scientific reports

Check for updates

OPEN

Enhancing disease clustering through symptom-based analysis and large language model interpretations

Efe Onojete[1], Ebuka Ibeke[1,6]✉, Chinedu Pascal Ezenkwu[1,6], Celestine Iwendi[2,6] & Imed B. Dhaou[3,4,5,6]✉

Humans face various diseases that are mainly caused by environmental conditions and living habits. These diseases exhibit several symptoms and can share a relationship based on their symptoms. The identification and interpretation of these groups of symptom-based diseases can aid in developing treatment plans for a new outbreak of disease. This research explores the intersection of machine learning and healthcare, specifically focusing on the enhancement of disease classification through symptom-based cluster analysis. By leveraging unsupervised machine learning algorithms, patterns and relationships within diverse symptom datasets were identified, revealing novel associations and subtypes in disease manifestation. The integration of a Large Language Model (LLM), specifically OpenAI's Generative Pretrained Transformer(GPT), played a pivotal role in interpreting and communicating the complex outputs of the machine learning process. The results indicated a significant improvement in defining distinct clusters based on the relationship between diseases and symptoms, with GPT-4o providing simplified explanations that bridge the gap between machine-generated insights and healthcare professional's understanding. The study's findings offer a more profound understanding of the distinctive features characterising the different clusters of diseases generated by the machine learning models.

Keywords Machine learning, Diseases, Large language model, Symptoms, Interpretability, Clustering, Unsupervised learning

The healthcare field produces extensive and varied data, which machine learning algorithms can leverage to detect new illnesses and optimize treatment plans[1]. Deep learning (DL), when trained on high-quality data, has significantly advanced clinical diagnostics and facilitated disease clustering[2]. One example is symptom-based clustering, which can enhance diagnostic accuracy and support personalized patient care[3].

Diseases with overlapping symptoms pose significant challenges for accurate clinical diagnosis, a problem that can be mitigated through coordinated care and collaboration between multidisciplinary teams[4].

Traditionally, physical exams or laboratory tests are used to identify diseases. This process can be complicated and sometimes inaccurate, as many diseases share similar symptoms[5]. ML-enabled techniques help to discover new disease subtypes and understand the diversity of the patient population by uncovering hidden patterns within complex data sets[6]. Symptom-based cluster analysis is an effective technique for providing precise and targeted medical information[7]. However, interpreting these complex models poses a unique challenge. Watson[8] argued that while clustering algorithms efficiently reveal connections, converting these clusters and patterns into meaningful medical insights is difficult.

[1]School of Computing, Engineering and Technology, Robert Gordon University, Garthdee Road, Garthdee, Aberdeen AB10 7AQ, United Kingdom. [2]Centre of Intelligence of Things, School of Creative Technologies, University of Greater Manchester, Bolton BL3 5AB, United Kingdom. [3]Department of Computer Science, Hekma School of Engineering, Computing and Design, Dar Al-Hekma University, Jeddah, Saudi Arabia. [4]Department of Computing, University of Turku, Turku, Finland. [5]Department of Technology, Higher Institute of Computer Sciences and Mathematics, University of Monastir, Monastir, Tunisia. [6]Ebuka Ibeke, Chinedu Pascal Ezenkwu, Celestine Iwendi, Imed B. Dhaou have contributed equally to this work. ✉email: e.ibeke@rgu.ac.uk; imed.bendhaou@utu.fi

Recently, large language models (LLM) have emerged as effective tools in multiple fields, including healthcare[9], education, and chip design[10]. In clinical diagnosis, they can improve the interpretation of groups by providing insight into the relationships between symptoms and diseases, thereby supporting more accurate and meaningful analysis[11].

This paper presents a novel approach that integrates unsupervised machine learning techniques with LLM, specifically GPT-4o, to improve the clustering and interpretation of disease in healthcare. The key contributions of this study include:

1. Bridging Machine Learning Outputs and Clinical Interpretability: The integration of LLM for interpreting clustering results addresses a critical gap in the literature—translating complex, unsupervised machine learning outputs into clinically meaningful insights. This contribution enhances the usability of clustering techniques for healthcare professionals, which is often cited as a limitation in existing studies.
2. Comprehensive Performance Analysis of Clustering Algorithms: The paper systematically evaluates multiple clustering algorithms (K-means, Fuzzy C-Means, Hierarchical Clustering, and DBSCAN) using a broad range of metrics, contributing to a deeper understanding of their comparative effectiveness in disease clustering. This fills a gap in literature studies where comparative analyses are often underexplored.
3. Exploration of Symptom-Based Disease Clustering: The study expands on prior research by focusing on symptom-based disease clusters, uncovering new relationships and subtypes in disease manifestation. This contribution adds to the growing body of literature on leveraging symptom co-occurrence for improved disease subtyping and diagnosis.

The rest of the paper is organized as follows. Sect. "Literature review" examines curent published technique for disease clustiring and puts our work in context o exising techniques. Section "Methodology" presents the experiment and results. Section "Limitations and future work" highlights the limitations of the present approach. Section "Conclusion and recommendation" presents the Conclusion and Recommendation.

Literature review

Several existing research has focused on evaluating individual symptoms in patients with chronic conditions. However, the approaches did not consider that symptoms rarely occur alone - they often co-occur with one another. Due to the preceding reasons, there has been a developing attention towards co-occurring symptom clusters as an alternative means of identifying diseases. Lin[12] analysed symptom data involving gastric cancer patients. The study identifies five different clusters of symptoms, each showing unique clinical characteristics and survival outcomes. Mousavi et al.[13] utilised machine learning in uncovering subgroups of irritable bowel syndrome (IBS), revealing their associated co-occurring symptom clusters, demonstrated distinct patterns associated with different disease severities and treatment responses. Byale et al.[14] applied clustering on a high-dimensional large dataset of IBS patients. The authors identified seven distinct disease subtypes with specific symptom profiles, treatment responses, and prognosis. Nikolaou et al.'s[2] systematic review on chronic obstructive pulmonary disease (COPD) phenotypes and machine learning cluster analysis reports the potential of machine learning and cluster analysis in identifying distinct COPD subgroups with distinct clinical presentations and outcomes. Qiu et al.[15] developed a model for automated cardiovascular disease (CVD) detection from electrocardiogram (ECG) data, combining convolutional and recurrent neural networks with language models (LLMs) pre-trained on ECG and medical case data to classify eight heart diseases. Their approach achieved high diagnostic accuracy by capturing spatial–temporal interactions, demonstrating how LLMs can enhance the interpretability of complex medical models. Similarly, Yin et al.[16] pre-trained and fine-tuned an LLM to support graph neural networks in biomedical signal processing, improving the prediction of electronic and functional properties of organic molecules by revealing complex relationships in molecular structures.

Leveraging algorithms trained on large datasets containing information about patients' medical history including demographic factors like age or sex, has been demonstrated to be able to improve the accuracy in identifying diseases based on clustering patterns[17], making these algorithms significant in symptom-based cluster analysis of diseases for the identification and classification of similar symptoms among patients[5]. Several researchers have adopted various algorithms such as K-means clustering, hierarchical clustering, fuzzy C-Means (FCM), and density-based spatial clustering applications with noise (DBSCAN) to identify commonalities within symptom datasets[18]. Fuzzy C-Means has emerged as a powerful tool for medical diagnosis, showing great ability in handling uncertaintities in data[19]. This and similar methods can help to develop targeted treatments for specific symptom clusters rather than blanket solutions for broad disease categories, creating an opportunity for personalised care and improved patient outcomes. Hierarchical clustering has been an increasingly popular clustering technique in medicine for grouping diseases based on their symptoms, supporting the development of tailored treatment plans[20,21]. Nicolet et al.[22] argues that these automated data analytics can help to effectively recognise high-risk populations while reducing unnecessary procedures for those not susceptible, thereby addressing overloading the healthcare, and contributing to the management of pandemics such as COVID-19 or annual influenza outbreaks[23]. For example, K-Means clustering has been utilised in analysing and grouping symptoms of diseases based on age, sex and congenital diseases for understanding the causes of death due to Covid-19 due to Indonesian Navy personnel and their families[24]. Gousia and Shaima[25] presents a hybrid approach combining convolutional neural networks (CNN) and DBSCAN clustering in addressing challenges related to SARS-CoV with 97.35%.

However, none of these research papers considered the interpretability of the clusters, which can be difficult for humans to contextualise and make sense of on their own[26]. As such, the work entailed by this paper contributes to the increasing focus on interpretable machine learning methods and tools in the health domain. Leveraging large -language models (LLMs) such as GPT, human-level interpretation can provided for

understanding symptom clusters. LLMs can generate plain-language explanations of disease groups identified through clustering or other unsupervised machine learning techniques[9], highlighting the prevalence, patient profiles, symptoms, comorbidities, and other descriptive factors that characterize specific disease clusters. For disease clusters with nuanced relationships or connections, LLM can suggest hypotheses analysing these diseases and their complex relationships. Such hypothesis generation can help to clearly define conjectures, guiding investigation into the causal factors behind clustered conditions[11].

Overall, given the limited exploration of addressing interpretability of symptom clusters using LLMs in this domain, this paper seeks to leverage OpenAI's GPT-4o for this task. GPT-4o's natural language processing capabilities can provide explanatory role, contributing to addressing the research gap.

Methodology

This section describes the methodology and technique applied to achieve the objective of the research. The experiments were conducted in three key phases: Data Description, Machine Learning Algorithms Clusters, and Application of GPT-4o.

Data description

The dataset used in this study is n data collated by Zhou et al.[27] in their study of "Human symptoms–disease network" to investigate the connection between clinical manifestations of diseases and their underlying molecular interactions. The dataset was originally curated using systemized nomenclature of medicine–clinical terms (SNOMED-CT). SNOMED-CT are coded terms entered into electronic health records (EHRs) to capture, record, and share clinical data for use by healthcare organisations. The dataset obtained was found relevant for this study as it contained documentation on different conditions across disciplines like cardiology, neurology, immunology, etc. This expansive scope of diseases provides a strong foundation for enabling the discovery of subtype clusters through symptom analysis. The dataset is categorical data that contains 3,011 rows, 2 columns, 168 missing values, and 15 duplicate rows. The 2 columns provided information about 2602 records of disease-symptom relationships. The disease column had 1769 different disease categories and the symptom column had 833 distinct symptom categories. A random sample of 5 rows capturing the disease and symptoms column is displayed in Table 1 below.

This study considered two methods (Deletion or Imputation) to deal with the missing values and chose deletion because the missing values were missing completely at random (MCAR) as they were not dependent on any other variables in the dataset. The data was thereby transformed afterwards using the one-hot encoding technique to turn categorical variables into numerical values. Applying one-hot encoding to the dataset led to increased dimensionality, as 833 columns were created for each category. This disadvantage was handled using the principal component analysis (PCA). PCA is a widely used dimensionality reduction technique for reducing the number of variables in a dataset while retaining the majority of the information in the original dataset. PCA transforms the original features into a new set of uncorrelated variables called principal components (PCs), which are linear combinations of the original features[28]. To choose the best number of clusters, it is necessary to find the optimal K-value. The elbow method - a technique to find the best number of clusters by identifying the point where adding more clusters stops significantly improving the fit -is one way of finding the optimal K-value. Figure 1 shows a graphical output of finding the optimal K-value; it can be seen that the elbow is at K = 4. Thus, 4 clusters are the best number for the clustering calculations. As well as the distortion score, the time of convergence of K-means for a given K-value has been demonstrated in Figure 1.

Although the current paper has leveraged the elbow method, one of the most popular methods, in identifying the optimal number of clusters within the dataset[29], several other criteria can be explored to determine the optimal K – such as the Average Silhouette Width, Gap Statistic or the Calinski-Harabasz. The elbow method has been shown to be effective in some applications of clustering in medicine compared to the Silhouette method[30].

The hyperparameters for each algorithm have been tuned using the halving random grid search[31] to ensure their fair comparison. The halving random grid search benefits from the strengths of random and grid search in hyperparameter tuning, making it a good hyperparameter tuning approach, both in terms of performance and speed.

Results and discussion of findings

The study involved evaluating the performance of the various clustering algorithms (K-means, Fuzzy C means, Hierarchical, and DBSCAN) across 10 unsupervised learning evaluation metrics (Adjusted Rand Index (ARI); Calinski Harabasz index (CHI); Davies Bouldin Index (DBI); Fowlkes Mallows Index (FMI); Adjusted Mutual Information (AMI); Normalized Mutual Information (NMI); Homogeneity Score (HS); Completeness Score (CS); V-Measure Score (VMS); and Silhouette Index (SI)), covering both label- and shape-based metrics, as

Diseases	Symptoms
Urticaria Disorder	Cold Reflex Urticaria
Hemiplegia or Hemiparesis	Hemiplegia Disorder
Disorder of Scalp	Itchy Scalp
Diarrhea due to Staphylococcus	Loose Stool
Benign Neonatal Familial Convulsions	Seizure

Table 1. 5 Random Samples of Diseases and their Symptoms.

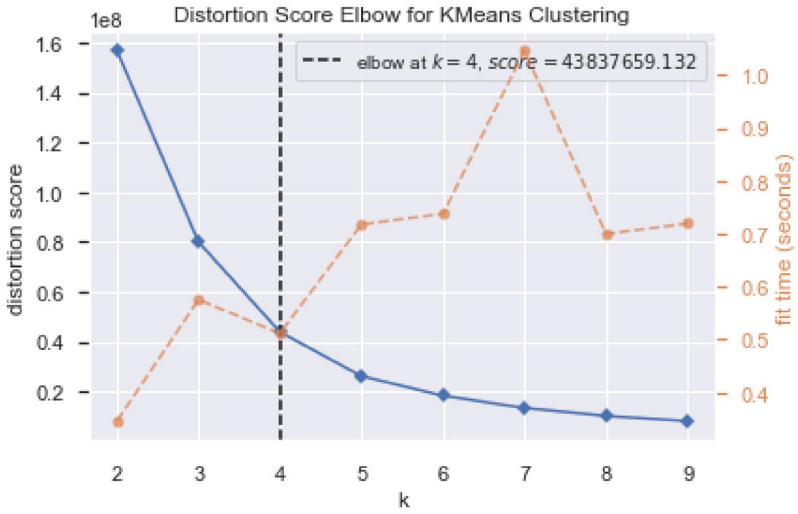

Figure 1. Elbow graph of finding optimal K-value: Distortion Score and Fit Time Analysis.

Metric	Metric Type	K-means	Fuzzy c means	Hierarchical Clustering	DBSCAN
AMI	Label-based	**0.141**	**0.141**	0.137	0.029
ARI	Label-based	**0.005**	**0.005**	0.004	0.001
CHI	Shaped-based	**13631**	13533	11575	3.591
CS	Shaped-based	**1**	**1**	**1**	0.854
DBI	Shaped-based	0.537	0.532	**0.509**	6.039
FMI	Label-based	**0.055**	**0.055**	0.053	0.031
HS	Label-based	**0.192**	0.191	0.187	0.085
NMI	Label-based	**0.322**	0.321	0.315	0.155
SI	Shaped-based	**0.560**	**0.560**	0.552	-0.145
VMS	Label-based	**0.322**	0.321	0.315	0.155

Table 2. Performance of the Different Clustering Algorithms.

shown in Table 2, to mitigate the bias associated with specific shape- and density-related assumptions of the different algorithms. It is through the detailed explanation of the metrics below that meaningful insights can be gleaned from the model evaluation results:

- Adjusted Rand Score: Compares the similarity between two different cluster label assignments to the same data set to measure accuracy. Values closer to 1 indicate greater similarity between the clustering solutions.
- Calinski-Harabasz Score: Evaluates the cluster validity based on the ratio of between-cluster dispersion to within-cluster dispersion. Higher scores indicate clusters are dense and well-separated.
- Davies-Bouldin Score: Calculates the average similarity between each cluster and its most similar counterpart. Lower values indicate tighter, more distinct clusters, while higher scores signal greater cluster overlap.
- Fowlkes-Mallows Score: Compares cluster assignments to external benchmark classifications to determine accuracy. Higher values signify greater agreement between the cluster labels and external labels.
- Adjusted Mutual Information Score: Compares two clusterings by calculating normalized mutual information adjusted for chance. Higher scores indicate greater shared information between the two clusterings.
- Normalized Mutual Information Score: An unadjusted variant of AMI that calculates mutual information between clusters without accounting for random chance agreement. Also rates cluster correspondence.
- Homogeneity Score: Quantifies the extent that which clusters only contain a single class/external label. Higher values reflect greater homogeneity.
- Completeness Score: Measures if external labels are concentrated into a single dominant cluster rather than scattered across multiple clusters. Higher is better.
- V-Measure Score: The harmonic mean combining homogeneity and completeness to evaluate overall external cluster label correspondence. Higher reflects clusters and labels better match.
- Silhouette Score: Calculates cluster cohesion and separation by comparing intra-cluster distances to distances with other clusters. Scores range from -1 to 1, with higher values indicating appropriate clustering.

The results presented in Table 2 indicate that the K-means clustering model outperformed the other algorithms, achieving the highest silhouette score of 0.56, a completeness score of 1.0, and the top Calinski-Harabasz index.

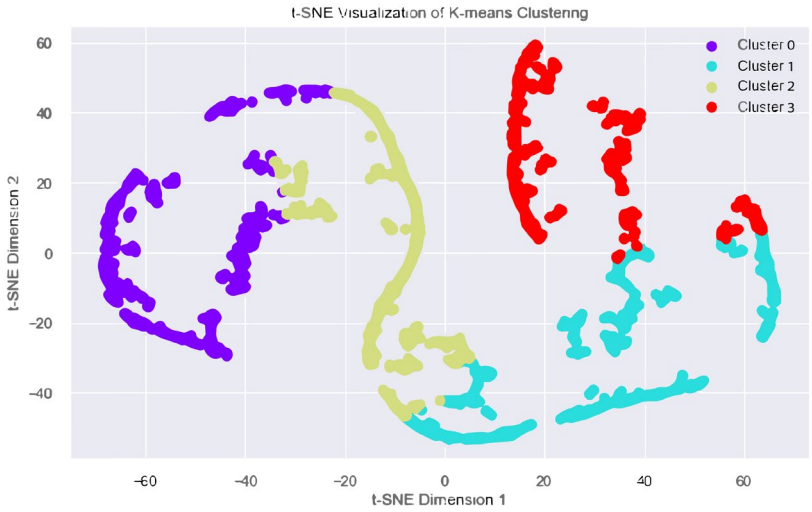

Figure 2. K-means clusters showing different disease groups.

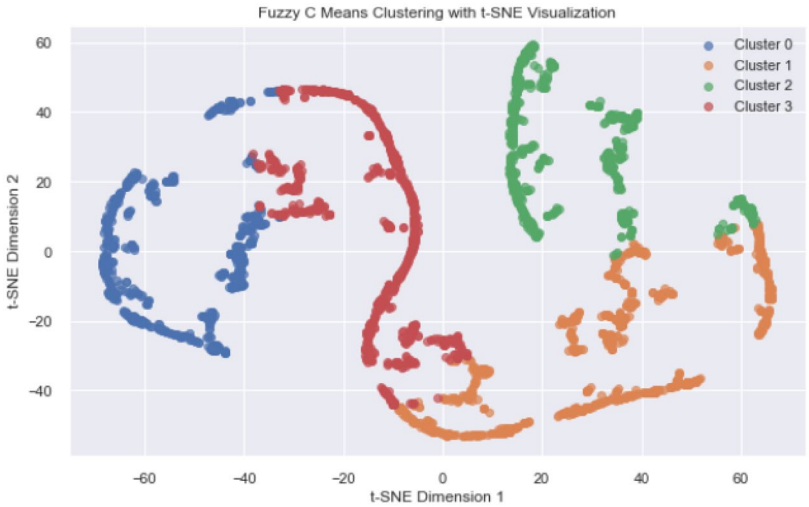

Figure 3. Disease clusters from the Fuzzy C-Means clustering algorithm.

These metrics collectively highlight the effectiveness of K-means in accurately identifying and separating disease subgroups, suggesting that it was particularly well-suited for capturing the underlying structure of the dataset. DBSCAN produced a negative Silhouette Score of -0.145 and low CHI of 3.591. This is due to the varying density and high dimensionality of the dataset. DBSCAN struggles when clusters are not well-separated or differ significantly in density, leading to misclassification of dense regions and excessive noise points.

Machine learning algorithms clusters

The K-means algorithm effectively partitioned the dataset into four distinct clusters (0 to 3), as shown in Figure 2. These clusters each exhibit unique characteristics and groupings of data points. The figure provides a clear illustration of how the algorithm organized the diseases based on underlying patterns in the data. This clustering outcome highlights the algorithm's ability to uncover structure within the dataset, offering valuable insights into the relationships and similarities among the grouped conditions.

The Fuzzy C-Means (FCM) model demonstrated strong performance in defining clusters, as reflected by its evaluation metrics: a completeness score of 1.0, a Calinski-Harabasz index of 13,533, and a silhouette index of 0.560. These results suggest that the FCM algorithm was highly effective in grouping data points from the same true class while maintaining well-separated and compact clusters. Figure 3 visually illustrates the clustering outcome, revealing four distinct clusters labeled Cluster 0 through Cluster 3. The FCM algorithm is particularly well-suited for datasets with overlapping or ambiguous boundaries, as it assigns degrees of membership to each data point rather than forcing hard assignments. This visual representation provides valuable insight into how the model captures the nuanced, fuzzy relationships within the data.

Similarly, the Hierarchical clustering algorithm also achieved a completeness score of 1.0, indicating perfect alignment with the true class labels—no class was split across multiple clusters as shown in Figure 4. Its Calinski-Harabasz index and silhouette index were 11,575 and 0.552, respectively, suggesting that while the clusters were slightly less compact and well-separated than those produced by FCM, the overall structure was still robust and meaningful.

DBSCAN achieved a completeness score of 0.854, which is relatively close to 1. This indicates that the algorithm performed well in preserving the integrity of the true class labels during clustering. In other words, DBSCAN was effective in grouping data points that belong to the same actual class, suggesting that it successfully captured the natural structure present in the dataset. However, DBSCAN has a negative silhouette index score of -0.145. This suggests that the algorithm struggled to effectively separate the diseases into well-defined clusters. Although a Silhouette index score of -0.145 clearly indicates a poor clustering outcome, the underlying issue may lie in the structure of the dataset itself. With 833 features, the dataset is highly dimensional, which likely introduced challenges associated with the curse of dimensionality, making it difficult for the algorithm to identify meaningful patterns. The curse of dimensionality states that as the number of features increases, the amount of data representing the relationship between them expands exponentially, making it harder for density-based algorithms, such as DBSCAN, to learn effectively. This problem could be mitigated by feature selection, but the 833 unique features used in the development of the model were important. Also, the CHI had a low score of 3.591, indicating that the clusters identified by the DBSCAN algorithm were not well-separated or well-defined. This means that the data points within each cluster were not tightly grouped, and the distance between the clusters was not sufficiently large. Although the result is discouraging, it serves as an opportunity for further investigation. One of the standout features of the DBSCAN algorithm is its ability to effectively identify and separate noise from the core data structure[32]. In Figure 5, this noise is visually represented in grey.

Challenges in interpreting disease subgroups from the clustering algorithms

In this section, the major challenge encountered in gleaning meaningful interpretations of the subgroups generated by the algorithm is the lack of inherent subgroup labelling. i.e., while the algorithms identified the distinct clusters in the highly dimensional data, they could not assign intuitive labels or explanations describing the key features that differentiate the subgroups of diseases. The 184 seizure-based diseases were separated into 4 clusters with Cluster 0, Cluster 1, Cluster 2, and Cluster 3 having 40, 55, 42, and 47 diseases, respectively. For example, Table 3 shows the output of the generated subgroups of diseases having 'Seizure' as the primary symptom with no interpretation of the key distinguishing features between the subgroups.

However, LLM can enhance the interpretability of the machine learning models by providing natural language explanations for the identified subgroups or clusters. By integrating GPT-4o into the analysis, the model can generate intuitive and contextually relevant labels or descriptions for the key features of the subgroups. This not only makes the results more accessible to non-experts but also facilitates a deeper understanding of the factors driving differentiation within the clusters. GPT-4o's natural language generation capabilities enable it to bridge the gap between complex machine-generated outputs and human comprehension, offering valuable insights into the characteristics and significance of the identified subgroups.

Application of large language module (model: GPT-4o)

This section explores the practical application of GPT to discern the distinct and unique characteristics of the diseases grouped in each cluster having 'seizure' as the primary symptom.

Elsborg and Salvatore[33] analysed biomarkers at the single-cell level to improve understanding. They demonstrated the usability of using LLM models to simplify gene signatures, facilitating their interpretability

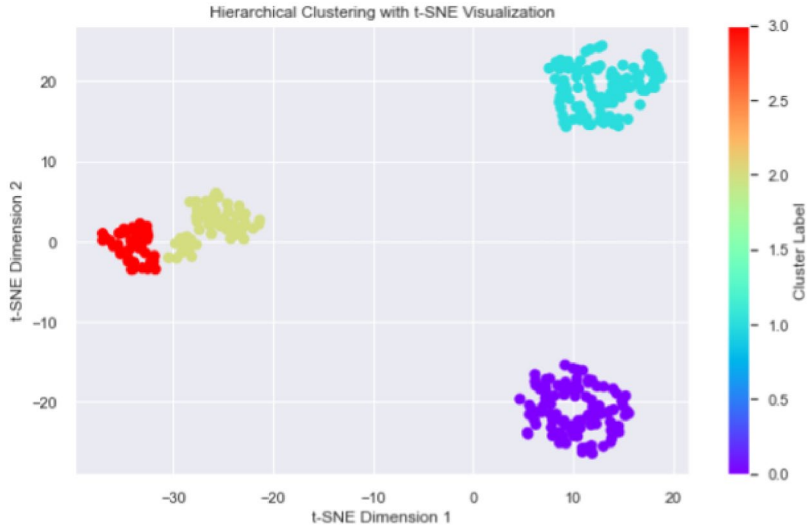

Figure 4. Disease clusters from the Hierarchical clustering algorithm.

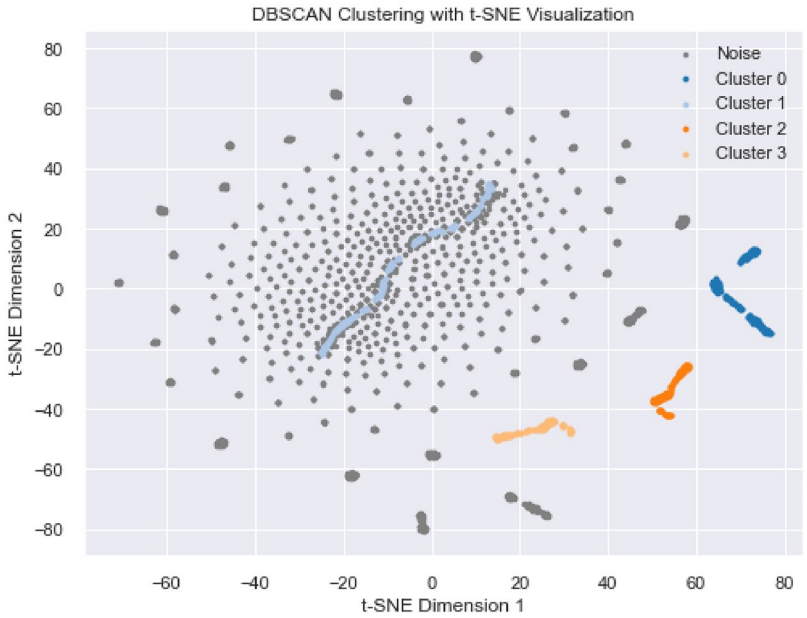

Figure 5. DBSCAN clustering algorithm Disease clusters.

Diseases	Clusters
Febrile Convulsion	2
Eclampsia Disorder	2
Epileptic Seizure	2
Seizure Disorder	1
Localisation-Related Epilepsy	3
Jacksonian, Focal, or Motor Epilepsy	3
Benign Rolandic Epilepsy	0
Epilepsy Characterized by Intractable Complex	2
Refractory Generalized Nonconvulsive Epilepsy	1
Chronic Progressive Epilepsia Partialis	0

Table 3. Subgroups of Seizure-Based Diseases.

and explaining underlying molecular disease mechanisms. This section explores the practical application of GPT-4o to discern the distinct and unique characteristics of the diseases grouped in each of the clusters having 'seizure' as the primary symptom. This aligns with findings by Rao et al.[34], who used LLMs to accurately decode the structural distinctions of proteins from sequence data alone. The workflow shown in Figure 6 illustrates the integration of GPT-4o with unsupervised machine learning algorithms for disease clustering. The interpretation stage leveraged GPT's natural language understanding capabilities to generate human-readable descriptions and underlying connections of the disease clusters, highlighting the key characteristics and insights that make each cluster unique, demonstrated by Savage[35] in his research to reveal how genes and diseases are connected using LLM to make complex concepts and terminologies easy to understand. To interpret and generate meaningful characteristics and differences of each cluster, a function integrating OpenAI's GPT-4o was coded to extract and transform the diseases having seizure as their primary symptom. Table 4 presents the prompt presented to the LLM. The prompt consists of a role for the LLM to *"act as a medical professional"* and instructions to *"thoroughly go through the clusters and highlight the unique characteristics and differences between each cluster"*.

Figure 7 displays a visual representation of the clusters of seizure-based diseases obtained from the best-performing machine learning algorithm (K-means). Although the diseases were all clustered based on symptoms, it was intriguing to dive deeper and explore why diseases sharing the same symptom belonged to other clusters. The discussion focused on the interpretation of seizure-based diseases because seizure was the most occurring symptom, appearing in 184 diseases in the dataset.

The clusters revealed by the GPT model each consist of seizure disorder and epilepsy-related terms. However, the associated factors, seizure characteristics, and types of conditions are unique across the clusters. This corresponds to the work of Cui et al.[36], who found LLMs particularly effective in extracting critical biological insights from cell-type clusters. The key unique differences between the four clusters are discussed below:

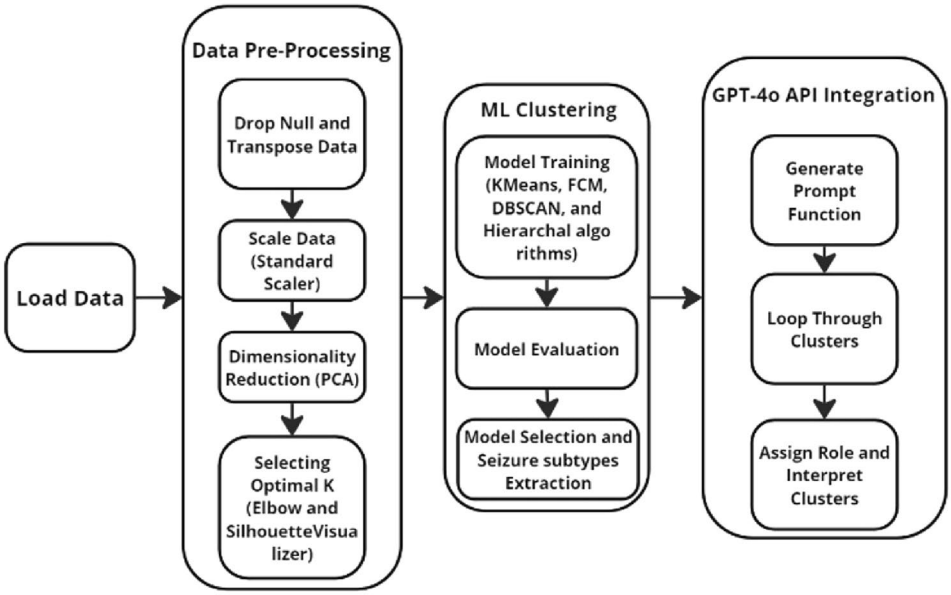

Figure 6. Model Pipeline.

Prompt
"Act as a medical professional"
"You are provided with a list of diseases grouped by clusters. Thoroughly analyse the clusters and highlight the unique characteristics and differences between each cluster. Avoid making any definitive diagnosis. Instead, describe patterns, associations, and distinguishing features that may be clinically relevant."

Table 4. LLM Interpretation Prompt Engineering.

- **Cluster 0 (Green):**
  - Childhood Focus: This cluster primarily includes seizure types and epilepsy syndromes associated with infancy and childhood, with many of the conditions affecting paediatric populations.
  - Alcohol-Related Diseases: These are seizures triggered by alcohol usage, addiction, or withdrawal. This shows a potential link between alcohol and seizures.
- **Cluster 1 (Lemon Green):**
  - Diverse Age Range: This group includes epilepsy syndromes that span a wide age range—from childhood through adulthood.
  - Myoclonic Features: Many of the conditions here involve myoclonic seizures, which are sudden, brief muscle jerks.
  - Focal Localization: Several disorders also have a focal origin, meaning seizures start in specific brain regions like the operculum, lateral temporal lobe, or parietal lobe.
  - Specific Triggers: Some of these epilepsies in cluster 1 are triggered by specific sensory stimuli, such as music or certain smells.
- **Cluster 2 (Yellow):**
  - Reflex Epilepsy: Cluster 3 is notable for reflexes, where seizures are triggered by specific actions or stimuli—like reading, writing, or even touch.
  - Drug-Resistant Epilepsy: Many of the conditions in this group are also drug-resistant, meaning they don't respond well to standard treatments.
- **Cluster 3 (Red):**
  - Febrile, Eclampsia-Related Diseases and Obstetric-Related Diseases: This cluster includes epilepsies linked to fever (febrile seizures), pregnancy-related conditions like eclampsia, and postpartum seizures.
  - Gelastic Seizure: This cluster also features gelastic seizures, which involve sudden, uncontrollable laughter.
  - Drug-Induced and Metabolic Seizures: Additionally, there are conditions caused by drugs or metabolic imbalances, such as seizures from drug withdrawal or electrolyte disturbances.

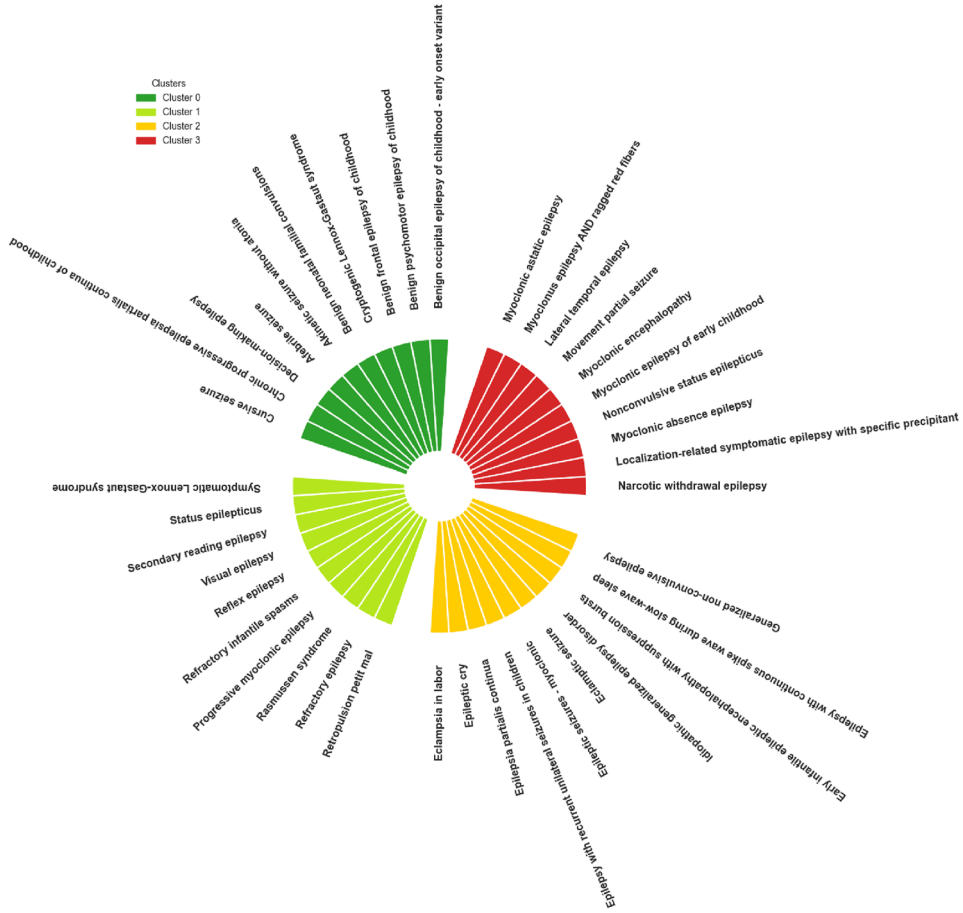

Figure 7. Seizure-based disease clusters.

Cluster	Key Features	ILAE/ICD-11 Subtype Alignment
1 (Green)	Childhood-onset, alcohol-related	- Genetic Epilepsy with Febrile Seizures+ - Alcohol-Related Seizures (ICD-11 8A65)
2 (Lemon Green)	Myoclonic, focal localisation	- Juvenile Myoclonic Epilepsy (ILAE G40.3) - Temporal Lobe Epilepsy (ILAE G40.2)
3 (Yellow)	Reflex epilepsy, drug-resistant	- Reflex Epilepsy (ILAE G40.83) - Drug-Resistant Epilepsy (ILAE G40.8)
4 (Red)	Febrile/eclampsia-related, gelastic	- Febrile Seizures (ICD-11 8A62) - Eclampsia (ICD-11 JA24) - Hypothalamic Hamartoma (ICD-11 8A62.0)

Table 5. Seizure-Based Cluster Classification and ILAE/ICD-11 Alignment.

Each cluster highlights different aspects of epilepsy—like age of onset, seizure type, triggers, and underlying causes. The K-means clustering algorithm indicates that while the only common thread across all clusters is the symptom "seizure," the diversity shows the complexity of epilepsy. More research is needed to uncover deeper connections and shared mechanisms within each group. Therefore, further investigation is required to explore other shared mechanisms and symptoms for all clusters. However, the seizure-based clusters have been matched to the ILAE/ICD-11 subtypes as shown in Table 5. The table presents a classification of seizure clusters based on shared clinical features, linking each to corresponding ILAE and ICD-11 diagnostic subtypes. This alignment aids in understanding the likely aetiology and presentation of different seizure profiles.

Statistical analysis

The Friedman test was conducted to evaluate differences among algorithm performances, yielding a test statistic of 21.26 with a highly significant p-value of 0.000093. Consequently, the null hypothesis that there are no significant differences among the algorithm performances was rejected. To further explore where these differences occur, the Nemenyi post-hoc test was employed. This analysis compares the average ranks of each algorithm against a critical distance (CD), which, at an α-level of 0.05, was calculated to be 1.483. The computed average ranks for the algorithms were as follows: kMeans (1.519), Fuzzy C-Means (1.819), Hierarchical clustering (2.767), and DBSCAN (4.000). Pairwise comparisons indicated that DBSCAN exhibited significantly inferior performance compared to both kMeans (difference = 2.481) and FCM (difference = 2.181), as these differences surpassed the critical distance. Conversely, the difference between DBSCAN and Hierarchical clustering (difference =

1.233) was not statistically significant. There is no statistically significant difference among the top-performing algorithms - kMeans, Fuzzy C-Means, and Hierarchical clustering. Specifically, comparisons between kMeans and Hierarchical clustering (1.248), Fuzzy C-means and Hierarchical Clsutering (0.948), and kMeans and Fuzzy C-means (0.3) all yielded differences below the critical distance. In summary, DBSCAN's performance was significantly inferior compared to both kMeans and Fuzzy C-Means.

Limitations and future work

GPT-4o is a powerful large language model, but it may lack the specialised domain knowledge required for accurate interpretation of unsupervised machine learning outputs in healthcare since unsupervised machine learning often produces naturally ambiguous outputs, with complex patterns and relationships. LLMs may struggle to accurately capture and represent the inherent uncertainty and ambiguity present in such outputs, leading to potentially misleading or overconfident interpretations.

Although LLMs can provide a less complex interpretation and distinction of the clusters, future work should aim at incorporating healthcare professionals and domain experts into the interpretation workflow, allowing them to review, validate, and provide feedback on GPT's interpretations. This collaborative approach can lead to more accurate, clinically relevant, and trustworthy interpretations of unsupervised machine learning outputs, ultimately enhancing the effectiveness and adoption of these models in the healthcare domain. Also, this study is only focused on seizure-based diseases; a future direction is to extend this approach to diverse symptom types (e.g., gastrointestinal, dermatological).

Although we have examined the clusters, confirming that the GPT classifications align with ILAE and ICD-11 diagnostic subtypes, we are not able to perform medical professionals' evaluation, which warrants future research. Furthermore, because seizures were the most common symptom in the 184 diseases, most of the discussions have been based on seizures, which requires further studies to focus on other symptoms.

Conclusion and recommendation

In conclusion, this study provided valuable insights into symptom-based cluster analysis of diseases, offering a more nuanced understanding of disease subtypes. The models were evaluated across 10 unsupervised machine learning evaluation metrics and K-means was the best-performing model. Leveraging the generative power of GPT-4o significantly enhanced the interpretability of the identified subgroups of diseases by providing distinctive characteristics between clusters. Building upon the current findings, future research could explore the integration of diverse data sources, including genetic and imaging data to create a more comprehensive and multi-model approach to disease clustering. While the HSDN dataset provides a valuable foundation for symptom-based clustering, it lacks critical demographic (e.g., age, sex) and temporal (e.g., date of symptom onset) information. This absence limits our ability to explore clinically relevant subgroup patterns, such as age-specific symptom clusters or comorbidity trends across time. Consequently, the findings may not fully capture population-level heterogeneity or the dynamic nature of disease progression. Moreover, the dataset may inherently underrepresent rare diseases or contain ambiguities in symptom reporting, which could influence clustering outcomes and model generalisability. These limitations highlight the need for future work with richer, multi-dimensional datasets that combine symptoms with demographic, clinical, and longitudinal data to improve both interpretability and clinical relevance.

In addition to domain knowledge limitations, LLMs such as GPT are prone to hallucinations. This is particularly concerning in clinical contexts, where misinformation can have serious consequences. Moreover, LLM usage may involve handling sensitive patient data, raising privacy and regulatory issues, especially in relation to GDPR and HIPAA. Finally, while LLM-generated outputs can aid clinical reasoning, they must not replace expert judgment. As such, the deployment of LLMs in clinical settings should be approached with caution, with strong validation protocols, data protection measures, and human oversight mechanisms in place.

Data Availability

The data for this study was sourced from Zhou et al. (2014), who compiled the Human Symptoms–Disease Network dataset. The dataset is publicly available at https://github.com/dhimmel/hsdn.

Received: 25 February 2025; Accepted: 15 September 2025
Published online: 21 October 2025

References

1. Singh, P., Singh, N., Singh, K.K. & Singh, A. Diagnosing of disease using machine learning. *Machine learning and the internet of medical things in healthcare* 89–111 (2021).
2. Nikolaou, V., Massaro, S., Fakhimi, M., Stergioulas, L. & Price, D. Copd phenotypes and machine learning cluster analysis: A systematic review and future research agenda. *Respir. Med.* **171**, 106093 (2020).
3. Oh, S. H., Lee, S. J. & Park, J. Effective data-driven precision medicine by cluster-applied deep reinforcement learning. *Knowl.-Based Syst.* **256**, 109877 (2022).
4. Geerlings, A. D. et al. Case management interventions in chronic disease reduce anxiety and depressive symptoms: A systematic review and meta-analysis. *PLoS ONE* **18**(4), 0282590 (2023).
5. Shaheen, M. Y. Adoption of machine learning for medical diagnosis. *ScienceOpen preprints* (2021).
6. Pillai, R., Oza, P. & Sharma, P. Review of machine learning techniques in health care. 103–111 (2020).
7. Maohua, L. & Zéman, Z. The application of cluster analysis in economics science (2016).
8. Watson, D. S. Interpretable machine learning for genomics. *Hum. Genet.* **141**(9), 1499–1513 (2022).
9. Yu, P., Xu, H., Hu, X. & Deng, C. Leveraging generative ai and large language models: a comprehensive roadmap for healthcare integration. *Healthcare* **11**, 2776 (2023).

10. Ben Dhaou, I. et al. Deep learning and generative ai for monolithic and chiplet soc design and verification:: A survey. *Found. Trends® Electron. Des. Autom.* **14**(4), 245–294 (2025)
11. Dave, T., Athaluri, S. A. & Singh, S. Chatgpt in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front. Artif. Intel.* **6**, 1169595 (2023).
12. Lin, Y. Common and co-occurring symptoms experienced by patients with gastric cancer. *Number 2/March 2020***47**(2), 187–202 (2020).
13. Mousavi, E., Hassanzadeh Keshteli, A., Sehhati, M., Vaez, A. & Adibi, P. Exploring new subgroups for irritable bowel syndrome using a machine learning algorithm. *Sci. Rep.* **13**(1), 18483 (2023).
14. Byale, A. et al. High-dimensional clustering of 4000 irritable bowel syndrome patients reveals seven distinct disease subsets. *Clin. Gastroenterol. Hepatol.* (2022).
15. Qiu, X., Wang, H., Tan, X. & Jin, Y. Cvdllm: Automated cardiovascular disease diagnosis with large-language-model-assisted graph attentive feature interaction. *IEEE Trans. Artif. Intel.* (2025).
16. Yin, Z. et al. A novel approach to unlocking the synergy of large language models and chemical knowledge in biomedical signal applications. *Biomed. Signal Process. Control* **103**, 107388 (2025).
17. Harris, C. S. et al. Advances in conceptual and methodological issues in symptom cluster research: A 20-year perspective. *Adv. Nurs. Sci.* **45**(4), 309–322 (2022).
18. Liu, Q. et al. Symptom-based patient stratification in mental illness using clinical notes. *J. Biomed. Inform.* **98**, 103274 (2019).
19. Anggraeni, W., Yuniarno, E. M., Rachmadi, R. F. & Purnomo, M. H. et al. Fuzzy c-means and social network analysis combination for better understanding the patient-based spread of dengue fever with climate and geographic factors. *Int. J. Intel. Eng. Syst.***15**(3) (2022).
20. Newcomer, S. R., Steiner, J. F. & Bayliss, E. A. Identifying subgroups of complex patients with cluster analysis. *Am. J. Manag. Care* **17**(8), 324–332 (2011).
21. Melcer, T., Zouris, J., MacGregor, A., Crouch, L. D., Sheu, C. R. & Galarneau, M. Cluster analysis of outpatient prescription medications after combat-related amputations: A retrospective study. *Archives of Physical Medicine and Rehabilitation* (2025).
22. Nicolet, A. et al. Exploring patient multimorbidity and complexity using health insurance claims data: A cluster analysis approach. *JMIR Med. Inform.* **10**(4), 34274 (2022).
23. Miller, A. C., Arakkal, A. T., Koeneman, S. H., Cavanaugh, J. E. & Polgreen, P. M. A clinically-guided unsupervised clustering approach to recommend symptoms of disease associated with diagnostic opportunities. *Diagnosis* **10**(1), 43–53 (2022).
24. Suharjo, B. & Utama, M. S. Y. K-means cluster analysis of sex, age, and comorbidities in the mortalities of covid-19 patients of indonesian navy personnel. *JISA (Jurnal Informatika dan Sains)* **4**(1), 17–21 (2021).
25. Habib, G. & Qureshi, S. Convolutional neural networks (cnn) and dbscan clustering for sars-cov challenges: Complete deep learning solution. In: *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022***2**, 473–491 (2022). Springer
26. ElShawi, R., Sherif, Y., Al-Mallah, M. & Sakr, S. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Comput. Intell.* **37**(4), 1633–1650 (2021).
27. Zhou, X., Menche, J., Barabási, A.-L. & Sharma, A. Human symptoms-disease network. *Nat. Commun.* **5**(1), 4212 (2014).
28. Vats, D. & Sharma, A. Dimensionality reduction techniques: Comparative analysis. *J. Comput. Theor. Nanosci.* **17**(6), 2684–2688 (2020).
29. Shi, C. et al. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP J. Wirel. Commun. Netw.* **2021**(1), 31 (2021).
30. Juanita, S. & Cahyono, R. D. K-means clustering with comparison of elbow and silhouette methods for medicines clustering based on user reviews. *Jurnal Teknik Informatika (JUTIF)* **5**(1), 283–289 (2024).
31. Nagaraj, B. & Malagi, K.B. Boosting the accuracy of optimisation chatbot by random forest with halving grid search hyperparameter tuning. *ICTACT Journal on Soft Computing***13**(3) (2023).
32. Heidari, S., Alborzi, M., Radfar, R., Afsharkazemi, M. A. & Rajabzadeh Ghatari, A. Big data clustering with varied density based on mapreduce. *J. Big Data* **6**, 1–16 (2019).
33. Elsborg, J. & Salvatore, M. Using llm models and explainable ml to analyse biomarkers at single cell level for improved understanding of diseases. *bioRxiv* 2023–08 (2023).
34. Rao, R. M. et al. Msa transformer. In: *International Conference on Machine Learning* 8844–8856 (2021). PMLR
35. Savage, N. Drug discovery companies are customizing chatgpt: here's how. *Nat. Biotechnol.* **41**(5), 585–586 (2023).
36. Cui, H. et al. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nat. Methods* 1–11 (2024).

Acknowledgements
We thank all the reviewers for their time and consideration in reviewing our article and helping improve its quality.

Author contributions
Conceptualisation, EO, EI. Investigation and methodology, EO, EI. Project administration, EO, CPE. Resources, CI, IBD. Supervision, EI, CPE, CI, IBD. Writing of the original draft, EO, EI. Writing of the review and editing, EO, CI, IBD. Software EO, CPE, CI. Validation, EO, EI, CPE. Formal analysis, EO, EI, IBD. Data curation, EO, CPE. Visualization, EO, EI.

Funding
There was no funding to complete this research.

Declarations

Ethics approval and consent to participate
No ethics approval or consent to participate was needed for this article.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to E.I. or I.B.D.