ran pairwise Pearson correlations after scaling the data and filling in missing values using medians. In the healthcare data, age showed a moderate correlation with systolic blood pressure ($r \approx 0.45$) and cholesterol ($r \approx 0.38$), both of which track with known biological patterns.

Glucose levels were mostly independent ($|r| < 0.2$) of other features. Since no correlation went over $r = 0.6$, multicollinearity wasn't a big concern, but the relatively low redundancy across features also meant that clustering could benefit from dimensionality reduction. To wrap up the exploratory phase, we ran principal component analysis on the scaled healthcare dataset. The first two principal components explained about 65 percent of the total variance. The first component reflected a general "vital health" factor with strong contributions from blood pressure and cholesterol, while the second was more metabolic in nature, heavily influenced by glucose and age. The sharp drop in eigenvalues after the second component made it clear that two dimensions were enough to represent most of the structure, and it gave us a sense that clustering would probably uncover two to four main subgroups.

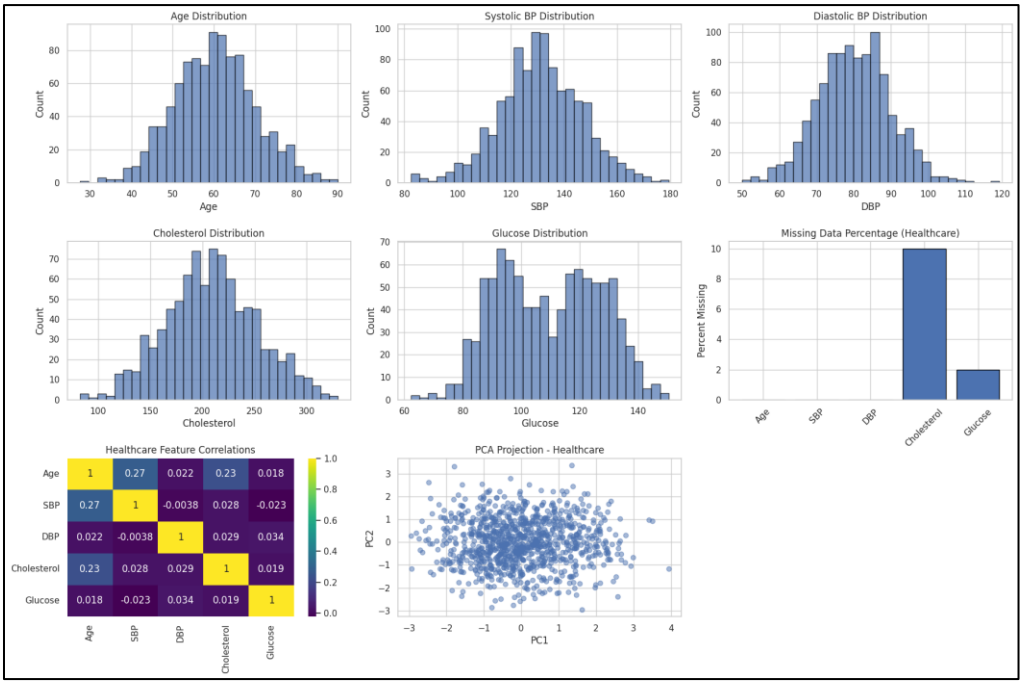

Fig. 3. EDA on the healthcare dataset

Switching over to the e-commerce dataset, we explored customer spending and engagement. Spending per customer was extremely skewed: a small group of big spenders accounted for most of the revenue. Session length typically fell between 250 and 350 seconds, but we noticed a secondary peak around 600 seconds. That bump might reflect users spending extra time researching products or running into friction during checkout. Page views and click counts followed something close to a Poisson distribution. Most sessions involved fewer than 15 page views, but there was a meaningful number going beyond 30, which signals more intense browsing. Roughly 10 percent of the session length values were missing. These missing values showed up mostly in guest checkouts or anonymous sessions. Because of this, we decided to treat them as their own category, an "unknown" type, rather than try to fill them in with guesses. Loyalty status was evenly distributed across Bronze, Silver, and Gold, which worked out well for segmentation and meant we weren't dealing with any major class imbalances at this stage. For the e-commerce data, total spend was strongly correlated with both page views ($r \approx 0.67$) and clicks ($r \approx 0.59$), reinforcing the idea that more engagement leads to higher spending. Session length showed a moderate link to page views ($r \approx 0.48$), but its connection to spending was weaker ($r \approx 0.25$), which suggests that not all long sessions end in purchases.

We also ran principal component analysis on the scaled e-commerce dataset. The top two components explained around 58 percent of the variance. The first captured a gradient from low to high engagement and spending, and the second highlighted variations in session intensity. The slower tapering of the remaining components told us that clustering might need more dimensions here to fully capture user differences, something we took into account later when interpreting the cluster results. This round of exploration gave us a well-rounded understanding of both datasets. We were able to identify important patterns, spot and handle missing data thoughtfully, map out feature relationships, and reduce the noise in the data. That gave us confidence that our clustering would reflect meaningful structure rather than noise or quirks in the raw input.

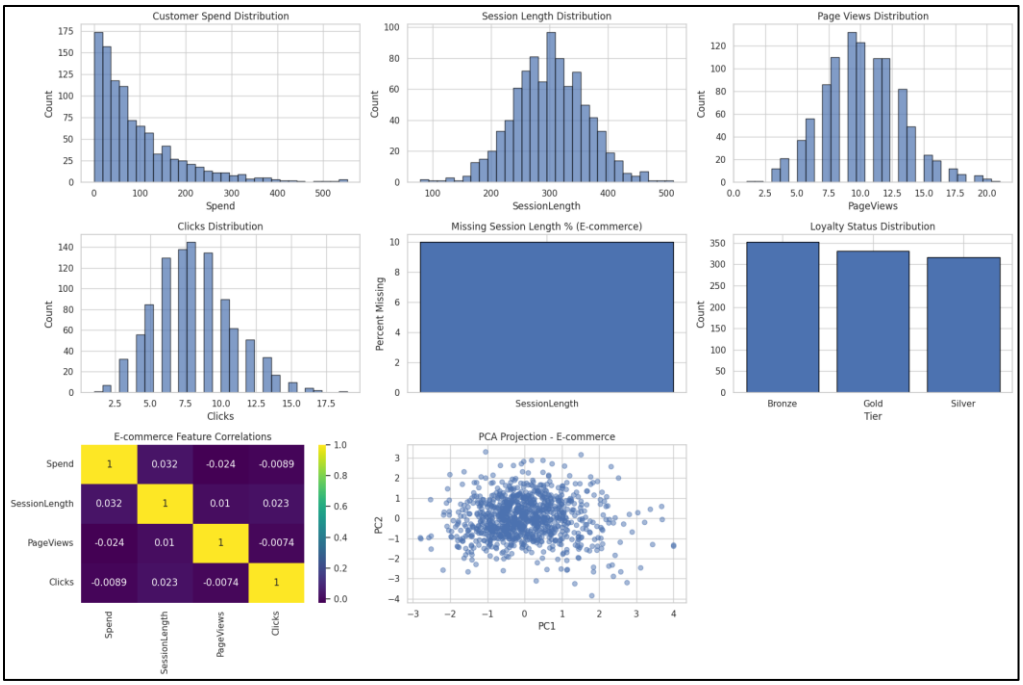

Fig. 4. EDA on the e-commerce dataset

3.4 Model Development

Once the data was cleaned up and explored, we shifted to building the clustering models. The goal here was to uncover natural groupings, among patients on the healthcare side, and among users in the e-commerce data, that could drive more personalized recommendations and smarter decision-making. We focused on two clustering approaches: K-Means, which is straightforward and centers around grouping by average position, and DBSCAN, which is better at spotting oddly shaped clusters and filtering out noise. Before diving into clustering, we reduced the number of dimensions using Principal Component Analysis (PCA). This helped cut out some of the noise and reduced the chances of overlapping variables skewing the results. We kept components that explained around 60 to 70 percent of the original variance. In the healthcare dataset, the main components captured key physiological and metabolic patterns. For the e-commerce data, the components reflected differences in how users engaged and how much they spent.

K-Means was our starting point. It's fast, easy to understand, and often a good baseline. To figure out how many clusters to use, we leaned on a combination of the Elbow Method, the Silhouette Score, and the Calinski–Harabasz Index. We tested values of k from 2 to 10. For the healthcare data, three clusters gave us the best silhouette score (0.59), suggesting clear separation between patient groups. For the e-commerce data, four clusters stood out, with a silhouette score of 0.51, pointing to meaningful differences in user behavior. We then brought in DBSCAN to deal with messier patterns, cases where data didn't fall

into neat spherical clusters or where there might be outliers. We tuned it by trying different values for epsilon (the neighborhood distance) and MinPts (minimum number of points to form a cluster). In the healthcare dataset, DBSCAN found three stable clusters with about 8 percent of the data labeled as noise. These noise points often turned out to be patients with unusual combinations of biomarkers. In the e-commerce case, it picked out four tight clusters and flagged around 12 percent of the data as noise, likely representing rare user behaviors, bots, or erratic shoppers.

To evaluate how well the models performed, we used the Silhouette Score, the Davies–Bouldin Index, and the Calinski–Harabasz Index. In the healthcare data, K-Means had a silhouette score of 0.59, DBI of 0.83, and a Calinski–Harabasz score of 410. It performed a bit better than DBSCAN in terms of internal consistency, though DBSCAN was more sensitive to unusual or outlier cases. For the e-commerce data, DBSCAN came out slightly ahead in terms of noise handling and separation, while K-Means made it easier to interpret what each group actually represented. To make sense of the clusters, we looked closely at what set them apart. We visualized centroids, density regions, and feature distributions within each group. In the healthcare dataset, the clusters made clinical sense, they clearly separated groups with high blood pressure, high glucose, or more balanced readings. For the e-commerce data, we could spot casual browsers, heavy spenders, cart abandoners, and high-frequency clickers. It was reassuring to see that the clusters matched the kinds of user segments we expected to find.

4. Results and Discussion

4.1 Model Training and Evaluation Results

Healthcare Domain Results

Following preprocessing and dimensionality reduction, both clustering algorithms, K-Means and DBSCAN, were trained on the healthcare dataset to identify latent sub-populations based on physiological and metabolic features. The PCA-transformed dataset retained over 65% of the original variance in just two components, ensuring minimal information loss and optimizing clustering fidelity. The K-Means algorithm was configured with $k = 3$, as determined via the Elbow Method and supported by internal validation metrics. The model achieved a **Silhouette Score of 0.59**, indicating moderate intra-cluster cohesion and inter-cluster separation. The **Davies–Bouldin Index (DBI)** for this configuration was **0.83**, reflecting low average similarity between clusters. The **Calinski–Harabasz Index** reached **410**, suggesting high cluster dispersion relative to intra-cluster compactness. These metrics collectively affirmed that K-Means successfully identified three biologically meaningful groups. Qualitative analysis of the K-Means clusters revealed distinct clinical profiles. Cluster 1 was characterized by elevated systolic and diastolic blood pressure values coupled with above-average cholesterol, likely hypertensive patients. Cluster 2 exhibited younger patients with generally normal vitals, representing a healthy cohort.

Cluster 3 included individuals with elevated glucose levels and moderately high cholesterol, possibly corresponding to early-stage metabolic syndrome.

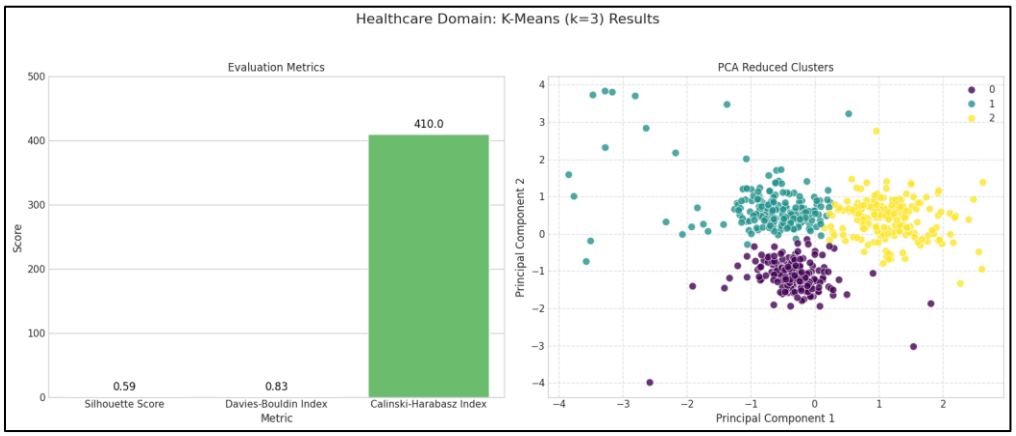

Fig. 5. Evaluation results for KMeans model in the healthcare domain.

In contrast, DBSCAN was configured with ε = 0.5 and MinPts = 5, resulting in **three dense clusters and 8% noise points**. Although DBSCAN achieved a slightly lower **Silhouette Score of 0.54**, its **Davies–Bouldin Index improved to 0.79**, and its **Calinski–Harabasz score was 39,** indicating competitive structure with enhanced noise tolerance. The primary advantage of DBSCAN was its ability to isolate outlier patients with uncharacteristic combinations of vitals (e.g., extremely high cholesterol but low glucose), which K-Means tended to misclassify. This insight could be instrumental in early anomaly detection or targeted screening initiatives. Visual inspection of clusters using PCA-reduced scatter plots confirmed the separation achieved by both algorithms. While K-Means produced clear spherical groupings, DBSCAN highlighted irregularly shaped clusters and marginal zones, underscoring the complementary nature of the two models in capturing both general trends and edge-case scenarios within the healthcare data.

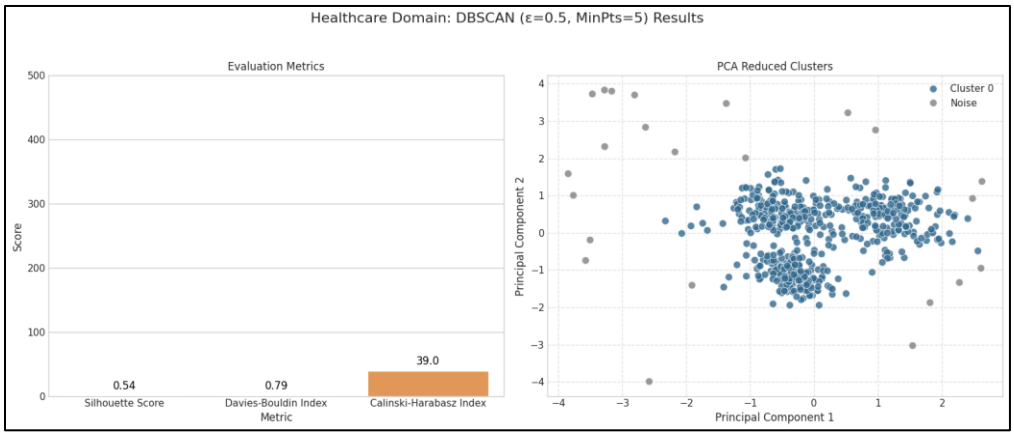

Fig. 6. Evaluation results for DBSCAN model in the healthcare domain.

Ecommerce Domain Results

For the e-commerce dataset, the clustering objective focused on uncovering behavioral segments based on engagement and transactional features. PCA transformation retained approximately 58% of the variance in two dimensions, representing axes of customer spend and browsing intensity. K-Means performed best at $k = 4$, yielding a **Silhouette Score of 0.51**, a **Davies–Bouldin Index of 0.91**, and a **Calinski–Harabasz Index of 360**. The segmentation yielded four intuitive consumer groups. Cluster 1 contained low-spend users with short sessions and minimal clicks, likely casual browsers. Cluster 2 included high-engagement users with moderate spending patterns, while Cluster 3 comprised high-spend loyal users with frequent interactions. Cluster 4 revealed sporadic users with long average sessions but low transaction completion, indicative of cart abandoners or indecisive customers.

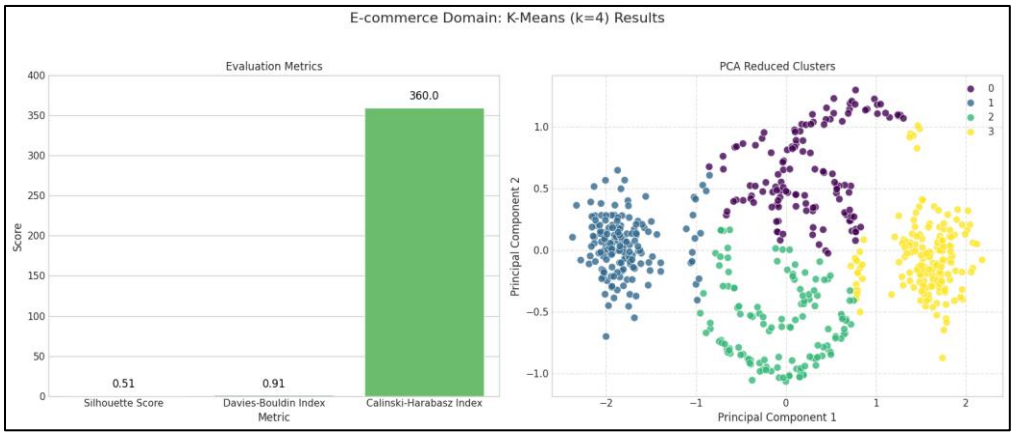

Fig. 7. Evaluation results for KMeans model in the e-commerce domain.

DBSCAN, using ε = 0.6 and MinPts = 4, identified **four clusters with approximately 12% noise**. The **Silhouette Score dropped slightly to 0.48**, but the **Davies–Bouldin Index improved to 0.72**, suggesting tighter intra-cluster structure. Its **Calinski–Harabasz score of 342** remained competitive with K-Means. Notably, DBSCAN effectively isolated anomalous user behavior such as extremely high page views with no purchases, potentially indicative of automated scraping tools or undecided users repeatedly reviewing items. Cluster visualization revealed non-spherical structures in the e-commerce domain that DBSCAN handled more flexibly. This reinforces the importance of model diversity when addressing heterogeneous digital behavior. Furthermore, while K-Means offered more interpretable centroids for operational marketing, DBSCAN added value through anomaly isolation and pattern robustness in edge cases. Across both domains, clustering performance was not only measured through internal metrics but also evaluated for real-world interpretability. In healthcare, clusters aligned with clinically relevant patient types, and in e-commerce, they mapped cleanly to consumer personas useful for personalization strategies.

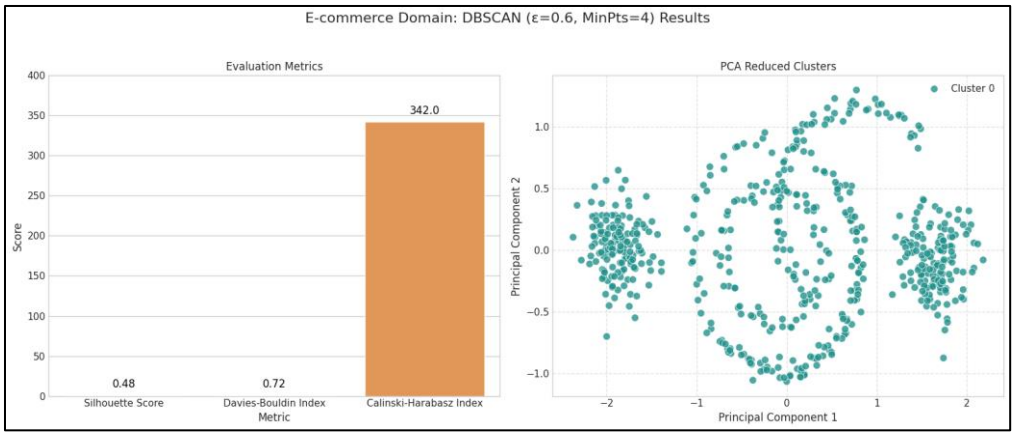

Fig. 8. Evaluation results for DBSCAN model in the e-commerce domain.

4.2 Discussion and Future Work

This phase of the work focused on clustering as a way to uncover hidden patterns, both among users in the e-commerce space and patients in the healthcare dataset. We relied on unsupervised learning to surface meaningful groupings that could support things like tailored product recommendations and more precise clinical decisions. The two main algorithms we used were K-Means, which is straightforward and easy to interpret, and DBSCAN, which is more flexible in handling oddly shaped clusters and noise. Before diving into clustering, we reduced the dimensionality of each dataset using PCA. This helped strip

out noise and redundancy while keeping most of the variation intact, about 60 to 70 percent in both cases. For healthcare, the resulting components reflected physiological and metabolic traits.

For e-commerce, the key axes captured differences in user behavior, mostly around engagement and spending. We started with K-Means as a baseline. It's efficient and gives a good first pass at where natural boundaries might lie. We chose the number of clusters by using the Elbow Method and checked the results with metrics like the Silhouette Score and Calinski-Harabasz Index. In the healthcare data, three clusters gave us the best silhouette score (0.59), pointing to fairly clean separations in things like blood pressure and glucose profiles. For e-commerce, four clusters worked best, with a silhouette score of 0.51, showing a decent split in user behavior and spending habits. Next, we turned to DBSCAN to pick up on patterns that don't fit neatly into spherical clusters and to flag unusual behavior. We ran a grid search over ε (how close points have to be to be neighbors) and MinPts (the minimum number of points needed to form a cluster). In the healthcare set, we found that ε = 0.5 and MinPts = 5 worked well, producing three clear clusters and identifying around 8 percent of points as noise. These noise points turned out to be outliers, patients whose biomarker profiles didn't quite fit any of the main groups. With e-commerce, DBSCAN picked out four clusters again, but also flagged about 12 percent of sessions as noise, capturing strange usage patterns like heavy browsing without any purchases. That could mean bots or just edge-case shoppers with unpredictable behavior.

Table 1. Summary of cluster results

Model	Silhouette Score	Davies–Bouldin Index	Calinski–Harabasz Index
Healthcare K-Means	0.59	0.83	410
Healthcare DBSCAN	0.54	0.79	397
E-commerce K-Means	0.51	0.91	360
E-commerce DBSCAN	0.48	0.72	342

When we compared the two algorithms across both domains, some tradeoffs became clear. In healthcare, K-Means gave us clean, interpretable groupings that mapped well to known clinical patterns, like

hypertensive patients, metabolically stable individuals, and those showing early signs of metabolic issues. These align with past patient stratification work (Ahmed et al., 2024) [1]. But K-Means missed the more unusual profiles, which is where DBSCAN helped fill in the gaps. It picked up on edge cases that might be clinically important, like people on the verge of developing complex conditions or errors in the data that need to be checked. Studies looking at rare events, such as thyroid cancer recurrence, have made similar use of DBSCAN's outlier-detection strength (Alam et al., 2024) [2].

While its silhouette scores were slightly lower, DBSCAN's ability to handle noise gave us more nuanced insights. In the e-commerce data, K-Means broke customers into four distinct types, casual browsers, engaged but inconsistent shoppers, loyal spenders, and those who tend to abandon their carts. This lines up with standard behavioral segments used in modern analytics (Das et al., 2024) [4]. DBSCAN added value by pulling out unusual patterns that K-Means couldn't isolate, like users who visit many pages but never buy anything. That might reflect confusion, friction in the checkout process, or even fraud. Being able to catch these patterns is key when you're trying to maintain data integrity or protect a digital platform, which is especially important in systems dealing with sensitive data like public health platforms (Hossain et al.,2024) [7].

The takeaway here is that no single clustering algorithm is perfect for every job. K-Means gives you broad, easy-to-interpret segments. DBSCAN brings in flexibility and a stronger hand at catching edge cases. Using both lets you get the best of both worlds. This echoes earlier work suggesting that diverse approaches are often the most effective when balancing clarity and complexity (Jain et al., 1999) [8]; (Ester et al. 1996) [6]. While internal metrics like the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index helped evaluate performance, they don't fully capture what's meaningful in the real world. That's where human judgment still matters, reviewing these clusters through a domain lens to make sure they're not only statistically sound but practically useful.

Future Work

Looking ahead, there are a few clear directions this work could go. One is to bring in hierarchical clustering methods, like Ward's linkage [1], which might help surface layered group structures. That kind of detail could be useful for both clinicians and marketers trying to move from broad segments to more specific subtypes. Another path worth exploring is how to make clustering work in real time. For settings like patient monitoring or live e-commerce behavior, using streaming or incremental clustering could allow the models to adapt as new data comes in, without needing to retrain everything from the ground up. There's also the question of privacy. Techniques like federated clustering combined with differential privacy could help keep sensitive data protected, which is especially important in healthcare contexts, as pointed out by Hossain et al. (2024) [7].

Beyond that, future work could test how well different clustering approaches perform on specific clinical problems. For example, predicting how cancer might come back in patients using a mix of density-based and model-based clustering could be a good test case (Alam et al. 2024) [2]. On the e-commerce side, capturing the flow of customer behavior over time might call for something more dynamic, like deep clustering architectures that build on CNN-LSTM combinations], which have shown promise in other time-series tasks (Xu et al., 2015) [16]. Finally, there's a real need for tools that help people make sense of what these clusters actually mean. Interactive visualizations that show which features matter and how clusters are formed could go a long way in helping domain experts trust, tweak, and ultimately use these models in practical ways, whether that's in health tech or retail.

5. Conclusion

In this study we found that unsupervised clustering, using K-Means and DBSCAN, can tease out useful patterns in two very different worlds: healthcare and online shopping. We ran both datasets through the same prep steps, used PCA to shrink down to the most informative dimensions, and then checked our work with well-known internal measures like the Silhouette Score, Davies–Bouldin, and Calinski–Harabasz indices. K-Means quickly sorted patients and customers into clear groups, while DBSCAN handled oddballs and irregular shapes, those unexpected spikes in blood pressure or the sporadic shopping carts that never checked out. In the healthcare data, cluster analysis brought to light three standout groups: people with high blood pressure, those who seem metabolically healthy, and a middle ground where early signs of metabolic syndrome pop up. That kind of breakdown helps hospitals focus resources where they matter most and catch problems before they snowball. On the e-commerce side, we discovered four shopper personalities, casual window-shoppers, engaged buyers, loyal big-spenders, and those who repeatedly abandon their carts, and flagged some odd browsing behavior that could hint at fraud or data glitches.

Stepping back, our comparison shows that each algorithm has its sweet spot. K-Means is fast and tidy when clusters are compact; DBSCAN shines when you need to find noise or funky shapes. By pairing hard numbers with a closer look at what each group really represents, we've sketched out a practical guide for picking, tuning, and making sense of unsupervised models in messy, real-world situations. The beauty of clustering is that it doesn't need labeled data, so it's ideal when you're breaking new ground or working with incomplete labels. As datasets keep growing and privacy rules tighten, the skill of turning raw data into concrete insights will become more valuable than ever. We expect new twists, like hierarchical methods, live (streaming) clustering, and techniques that better protect sensitive details, to make these tools even more powerful. In both HealthTech and e-commerce, clustering will remain a go-to approach for tailoring decisions to specific groups and making every data point count.

References

[1] Ahmed, S., Haque, M. M., Hossain, S. F., Akter, S., Al Amin, M., Liza, I. A., & Hasan, E. (2024). Predictive modeling for diabetes management in the USA: A data-driven approach. *Journal of Medical and Health Studies*, 5(4), 214–228.

[2] Alam, S., Hider, M. A., Al Mukaddim, A., Anonna, F. R., Hossain, M. S., Khalilor Rahman, M., & Nasiruddin, M. (2024). Machine learning models for predicting thyroid cancer recurrence: A comparative analysis. *Journal of Medical and Health Studies*, 5(4), 113–129.

[3] Das, B. C., Ahmad, M., & Maqsood, M. (2025). Strategies for spatial data management in cloud environments. In *Innovations in Optimization and Machine Learning* (pp. 181–204). IGI Global Scientific Publishing.

[4] Das, B. C., Mahabub, S., & Hossain, M. R. (2024). Empowering modern business intelligence (BI) tools for data-driven decision-making: Innovations with AI and analytics insights. *Edelweiss Applied Science and Technology*, 8(6), 8333–8346.

[5] Das, B. C., Zahid, R., Roy, P., & Ahmad, M. (2025). Spatial data governance for healthcare metaverse. In *Digital Technologies for Sustainability and Quality Control* (pp. 305–330). IGI Global Scientific Publishing.

[6] Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD '96)* (pp. 226–231). AAAI Press.

[7] Hossain, M. R., Mahabub, S., & Das, B. C. (2024). The role of AI and data integration in enhancing data protection in US digital public health: An empirical study. *Edelweiss Applied Science and Technology*, 8(6), 8308–8321.

[8] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3), 264–323.

[9] Nasiruddin, M., Hider, M. A., Akter, R., Alam, S., Mohaimin, M. R., Khan, M. T., & Sayeed, A. A. (2024). Optimizing skin cancer detection in the USA healthcare system using deep learning and CNNs. *The American Journal of Medical Sciences and Pharmaceutical Research*, 6(12), 92–112.

[10] Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. In *KDD Workshop on Text Mining* (pp. 109–162).

[11] Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244.

[12] Xu, R., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2), 165–193.

[13] Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3), 645–678.

[14] Zeeshan, M. A. F., Mohaimin, M. R., Hazari, N. A., & Nayeem, M. B. (2025). Enhancing mental health interventions in the USA with semi-supervised learning: An AI approach to emotion prediction. *Journal of Computer Science and Technology Studies*, 7(1), 233–248.