scientific reports

Check for updates

OPEN

Variational graph autoencoder for reconstructed transcriptomic data associated with NLRP3 mediated pyroptosis in periodontitis

Pradeep K. Yadalam[1], Prabhu Manickam Natarajan[2]✉ & Carlos M. Ardila[3]✉

The NLRP3 inflammasome, regulated by TLR4, plays a pivotal role in periodontitis by mediating inflammatory cytokine release and bone loss induced by *Porphyromonas gingivalis*. Periodontal disease creates a hypoxic environment, favoring anaerobic bacteria survival and exacerbating inflammation. The NLRP3 inflammasome triggers pyroptosis, a programmed cell death that amplifies inflammation and tissue damage. This study evaluates the efficacy of Variational Graph Autoencoders (VGAEs) in reconstructing gene data related to NLRP3-mediated pyroptosis in periodontitis. The NCBI GEO dataset GSE262663, containing three samples with and without hypoxia exposure, was analyzed using unsupervised K-means clustering. This method identifies natural groupings within biological data without prior labels. VGAE, a deep learning model, captures complex graph relationships for tasks like link prediction and edge detection. The VGAE model demonstrated exceptional performance with an accuracy of 99.42% and perfect precision. While it identified 5,820 false negatives, indicating a conservative approach, it accurately predicted 4,080 out of 9,900 positive samples. The model's latent space distribution differed significantly from the original data, suggesting a tightly clustered representation of the gene expression patterns. K-means clustering and VGAE show promise in gene expression analysis and graph structure reconstruction for periodontitis research.

Keywords Periodontitis, Variational autoencoders, Inflammasome, K means clustering

Periodontitis is a chronic inflammatory disease characterized by the destruction of the supporting structures of the teeth, particularly the periodontal ligament and alveolar bone[1,2]. It is primarily caused by an aberrant immune response to dysbiotic oral microbiota, leading to inflammation, tissue destruction, and tooth loss. One of the critical pathways implicated in the pathogenesis of periodontitis is the activation of the NLRP3 inflammasome[3], which plays a pivotal role in the inflammatory responses associated with various chronic diseases, including periodontitis.

The body's innate immune system uses pattern recognition receptors (PRRs)[4,5] to quickly identify and respond to threats like pathogens and cellular imbalances. Nucleotide-binding and leucine-rich repeat (NLR) receptors play a crucial role in the innate immune system by activating the inflammasome, a multi-protein complex that processes and releases proinflammatory cytokines like IL-1 and IL-18[6]. Dysregulation of the inflammasome has been linked to various autoinflammatory disorders. Recent studies highlight the role of the inflammasome complex in periodontal tissue immunology, highlighting the need for a balance between inflammation and cell death for tissue regeneration. The NLRP3 inflammasome is more prevalent in patients with chronic and advanced periodontitis, requiring TLR4 protein control for effective management. Osteoclastogenesis relies on the RANK/RANKL/OPG axis. The NLRP3 inflammasome mediates the release of inflammatory cytokines and affects bone loss caused by *Porphyromonas gingivalis*, with NLRP3-KO mice showing higher OPG and lower RANKL levels[6–8].

Periodontal disease often leads to hypoxic conditions due to inflammation, tissue damage, and reduced blood flow. These conditions, especially during acute exacerbations, promote the survival of anaerobic bacteria and exacerbate the inflammatory response. To survive, cells activate adaptive mechanisms, such as AMPK

[1]Department of Periodontics, Saveetha Dental College, Saveetha Institute of Medical and Technology Sciences, SIMATS, Saveetha University, Chennai 600077, Tamil Nadu, India. [2]Department of Clinical Sciences, Center of Medical and Bio-allied Health Sciences and Research, College of Dentistry, Ajman University, Ajman 346, United Arab Emirates. [3]Department of Basic Sciences, Faculty of Dentistry, Universidad de Antioquia U de A, Medellín 050010, Colombia. ✉email: prabhuperio@gmail.com; martin.ardila@udea.edu.co

activation[9,10], which maintains cellular energy homeostasis and promotes mitochondrial biogenesis, glucose uptake, and fatty acid oxidation. AMPK also exhibits anti-inflammatory properties. The NLRP3 inflammasome is a crucial sensor of cellular stress and danger signals, triggering the cleavage of pro-caspase-1 and the secretion of proinflammatory cytokines. In periodontitis, it is activated by bacterial components and host-derived signals, leading to pyroptosis. Pyroptosis, a form of programmed cell death, can be detrimental in chronic inflammatory diseases like periodontitis. It exacerbates the inflammatory response and tissue damage, perpetuating the cycle of periodontal inflammation and destruction[11].

Understanding and recreating the intricate gene networks in complex diseases is crucial for unraveling disease mechanisms and identifying potential therapeutic targets. In recent years, advancements in high-throughput omics technologies, such as transcriptomics, have enabled the generation of vast amounts of data that capture the molecular state of diseases like periodontitis[12]. Reconstructed gene data, derived from comparative genomics and evolutionary studies, is crucial in various fields of biological research and periodontal personalized medicine. It helps understand evolutionary relationships, functional annotation, genomic studies, disease research, synthetic biology, conservation biology, enhanced genomic resources, and insights into complex traits. Reconstructed gene data also contributes to large genomic databases, promotes open science, and enables the study of complex traits influenced by multiple genes. Deep Learning (DL) approaches are useful for integrated multi-omics analysis of cancer data, but high-dimensional data can be imbalanced. One study used VAE and its improved version, Maximum Mean Discrepancy VAE (MMD-VAE)[13], to classify transcriptional subtypes of ovarian cancer with 93.2–95.5% and 87.1–95.7% accuracy, respectively, and application of VAE is not much done in periodontal disease.

One promising approach for reconstructing gene networks from transcriptomic data is deep learning models, specifically Variational Graph Autoencoders (VGAEs). VGAEs[14] leverage the power of graph encoding and decoding, combined with variational inference, to learn latent representations of genes and their interactions in a data-driven manner. Training VGAEs on gene expression profiles makes it possible to infer regulatory interactions and reconstruct gene regulatory networks.

Therefore, in this study, we aim to evaluate the performance of the VGAE model in reconstructing gene data involved in NLRP3-mediated pyroptosis in periodontitis. We will utilize publicly available transcriptomic datasets that include gene expression profiles from samples related to NLRP3-mediated pyroptosis in periodontitis. The evaluation will enhance the VGAE model's effectiveness and understanding of NLRP3-mediated pyroptosis in periodontitis, potentially aiding in developing targeted therapeutic strategies.

Materials and methods

Figure 1 shows the methodology used in this study. The study utilized a stepwise methodology to investigate the relationship between transcriptomics data, NLRP3-mediated pyroptosis, and periodontitis. Initially, transcriptomics data were collected and preprocessed to ensure high-quality input for further analysis. The processed data were then fed into a Variational Graph Autoencoder (VGAE), which was employed to model and integrate complex gene expression relationships, particularly focusing on pathways linked to NLRP3-mediated pyroptosis. The analysis aimed to uncover significant gene interactions and their role in the pyroptosis pathway. The findings were subsequently correlated with periodontitis to determine the extent of their association, highlighting key biomarkers and pathways involved. This comprehensive methodological approach provided insights into the mechanistic link between NLRP3-mediated pyroptosis and periodontitis, bridging transcriptomic alterations with disease outcomes.

Preparation of the dataset

Data were obtained using the NCBI GEO dataset-GSE262663[15], and this dataset comprises six samples that examine the effects of hypoxia on pyroptosis. It includes THP-1 cells with three samples for both hypoxic and control conditions. The data were subjected to k-means clustering for further analysis.

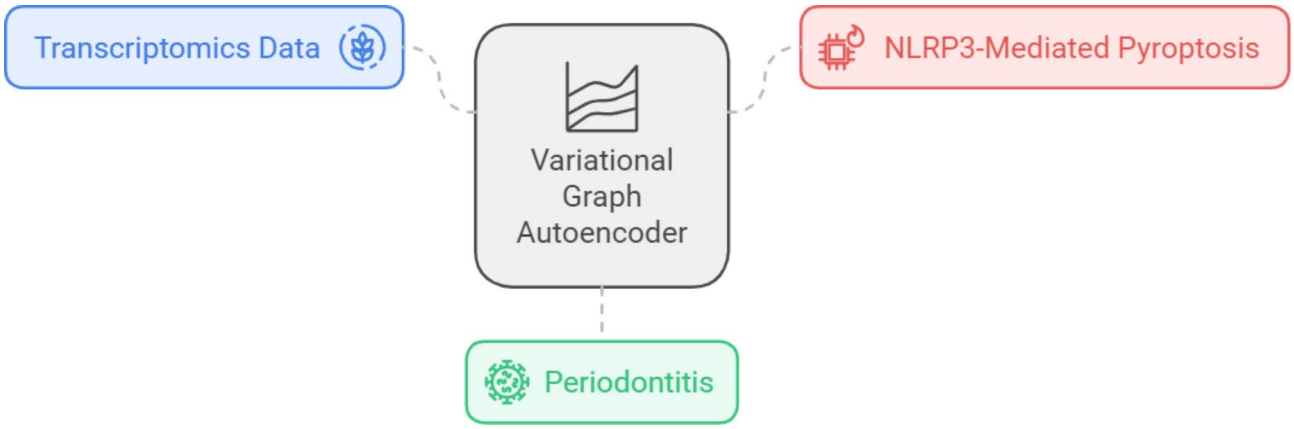

Fig. 1. Study methodology.

Figure 2 presents a detailed flowchart illustrating the stepwise methodology implemented in the study. The workflow begins with dataset preparation, where transcriptomics data from the NCBI GEO Dataset (GSE260263) is collected and preprocessed to ensure quality and compatibility for further analysis. From this starting point, the methodology branches into two main analytical approaches: unsupervised K-Means clustering and the Variational Graph Autoencoder (VGAE) framework.

The K-Means clustering process follows a structured sequence that includes data collection, preprocessing, feature selection, dimensionality reduction, and implementation. These steps are designed to refine the data

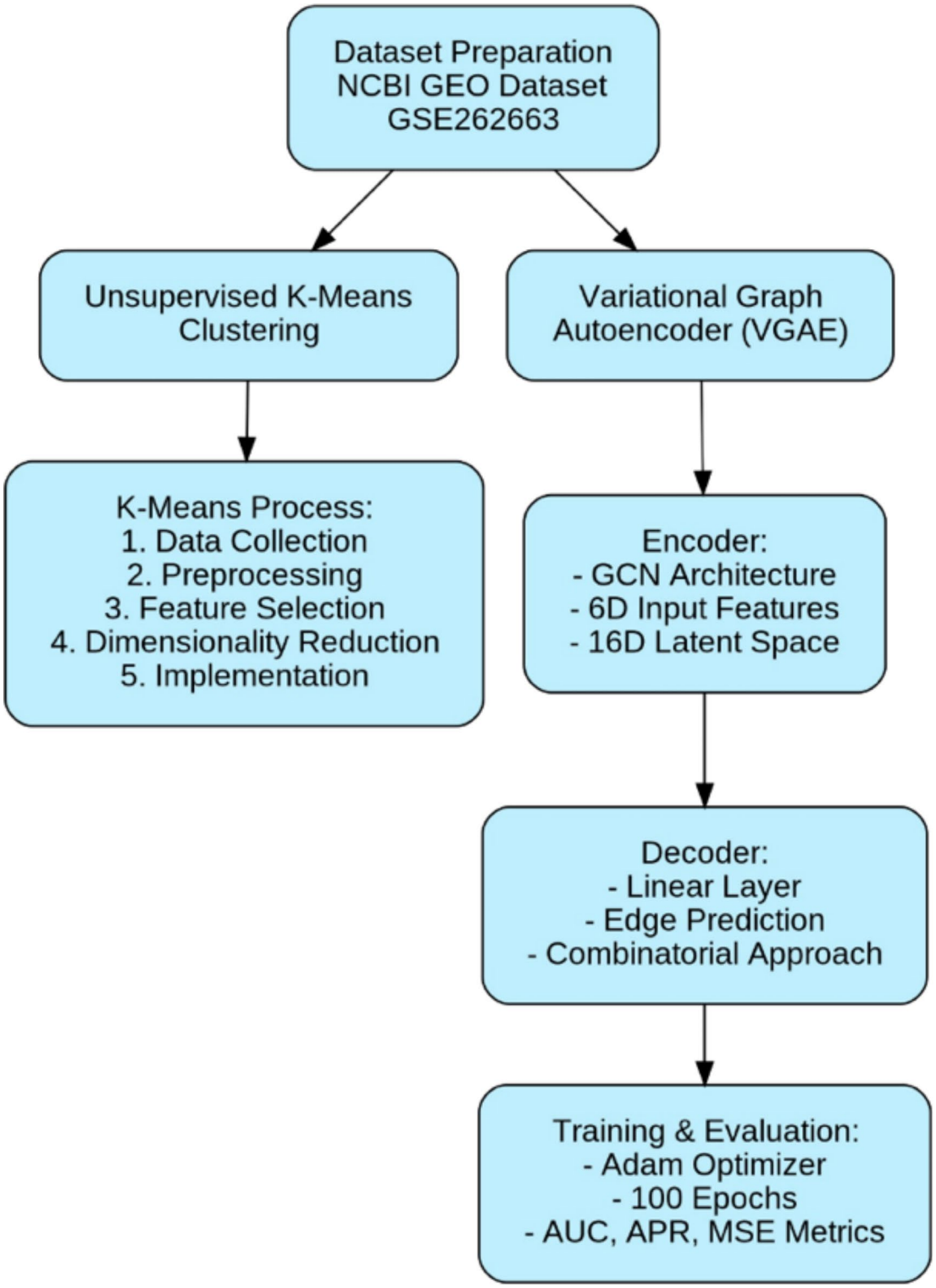

Fig. 2. Workflow of the study design.

and extract key features that facilitate downstream analysis. In parallel, the VGAE method operates through an encoder-decoder architecture. The encoder employs a Graph Convolutional Network (GCN) with 60 input features and compresses the data into a 16-dimensional latent space, enabling the identification of latent relationships within the graph data. The decoder then reconstructs these relationships through a linear layer, focusing on edge prediction and utilizing a combinatorial approach to uncover significant connections within the dataset.

Finally, the outputs from both analytical streams are integrated into the training and evaluation phase. The Adam Optimizer is used to refine model performance, while key evaluation metrics such as Area Under the Curve (AUC), Average Precision Rate (APR), and Relative Squared Error (RSE) are calculated to assess predictive accuracy and overall model robustness. This flowchart effectively conveys the sequential workflow, illustrating the comprehensive and methodical approach used to process, analyze, and evaluate the transcriptomics data.

Unsupervised K-means clustering for data preparation

In omics studies, unsupervised K-means clustering is used to identify natural groupings in high-dimensional biological data without prior labels. The process involves data collection, preprocessing, feature selection, dimensionality reduction, and implementation. The choice of the number of clusters is crucial, and the iterative clustering process involves assigning each data point to the nearest centroid using a distance metric. K-means clustering can be integrated with other analysis methods to refine insights or validate results. This structured architecture allows researchers to draw meaningful biological conclusions and advance our understanding of various biological systems. This study involves initializing K clusters from the data, randomly selecting data points as cluster centroids, assigning them to the closest cluster, updating the cluster centroids, and repeating steps until convergence is reached. The final output represents the K-means clustering solution, with each data point assigned to the cluster with the closest centroid and the centroids representing the mean coordinates of the data points within each cluster.

Variational graph autoencoder (VGAE) model architecture

The Variational Graph Autoencoder (VGAE)[12] is a generative model designed for graph-structured data that combines graph convolutional networks (GCNs) with variational inference. VGAE uses an encoder to learn node embeddings and a decoder to predict edges, effectively capturing complex graph relationships for tasks like link prediction and community detection. The Variational Graph Autoencoder (VGAE) is a generative model designed to effectively capture the latent structures within graph-structured data by leveraging both a probabilistic framework and graph convolutional techniques. VGAE was implemented for analysis using Google Colab in a Python environment. The architecture consists of two primary components: an encoder and a decoder.

Encoder architecture

The encoder employs a Graph Convolutional Network (GCN) to process the input graph through multiple layers, extracting node features and incorporating neighboring information, thereby aggregating node features across the local neighborhood and effectively capturing relational patterns.

Input features

Node representation in a graph is determined by a 6-dimensional feature vector, which may include node degree, labels, or other domain-specific attributes affecting the graph's topology.

Latent space dimension

The model's 16 latent space dimensions balance expressiveness and computational efficiency, allowing it to represent complex relationships and node variations and prevent overfitting.

Decoder architecture

The decoder is a linear layer that accepts latent node representations from the encoder, calculating the likelihood of an edge between nodes. Edge prediction is a combinatorial approach that evaluates potential edges once.

Training strategy

The VGAE model uses the Adam optimizer for parameter updates, utilizing adaptive learning rate features and training over 100 epochs to minimize a composite loss function. The reconstruction loss measures the discrepancy between actual graph edges and decoder predictions, while the Kullback-Leibler Divergence measures the difference between learned and prior Gaussian distributions. The VGAE model's performance is evaluated using Area Under the Curve (AUC), Average Precision, Mean Squared Error (MSE), and Graph Sparsity. AUC evaluates the trade-off between true and false positive rates, while APR summarizes the precision-recall curve. MSE quantifies the average squared difference between predicted and actual edges, while Graph Sparsity evaluates the model's capacity to discern significant connections while avoiding over-connection. The VGAE model uses GCN architecture and probabilistic underpinnings to provide valuable insights into graph data structure, pattern recognition, community detection, and link prediction. It captures intricate graph relationships while maintaining generalizability and robust performance across diverse tasks (Fig. 2).

Results

K-means clustering

The K-means clustering algorithm identified three distinct clusters, each exhibiting unique gene expression patterns across the samples. Cluster 0, the largest cluster comprising 25,948 genes, likely represents genes with consistent expression patterns across all samples. In contrast, Cluster 1, the smallest cluster with 2,222 genes,

appears to be enriched with genes showing higher expression levels in samples H4 and H5. Cluster 2, containing 7,630 genes, may represent elevated expression in samples C1, C2, and C3.

The K-means clustering algorithm successfully partitioned the dataset into distinct groups based on their gene expression features, minimizing within-cluster variance while maximizing between-cluster variance. This iterative process involved initialization, assignment, update, and iteration steps to achieve the optimal clustering solution.

The effectiveness of the K-means clustering algorithm is visualized in Fig. 3, which displays a scatter plot of the gene expression data. The x-axis and y-axis represent the First and Second Principal Components, respectively. The data points are colored according to their assigned cluster, revealing well-separated groups. The color gradient suggests a continuous variable, possibly indicating density. The distinct clustering pattern indicates that the K-means algorithm successfully identified distinct groups of genes with similar expression profiles.

To determine the optimal number of clusters, we employed the elbow method, a heuristic used in cluster analysis (Fig. 4). The elbow method plots the explained variation against the number of clusters, and the point of inflection (elbow) indicates the optimal number of clusters. In Fig. 4, the Within-Cluster Sum of Squares (WCSS) decreases as clusters increase, with a pronounced elbow at 3 or 4 clusters. The curve flattens around this point, suggesting that adding more clusters results in only a marginal decrease in WCSS. This indicates that 3 or 4 clusters is the optimal number of clusters for our gene expression data.

The K-means clustering algorithm yielded three distinct clusters: Cluster 0 (25,948 genes), Cluster 1 (2,222 genes), and Cluster 2 (7,000 genes). The size of each cluster can be interpreted as its biological relevance to the functions or expressions of genes in each cluster. Cluster 0, comprising the largest number of genes, indicates a broad range of shared gene expressions. Cluster 1, with 2,222 genes, represents a specialized subset with unique

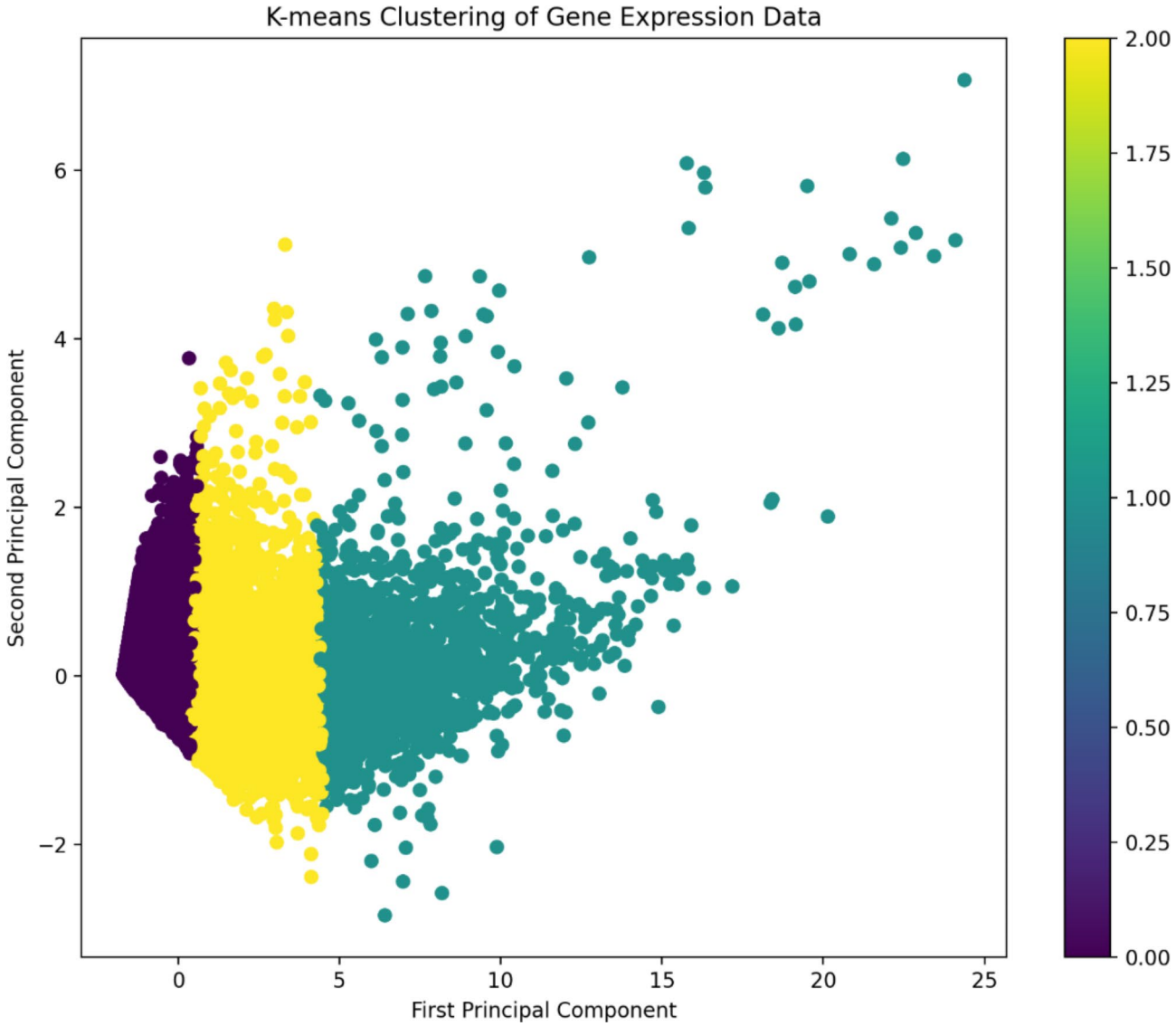

Fig. 3. K-means clustering of gene expression data.

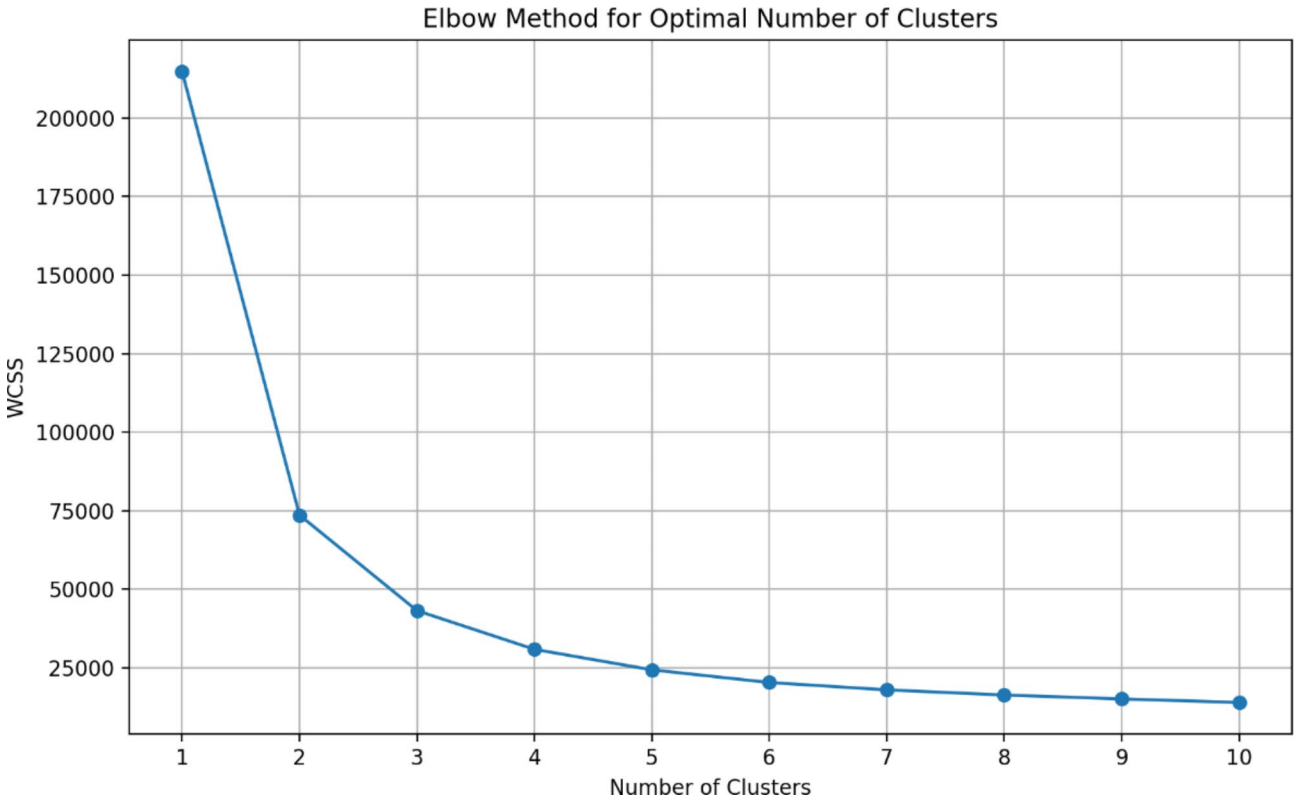

Fig. 4. Elbow method for determining optimal number of clusters.

features. Cluster 2, having a moderate number of genes, captures specific patterns but is not as extensive as Cluster 0.

Key Pyroptosis genes, including GSDMD, CASP1, NLRP3, IL1B, and IL18, exhibit moderate expression levels across samples. Cluster 0 contains housekeeping genes and moderate to low expression of pyroptosis-related genes, while Cluster 1 is rich in highly expressed genes and contains several pyroptosis-related genes. The high expression of CASP1 indicates active inflammatory signaling, while GSDMD expression suggests potential pyroptotic activity. Differential expression suggests distinct pyroptosis pathway regulatory states. The clustering pattern indicates two distinct cellular states: baseline (Cluster 0) with normal homeostatic functions and activated (Cluster 1) with enhanced pyroptosis-related gene expression, suggesting active inflammatory signaling and potential pyroptotic activity.

VGAE model results

Training results

The Graph VAE model underwent 100 training epochs, consistently decreasing the loss metric over time. Notable loss values recorded during training included 1.4689, 1.4026, 1.3939, 1.3927, and 1.3912, respectively. The model generated a torch. Size([1000, 8]) dataset with 1000 samples, each with eight features. The original data's mean values and standard deviations were similar to the generated data, indicating that the model successfully learned from the provided data.

Performance evaluation

The model demonstrated exceptional performance with an accuracy of 99.42% and perfect precision, as indicated by a confusion matrix showing no false positives and a high number of true negatives (990,100). However, it identified only 4,080 true positives out of 9,900 positive samples, resulting in 5,820 false negatives. This reflects a conservative prediction approach, prioritizing the avoidance of false positives while missing some positive instances. Given the imbalanced dataset composed primarily of negative samples, the high accuracy may be misleading, highlighting the importance of considering precision, recall, and F1-score for a more comprehensive evaluation of the model's effectiveness in predicting rare positive cases.

Generative model evaluation

The generative model produced data with a distribution that differed significantly from the original data. The generated data had smaller standard deviations and means, indicating a tightly clustered origin in the latent space.

Reconstructed metrics

Evaluating the VGAE model on the sampled data provided insights into the model's performance in reconstructing the graph structure. The reconstructed metrics indicated that the model achieved an AUC (Area Under the Curve) score of 0.4766, suggesting that the model's ability to distinguish between connected and non-connected node pairs was slightly better than random guessing. The Average Precision (AP) score was 0.0090, indicating that the model struggled to predict the presence of edges in the graph accurately. This low AP score reflects the challenges in capturing the true positive connections within the graph.

Addressing class imbalance is crucial for improving VGAE performance on imbalanced datasets. Corrective strategies like re-sampling and synthetic data augmentation can mitigate these effects.

Figure 5 illustrates the Epoch Loss Curve, a graphical representation of the loss value during model training over epochs. The curve starts at a higher loss value and shows a steep decline initially, indicating effective learning. After the initial drop, the curve flattens out, suggesting stabilizing performance. The overall trend indicates that the model is improving as training progresses, with the loss decreasing significantly in the first few epochs.

Figure 6 shows a side-by-side comparison of the Original Adjacency Matrix and the Reconstructed Adjacency Matrix. The Original Matrix represents the initial connections in the network, characterized by a predominantly white color scheme and large dimensions. In contrast, the Reconstructed Matrix, derived from the VGAE model's analysis, features a stark contrast with black areas indicating stronger connections. Notably, the original matrix appears sparse and less structured, while the reconstructed matrix exhibits a clearer pattern of connections, possibly indicating a more organized relationship structure.

The Mean Squared Error (MSE) between the true and predicted adjacency matrices is 0.3850, representing the average squared difference between actual and predicted edge probabilities. This value highlights the discrepancies in the model's reconstruction of the graph.

The graph sparsity, calculated as 0.9901, indicates that the reconstructed graph is highly sparse, with most possible edges absent. This sparsity is consistent with the original graph's structure, exhibiting a sparsity of 0.9901.

The original graph consists of 1000 nodes and 9900 edges. The graph's sparsity reflects the proportion of absent edges relative to the total possible edges in a fully connected graph. The high sparsity in real-world networks is a common characteristic, where connections between nodes are relatively rare.

A visual comparison of the original and reconstructed adjacency matrices (Fig. 5) provides insight into the model's performance. The original adjacency matrix represents true node connections, while the reconstructed matrix shows predicted connections, highlighting areas where the model effectively captures the graph structure.

While the VGAE model captures some aspects of the graph structure, it requires improvement in accurately reconstructing graph connectivity. Further tuning of parameters, architecture, or training processes may enhance performance in capturing data patterns.

Discussion

NOD-like receptors (NLRs)[16] represent a family of intracellular sensors that enable the inflammatory protease caspase-1 to generate proteolytic cleavage and release of pro-interleukin (IL)-1β and to mediate pyroptosis. *Porphyromonas gingivalis* can affect the glycemic level and induce NLRP3 inflammasome-related IL-1β secretion

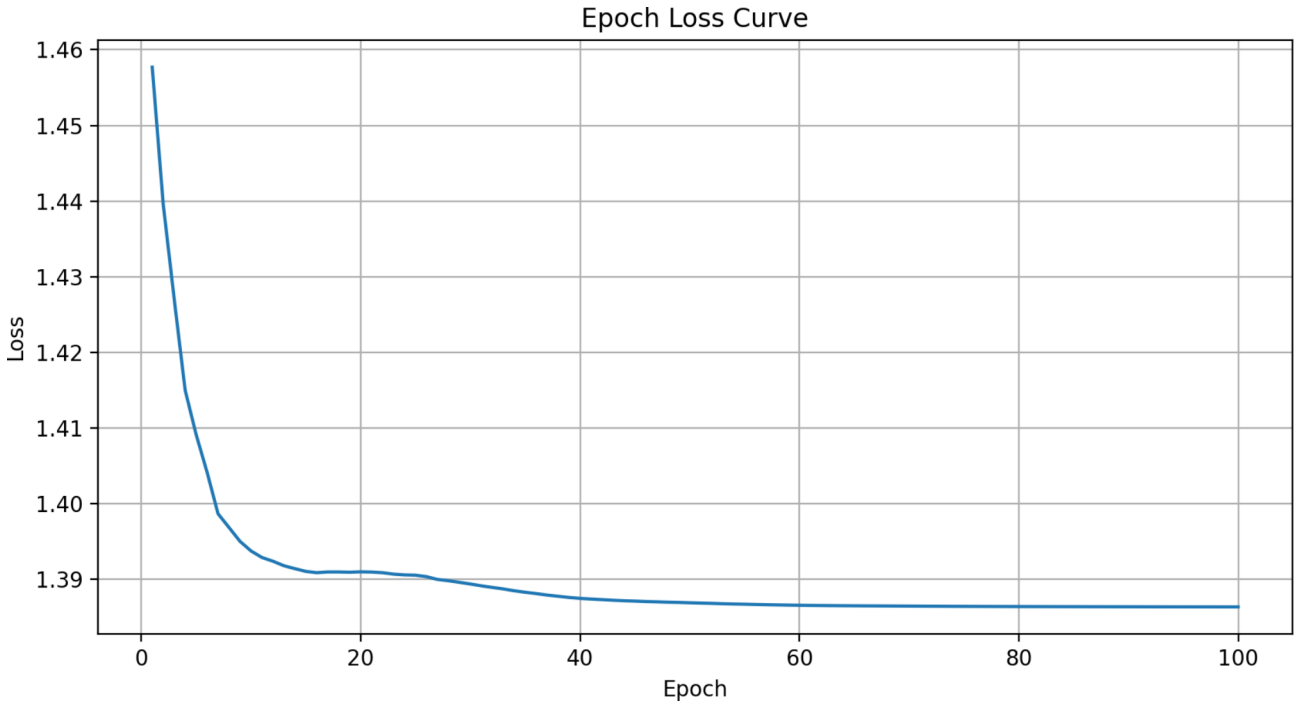

Fig. 5. Epoch loss curve for graph VAE model training.

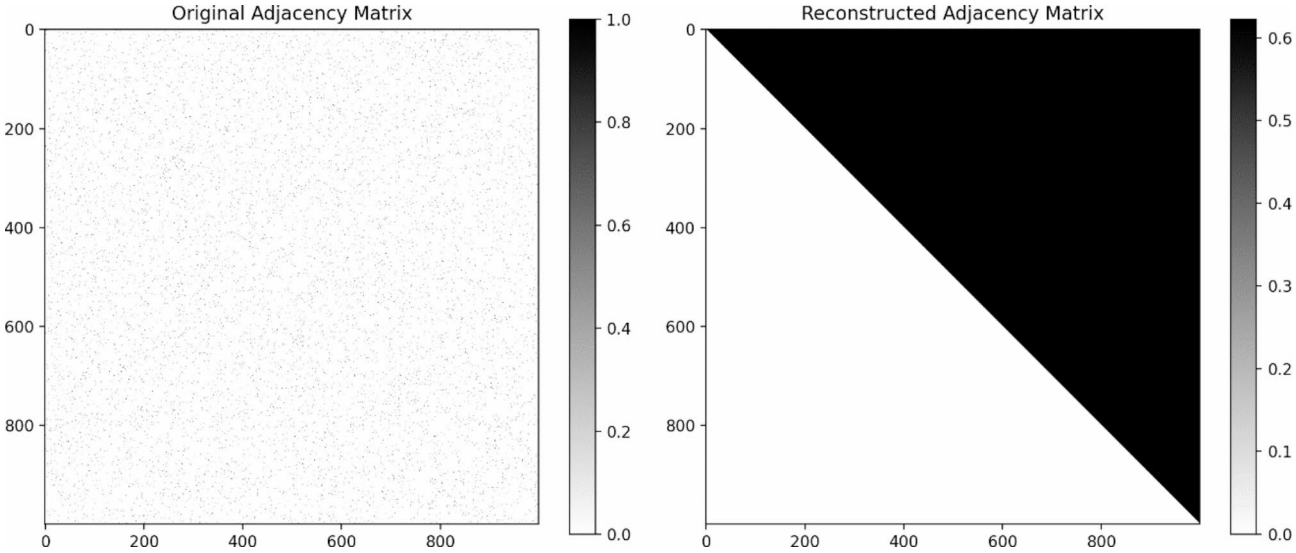

Fig. 6. Comparison of original and reconstructed adjacency matrices.

and pyroptotic cell death. Pyroptosis manifests distinct features from apoptosis and necrosis, in which the plasma membrane ruptures and disintegrates rapidly, releasing intracellular contents into the extracellular milieu[17,18]. This study has identified that crucial genes are ABCG1, RGS16, SSTR2, and NPTX1, which are involved in pyroptosis in periodontitis and are crucial for maintaining cellular membrane integrity and macrophage metabolic functions. These genes play a biologically significant role in preserving cellular membrane integrity and regulating the metabolic functions of macrophages. Ribonuclease activity, essential for RNA metabolism and degradation, influences cytokine production and immune response. Abnormalities in ribonuclease activity could exacerbate sustained inflammation, a hallmark of periodontitis.

The cytoprotective gene Nrf2 plays a role in activating the antioxidant element ARE, increasing the expression of antioxidant enzymes HO-1 and NQO-1[16–19], and removing overproduced ROS to defend against oxidative stress and inflammatory damage. Zhenxing Zhao showed that by establishing a model of experimental diabetes mellitus-periodontitis in rats, we found that IL-1β and gasdermin D were highly expressed, leading to aggravated periodontal tissue destruction. Hyperglycemia triggers IL-1β production[20–23] and pyroptosis in macrophages, suggesting potential therapeutic strategies for diabetes mellitus-periodontitis by modifying autophagy activity and targeting the ROS-inflammasome pathway. These studies motivated us to retrieve this genomic dataset for our generative AI model analysis[16,24–26] (Figs. 1 and 2).

This study used unsupervised K-means clustering (Figs. 3 and 4), which groups similar samples based on their omics profiles, facilitating the discovery of disease subtypes for personalized periodontal medicine. It also reduces noise and improves signal detection in high-throughput omics datasets. K-means is a flexible, scalable, and adaptable omics data analysis tool with interdisciplinary applications beyond genetics, including environmental omics and microbiomics. The challenges faced include the subjective selection of K, assumptions about cluster size, and sensitivity to outliers. This study showed K-means clustering results in three clusters: Cluster 0 (25,948 genes), Cluster 1 (2,222 genes), and Cluster 2 (7,000 genes). Each cluster's size indicates biological relevance, with Cluster 0 having a larger number, indicating shared gene expressions[27].

VAEs provide an unsupervised approach for generating meaningful latent representations of integrated data, which can be exploited for analysis or deployed on other heterogeneous data sets. Previous work relates to employing VAEs for constructing latent representations and analyzing transcriptomic cancer data from TCGA, reducing dimensionality while identifying patterns and genes relevant to different cancer types and subtypes. One previous study has shown that Attention VGAE is a novel spatial transcriptomics technique that effectively captures spatial domain information, addressing low-quality gene expression calibration and balancing local and global structures. A multi-head attention system uses a graph convolutional neural network framework to accurately model tissue microenvironments, enhancing spatial domain detection, gene expression patterns, and genomic research and one recent study on variational autoencoders in cell-to-cell communications VGAE-CCI that identifies cell communication in tissues effectively uses incomplete data and has demonstrated superior results compared to other methods[28]. These studies motivated us to use the VGAE model for periodontal inflammasome data[29–32]. Reconstruction of the genomic data will help solve crucial complex biological problems related to pyroptosis induced by nlrp3-based periodontitis.

Our VGAE model results showed a 99.42% accuracy and perfect precision but only identified 4,080 true positives out of 9,900 positive samples, resulting in 5,820 false negatives (Figs. 5 and 6). This conservative prediction approach may be misleading due to the imbalanced dataset. The generative model produces data with smaller standard deviations and means, indicating a tightly clustered origin. The VGAE model's performance in reconstructing the graph structure is slightly better than random guessing but struggles to accurately predict edges, similar to previous studies[33–35]. XOmiVAE, a deep learning model that reveals gene and latent dimension contributions for classification predictions, correlates between genes and dimensions, and explains supervised

and unsupervised clustering results, showing potential for drug-omics to achieve greater accuracy and VGAE-CCI is a deep learning tool that effectively and reliably detects cell communication in tissues, even with incomplete data, proving its effectiveness in tests[33,36]. While supervised learning aims at learning an embedded representation of the input, the VAEs focus on learning the underlying distribution of the input data, allowing for data generation[18,37]. scGMM-VGAE model enhances cell clustering performance in single-cell RNA sequencing data by combining a Gaussian mixture model and variational graph autoencoder, outperforming four baseline methods in generating latent representations, especially in identifying rare cell types in omics data[38]. These studies highlight the importance of VGAE, which has its limitations and advantages.

K-means clustering and the VGAE model[28,31–33] are effective for gene expression analysis and graph structure reconstruction but have limitations like spherical cluster assumption, initialization sensitivity, K choice, and scalability. Future K-means directions include improved initialization techniques, domain knowledge, cluster quality metrics, distance metrics, and dimensionality reduction. However, K-means assumes spherical clusters, which may not be applicable in real-world data, and results can be sensitive to initialization, making comparisons challenging. For the VGAE model, future directions include model architecture enhancements, hybrid approaches, improving edge prediction, transfer learning, and hyperparameter optimization. However, the VGAE model faces challenges such as imbalanced data, sensitivity to input features, reconstruction error, and interpretability. To address these limitations, focus on improving initialization techniques, incorporating domain knowledge, adjusting distance metrics, and enhancing the VGAE model's performance. Additionally, addressing the limitations of both models could enhance their utility in biological research.

The Variational Graph Autoencoder model can be improved by combining it with clustering algorithms. Hybrid models can be developed using DBSCAN, Hierarchical Clustering, and Self-Organizing Maps. DBSCAN enhances positive case identification by identifying dense regions. At the same time, hierarchical clustering generates a tree of clusters for a nuanced data structure, and self-organizing maps preserve topological properties for well-organized representation for improving recall and precision.

Oversampling techniques like SMOTE increase positive samples by duplicating or creating synthetic copies, which is particularly effective in generating new synthetic examples based on existing minority class instances. Undersampling reduces negative samples but may lead to data loss. Combined sampling, combining oversampling and undersampling, maintains dataset integrity while achieving balance. Synthetic data generation uses generative models, graph data enhancement uses noise and perturbation, and cost-sensitive learning penalizes minority class misclassifications by focusing on less frequent positive samples. Ensemble methods combine multiple models and transfer learning on imbalanced datasets, providing better insights into performance under class imbalance using metrics like F1-score, Precision-Recall AUC, or Balanced Accuracy. Implement methods by applying re-sampling to training sets, monitoring model complexity and domain relevance, and using validation sets to ensure synthetic data accurately represents distribution and relationships.

Variational Graph Autoencoders in periodontitis research offer significant benefits in understanding the disease's genetic and biological foundations by reconstructing synthetic data. VGAEs are powerful tools for capturing complex relationships in multi-dimensional datasets and identifying gene networks concerning pyroptosis. This study model can be beneficial for future researchers to explore further and contribute to the scientific community regarding periodontal disease by generating new models and data. This study contributes to scientific knowledge sharing, fostering interdisciplinary networks, and raising public awareness of the genetic aspects of periodontal pyroptosis.

Conclusion

The study highlights the potential of advanced analytical AI methods like K-means clustering and Variational Graph Autoencoders (VGAEs) in understanding NLRP3 inflammasome-related pyroptosis in periodontal disease. K-means clustering identifies gene expression profiles, enabling personalized treatment. VGAE model achieves high accuracy and precision but requires refinement for better interpretability and performance. Future research should improve initialization techniques, incorporate domain-specific knowledge, and explore hybrid approaches to enhance the analytical capabilities of K-means clustering and the VGAE model in analyzing gene expression data and reconstructing graph structures. This could lead to innovative therapeutic strategies targeting disease mechanisms, improving patient outcomes in periodontal inflammation.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 16 October 2024; Accepted: 10 January 2025
Published online: 14 January 2025

References

1. Boonyaleka, K. et al. Fusobacterium nucleatum infection activates the noncanonical inflammasome and exacerbates inflammatory response in DSS-induced colitis. *Eur. J. Immunol.* **53**, e2350455 (2023).
2. Ribeiro, C. C. C. et al. Systemic circulating inflammatory burden and periodontitis in adolescents. *Clin. Oral Investig.* **25**, 5855–5865 (2021).
3. Wang, Y. et al. AMPK/mTOR/p70S6K axis prevents apoptosis of Porphyromonas gingivalis-infected gingival epithelial cells via bad(Ser136) phosphorylation. *Apoptosis* **28**, 1012–1023 (2023).
4. Xu, X. et al. Pyroptosis in periodontitis: from the intricate interaction with apoptosis, NETosis, and necroptosis to the therapeutic prospects. *Front. Cell. Infect. Microbiol.* **12**, 953277 (2022).
5. Zhao, P. et al. Hyperglycaemia-associated macrophage pyroptosis accelerates periodontal inflamm-aging. *J. Clin. Periodontol.* **48**, 1379–1392 (2021).

6. Zhao, Z. et al. Hyperglycemia aggravates periodontitis via autophagy impairment and ROS-inflammasome-mediated macrophage pyroptosis. *Int. J. Mol. Sci.* **24**, e7375 (2023).
7. Liu, H., Liu, Y., Fan, W. & Fan, B. Fusobacterium nucleatum triggers proinflammatory cell death via Z-DNA binding protein 1 in apical periodontitis. *Cell. Commun. Signal.* **20**, 196 (2022).
8. Ozkocer, O. Immunohistochemical analysis with apoptosis and autophagy markers in periodontitis and peri-implantitis: clinical comparative study. *J. Periodontal Res.* **58**, 456–464 (2023).
9. Li, X. et al. Silibinin attenuates experimental periodontitis by downregulation of inflammation and oxidative stress. *Oxid Med Cell Longev*. 5617800 (2023).
10. Hoare, A., Soto, C., Rojas-Celis, V. & Bravo, D. Chronic inflammation as a link between periodontitis and carcinogenesis. *Mediators Inflamm*. 1029857 (2019).
11. Pan, S. et al. Identification of ferroptosis, necroptosis, and pyroptosis-associated genes in periodontitis-affected human periodontal tissue using integrated bioinformatic analysis. *Front. Pharmacol.* **13**, 1098851 (2022).
12. Benkirane, H., Pradat, Y., Michiels, S. & Cournède, P. H. CustOmics: a versatile deep-learning based strategy for multi-omics integration. *PLoS Comput. Biol.* **19**, e1010921 (2023).
13. Hira, M. T. et al. Integrated multi-omics analysis of ovarian cancer using variational autoencoders. *Sci. Rep.* **11**, 6265 (2021).
14. Withnell, E., Zhang, X., Sun, K. & Guo, Y. XOmiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data. *Brief. Bioinform*. **22**, bbab454 (2021).
15. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2013).
16. Xu, Y. et al. BML-111 inhibits H2O2-induced pyroptosis and osteogenic dysfunction of human periodontal ligament fibroblasts by activating the Nrf2/HO-1 pathway. *BMC Oral Health*. **24**, 40 (2024).
17. Hashim, N. et al. NLRP3 inflammasome in autoinflammatory diseases and periodontitis advance in the management. *J. Pharm. Bioallied Sci.* **16**, S1110–S1119 (2024).
18. Zhou, X. et al. Metformin ameliorates the NLPP3 inflammasome mediated pyroptosis by inhibiting the expression of NEK7 in diabetic periodontitis. *Arch. Oral Biol.* **116**, 104763 (2020).
19. Wang, L., Pu, W., Wang, C., Lei, L. & Li, H. Microtubule affinity regulating kinase 4 promoted activation of the NLRP3 inflammasome-mediated pyroptosis in periodontitis. *J. Oral Microbiol.* **14**, 2015130 (2022).
20. Wang, Z. et al. FoxO1 knockdown inhibits RANKL-induced osteoclastogenesis by blocking NLRP3 inflammasome activation. *Oral Dis.* **30**, 3272–3285 (2024).
21. Wang, Z. et al. Artesunate ameliorates ligature-induced periodontitis by attenuating NLRP3 inflammasome-mediated osteoclastogenesis and enhancing osteogenic differentiation. *Int. Immunopharmacol.* **123**, 110749 (2023).
22. Hu, A., Xiao, F., Wu, W., Xu, H. & Su, J. LincRNA-EPS inhibits caspase-11 and NLRP3 inflammasomes in gingival fibroblasts to alleviate periodontal inflammation. *Cell. Prolif.* **57**, e13539 (2024).
23. Jiang, X. et al. Dioscin alleviates periodontitis by inhibiting NLRP3 inflammasome activation via regulation of K + homeostasis and mitochondrial function. *Int. J. Biol. Sci.* **20**, 1375–1388 (2024).
24. Xu, S. et al. Mesenchymal stem cells and their extracellular vesicles in bone and joint diseases: targeting the NLRP3 inflammasome. *Hum. Cell*. **37**, 1276–1289 (2024).
25. Zhou, X. et al. Inhibition of METTL3 alleviates NLRP3 inflammasome activation via increasing ubiquitination of NEK7. *Adv. Sci. (Weinh)*. **11**, e2308786 (2024).
26. Rusetskaya, N. Y., Loginova, N. Y., Pokrovskaya, E. P., Chesovskikh, Y. S. & Titova, L. E. Redox regulation of the NLRP3-mediated inflammation and pyroptosis. *Biomed. Khim.* **69**, 333–352 (2023).
27. Huang, L. et al. Deep learning methods for omics data imputation. *Biology (Basel)*. **12**, 1297 (2023).
28. Lei, L. et al. Attention-guided variational graph autoencoders reveal heterogeneity in spatial transcriptomics. *Brief. Bioinform*. **25**, bbae173 (2024).
29. Simidjievski, N. et al. Variational autoencoders for cancer data integration: design principles and computational practice. *Front. Genet.* **10**, 1205 (2019).
30. Rong, Z. et al. MCluster-VAEs: an end-to-end variational deep learning-based clustering method for subtype discovery using multi-omics data. *Comput. Biol. Med.* **150**, 106085 (2022).
31. Eltager, M. et al. Benchmarking variational autoencoders on cancer transcriptomics data. *PLoS One*. **18**, e0292126 (2023).
32. Rivero-Garcia, I., Torres, M. & Sánchez-Cabo, F. Deep generative models in single-cell omics. *Comput. Biol. Med.* **176**, 108561 (2024).
33. Allesøe, R. L. et al. Discovery of drug-omics associations in type 2 diabetes with generative deep-learning models. *Nat. Biotechnol.* **41**, 399–408 (2023).
34. Marino, J. Predictive coding, variational autoencoders, and biological connections. *Neural Comput.* **34**, 1–44 (2021).
35. Ranjbari, S. & Arslanturk, S. Integration of incomplete multi-omics data using knowledge distillation and supervised variational autoencoders for disease progression prediction. *J. Biomed. Inf.* **147**, 104512 (2023).
36. Zhang, T. et al. VGAE-CCI: variational graph autoencoder-based construction of 3D spatial cell-cell communication network. *Brief. Bioinform*. **26**, bbae619 (2024).
37. Kalafut, N. C., Huang, X. & Wang, D. Joint variational autoencoders for multimodal imputation and embedding. *Nat. Mach. Intell.* **5**, 631–642 (2023).
38. Lin, E. et al. scGMM-VGAE: a gaussian mixture model-based variational graph autoencoder algorithm for clustering single-cell RNA-seq data. *Mach. Learn. : Sci. Technol.* **4**, 035013 (2023).

Acknowledgements

We would like to thank the Center of Medical and Bioallied Health Sciences and Research, Ajman University, Ajman, UAE.

Author contributions

Conceptualization, P. Y, P. N. and C. A.; Data curation, P. Y, P. N. and C. A.; Formal analysis, P. Y, P. N. and C. A.; Funding acquisition, P. N.; Investigation, P. Y, P. N. and C. A.; Methodology, P. Y, P. N. and C. A.; Project administration, P.Y.; Resources, P. Y, P. N.; Software, P. Y.; Supervision, P. Y, and C. A.; Validation, P. Y, P. N. and C. A.; Visualization, P. Y, P. N. and C. A.; Writing – original draft, P. Y, P. N. and C. A.; Writing – review & editing, . All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.M.N. or C.M.A.