IRSTI 28.23.15 https://doi.org/10.26577/jpcsit4120268

I. Tokhtakhunov[1, 2], M. Nurtas[1, 3*]

[1]International Information Technology University, Almaty, Kazakhstan
[2]School of Digital Technologies, Narxoz University, Almaty, Kazakhstan
[3]Al-Farabi Kazakh National University, Almaty, Kazakhstan
*e-mail: maratnurtas@gmail.com

NONLINEAR DIMENSIONALITY REDUCTION FOR LOOKALIKE AUDIENCE DETECTION USING MANIFOLD LEARNING AND AUTOENCODER-BASED REPRESENTATIONS

Abstract. Identifying users with similar behavioral characteristics is a critical task in modern targeted advertising and customer analytics systems. High-dimensional tabular datasets describing user activity often contain complex nonlinear relationships that cannot be effectively captured by traditional linear dimensionality reduction techniques. This study investigates representation learning approaches for constructing scalable look-alike audience detection systems using large-scale telecommunications data. Classical dimensionality reduction techniques, including Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE), are first analyzed as baseline methods for exploring the structure of high-dimensional data. While PCA performs linear projections that preserve global variance and t-SNE reveals local neighborhood structures through nonlinear embedding, these methods are primarily designed for visualization and exploratory analysis and do not provide scalable parametric mappings for new data samples. To address these limitations, a representation learning framework based on autoencoders is proposed for generating compact latent embeddings of users. The model is trained on a large-scale anonymized telecommunications dataset containing behavioral, demographic, device-related, and service usage attributes. Embeddings are learned for multiple feature entities and concatenated into a unified user representation that integrates heterogeneous behavioral information. User similarity is then computed using cosine similarity in the latent space, enabling efficient identification of look-alike audiences. The proposed system is evaluated using clustering metrics and multiple independent validation tasks with external target variables to ensure unbiased performance estimation. Experimental results demonstrate that autoencoder-based embeddings produce a more structured latent space and improve both similarity-based retrieval and downstream classification performance compared to classical dimensionality reduction techniques. The findings highlight the effectiveness of deep representation learning for high-dimensional tabular data in real-world recommendation and targeted advertising systems.

Keywords: dimensionality reduction, manifold learning, t-distributed stochastic neighbor embedding (t-SNE), autoencoder, representation learning, lookalike audience modeling, tabular data.

1. Introduction

The rapid growth of digital platforms and online services has led to the generation of massive volumes of high-dimensional user data. Such datasets often contain heterogeneous attributes describing user behavior, demographic characteristics, device information, and interaction histories. Analyzing these data in order to identify patterns and similarities between users has become a central task in modern data-driven systems, including recommendation engines, customer analytics platforms, and targeted advertising technologies [1].

One of the fundamental challenges when working with high-dimensional datasets is the so-called curse of dimensionality, where the increasing number of features makes it difficult to capture meaningful relationships between observations. As dimensionality grows, data points become sparse in the feature space, which negatively affects the ability of machine learning algorithms to identify informative structures. Traditional machine learning methods often struggle to operate effectively in such environments due to increased computational complexity and the presence of redundant or correlated variables. As a result, dimensionality reduction techniques have become an essential tool for transforming high-dimensional data into compact and informative representations.

Classical dimensionality reduction methods, such as Principal Component Analysis (PCA), assume linear relationships between variables and

project data onto directions that maximize variance [2]. Although these techniques are computationally efficient and widely used in practice, they are often unable to capture nonlinear structures that frequently arise in real-world datasets. To address this limitation, a class of algorithms known as manifold learning methods has been developed. These approaches assume that high-dimensional data points lie on or near a lower-dimensional manifold embedded within the original feature space.

Among the widely used techniques in this category is t-distributed Stochastic Neighbor Embedding (t-SNE), a nonlinear dimensionality reduction method designed to preserve local neighborhood relationships between data points when projecting them into a lower-dimensional space [3]. The algorithm converts pairwise distances into probability distributions and minimizes the divergence between similarity distributions in the original and embedded spaces. Due to its ability to reveal cluster structures in complex datasets, t-SNE has become a popular tool for visualization and exploratory data analysis. However, despite its effectiveness for visualization tasks, t-SNE does not provide a parametric mapping function and can be computationally expensive when applied to large datasets.

In recent years, deep learning–based representation learning approaches have emerged as a powerful alternative for nonlinear dimensionality reduction. In particular, autoencoders provide a neural network architecture capable of learning compressed latent representations of high-dimensional data [4]. By training the model to reconstruct the input data through a bottleneck layer, autoencoders learn compact embeddings that capture the most informative structures and nonlinear relationships present in the dataset [5]. These embeddings can subsequently be used as feature representations for a variety of downstream tasks, including classification, clustering, and similarity-based retrieval [6].

In the context of targeted advertising and customer analytics, dimensionality reduction plays a crucial role in constructing compact representations of users that enable similarity analysis between individuals. One important application is look-alike audience detection, where the goal is to identify users who exhibit behavioral characteristics similar to those of a reference group. Such systems are widely used in marketing platforms to expand target audiences for advertising campaigns.

A key requirement for practical look-alike systems is the ability to operate as generalized services capable of handling diverse targeting tasks. In production environments, different Business-to-Business (B2B) clients may submit audience expansion requests based on different behavioral signals or campaign objectives [1]. Therefore, the representation learning framework must remain independent of any specific target variable and instead capture general behavioral patterns of users that can support a wide range of downstream prediction tasks.

The objective of this study is to investigate dimensionality reduction and representation learning techniques for high-dimensional tabular data and to analyze their applicability in scalable look-alike audience detection systems. In particular, the study examines classical dimensionality reduction approaches, including PCA and t-SNE, and compares them with deep learning–based representation learning methods based on autoencoders. The proposed framework learns latent embeddings from large-scale anonymized telecommunications data and uses them to compute similarity between users. Experimental results demonstrate that neural network–based embeddings provide more structured latent representations and improve similarity-based user analysis compared to classical dimensionality reduction techniques.

2. Materials and Methods

This section describes the dataset, preprocessing procedures, and dimensionality reduction techniques used in the study. Particular attention is given to nonlinear representation learning methods and their application to high-dimensional tabular data. Classical dimensionality reduction approaches, including Principal Component Analysis (PCA) and nonlinear manifold learning techniques such as t-distributed Stochastic Neighbor Embedding (t-SNE), are considered alongside neural network–based representation learning methods based on autoencoders.

The objective of the proposed methodological framework is to transform high-dimensional user data into compact latent representations that preserve the most informative structural properties of the original dataset. These representations enable efficient similarity computation between users and support scalable look-alike audience detection. By comparing linear, nonlinear manifold learning, and

deep learning–based dimensionality reduction techniques, the study evaluates their ability to capture meaningful patterns in large-scale user datasets.

2.1. Dataset Description

The experiments were conducted using a large-scale anonymized dataset collected from a telecommunications and digital services platform during regular service operation. The dataset contains aggregated user-related attributes derived from multiple sources, including subscriber profiles, device characteristics, tariff plans, and behavioral activity indicators [7].

To ensure compliance with privacy and data protection regulations, all records used in the study were fully anonymized and did not contain any personally identifiable information. The dataset represents aggregated statistical indicators rather than raw user-level events, which further ensures the protection of sensitive information.

Initially, the raw dataset contained 2814 features describing different aspects of user behavior and system interaction. These attributes were derived from multiple heterogeneous data sources and included demographic indicators, device-related characteristics, service usage statistics, network activity measures, and other aggregated behavioral signals.

After applying data preprocessing procedures, the final dataset used in the experiments consisted of 948 features. The details of the preprocessing pipeline and feature transformation steps are described in the following subsection.

For experimental evaluation, the dataset was divided into training and validation subsets. The training data were used to learn the dimensionality reduction models and latent representations, while the validation subset was used to assess the quality of the resulting embeddings and their suitability for similarity-based user analysis [7].

2.2. Data Preprocessing

Before applying dimensionality reduction methods, several preprocessing steps were performed to prepare the dataset for analysis and ensure the consistency of the input features.

First, categorical variables were transformed into numerical representations using one-hot encoding. This transformation enables machine learning algorithms to process categorical attributes by converting each category into a separate binary feature.

Second, missing values were handled using statistical imputation techniques. Depending on the distribution and semantic interpretation of the variables, missing entries were replaced using mean, median, or mode estimates. This approach allowed the preservation of the overall dataset structure while minimizing information loss.

Third, redundant features were identified through pairwise correlation analysis. Highly correlated variables were removed in order to reduce multicollinearity and eliminate redundant information in the dataset. This step significantly reduced the dimensionality of the feature space while preserving the most informative characteristics of the data.

As a result of the preprocessing pipeline, the number of features was reduced from the initial 2814 attributes to 948 features used in the subsequent experiments.

Finally, numerical variables were normalized using min–max scaling, which transforms feature values into the range [0, 1]. This normalization ensures that all variables contribute proportionally during model training and prevents features with larger numerical ranges from disproportionately influencing the learning process [8].

2.3. Dimensionality Reduction Techniques

Dimensionality reduction plays a critical role in the analysis of high-dimensional datasets. Its primary objective is to transform the original feature space into a lower-dimensional representation while preserving the most informative structural properties of the data. By reducing the number of variables while retaining essential information, dimensionality reduction improves computational efficiency and facilitates the discovery of latent patterns within complex datasets.

In this study, both linear and nonlinear dimensionality reduction methods are considered. Linear approaches such as Principal Component Analysis (PCA) provide a simple and computationally efficient way to reduce dimensionality by projecting the data onto directions that maximize variance. However, linear techniques assume linear relationships between variables and may fail to capture complex nonlinear structures frequently present in real-world datasets.

To address this limitation, nonlinear dimensionality reduction techniques based on manifold learning are also examined. These methods assume that high-dimensional observations

lie on a lower-dimensional manifold embedded within the original feature space and attempt to preserve local neighborhood relationships between data points. One of the most widely used approaches in this category is t-distributed Stochastic Neighbor Embedding (t-SNE), which models pairwise similarities between samples using probability distributions [9].

In addition to classical dimensionality reduction techniques, this study also investigates deep learning–based representation learning using autoencoders. Unlike traditional manifold learning algorithms, autoencoders learn a parametric nonlinear transformation that maps the original feature space into a compact latent representation through neural network architectures.

2.3.1 Nonlinear Dimensionality Reduction and Manifold Learning

Manifold learning methods are based on the assumption that high-dimensional data points lie on or near a lower-dimensional manifold embedded within the original feature space. Although observations may be represented by a large number of variables, their intrinsic dimensionality can often be significantly smaller. The objective of manifold learning algorithms is therefore to identify this hidden structure and represent the data in a lower-dimensional space while preserving meaningful relationships between observations [9].

Unlike linear dimensionality reduction techniques, manifold learning methods attempt to capture nonlinear relationships between variables by preserving local neighborhood structures or pairwise similarities between data points. These approaches are particularly useful for analyzing complex datasets where the underlying structure cannot be adequately described using linear projections.

In practice, manifold learning techniques are widely applied for exploratory data analysis and visualization of high-dimensional datasets. By mapping data into a lower-dimensional space, these methods allow researchers to observe clustering patterns, identify latent structures, and better understand relationships between observations.

One of the most widely used nonlinear dimensionality reduction algorithms is t-distributed Stochastic Neighbor Embedding (t-SNE), which models pairwise similarities between observations using probability distributions and attempts to preserve local neighborhood structures in the embedded space [10].

2.3.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear dimensionality reduction technique designed to preserve local neighborhood relationships between data points when projecting high-dimensional data into a lower-dimensional space. The method converts pairwise distances between observations into probability distributions that represent similarities between data points.

In the high-dimensional space, the similarity between two data points is defined using a Gaussian distribution:

$$\boldsymbol{p_{ij}} = \frac{\exp\left(-\frac{||\boldsymbol{x_i} - \boldsymbol{x_j}||^2}{\boldsymbol{2\sigma_i^2}}\right)}{\sum_{\boldsymbol{k \neq l}} \boldsymbol{\exp}\left(-\frac{||\boldsymbol{x_k} - \boldsymbol{x_l}||^2}{\boldsymbol{2\sigma_k^2}}\right)} \tag{1}$$

where $x_i$and $x_j$represent data points in the original feature space and $\sigma_i$is the variance of the Gaussian distribution controlling the neighborhood size around point x_i [10].

In the low-dimensional embedding space, similarities between points are modeled using a Student's t-distribution with one degree of freedom:

$$\boldsymbol{q_{ij}} = \frac{\left(\mathbf{1} + ||\boldsymbol{y_i} - \boldsymbol{y_j}||^2\right)^{-\mathbf{1}}}{\sum_{\boldsymbol{k \neq l}} (\mathbf{1} + ||\boldsymbol{y_k} - \boldsymbol{y_l}||^2)^{-\mathbf{1}}} \tag{2}$$

where $\boldsymbol{y_i}$and $\boldsymbol{y_j}$represent the coordinates of the corresponding points in the embedded space.

The t-SNE algorithm minimizes the Kullback–Leibler divergence between the similarity distributions in the high-dimensional and low-dimensional spaces:

$$\boldsymbol{KL(P||Q)} = \sum_{\boldsymbol{i}} \sum_{\boldsymbol{j}} \boldsymbol{p_{ij}} \boldsymbol{\log} \frac{\boldsymbol{p_{ij}}}{\boldsymbol{q_{ij}}} \tag{3}$$

By minimizing this divergence, t-SNE attempts to preserve local neighborhood relationships between data points, allowing clusters and local structures in the data to become more visible in the embedded representation [11].

Although t-SNE is highly effective for visualizing complex high-dimensional datasets, it

does not learn an explicit mapping function from the original feature space to the embedding space. As a result, the algorithm is primarily used for exploratory analysis and visualization rather than for scalable representation learning in production systems.

2.3.3 Autoencoder Representation Learning

Autoencoders represent a class of neural network architectures designed for unsupervised representation learning and nonlinear dimensionality reduction. Unlike classical manifold learning algorithms, autoencoders learn a parametric nonlinear mapping between the original high-dimensional feature space and a compact latent representation through a neural network model [12].

The architecture of an autoencoder consists of two main components: an encoder and a decoder. The encoder transforms the original input feature vector into a lower-dimensional latent representtation, while the decoder attempts to reconstruct the original input from this compressed representation. The objective of the model is to learn a latent embedding that captures the most informative structural properties of the data while minimizing the loss of information during compression [13].

Let $x \in \mathbb{R}^d$denote the original input feature vector. The encoder network maps the input vector into a lower-dimensional latent representation $z \in \mathbb{R}^k$, where $k < d$. This transformation can be expressed as:

$$\boldsymbol{z = f_\theta(x)} \tag{4}$$

where $f_\theta(\boldsymbol{x})$ represents the nonlinear transformation defined by the encoder network with parameters θ [14]. The decoder then reconstructs the original input from the latent representation:

$$\boldsymbol{\hat{x} = g_\phi(z)} \tag{5}$$

where $g_\phi(\boldsymbol{z})$denotes the decoding function parameterized by the network weights ϕ, and $\hat{x}$represents the reconstructed input vector.

During training, the model minimizes the reconstruction error between the original input vector and its reconstructed version. In this study, the Mean Squared Error (MSE) loss function was used:

$$\boldsymbol{L = \frac{1}{N}\sum_{i=1}^{N}\left||x_i - \hat{x}_i|\right|^2} \tag{6}$$

where Ndenotes the number of training samples [15].

By minimizing the reconstruction loss, the autoencoder learns to capture nonlinear relationships between variables and to encode the most informative features of the dataset into a compact latent representation. These latent vectors serve as embedding representations that can be used for similarity analysis, clustering, and downstream machine learning tasks [16].

The architecture of the autoencoder used in this study is illustrated in Figure 1.

The diagram presents the full processing pipeline, beginning with the raw dataset containing 2814 features. After feature preprocessing and selection, the input dimensionality is reduced to 948 features, which are used as the input to the neural network model. The encoder network then compresses these features into a lower-dimensional latent embedding, which represents a compact representation of user characteristics. The decoder network subsequently reconstructs the original feature representation from the latent space, allowing the model to learn informative nonlinear structures in the data [7].

The training process of the autoencoder was monitored using the reconstruction loss on both training and validation datasets. Figure 2 illustrates the learning dynamics of the model during the training procedure.

The training loss gradually decreases as the network learns compact latent representations of the input features. The validation loss follows a similar trend and stabilizes after approximately 400 epochs, indicating convergence of the model and the absence of significant overfitting.

Based on the observed training dynamics, the model demonstrates stable convergence behavior. After approximately 400 epochs the improvement in reconstruction loss becomes marginal, indicating that the latent representation has captured the dominant structural patterns present in the dataset.

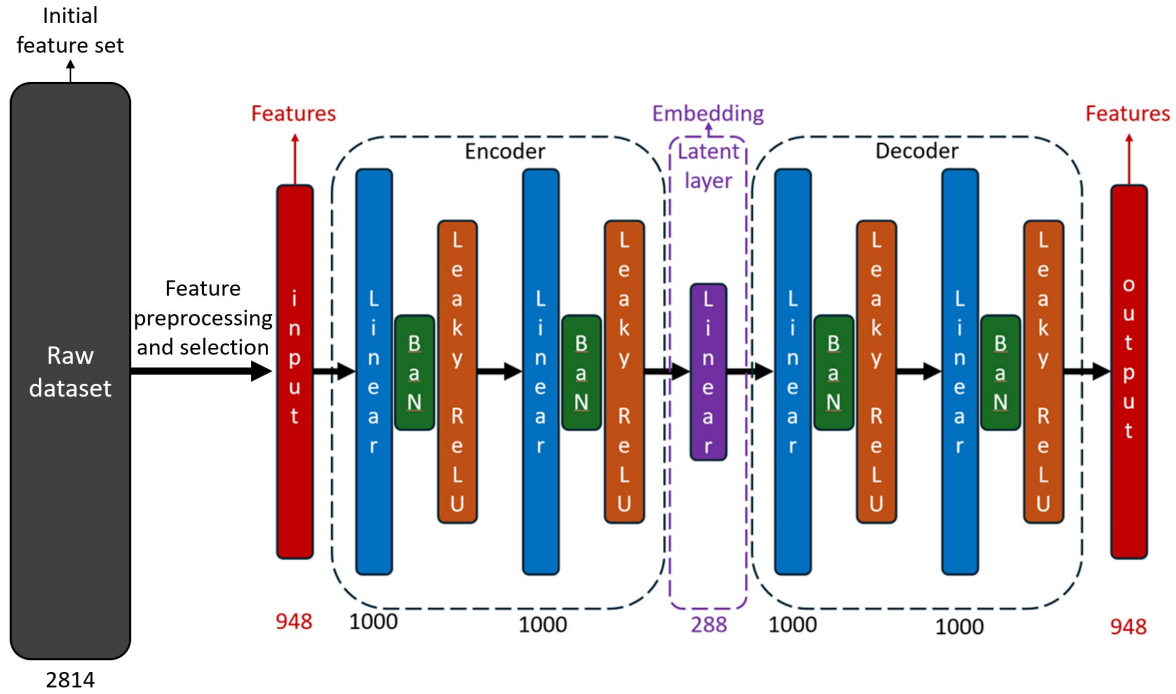

Figure 1. Architecture of the autoencoder-based representation learning model and feature preprocessing pipeline

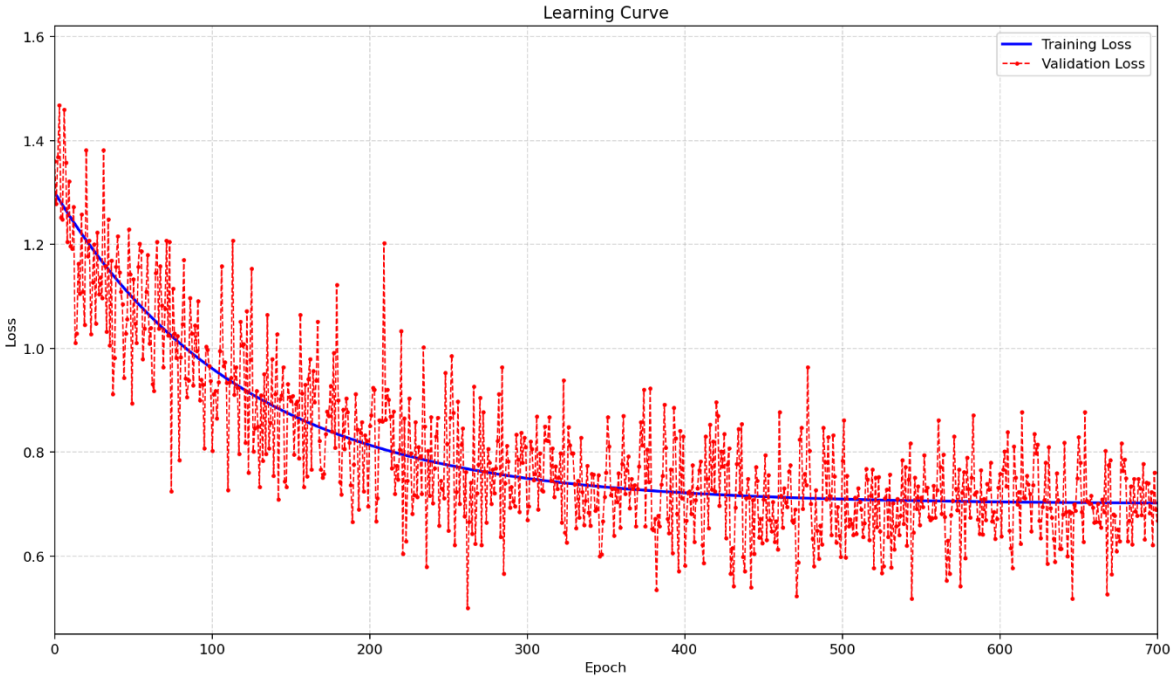

Figure 2. Learning curves of the autoencoder model during training

2.4 Multi-Entity Embedding Representation

The dataset used in this study contains heterogeneous groups of features describing different aspects of user behavior. These feature groups can be interpreted as entities representing distinct domains of information, including subscriber profiles, device characteristics, tariff plans, and network activity patterns [1].

To capture these heterogeneous characteristics more effectively, embeddings were learned separately for each entity. For each feature group, a dedicated autoencoder model was trained to generate a latent representation of the corresponding entity-specific feature space.

The resulting entity embeddings were then concatenated to form a unified representation of each subscriber:

$$z_{user} = [z_{sub}, z_{device}, z_{tariff}, z_{network}] \quad (7)$$

where each component corresponds to the latent representation learned from the respective entity feature group.

This concatenation strategy allows the model to integrate information from multiple behavioral domains while preserving the semantic structure learned within each entity. The resulting unified embedding provides a comprehensive representation of user characteristics that can be used for downstream similarity-based tasks [9].

2.5 Cosine Similarity

To identify users with similar behavioral characteristics, similarity between embedding vectors was computed using cosine similarity. Cosine similarity measures the angular distance between two vectors and is defined as:

$$sim(x, y) = \frac{x \cdot y}{||x|| \, ||y||} \quad (8)$$

where x and y represent embedding vectors corresponding to different users [17].

Unlike Euclidean distance, cosine similarity focuses on the orientation of vectors rather than their magnitude. This property makes it particularly suitable for comparing high-dimensional embeddings where the direction of the vector encodes semantic relationships between observations.

In the context of look-alike audience detection, users with higher cosine similarity values are considered behaviorally similar. By computing cosine similarity between the embedding vector of a reference user group and the embedding vectors of the entire subscriber base, it becomes possible to identify candidate users exhibiting similar behavioral patterns.

2.6 Modular System Design for Embedding-Based Look-Alike Model

To support scalable deployment in real-world data environments, the proposed representation learning framework was implemented using a modular system architecture. The system integrates data storage, preprocessing, model training, and evaluation components into a unified pipeline designed for large-scale subscriber datasets.

The overall architecture of the embedding-based look-alike modeling framework is illustrated in Figure 3.

The system combines multiple modules responsible for feature collection, preprocessing, representation learning, and model evaluation. This modular design allows different components of the pipeline to be developed and updated independently while maintaining the overall integrity of the system.

At the data storage level, the Hadoop Distributed File System (HDFS) serves as the primary repository for raw and processed data. Feature datasets are versioned using Data Version Control (DVC), which ensures reproducibility of experiments and enables consistent tracking of dataset modifications across different model training runs [7].

The feature preparation stage includes data collection, preprocessing, and transformation of raw subscriber data into structured feature datasets. These datasets are divided into training and validation subsets and stored as structured files that serve as inputs for the representation learning models.

Model training and experimentation are managed using MLflow, which provides experiment tracking, parameter logging, and model version management. In addition, the MinIO object storage system is used to store experiment artifacts and trained model checkpoints, ensuring reliable storage and accessibility of experimental results.

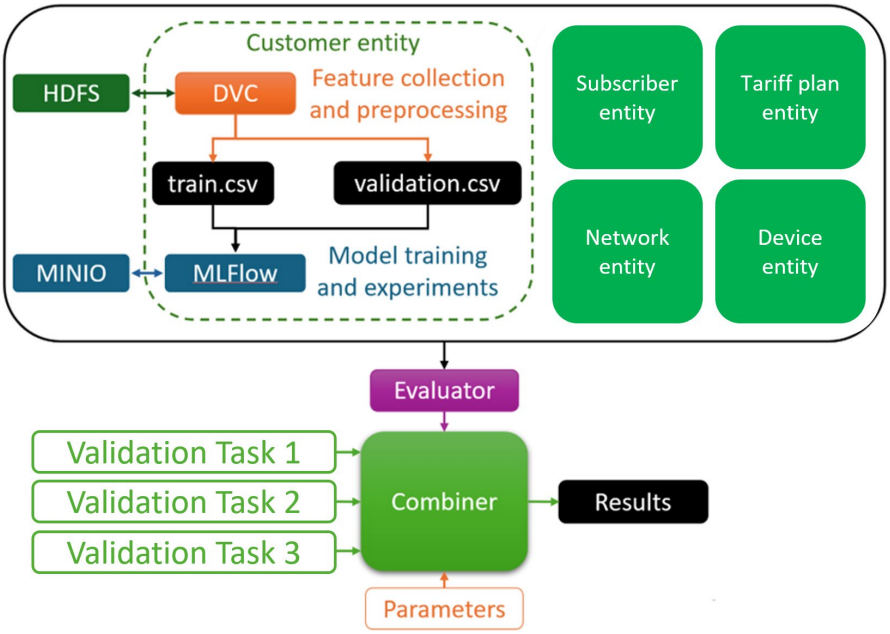

Figure 3. Modular architecture of the embedding-based look-alike modeling framework

The system supports entity-based feature representation, where different groups of features correspond to separate informational domains describing subscriber behavior. These entities include subscriber-related attributes, device characteristics, network usage statistics, and tariff plan information. By separating these domains into distinct entities, the framework enables flexible representation learning and facilitates the construction of multi-entity embedding vectors.

During the evaluation stage, the learned embeddings are processed by an evaluation module that measures their effectiveness across multiple validation tasks. These tasks may correspond to different prediction scenarios, including both binary and multiclass classification problems. The evaluation framework is designed to assess the robustness of the learned representations across different targets and application contexts.

To produce final performance estimates, the results from multiple validation tasks are aggregated by a combining module. This component collects evaluation metrics obtained from different validation scenarios and computes aggregated performance indicators that summarize the effectiveness of the learned embeddings.

Such a modular system architecture enables scalable experimentation and facilitates the deployment of embedding-based similarity models in real-world marketing and recommendation systems. By separating data processing, model training, and evaluation into independent components, the framework allows efficient experimentation with different representation learning strategies and similarity metrics while maintaining reproducibility and computational scalability.

3. Result

The performance of the dimensionality reduction techniques and the proposed representation learning approach was evaluated using the anonymized high-dimensional dataset described in the previous section. The experiments focused on analyzing the structure of the learned embeddings and comparing the effectiveness of different dimensionality reduction methods.

To visually assess the structure of the reduced feature space, the high-dimensional data were projected into a three-dimensional representation using Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and the autoencoder-based latent representtation. These techniques provide different perspectives on the structure of the data, ranging from linear projections to nonlinear manifold-based embeddings and deep learning–based representations.

The resulting projections of the combined entities are illustrated in Figure 4. The PCA projection shows a sparse distribution of data points due to the linear nature of the method, which captures only the directions of maximum variance in the dataset. The t-SNE embedding reveals local neighborhood structures and clustering patterns by preserving similarities between nearby observations in the high-dimensional space. In contrast, the autoencoder-based representation produces a more structured latent space, indicating that the neural network is able to capture complex nonlinear relationships between features.

These results suggest that representation learning methods based on autoencoders can provide more informative embeddings for high-dimensional tabular data compared to classical dimensionality reduction techniques. The learned latent representations are therefore more suitable for similarity-based user analysis and lookalike audience detection.

The clustering structure of user groups in the reduced feature space is further illustrated in Figure 5. The visualization highlights the distribution of different user categories identified by SIM and eSIM configurations, different colors represent distinct user categories based on SIM and eSIM configurations.

The t-SNE projection demonstrates the preservation of local neighborhood relationships between users; however, the clusters remain relatively dispersed and partially overlapping. This behavior is typical for manifold-based visualization methods that primarily focus on preserving local similarity rather than learning a globally structured representation.

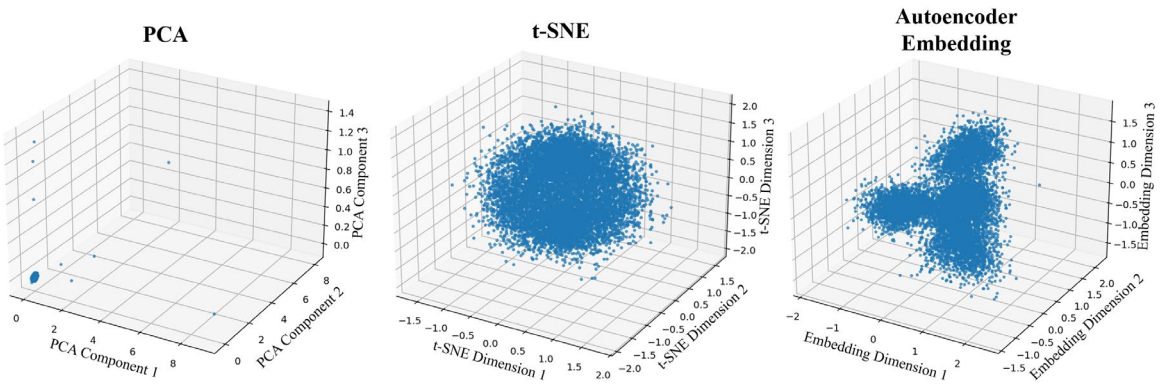

Figure 4. Comparison of dimensionality reduction techniques applied to the dataset using PCA, t-SNE, and autoencoder-based embeddings

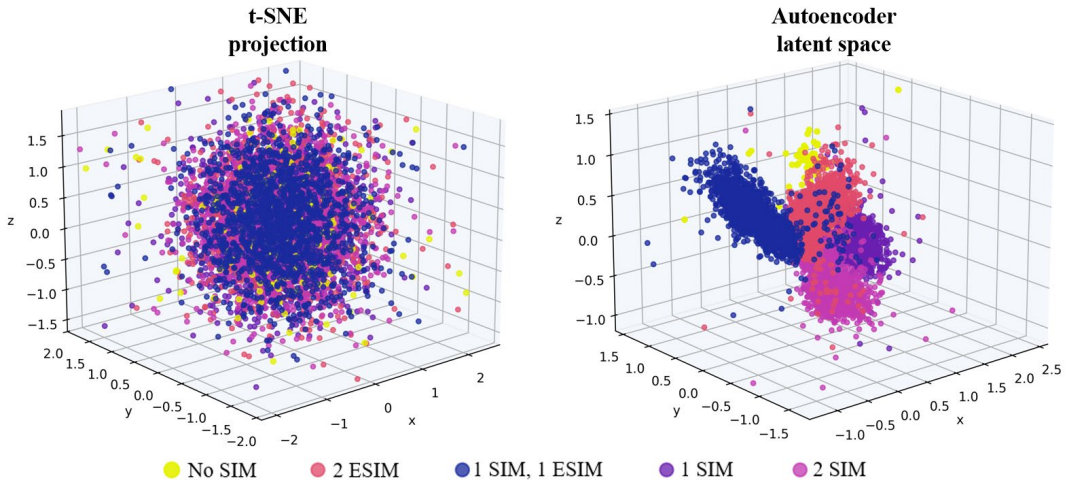

Figure 5. Visualization of user groups in the reduced feature space obtained using t-SNE and autoencoder embeddings

In contrast, the autoencoder-based embedding produces a more structured latent space in which user groups form more compact and distinguishable clusters. This result indicates that the neural network is capable of capturing complex nonlinear relationships between user attributes and encoding them into informative latent representations.

Such structured embeddings are particularly beneficial for downstream similarity-based tasks, including lookalike audience detection and user segmentation.

To further evaluate the quality of the learned representations, additional validation experiments were conducted using clustering and nearest-neighbor classification metrics. The embeddings produced by PCA, t-SNE, and the autoencoder model were compared using several widely used clustering evaluation metrics, including Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Score.

In addition, a k-Nearest Neighbors (kNN) classifier was trained on the resulting embeddings to assess their effectiveness for similarity-based classification tasks. The evaluation was conducted using several independent validation datasets representing different prediction scenarios, including both binary and multiclass classification tasks. Importantly, the target variables used in these validation datasets were obtained from external sources and were not present in the original dataset used for training the dimensionality reduction models. This design eliminates the possibility of target leakage and ensures that the learned embeddings do not implicitly encode information about the evaluation targets [18].

All evaluation metrics reported in the experiments represent averages across multiple validation datasets. This evaluation protocol provides a more reliable estimate of the generalization ability of the learned representations across different prediction tasks. Such robustness is particularly important for lookalike audience modeling systems, where the objective of the model is not tied to a single predefined target variable.

In real-world production environments, lookalike audience services operate as generalized tools for B2B clients. The specific prediction task and target behavior may vary significantly between campaigns, and the model must be capable of identifying relevant similarities between users regardless of the specific target definition. Therefore, averaging performance metrics across multiple external validation tasks provides a realistic assessment of how well the learned embeddings can support diverse downstream applications.

The results presented in Table 1 indicate clear differences in the quality of the learned representations. PCA demonstrates the lowest clustering quality across all metrics, which can be explained by the linear nature of the method that limits its ability to capture complex nonlinear relationships in the data [19].

Table 1. Comparison of dimensionality reduction techniques on validation tasks

Method	Silhouette Score	Davies–Bouldin	Calinski–Harabasz	kNN Accuracy	kNN F1
PCA	0.21	1.84	410	0.58	0.56
t-SNE	0.34	1.21	620	0.66	0.63
Autoencoder	**0.48**	**0.79**	**1020**	**0.74**	**0.71**

The t-SNE method shows improved clustering structure compared to PCA due to its ability to preserve local neighborhood relationships in the data. However, since t-SNE is primarily designed for visualization rather than representation learning, its performance on downstream tasks remains limited.

The autoencoder-based representation achieves the best performance across all evaluation metrics. In particular, it produces the highest Silhouette Score and Calinski–Harabasz Score while also minimizing the Davies–Bouldin Index, indicating more compact and well-separated clusters. Furthermore, the kNN classification results demonstrate that the autoencoder embeddings preserve meaningful similarity relationships between users.

These findings confirm that neural network–based representation learning provides a more informative latent feature space for high-

dimensional tabular data compared to classical dimensionality reduction techniques.

In the next stage of the study, the learned representations were used to construct a similarity-based lookalike audience detection framework. For each entity in the dataset, including subscriber attributes, device characteristics, tariff information, and network-related features, separate embeddings were generated using the trained autoencoder models. These entity-level embeddings were then concatenated to form a unified latent representation describing each user.

User similarity was computed using cosine similarity between the resulting embedding vectors. Cosine similarity measures the angular distance between vectors in the latent space and is widely used in representation learning tasks because it focuses on the orientation of vectors rather than their magnitude. This property makes it particularly suitable for comparing high-dimensional embeddings where the direction of the vector encodes semantic relationships between observations.

The concatenated entity embeddings therefore form a compact representation of user behavior across multiple data domains. Similar users can then be identified by measuring cosine similarity between their corresponding embedding vectors.

After evaluating the structural quality of the learned embeddings, the next experiment focuses on their practical applicability in the lookalike audience detection tasks. Classification metrics were calculated for several baseline machine learning models as well as embedding-based similarity approaches [20]-[22]. The results are summarized in Table 2.

Table 2. Provide a concise caption for each table, explaining its content and relevance

Model	CR	ROC AUC	Lift Top 1	Precision	Recall
SVM	0.13	0.64	4.9	0.54	0.55
Random Forest	0.15	0.66	5.4	0.60	0.54
LightGBM	0.19	0.69	6.6	0.64	0.56
Cosine similarity with embeddings	0.21	0.70	7.3	0.67	0.61
Cosine similarity with concatenated embeddings	0.31	0.76	11.7	0.73	0.70

The results demonstrate that embedding-based similarity methods significantly outperform traditional machine learning classifiers in the lookalike detection task. While classical models achieve moderate performance levels, the use of learned embeddings improves all evaluation metrics.

In particular, the cosine similarity approach applied to concatenated entity embeddings achieves the highest performance across all metrics. The Lift Top 1 metric increases from 6.6 for the best baseline model (LightGBM) to 11.7, indicating a substantial improvement in identifying the most relevant users within the target audience. Similarly, both precision and recall values increase, reflecting better identification of users with similar behavioral characteristics [23].

These results suggest that representation learning techniques provide more informative feature spaces for similarity-based analysis compared to traditional machine learning models operating directly on high-dimensional tabular data.

4. Discussion

The results presented in the previous section demonstrate the effectiveness of nonlinear representation learning methods for analyzing high-dimensional tabular datasets. The comparison between linear and nonlinear dimensionality reduction techniques highlights the limitations of traditional approaches such as Principal Component Analysis when applied to complex datasets containing heterogeneous user attributes.

Visualization experiments reveal clear structural differences between the examined dimensionality reduction methods. Linear projections produced by PCA tend to distribute observations more sparsely across the reduced space. This behavior is expected because PCA preserves directions of maximum global variance rather than capturing the intrinsic structure of the data. As a result, complex nonlinear relationships between features may remain hidden in the projected representation.

In contrast, nonlinear methods produce more structured representations of the data. The t-distributed Stochastic Neighbor Embedding (t-SNE) method improves the visualization of local neighborhood structures by preserving pairwise similarities between nearby observations. This allows clusters of similar users to become more visible in the embedded space. However, despite its effectiveness for exploratory visualization, t-SNE does not learn a parametric mapping function and therefore cannot be directly applied to new data samples without recomputing the embedding. This limitation restricts its applicability in large-scale production systems.

The autoencoder-based approach addresses these limitations by learning a parametric nonlinear transformation from the original feature space to a compact latent representation. The experimental results demonstrate that autoencoder embeddings produce a more structured latent space compared to classical dimensionality reduction techniques. This representation allows the model to capture complex nonlinear interactions between features and preserve meaningful relationships between users.

The evaluation results presented in Table 1 further confirm the advantages of learned embeddings. Autoencoder representations outperform PCA and t-SNE across clustering and nearest-neighbor classification metrics, indicating improved cluster separation and better preservation of local similarity relationships in the latent space.

The results summarized in Table 2 demonstrate the practical benefits of embedding-based similarity methods for look-alike audience detection. Traditional machine learning models trained directly on high-dimensional feature vectors achieve moderate predictive performance. In contrast, similarity-based approaches operating on learned embeddings show significantly improved results across multiple evaluation metrics.

In particular, the use of cosine similarity applied to concatenated multi-entity embeddings provides the highest performance among the evaluated methods. The concatenation of embeddings from multiple entities enables the model to integrate heterogeneous information describing different aspects of user behavior, including subscriber attributes, device characteristics, tariff plans, and network usage patterns. This unified representation captures complementary information from multiple domains and allows more accurate identification of similar users in the latent space.

From an industrial perspective, the proposed approach provides an effective framework for scalable look-alike audience modeling. Unlike traditional campaign-specific classification models, the embedding-based framework produces generalized user representations that can support multiple prediction tasks. This property is particularly important in real-world marketing platforms where different B2B clients may require similarity analysis for diverse target behaviors.

Despite these advantages, several limitations should be acknowledged. First, the quality of learned embeddings depends strongly on the quality and diversity of the training data. If the dataset contains biased or incomplete information, the resulting representations may fail to capture certain behavioral patterns. Second, although neural network–based approaches are scalable once trained, the training process itself may require substantial computational resources when working with very large datasets.

Future research directions may include the exploration of alternative representation learning architectures for tabular data, including transformer-based models or hybrid embedding frameworks. In addition, further investigation of similarity metrics and embedding aggregation strategies may provide additional improvements for large-scale user similarity analysis in industrial environments.

5. Conclusions

This study investigated dimensionality reduction and representation learning techniques for analyzing high-dimensional tabular datasets in the context of look-alike audience detection. The research compared classical dimensionality reduction methods, including Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE), with a deep learning–based representation learning approach based on autoencoders.

The experimental results demonstrate that autoencoder-based models are capable of learning compact and informative latent representations of user data. Unlike classical dimensionality reduction techniques, which either rely on linear projections or are primarily designed for visualization, autoencoders learn a parametric nonlinear mapping between the original feature space and a lower-dimensional embedding space. This property allows the model to capture complex feature interactions

and preserve meaningful similarity relationships between users.

The evaluation results show that autoencoder embeddings provide improved clustering quality and higher performance in similarity-based classification tasks compared to PCA and t-SNE representtations. In addition, the use of cosine similarity applied to concatenated multi-entity embeddings enables effective identification of similar users across heterogeneous data sources. By integrating embeddings derived from different entities, the proposed approach constructs a unified user representation that captures multiple aspects of subscriber behavior.

From a practical perspective, the proposed framework supports scalable look-alike audience modeling for large telecommunications datasets. Unlike traditional campaign-specific models, the embedding-based approach produces generalized user representations that can be reused across multiple prediction tasks. This property makes the method particularly suitable for industrial applications such as automated recommendation systems and targeted advertising platforms serving multiple B2B clients.

Overall, the findings confirm that deep learning–based representation learning provides an effective solution for handling complex high-dimensional tabular data. Future research may focus on exploring alternative neural architectures for tabular representation learning, investigating advanced similarity metrics, and extending the proposed framework to additional domains involving large-scale heterogeneous datasets.

Author Contributions

Conceptualization, M.N; Methodology, M.N.; Software, I.T.; Validation, I.T. and M.N.; Formal Analysis, I.T. and M.N.; Investigation, I.T.; Resources, I.T. and M.N.; Data Curation, I.T.; Writing – Original Draft Preparation, I.T.; Writing – Review & Editing, M.N.; Visualization, I.T.; Supervision, M.N.; Project Administration, M.N.; Funding Acquisition, M.N.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. A. Altaibek, I. Tokhtakhunov, M. Nurtas, D. Kozhamzharova, and M. Aitimov, “The efficacy of autoencoders in the utilization of tabular data for classification tasks,” *Procedia Computer Science*, vol. 238, pp. 492–502, 2024, doi: 10.1016/j.procs.2024.06.052.

2. I. T. Jolliffe and J. Cadima, “Principal component analysis: A review and recent developments,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, Art. no. 20150202, 2016, doi: 10.1098/rsta.2015.0202.

3. L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

4. G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

5. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

6. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

7. I. Tokhtakhunov, M. Nurtas, A. Alex, E. Neftissov, S. P. I. Kazambayev, and L. Kirichenko, “Exploring autoencoder-based representations for tabular data classification,” *Engineered Science*, vol. 37, Art. no. 1703, 2025, doi: 10.30919/es1703.

8. I. Tokhtakhunov, A. Altaibek, and M. Nurtas, “Optimizing similar audience search in targeted advertising: Effectiveness of Siamese networks for autoencoder-based user embeddings,” *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23367–23375, 2025, doi: 10.48084/etasr.10527.

9. S. Roweis and L. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

10. J. B. Tenenbaum, V. de Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

11. L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform manifold approximation and projection for dimension reduction,” *arXiv preprint* arXiv:1802.03426, 2018.

12. A. Ng, “Sparse autoencoder,” *CS294A Lecture Notes*, Stanford University, Stanford, CA, USA, 2011.

13. Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013, doi: 10.1109/TPAMI.2013.50.

14. A. Sannigrahi, R. Walambe, and K. Kotecha, “Multi-head variational graph autoencoder framework for link prediction on citation graphs,” *Engineered Science*, vol. 34, Art. no. 1406, 2025, doi: 10.30919/es1406.

15. S. Abrar and M. D. Samad, “Perturbation of deep autoencoder weights for model compression and classification of tabular data,” *Neural Networks*, vol. 156, pp. 160–169, 2022, doi: 10.1016/j.neunet.2022.09.020.

16. H. Torabi, S. L. Mirtaheri, and S. Greco, “Practical autoencoder based anomaly detection by using vector reconstruction error,” *Cybersecurity*, vol. 6, Art. no. 1, 2023, doi: 10.1186/s42400-022-00134-9.

17. T. P. Rinjeni, A. Indriawan, and N. A. Rakhmawati, “Matching scientific article titles using cosine similarity and Jaccard similarity algorithm,” *Procedia Computer Science*, vol. 234, pp. 553–560, 2024, doi: 10.1016/j.procs.2024.03.039.

18. H. S. Lom, A. C. Thoo, W. M. Lim, and K. Y. Koay, “Advertising value and privacy concerns in mobile advertising: The case of SMS advertising in banking,” *Journal of Financial Services Marketing*, vol. 29, no. 3, pp. 1135–1153, 2024, doi: 10.1057/s41264-023-00263-3.

19. O. Rainio, J. Teuho, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” *Scientific Reports*, vol. 14, no. 1, Art. no. 6086, 2024, doi: 10.1038/s41598-024-56706-x.

20. F. Pedregosa *et al*., “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

21. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

22. C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.

23. P. Geetha, C. Naikodi, and L. Suresh, “Optimized deep learning for enhanced trade-off in differentially private learning,” *Engineering, Technology & Applied Science Research*, vol. 11, no. 1, pp. 6745–6751, 2021, doi: 10.48084/etasr.4017.

Information about Authors:

Il’murat Tokhtakhunov is a PhD candidate at the Department of Mathematical and Computer Modelling, International Information Technology University (Almaty, Kazakhstan) and a Senior Lecturer at the School of Digital Technologies, Narxoz University (Almaty, Kazakhstan). His research focuses on machine learning methods for high-dimensional tabular data analysis, representation learning, dimensionality reduction, and lookalike audience modeling for targeted advertising systems.

Marat Nurtas is an Associate Professor at the Department of Mathematical and Computer Modelling, International Information Technology University (Almaty, Kazakhstan) and a Leading Researcher at the Institute of Ionosphere. He received his PhD degree in Mathematical and Computer Modelling from Kazakh-British Technical University and holds a bachelor’s degree in Mathematics from Al-Farabi Kazakh National University. His research interests include scientific machine learning, deep neural networks, physics-informed neural networks, geophysical data analysis, earthquake prediction models, and machine learning applications in complex dynamical systems.

Submission received: 21 February, 2026.
Revised: 20 March, 2026.
Accepted: 20 March, 2026.