Artículo Original

Volumen 16, N° 32. Enero-Junio 2026
Pradeep Kumar y Col.
Pg. 83-119

Acta BIOCLÍNICA
ISSN Electrónico: 3105-403X

DOI:

AUTOENCODERS AND SYNTHETIC DATA FOR PREDICTING HUB GENES IN PERIAPICAL PAIN

Pradeep Kumar Yadalam[1]**; Carlos M. Ardila**[1,2]

1. Department of Periodontics,Saveetha Dental College and Hospital,Saveetha Institute of Medical and Technical Sciences,Saveetha University, Chennai- 600077,Tamil Nadu, India.
2. Basic Sciences Department, Faculty of Dentistry, Biomedical Stomatology Research Group, Universidad de Antioquia U de A, Medellín, 050010, Colombia.

EMAIL: martin.ardila@udea.edu.co

ABSTRACT

Introduction: Periapical pain, commonly associated with apical periodontitis, arises from inflammation and infection of the periapical tissues. **Objectives:** This study evaluated the application of stacked autoencoders and synthetic data generation in predicting hub genes related to periapical pain. **Methods:** Differential gene expression analysis was conducted using the GEO GSE237398 dataset, focusing on patients with pain responding to mechanical stimulation. The top 250 differentially expressed genes were analyzed, and an interactome was constructed using Cytoscape. CytoHubba was employed to identify the top 60 hub genes, emphasizing their regulatory significance. These hub genes were labeled and used to train

Received: July 31, 2025
Accepted: September 20, 2025

models based on stacked autoencoders and synthetic data generation. **Results:** The autoencoder model demonstrated a moderate accuracy of 68%, correctly classifying 68% of the test samples. The model's performance was reflected in a weighted average precision of 0.74, a recall of 0.68, and an F1 score of 0.69. Notably, the model performed well for the hub gene class but struggled with the non-hub class, likely due to dataset imbalance. In contrast, the synthetic data generation model exhibited precision scores of 0.93–1.00 and recall scores of 92%–100%, underscoring its high accuracy in generating synthetic data. **Conclusions:** This study highlights the potential of artificial intelligence–driven models in predicting hub genes associated with periapical pain. Though moderate in accuracy, the autoencoder model provides a foundation for further refinement. The random forest classifier, trained on original and synthetic data, showed high accuracy and reliability.

KEYWORDS: generative AI; periapical pain; hub genes; autoencoders; random forest; omics.

DIFFERENCES IN PHYSICAL CONDITION AND MOTOR COORDINATION BY RISK OF MORTALITY IN SCHOOLCHILDREN: PILOT STUDY

ABSTRACT

Introducción: El dolor periapical, comúnmente asociado con la periodontitis apical, surge de la inflamación e infección de los tejidos perirradiculares. Los genes hub son fundamentales para

Received: July 31, 2025
Accepted: September 20, 2025

el desarrollo de fármacos y la medicina de precisión. Los autoencoders, como MLP-SAE y SMALF, ofrecen herramientas prometedoras para predecir genes hub, mejorar conjuntos de datos y generar modelos que preservan la privacidad. **Objetivos:** Este estudio evaluó la aplicación de autoencoders apilados y generación de datos sintéticos en la predicción de genes hub relacionados con el dolor periapical. **Métodos:** Se realizó un análisis de expresión génica diferencial utilizando el conjunto de datos GEO GSE237398, centrado en pacientes con dolor que respondían a estímulos mecánicos. Se analizaron los 250 genes diferencialmente expresados más relevantes y se construyó un interactoma mediante Cytoscape. Se empleó CytoHubba para identificar los 60 principales genes hub, destacando su importancia regulatoria. Estos genes se etiquetaron y utilizaron para entrenar modelos basados en autoencoders apilados y generación de datos sintéticos. **Resultados:** El modelo de autoencoder mostró una precisión moderada del 68%, clasificando correctamente el 68% de las muestras de prueba. Su rendimiento se reflejó en una precisión promedio ponderada de 0.74, un recall de 0.68 y un F1-score de 0.69. Destacó su buen desempeño en la clase de genes hub, pero dificultades en la clase no hub, probablemente por desbalance del dataset. En contraste, el modelo de generación de datos sintéticos mostró precisiones de 0.93–1.00 y recalls del 92%–100%, evidenciando alta precisión en la generación de datos sintéticos. **Conclusiones:** Este estudio resalta el potencial de los modelos basados en inteligencia artificial para predecir genes hub asociados al dolor periapical. Aunque de precisión moderada, el

Received: July 31, 2025
Accepted: September 20, 2025

autoencoder sienta bases para mejoras futuras. El clasificador random forest, entrenado con datos originales y sintéticos, mostró alta precisión y confiabilidad.

PALABRAS CLAVE: inteligencia artificial generativa; dolor periapical; genes hub; autoencoders; random forest; ómicas.

INTRODUCTION

Toothache is very common and can significantly impact quality of life and productivity. Odontogenic pain is transmitted by the trigeminal nerve, with the dentine–pulp complex and periodontium being the most common sources [1]. Dental periapical pain, often linked to apical periodontitis, is a common endodontic condition caused by inflammation and infection of the periapical tissues. Treatment involves traditional and advanced methods, with the host immune response playing a crucial role [2]. Dental afferent nerve fibers enter the tooth via the apical foramina and carry signals from the pulp to the brainstem and thalamus. Afferent fibers are perceived as pain, and sympathetic efferent fibers play a role in hemodynamic control. The periodontium contains mechanoreceptors contributing to proprioception during mastication, whereas the pulp has fewer receptors, resulting in poorly localized pain. Pulpal nerve fibers include myelinated A-fibers and unmyelinated C-fibers, with A-fibers primarily responding to stimuli and producing sharp pain. The stimulation of

Received: July 31, 2025
Accepted: September 20, 2025

pulp afferent fibers causes the release of several peptides, leading to vasodilation, inflammation, and sensitization of nociceptors. This results in hyperalgesia, allodynia, and peripheral and central sensitization, causing increased pain perception. Periradicular tissues have abundant neurons originating from the trigeminal ganglion [3,4].

Apical periodontitis [5] with periapical pain is a complex inflammatory tooth infection characterized by bone loss, granulomatous tissue formation, and variable pain experiences among patients. Approximately 60% of patients exhibit mechanical allodynia, whereas 40% remain asymptomatic despite inflammation and bone loss. Differences in pain responses are noted between sexes, with women showing a 25% greater reduction in pain thresholds and a higher likelihood of persistent pain after treatment compared to men. These variations highlight the differential mechanisms in the trigeminal system and nociception-related sexual dimorphism [1,6]. The lack of relief from antimicrobial drugs suggests persistent nociceptive changes, indicating a need for deeper investigation into the mechanisms behind these pain phenotypes.

The transcriptomic hub genes [7] are a centralized platform for analyzing and interpreting transcriptomic data, which is crucial in understanding gene expression and regulation. One recent study used WGCNA to analyze microarray data from 13 inflamed and 11 normal pulps. The data were categorized into 22 modules, with the

Received: July 31, 2025
Accepted: September 20, 2025

dark grey module having the highest correlation with pulpitis. Five hub genes (*HMOX1*, *LOX*, *ACTG1*, *STAT3*, *GNB5*) were identified. RT-qPCR showed differences in expression levels of these genes in the inflamed dental pulp. Pulp capping using iRoot BP plus reversed the expression levels of these genes [8]. One previous genomics study identified seven functional categories of genes related to apical periodontitis, primarily associated with immune cell function [9,10]. Differences in gene expression were observed between male and female patients, with immune response genes being the most prevalent. Genes *C3*, *ERAP2*, and *CHIT1* were significantly upregulated in symptomatic women, suggesting potential therapeutic interventions [11]. Predicting hub genes is crucial because they play a significant role in molecular and cellular processes, influencing the behavior of biological systems. These genes are often key regulators or essential components of biological pathways, providing insights into underlying mechanisms [12–14]. A recent study used the Gene Expression Omnibus database to identify 843 differentially expressed genes related to pulpitis. A protein–protein interaction network was constructed, and the key functional subset was identified through GO and KEGG enrichment analyses. The intersection of these networks and pulpitis-related genes led to the discovery of 20 seed genes and 50 highly correlated genes. The top 50 genes were then screened, and four hub genes were identified as coexpressed with

Received: July 31, 2025
Accepted: September 20, 2025

chemokine-related genes and significantly upregulated in the pulpitis group [15].

An autoencoder [16,17] is a neural network trained to reconstruct input data, consisting of input, hidden, and output layers. It minimizes reconstruction loss and learns neural network weights without extra information [12]. Autoencoders are useful for predicting hub genes because they can learn complex representations of gene expression data. By stacking multiple layers, each layer can learn higher-level features and capture intricate patterns in the data. These features can capture the underlying structure or interactions among genes, which are crucial in the biological network. Hub genes are highly connected and pivotal in gene networks, providing insights into specific conditions or diseases. A recent study introduced a deep auto-encoder model, MLP-SAE, for predicting gene expression from single-nucleotide polymorphism genotypes. It outperformed existing methods such as LASSO and random forests on real genomic datasets from yeast, highlighting deep learning's suitability for genomic predictive modeling [18]. One previous study developed a computational framework called SMALF to predict unknown miRNA-disease associations. It used a stacked autoencoder to learn miRNA and disease latent features and XGBoost to predict associations. Cross-validation experiments showed that SMALF [19] achieved the best AUC value. Three case studies showed that SMALF was effective in identifying unknown miRNA-disease associations. Once trained, it can

Received: July 31, 2025
Accepted: September 20, 2025

predict hub genes on new, unseen data by encoding the input data into a lower-dimensional representation and decoding it back to reconstruct the original data [16,17]. Synthetic data for hub genes can be useful for several reasons, including data augmentation, validation and benchmarking, privacy preservation, and exploration of data characteristics [20,21]. It can explore unique data scenarios and hub gene behavior. No known studies have evaluated stacked autoencoders and created synthetic hub genes for periapical pain. This study evaluated the use of autoencoders and synthetic data generation to predict hub genes for periapical pain.

This study aimed to evaluate the effectiveness of autoencoders and synthetic data generation techniques in accurately predicting hub genes associated with periapical pain. The specific objectives included (1) developing an autoencoder model tailored to gene prediction, (2) generating synthetic data to enhance the dataset, (3) identifying key hub genes linked to periapical pain, (4) assessing model performance, and (5) validating predicted hub genes through biological or clinical experiments.. We aimed to improve model robustness, predict hub genes, and assess the model's predictive accuracy and robustness against standard machine-learning methods. We hypothesized that using autoencoders and synthetic data generation techniques would improve the predictive accuracy of hub genes linked to periapical pain. These genes will be

Received: July 31, 2025
Accepted: September 20, 2025

significant biomarkers for diagnosis and therapeutic targets.

2. Materials and Methods

2.1 Differential gene expression analysis

The study used GEO [22,23] accession number GSE237398. This is a dataset aimed to determine genomic differences between symptomatic and asymptomatic patients with pain on mechanical stimulation. The study investigated periapical pathology in 12 patients with symptomatic apical periodontitis, focusing on mechanical allodynia. Biopsies from six female and six male patients underwent endodontic microsurgery, with each sex having three symptomatic and three asymptomatic cases. The GEO R tool was used for differential gene expression analysis, obtaining the top 250 genes with a p-value of 0.05 and a fold change of 1.5, comparing male and female patients with and without pain.

2.2 Cytoscape and CytoHubba

Cytoscape [24] is a powerful software tool for visualizing and analyzing biological networks, including gene regulatory networks. It was used to build an interactome of the top 250 differentially expressed genes, representing the interactions between genes or proteins. This provides insights into the biological processes and pathways affected by these genes. After building the interactome, CytoHubba, a Cytoscape plugin, was used to identify the top 60 hub genes. Hub genes are highly connected within a network, indicating their potential importance and

Received: July 31, 2025
Accepted: September 20, 2025

centrality in the regulatory network. CytoHubba offers various methods for identifying hub genes, such as the maximum clique centrality (MCC) method. By identifying the top 60 hub genes using MCC, researchers can focus on genes that play a central role in the regulatory network associated with periapical pain.

2.3 Prediction of hub genes: dataset preparation

After selecting and identifying hub genes, these were labeled and subjected to training; 80% and 20% of the test data were subjected to stacked autoencoders and synthetic data generation, respectively.

2.4 Autoencoder hyperparameters and architecture

The autoencoder is an unsupervised deep-learning technique that compresses input data into a lower-dimensional representation and reconstructs it [15]. The process involves three dense layers with 32, 16, and eight neurons using ReLU activation, mirroring the encoder. Using the Adam optimizer, the autoencoder is trained to minimize the mean squared error (MSE) between the input and its reconstruction. The model is trained for 50 epochs with a batch size of 32, with 20% of the training data used for validation.

The encoder is then used to transform the original data into the learned eight-dimensional representation, which is used for the classification task. A logistic regression model performs the classification task using the encoded features. Data

Received: July 31, 2025
Accepted: September 20, 2025

preparation involves encoding the target variable (y) using LabelEncoder and splitting the data into training and testing sets. The model is trained on the encoded features of the training set, and predictions are made on the encoded features of the test set.

Several evaluation metrics are used to evaluate the performance of the classification model: accuracy, precision, recall, F1 score, and classification report. The stacked autoencoder was chosen for its compact data representation, potentially capturing important features that might not be immediately apparent in the original feature space. Logistic regression was chosen for its simplicity and interpretability, allowing us to assess how well the encoded features could linearly separate classes. This approach combines unsupervised feature learning (autoencoder) with supervised classification (logistic regression), attempting to leverage the autoencoder’s ability to capture complex patterns in the data for improved classification performance. The moderate performance of the final model suggests that although the autoencoder captured some useful information, room for improvement remains in either the feature learning process or the classification method.

The autoencoder features a three-layer encoder and a three-layer decoder, employing ReLU activation functions in each layer. Hyperparameters were fine-tuned through grid search and cross-validation. To mitigate the risk of overfitting, dropout layers with a rate of 0.4 and L2 regularization were included, along with

Received: July 31, 2025
Accepted: September 20, 2025

early stopping predicated on the improvement of validation loss.

2.5 Synthetic data generation architecture for creation of hub genes

The synthetic minority oversampling technique [20,25] for nominal and continuous features (SMOTENC) was used for synthetic data. It was particularly useful for datasets with numerical and categorical features. The process involved loading the original dataset, identifying categorical variables, and transforming them into a numerical format using a LabelEncoder. The SMOTENC algorithm was implemented using the imbalanced-learn library, with the categorical_features parameter set to specify which columns contained categorical data. SMOTENC was applied to the training data to generate synthetic samples, balancing the dataset. The SMOTENC algorithm creates synthetic examples in the feature space, handling continuous and categorical features.

Received: July 31, 2025
Accepted: September 20, 2025

The random forest classifier from scikit-learn was used in an analysis, initializing a model with 100 estimators. The data were split into training and testing sets using an 80/20 ratio. The model was trained using the fit method and used to predict on a test set. Performance metrics were calculated, and feature importance was extracted. The random forest classifier aggregates predictions from multiple decision trees, with randomness for decorrelation. The final prediction is made by averaging or taking a majority vote.

These combined methods—synthetic data generation using SMOTENC and classification using random forests—allowed us to address potential class imbalance issues in the original data and create a robust, high-performing classification model. The synthetic data helped augment the training set, potentially improving the model's generalization ability, and the random forest algorithm provided a powerful and interpretable classification method.

3. Results

The differential gene expression analysis results revealed 250 distinct genomic profiles in male and female patients, revealing potential biomarkers and pathways involved in pain perception and mechanical stimuli response. Figure 1 shows a volcano plot of the differential gene expression with downregulated and upregulated genes with red dots as upregulated and blue dots as

Received: July 31, 2025
Accepted: September 20, 2025

downregulated with |log2 FC| > 1.5 and p < 0.05

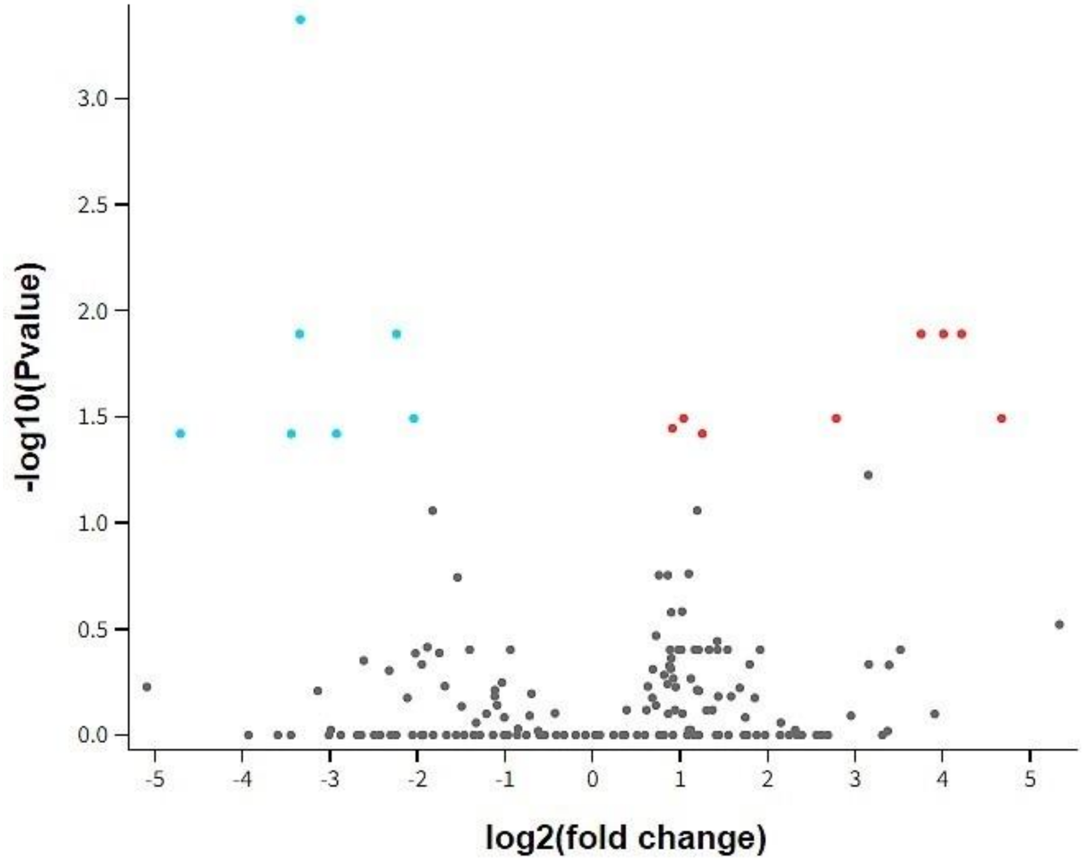


Figure 1. Volcano plot of differential gene expression.

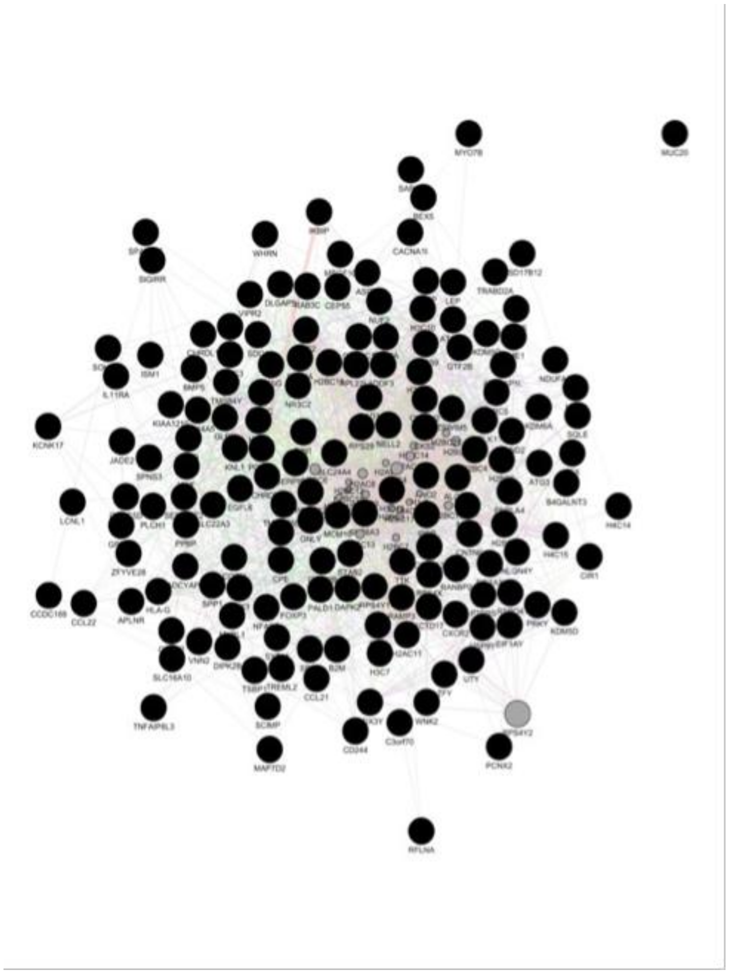
3.1 Network results

The network had 183 nodes and 2,070 edges, indicating a densely interconnected structure. It had an average number of neighbors of 17.901, suggesting rich interconnectivity. The network density was relatively low, suggesting the potential for

Received: July 31, 2025
Accepted: September 20, 2025

Acta BIOCLÍNICA
ISSN Electrónico: 3105-403X

more connections. The shortest path between nodes was 2.282 steps, suggesting efficient communication. The network had a maximum distance of 5 and a minimum of 3 from a central node to the farthest node. The clustering coefficient indicated a moderately high tendency for nodes to form tightly knit groups. Figure 2 shows the interactome of the top 250 genes associated with dental pain.



Received: July 31, 2025
Accepted: September 20, 2025

Figure 2. Interactome of the top 250 genes associated with dental pain.

3.2 Hub network

The hub network, consisting of 50 nodes and 614 edges, was densely interconnected, with an average of 24.56 neighbors per node. Its path length was 1.691, exhibiting high clustering with a coefficient of 0.872. The network's density was 0.501, with nearly half the possible connections realized. It had moderate heterogeneity and a centralization score of 0.243. It was entirely connected, with just one connected component. Figure 3 shows the top 60 hub genes identified using the MCC method from CytoHubba. The top identified hub genes included *H2BC4*, *H2AC11*, *BIRC5*, *H2BC11*, *PRKY*, *H1-5*, *NUF2*, *KDM6A*, and *H3C12*.

3.3 Autoencoder results

The model achieved an overall accuracy of 68%, indicating that it correctly classified 68% of the samples in the test set. This performance was moderate, suggesting that the encoded features from the autoencoder captured some meaningful information for classification. The average precision was approximately 0.74, the recall was 0.68, and the F1 score was approximately 0.69. The precision (0.74) indicates that the model had a relatively low false-positive rate, and the recall (0.68) means that the model correctly identified 68% of the actual positive cases on average. The F1 score

Received: July 31, 2025
Accepted: September 20, 2025

(0.69) is the harmonic mean of precision and recall, providing a single score that balances both metrics.

The classwise performance showed how the model performed for each class: hub class (precision: 0.48, recall: 0.73, F1-score: 0.58) and non-hub class (precision: 0.85, recall: 0.66, F1-score: 0.74). The model had higher recall but lower precision for the hub class, which was good at identifying hub samples (73% of actual hubs correctly identified), but also incorrectly classified some non-hubs as hubs (low precision of 0.48). The support values (15 for hub and 35 for non-hub) suggest that the dataset was imbalanced, with more non-hub samples than hub samples. Figure 4 shows the epoch loss curve of the autoencoder model. The image shows an autoencoder's training history graph, focusing on the MSE over several epochs. Key features include the x-axis (epoch) representing the number of training epochs and the y-axis (MSE) representing the error value. The model's training and validation losses decreased over time, indicating learning and improvement in training and unseen data.

Received: July 31, 2025
Accepted: September 20, 2025

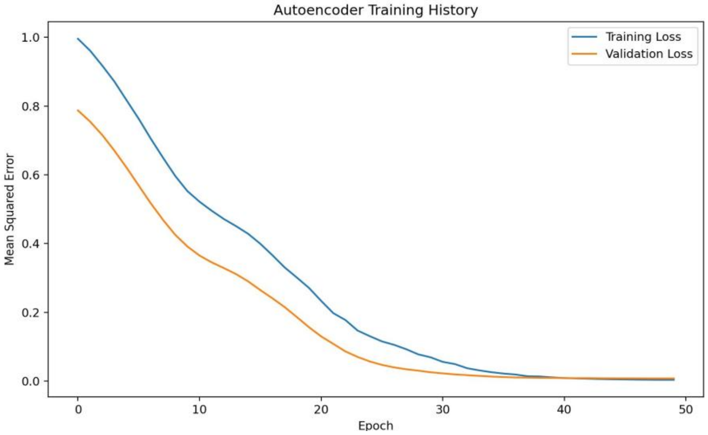


Figure 4. Epoch loss curve of autoencoder model.

3.4 Synthetic data generation: synthetic minority oversampling technique

The random forest classification report revealed consistently strong results across all classes. The model’s precision scores ranged from 0.93 to 1.00, indicating 93% to 100% accuracy in specific classes. Its recall scores were impressive, identifying 92% to 100% instances of each class. The F1 scores were also high. The model exhibited high accuracy across all classes, with most predictions falling on the diagonal and rare misclassifications represented by off-diagonal elements. This confirmed its ability to distinguish different classes. Its ROC curves and AUC values demonstrated the model's outstanding performance. The model maintained high performance across different sensitivity levels, making it robust and reliable for classification tasks. The strong performance suggests that the synthetic data generation process improved the model’s ability to make accurate predictions on unseen data. The proposed

Received: July 31, 2025
Accepted: September 20, 2025

model was evaluated on multiple datasets to assess its ability to reconstruct data and classify instances accurately. A comparison of the original data sets and their reconstructed versions, as shown in Figure 5, demonstrated the model's ability to effectively capture and reproduce the underlying data structure. The top row of Figure 5 displays the original data with varying y-axis values, and the corresponding reconstructed versions are displayed in the bottom row. The close resemblance between the original and reconstructed plots indicates that the model's reconstruction method successfully retained key features of the input data. This finding suggests that the model could preserve the essential characteristics of the data sets, which is crucial for further downstream analysis.

Received: July 31, 2025
Accepted: September 20, 2025

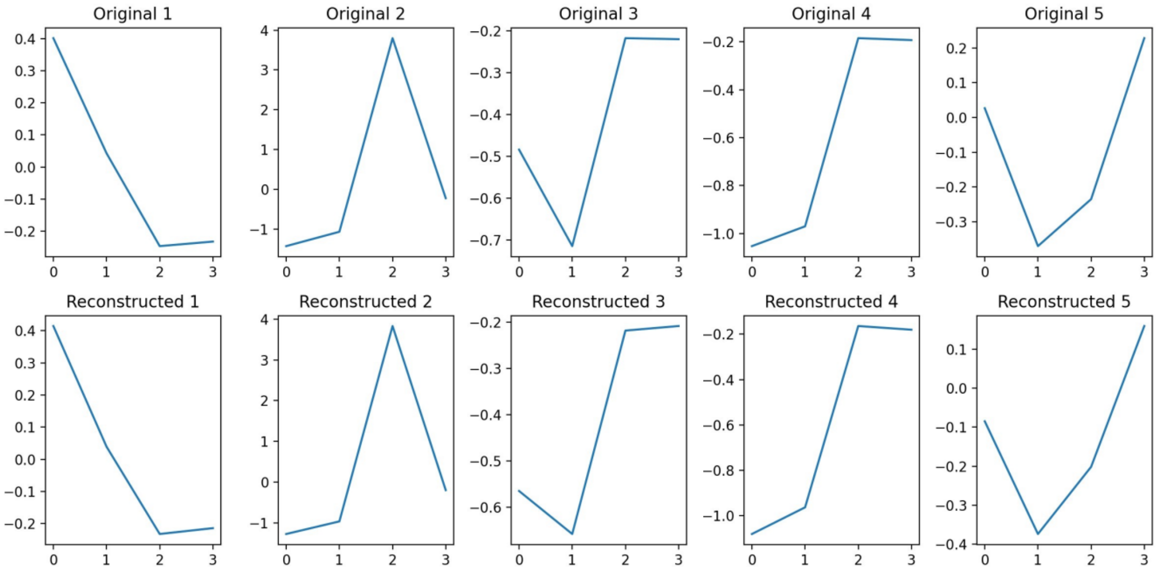


Figure 5. Line plots comparing original data sets with reconstructed versions.

A confusion matrix was generated to evaluate the classification performance, as presented in Figure 6.

Received: July 31, 2025
Accepted: September 20, 2025

ISSN Electrónico: 3105-403X

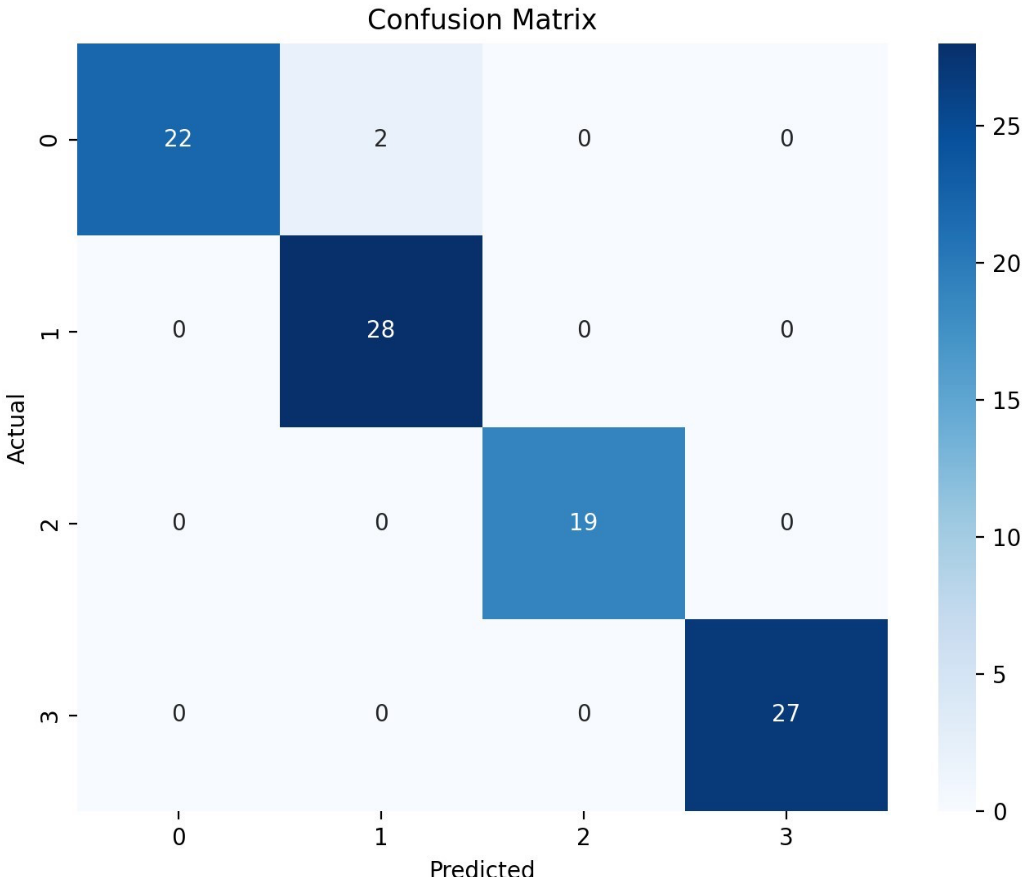


Figure 6. Confusion matrix.

The confusion matrix summarizes the model's ability to correctly classify instances across different classes, with actual and predicted values represented on the axes. Most instances fall along the diagonal, indicating that the model achieved high

Received: July 31, 2025
Accepted: September 20, 2025

accuracy in classifying the correct class labels. As indicated by off-diagonal cells, only a few misclassifications were observed, but these were minimal compared to the number of correctly classified instances. The color scale further emphasizes the model's strong performance, with darker shades indicating higher classification accuracy.

The results suggest that the model performed well in data reconstruction and classification tasks. The minimal discrepancies between the original and reconstructed data demonstrate the robustness of the reconstruction process, and the confusion matrix highlights the model's classification accuracy. These findings confirm the model's applicability to tasks requiring precise data reconstruction and accurate class prediction.

The random forest classifier, trained on original and synthetic data, demonstrated exceptional performance in classifying the target variable. The model achieved an impressive overall accuracy of 97.96%, indicating its proficiency in correctly predicting the classes across the dataset. This high accuracy suggests that the model successfully captured the underlying patterns and relationships within the data, including those introduced by the synthetic data generation process shown in Figure 7.

Received: July 31, 2025
Accepted: September 20, 2025

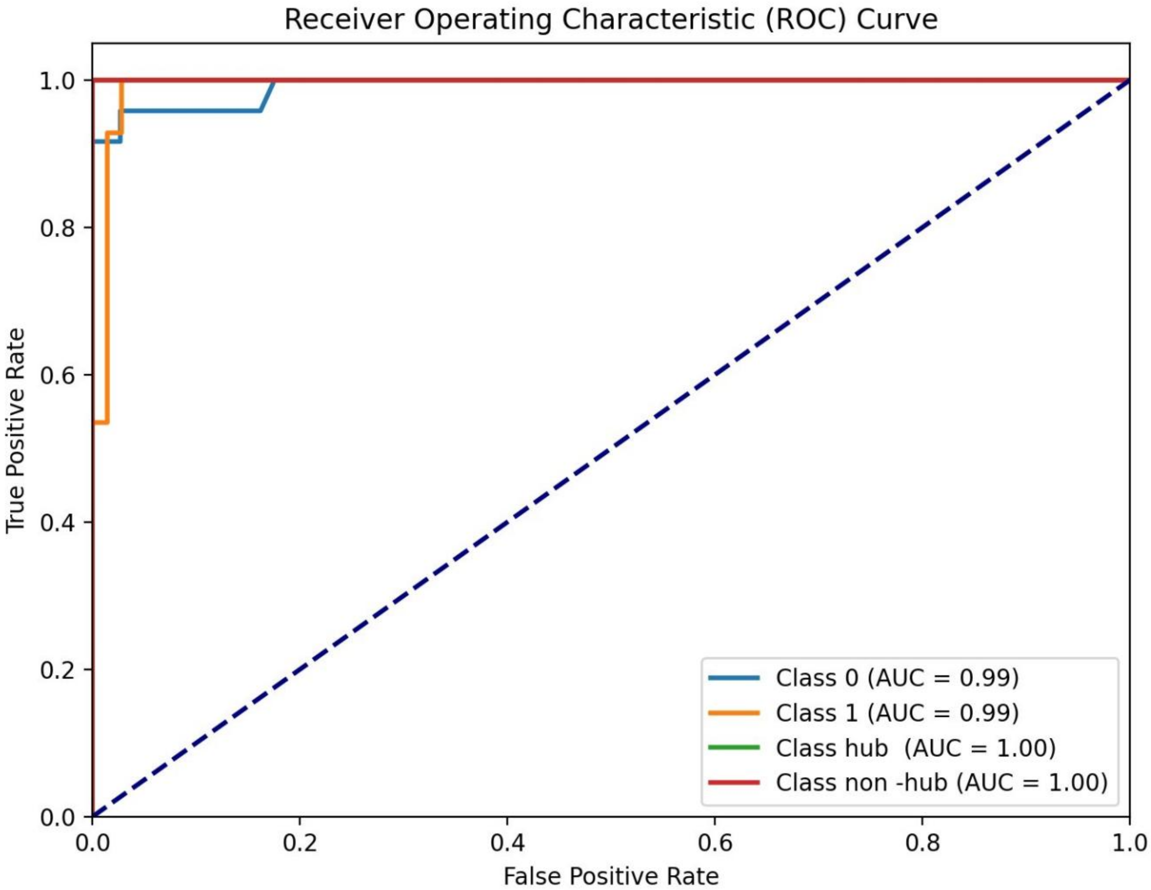


Figure 7. Receiver operating characteristic curve.

The receiver operating characteristic (ROC) curve is a graphical representation of a binary classification model's performance (Fig. 7). It includes an x-axis false-positive rate (FPR) and a y-axis true positive rate (TPR), plotted against the FPR at different thresholds. The graph includes curves for multiple classes, with an AUC of 1.00 for perfect classification and 1.00 for non-hub classification. The legend provides context

Received: July 31, 2025
Accepted: September 20, 2025

for the curves. The model performed well, particularly for hub and non-hub classes. The random forest classifier, trained on original and synthetic data, demonstrated exceptional performance in classifying the target variable. The model achieved an impressive overall accuracy of 97.96%, indicating its proficiency in correctly predicting the classes across the data set. This high accuracy suggests that the model successfully captured the underlying patterns and relationships within the data, including those introduced by the synthetic data generation process.

4. Discussion

Dental periapical pain, caused by inflammation or pulp infection, can be identified as a potential drug target. Studying the underlying mechanisms can provide insights into pain and inflammation, leading to the discovery of novel therapeutic targets to reduce both [1–3]. A deeper understanding of these mechanisms may also contribute to developing more effective therapies for managing this common dental condition. Periapical pain can be alleviated by targeting neurogenic inflammation, which involves the release of neuroactive substances by sensory nerve fibers. Identifying key immune mediators and signaling pathways can aid in developing drugs to modulate the immune response, reduce inflammation, and alleviate pain [10,11,26].

The activation of C-fibers (e.g., deep cavity preparation, high-intensity thermal or chemical stimuli) increases pulpal blood

Received: July 31, 2025
Accepted: September 20, 2025

flow due to the action of neurokinins, especially substance P, leading to neurogenic inflammation. Substance P receptors (NK1, NK2, NK3) mediate these effects, with NK1 being the primary receptor involved in physiological conditions and NK1/NK2 in pathological conditions. Substance P stimulates the release of inflammatory mediators, perpetuating the cycle and increasing pain sensitivity. This continuous cycle emphasizes the importance of developing therapeutic strategies to block substance P activity, which may provide significant relief to patients experiencing periapical pain. PGF2-α can affect the tissue response to PGE2, causing AMPc accumulation and increasing cGMP levels [2–4,6].

Top identified hub genes are essential for mitotic spindle organization, cytokinesis, DNA topological changes, and genomic integrity (Figures 1 and 2). They regulate apoptosis, cell cycle, and immune responses. Our study's gene ontology (GO) analysis further highlights the significance of these genes, showing key cellular locations and the importance of DNA binding and transcriptional regulation in modulating periapical pain. The GO cellular component data reveal cellular locations, DNA binding, and transcriptional regulation. The GO molecular function data explore molecular activities, and the KEGG 2021 human pathways data provide insights into autoimmune disorders, cancer, and cell signaling [27,28]. The genes *H2BC4*, *H2AC11*, *BIRC5*, *H2BC11*, *PRKY*, *H1-5*, *NUF2*,

Received: July 31, 2025
Accepted: September 20, 2025

KDM6A, and *H3C12* are involved in biological processes that may influence dental pain and nociception. *H2BC4* and *H2AC11* are part of the chromatin structure and may influence gene expression related to pain pathways. *PRKY*, *H1-5*, *NUF2*, *KDM6A*, and *H3C12* influence pain perception through stress responses, neuronal signaling, and cell division. Genes such as *BIRC5* and *KDM6A* could be used as diagnostic biomarkers for dental pain assessment, early detection of pathologies, treatment monitoring, and identifying patients at risk of chronic pain. This knowledge could lead to the development of novel pain drugs and treatments, but challenges include testing costs and implementation barriers.

Omics data can be analyzed using bioinformatics and systems biology to identify dysregulated molecular pathways in periapical pain. This integrative approach provides a more holistic understanding of the molecular mechanisms at play in this condition. Autoencoder-based integrative multi-omics data embedding (AIME) [7,15,29] is a deep-learning method that extracts low-dimensional nonlinear representations from omics data for integrative analysis, incorporating clinical confounding factors. Additionally, pharmacogenomics can provide insights into individual variations in drug response and susceptibility, enabling omics-based studies to identify therapeutic drugs for periapical pain, accelerating drug development by analyzing patient omics

Received: July 31, 2025
Accepted: September 20, 2025

profiles, identifying dysregulated pathways, and optimizing dosages. Data-driven methods generate synthetic data trained on current or observed data using generalized linear regression, nonlinear methods, and deep-learning techniques such as variational autoencoders and generative adversarial networks. Although these models enhance the identification of potential therapeutic pathways, limitations remain in data variability and the model's ability to generalize across different data sets. A previous study demonstrated the effectiveness of omicsGAN [18,29], a GAN model that integrates interaction networks and omics data sets, in improving cancer outcome classification and patient survival prediction. Synthetic data sets [15,17,30,31] also contain more significant features, resulting in better predictive performance. The effect of the interaction network on the quality of synthetic data was analyzed. Depending on input data quality and user acceptance, these can improve data utility but have limitations in reproducing outliers.

The autoencoder model achieved an overall accuracy of 68%, correctly classifying 68% of the samples in the test set. It had a moderate accuracy rate, with a weighted average precision of 0.74, recall of 0.68, and F1 score of 0.69 (Figure 3). This is similar to a previous study that applied a proposed framework to five cancer data sets from The Cancer Genome Atlas, showing that compression of input features improved survival patterns and silhouette scores [12,18,32]. Our model performed well for each class, with higher recall but lower

Received: July 31, 2025
Accepted: September 20, 2025

ACTA BIOCLÍNICA
ISSN Electrónico: 3105-403X

precision for the hub class. The model's precision scores ranged from 0.93 to 1.00, indicating 93% to 100% accuracy in specific classes (Figures 4–6). Its recall scores were impressive, identifying 92% to 100% instances of each class. Although these results are encouraging, the moderate accuracy underscores the need for further refinement and validation of the model to enhance its clinical applicability.

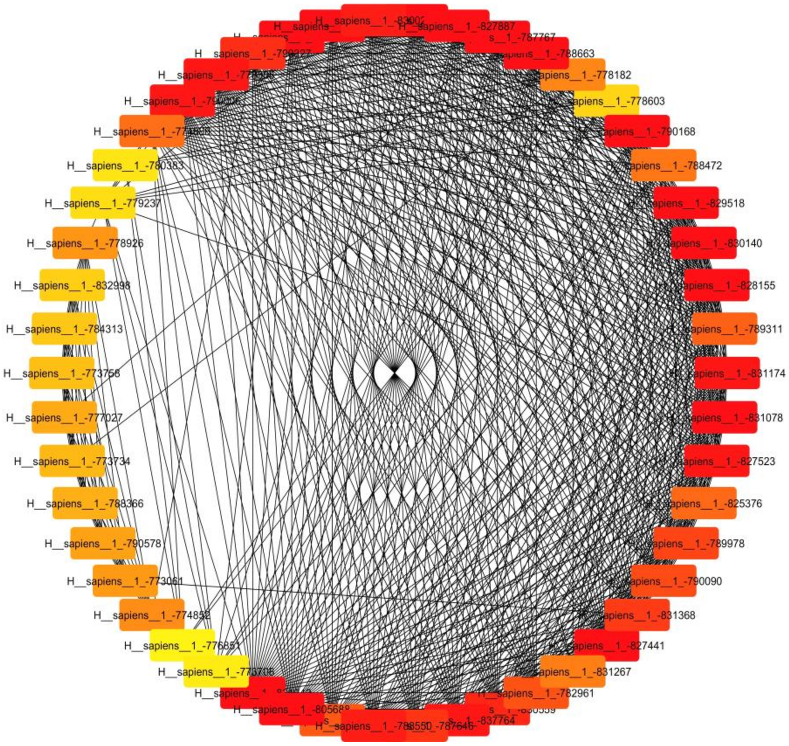


Figure 3. Top 60 hub genes identified.

Received: July 31, 2025
Accepted: September 20, 2025

Limitations of our study include the model’s moderate accuracy and predictive performance, which may be due to the variability in the quality of input data. Furthermore, although synthetic datasets improved some metrics, they may have introduced artifacts that affected the model’s ability to generalize to unseen data. Biological validation is required to confirm the relevance of the identified hub genes and the model’s findings in the context of periapical pain [17,19,31].

Future directions include feature selection techniques that identify the most relevant features, and oversampling or undersampling can address imbalanced data in predicting hub genes [26]. Class imbalance is crucial for improving model performance, especially when one class has more samples than another in predicting non-hub genes. Strategies to address this include resampling techniques such as oversampling the minority class, undersampling the majority class, or combining sampling. Modifying the loss function during training can increase the weights for the minority class, which can be specified in machine-learning libraries such as TensorFlow and PyTorch. Ensemble methods such as bagging or boosting can also adjust for class imbalance for non-hub genes. Data augmentation, anomaly detection, post-processing, metrics evaluation, cross-validation, and experimentation with alternative models can improve minority class training data diversity and model performance.

Received: July 31, 2025
Accepted: September 20, 2025

Ensemble methods, such as combining multiple models or boosting algorithms, can also enhance the model’s predictive accuracy. Hyperparameters can be fine-tuned using techniques such as grid search or Bayesian optimization. Domain expertise can be incorporated to refine the model. Generated hub genes can serve as diagnostic biomarkers for assessing pain severity, early detection of periapical pathologies, and assisting personalized treatment planning. They offer novel targets for drug development and gene-based pain therapies, aiding in risk assessment, postoperative care planning, and monitoring treatment responses. However, most applications are still in the research phase and require clinical validation before routine implementation in dental practice.

Further experimental validation is necessary to ensure that the model’s predictions are biologically meaningful and can be applied in clinical settings. The model’s generalizability and interpretation also need further validation through biological experiments [17,20,31]. The random forest classifier, trained on the combined original and synthetic data, has demonstrated exceptional classification capabilities. Its high accuracy, balanced performance across classes, and strong showing in various evaluation metrics all indicate a robust and reliable model.

Autoencoder models can be enhanced by adjusting their capacity, using advanced architectures such as convolutional

Received: July 31, 2025
Accepted: September 20, 2025

autoencoders, recurrent neural networks, and variational autoencoders, and experimenting with activation functions for better generation of hub genes. These modifications can enhance learning capacity, improve reconstruction quality, and capture more complex data patterns and distributions. Increasing the number of layers in both the encoder and decoder can capture more complex features for gene expression data. The successful integration of synthetic data from painomics training data has likely contributed to this strong performance, providing a larger and more balanced data set for training. These results suggest that this model could be confidently applied to similar classification tasks in the future, with a high expectation of accurate and reliable predictions for predicting and understanding hub genes' role in periapical pain.

5. Conclusions

A generative AI model was developed to predict hub genes in periapical pain, showing potential for understanding molecular mechanisms. The autoencoder model achieved moderate accuracy, but further improvements are needed. The random forest classifier, trained on both original and synthetic data, demonstrated high accuracy and reliability, with the successful integration of synthetic data enhancing its performance.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Received: July 31, 2025
Accepted: September 20, 2025

Author Contributions

PY, PW and CA designed the research study. PY, PW and CA performed the research. PY, PW and CA analyzed the data. PY, PW and CA wrote the manuscript. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript.

Ethics Approval and Consent to Participate: Not required

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

Funding: This research received no external funding

Conflict of Interest: The authors declare no conflict of interest.

REFERENCES

[1] de-Figueiredo FED, Lima LF, Lima GS, Oliveira LS, Ribeiro MA, Brito-Junior M, et al. Apical periodontitis healing and postoperative pain following endodontic treatment with a reciprocating single-file, single-cone approach: A randomized controlled pragmatic clinical trial. PLoS One. 2020;15(2):e0227347. doi: 10.1371/journal.pone.0227347

[2] Khan AA, Diogenes A. Pharmacological Management of Acute Endodontic Pain. Drugs. 2021;81(14):1627–43. doi: 10.1007/s40265-021-01564-4

[3] Khandelwal A, Jose J, Teja KV, Palanivelu A. Comparative evaluation of postoperative pain and periapical healing after root canal

Received: July 31, 2025
Accepted: September 20, 2025

treatment using three different base endodontic sealers – A randomized control clinical trial. J Clin Exp Dent. 2022;14(2):e144-e152. doi:10.4317/jced.59034.

[4] Cope AL, Francis N, Wood F, Chestnutt IG. Systemic antibiotics for symptomatic apical periodontitis and acute apical abscess in adults. Cochrane Database Syst Rev. 2018;9(9):CD010136. doi: 10.1002/14651858

[5] Schwendicke F, Göstemeyer G. Single-visit or multiple-visit root canal treatment: systematic review, meta-analysis and trial sequential analysis. BMJ Open. 2017;7(2):e013115. doi: 10.1136/bmjopen-2016-013115.

[6] García B, Larrazabal C, Peñarrocha M, Peñarrocha M. Pain and swelling in periapical surgery. A literature update. Med Oral Patol Oral Cir Bucal. 2008;13(11):E726-9.

[7] Selvarajoo K, Maurer-Stroh S. Towards multi-omics synthetic data integration. Brief Bioinform. 2024;25(3):bbae213. doi: 10.1093/bib/bbae213

[8] Zhang D, Zheng C, Zhu T, Yang F, Zhou Y. Identification of key module and hub genes in pulpitis using weighted gene co-expression network analysis. BMC Oral Health. 2023;23(1):2. doi: 10.1186/s12903-022-02638-9.

[9] Estrela C, Guedes OA, Silva JA, Leles CR, Estrela CR de A, Pécora JD. Diagnostic and clinical factors associated with pulpal and

Received: July 31, 2025
Accepted: September 20, 2025

periapical pain. Braz Dent J. 2011;22(4):306–11. doi: 10.1590/s0103-64402011000400008

[10] Sivakumar D, Ramli R. GABAergic signalling in modulation of dental pain. Eur J Pharmacol. 2022;924:174958. doi: 10.1016/j.ejphar.2022.174958

[11] Sacerdote P, Levrini L. Peripheral mechanisms of dental pain: the role of substance P. Mediators Inflamm. 2012;2012:951920. doi: 10.1155/2012/951920.

[12] Xie R, Wen J, Quitadamo A, Cheng J, Shi X. A deep auto-encoder model for gene expression prediction. BMC Genomics. 2017;18(9):845. doi: 10.1186/s12864-017-4226-0

[13] Jain N, Gupta A, N M. An insight into neurophysiology of pulpal pain: facts and hypotheses. Korean J Pain. 2013;26(4):347–55. doi: 10.3344/kjp.2013.26.4.347

[14] Närhi MVO. Nociceptors in the Dental Pulp. In: Schmidt RF, Willis WD, editors. Encyclopedia of Pain. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. p. 1418–20.

[15] Xin B, Lin Y, Tian H, Song J, Zhang L, Lv J. Identification of Pulpitis-Related Potential Biomarkers Using Bioinformatics Approach. Comput Math Methods Med. 2021;2021:1808361. doi: 10.1155/2021/1808361

[16] Tian T, Min MR, Wei Z. Model-based autoencoders for imputing discrete single-cell

Received: July 31, 2025
Accepted: September 20, 2025

RNA-seq data. Methods. 2021;192:112–9. doi: 10.1016/j.ymeth.2020.09.010.

[17] Sheng N, Huang L, Wang Y, Zhao J, Xuan P, Gao L, et al. Multi-channel graph attention autoencoders for disease-related lncRNAs prediction. Brief Bioinform. 2022;23(2): :bbab604. doi: 10.1093/bib/bbab604

[18] Madhumita, Paul S. Capturing the latent space of an Autoencoder for multi-omics integration and cancer subtyping. Comput Biol Med. 2022;148:105832. doi: 10.1016/j.compbiomed.2022

[19] Liu D, Huang Y, Nie W, Zhang J, Deng L. SMALF: miRNA-disease associations prediction based on stacked autoencoder and XGBoost. BMC Bioinformatics. 2021;22(1):219. doi: 10.1186/s12859-021-04135-2.

[20] Shi H, Wu C, Bai T, Chen J, Li Y, Wu H. Identify essential genes based on clustering based synthetic minority oversampling technique. Comput Biol Med. 2023;153:106523. doi: 10.1016/j.compbiomed.2022.106523

[21] Fu X, Chen Y, Tian S. DlncRNALoc: A discrete wavelet transform-based model for predicting lncRNA subcellular localization. Math Biosci Eng. 2023;20(12):20648–67. doi: 10.3934/mbe.2023913

[22] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--

Received: July 31, 2025
Accepted: September 20, 2025

update. Nucleic Acids Res. 2013;41(Database issue):D991-5. doi: 10.1093/nar/gks1193

[23] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10. doi: 10.1093/nar/30.1.207

[24] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504. doi: 10.1101/gr.1239303

[25] Hashimoto-Roth E, Surendra A, Lavallée-Adam M, Bennett SAL, Cuperlovic-Culf M. METAbolomics data Balancing with Over-sampling Algorithms (META-BOA): an online resource for addressing class imbalance. Bioinformatics. 2022;38(23):5326–7. doi: 10.1093/bioinformatics/btac649

[26] García B, Penarrocha M, Martí E, Gay-Escodad C, von Arx T. Pain and swelling after periapical surgery related to oral hygiene and smoking. Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 2007;104(2):271–6.

[27] Yadalam PK, Shenoy SB, Anegundi RV, Mosaddad SA, Heboyan A. Advanced machine learning for estimating vascular occlusion percentage in patients with ischemic heart disease and periodontitis. Int J Cardiol Cardiovasc Risk Prev. 2024;21:200291. doi: 10.1016/j.ijcrp.2024.200291

Received: July 31, 2025
Accepted: September 20, 2025

[28] Ardila CM, Yadalam PK, Minervini G. The potential of machine learning applications in addressing antimicrobial resistance in periodontitis. J Periodontal Res. 2024;59(5):1042-1043. doi: 10.1111/jre.13282.

[29] Ahmed KT, Sun J, Cheng S, Yong J, Zhang W. Multi-omics data integration by generative adversarial network. Bioinformatics. 2021;38(1):179–86. doi: 10.1093/bioinformatics/btab608

[30] Gold MP, LeNail A, Fraenkel E. Shallow Sparsely-Connected Autoencoders for Gene Set Projection. Pac Symp Biocomput. 2019;24:374–85.

[31] Hahn W, Schütte K, Schultz K, Wolkenhauer O, Sedlmayr M, Schuler U, et al. Contribution of Synthetic Data Generation towards an Improved Patient Stratification in Palliative Care. J Pers Med. 2022;12(8):1278. doi: 10.3390/jpm12081278.

[32] Staheli JP, Neal ML, Navare A, Mast FD, Aitchison JD. Predicting host-based, synthetic lethal antiviral targets from omics data. bioRxiv [Preprint]. 2023 Aug 16:2023.08.15.553430. doi: 10.1101/2023.08.15.553430. Update in: NAR Mol Med. 2024 Jan 23;1(1):ugad001. doi: 10.1093/narmme/ugad001.

Received: July 31, 2025
Accepted: September 20, 2025