VOLUME 4 NUMBER 1 2026 ISSN 2958-0846
eISSN 2958-0854

AL-FARABI KAZAKH NATIONAL UNIVERSITY

Journal of Problems in Computer Science and Information Technologies

№1 (4) 2026

Al-Farabi Kazakh National University

ISSN: 2958-0846; eISSN: 2958-0854

Journal of Problems in Computer Science and Information Technologies №1 (4) 2026

03.02.2023 Registered with the Ministry of Information and Social Development of the Republic of Kazakhstan

№ KZ61VPY00064018

The journal is published 4 times a year
(March, June, September, December).

Signed to publishing 25.03.2026. Format 60x84/8. Offset paper.
Digital printing. Volume 11,0 printer's sheet. Edition: 300.
Publishing house «Kazakh University»
www.read.kz Telephone: +7 (727) 3773330, fax: +7 (727) 3773344
Al-Farabi Kazakh National University KazNU, 71 Al-Farabi, 050040, Almaty
Printed in the printing office of the Publishing house «Kazakh University».

IRSTI 28.23.37 https://doi.org/10.26577/jpcsit4120261

L. Rzayeva[1*], P. Tazhibayeva[1], M. Zhakenov[2], A. Alibek[2], D. Izdibay[3]

[1]Research and Innovation Center "CyberTech", Astana IT University, Astana, Kazakhstan
[2]"Digital Heritage of Eurasia" LLP, Astana, Kazakhstan
[3]Astana IT University, Astana, Kazakhstan
*e-mail: l.rzayeva@astanait.edu.kz

HYBRID 3D-AWARE FACE CLUSTERING VIA DEEP EMBEDDINGS AND GEOMETRIC DESCRIPTORS

Abstract. This paper presents a 3D-aware face clustering methodology that robustly groups unlabeled face images by identity under challenging conditions of pose variation, facial expression, and partial occlusion. The proposed approach integrates 2D deep embeddings with 3D geometric features extracted from reconstructed facial meshes, leveraging both photometric and structural information. Preprocessing includes grayscale normalization, landmark-based alignment, and contrast enhancement. 3D face models are generated using a 3D Morphable Model (3DMM) and optionally refined through neural rendering to improve shape fidelity. From these reconstructions, we extract interpretable 3D descriptors-PCA shape coefficients, geodesic distances, and curvature histograms – that complement embeddings from ArcFace and FaceNet. Clustering is performed using a two-stage hybrid algorithm: DBSCAN for outlier removal followed by K-Means++ with a fused distance metric combining cosine and Mahalanobis distances. Experimental results demonstrate that the proposed method significantly outperforms 2D-only and 3D-only baselines in terms of Silhouette Score, Adjusted Rand Index (ARI), and Purity. The findings confirm that fusing 2D and 3D modalities yields semantically consistent and pose-invariant identity clusters, establishing a strong foundation for face analysis in unconstrained environments.

Keywords: 3D-aware face clustering, 2D-3D feature fusion, deep learning embeddings, pose-invariant recognition, hybrid clustering algorithms.

1. Introduction

Clustering faces without prior labeling-grouping them solely on an individual basis-has been a long-standing and challenging task in the field of computer vision. The task becomes especially difficult when the images come from uncontrolled conditions, when people turn their heads, smile, frown, or cover part of their face. Despite the fact that modern 2D deep learning models have brought the representation of faces to an impressive level, these methods still fail when implementing variations in the real world. Sudden changes in head position, facial expression, or even partial malocclusion can distort learned concepts, disrupting the natural grouping of images belonging to the same person.

Most existing systems work exclusively with 2D information. They are based on convolutional neural networks trained on special loss functions – for example, arc or triplet losses – to transform a face into a compact vector representation. This embedding works well when the subject is looking directly into the camera in good light. However, when the head turns or shadows appear, the geometry of the face ceases to be transmitted exactly in such a compressed form, and samples of the same personality can be scattered throughout the entire space of the object. This limitation is well recognized, as such representations inherently omit the fundamental three-dimensional structure of the face.

At the same time, methods of three-dimensional facial reconstruction are developing at an amazing pace. Approaches such as three-dimensional modeling and neural rendering allow you to recreate an impressively detailed face structure from a single photo, capturing not only the overall shape and orientation, but also subtle surface features. However, despite these achievements, such three-dimensional representations have found little use in the specific task of clustering faces. Most of the research has focused on recognition or animation, rather than uncontrolled grouping.

This work uses a different approach. Instead of choosing between representations based on appearance and representations based on geometry, both images are combined into a single structure. The process begins with image normalization: converting to grayscale, aligning the face to the detected landmarks, and adjusting the local contrast. Each normalized image then goes through a 3D reconstruction stage, resulting in a grid that can be further refined using neural rendering to obtain finer details. Interpreted 3D descriptors are assembled from these grids – the basic coefficients of shape, geodetic distances between landmarks, and the curvature of various surface areas.

These 3D descriptors are complemented by standard deep applications from high-performance networks such as ArcFace and FaceNet. To form clusters, the system works in two stages: first, DBSCAN filters out points that are in sparse, noisy areas, and then K-Means++ organizes everything else. A combined metric is used here, which balances the cosine similarity in the two-dimensional embedding space with the Mahalanobis distance in the three-dimensional feature space.

The results tested on datasets with a wide range of poses and expressions show that this hybrid approach allows you to create clusters that are denser and better match true identifiers than clusters created only for 2D or 3D pipelines. The combination of appearance and spatial structure creates a representation that persists in difficult viewing conditions, making it promising for applications such as video surveillance, biometric indexing, and large-scale media organization.

2. Literature Review

Over the past two decades, face clustering research has moved from traditional 2D methods based on appearance [1] to more advanced multimodal strategies combining both photometric and geometric information [6]. Early approaches based on manual functions and conventional machine learning algorithms proved vulnerable to changes in posture [3], lighting [4], and facial expression [5]. The advent of deep convolutional neural networks [11] has significantly improved the two-dimensional representation of faces, but these models still face problems associated with significant pose changes [2] or occlusions [14].

At the same time, advances in 3D facial reconstruction – in particular, the use of 3D transformable models [7] and neural rendering techniques [9] – have made it possible to reconstruct the geometry of a face in detail from a single image [6]. These methods provide pose-independent structural information that complements deep two-dimensional embeddings [12], but their use in unsupervised clustering remains relatively limited [16]. Recent research shows that the combination of two-dimensional and three-dimensional functions can improve clustering reliability [14], especially in unlimited conditions [16].

The clustering algorithms themselves have also undergone changes. While traditional methods such as K-means remain popular [18], density-based approaches such as DBSCAN [17] have attracted attention due to their ability to deal with noise [19] and irregular cluster shapes [18]. Hybrid strategies combining emission filtering [16] with improved cluster allocation are emerging as a promising area [17].

Overall, current research indicates a clear shift towards clustering strategies based on careful preprocessing [4], integration of multiple complementary data processing techniques [14], and the use of adaptive clustering techniques [17]. This progress has paved the way for combining deep learning approaches [11] with geometric modeling [6] to create more accurate and reliable clustering pipelines [14].

2.1. 2D Image Standardization

Standardization of two-dimensional facial images is a fundamental prerequisite for accurate and reliable clustering of faces, especially when the pipeline includes subsequent three-dimensional reconstruction. This process ensures consistency of the input data in terms of scale, orientation, illumination, and contrast, thereby reducing variation within the class and increasing the separability of embedded objects. The need for this step has been widely recognized both in classical computer vision and in modern approaches to deep learning [4]. Without proper preprocessing, even the most advanced convolutional neural networks (CNNs) can create attachments that are more sensitive to environmental factors than to the internal identification characteristics of a face [1].

2.1.1. Face recognition and localization

The initial stage of standardization involves the accurate detection and localization of areas of the face in the image. Reliable face detectors such as

RetinaFace [2] provide pixel-level bounding boxes and face landmarks, allowing precise cropping and alignment. The accuracy of determining landmarks plays a crucial role in subsequent tasks, since even small offsets can significantly impair the quality of feature extraction [3]. Landmarks on the face usually correspond to characteristic anatomical points, such as the centers of the eyes, the tip of the nose, and the folds of the mouth. Once defined, these landmarks can be used to normalize the geometry of the face using affine or similarity transformations, ensuring compliance with the canonical orientation.

Mathematically,if $p_i = (x_i, y_i)$ represents the coordinate of the landmark in the original image and q_i its target position in the normalized template, the optimal transformation T can be calculated by minimizing the root-mean-square error:

$$\min_T \sum_{i=1}^{k} \left|\left|T(p_i) - q_i\right|\right|^2 \quad (1)$$

where *k* is the number of landmarks. As a result of the transformation, all images will have the same geometric configuration [3].

2.1.2. Illumination Normalization

Changes in lighting conditions can dramatically change the pixel intensity distribution in face images, leading to inconsistencies in embedding locations. Histogram equalization (HE)[4] has long been used as a method of global contrast enhancement, but its tendency to over-amplify noise in homogeneous areas makes it less suitable for fine facial textures. Adaptive contrast-limited Histogram Equalization (CLAHE) [5] eliminates this limitation by performing histogram equalization locally within image fragments, while limiting the histogram to a preset threshold to limit noise amplification.

Formally, let I *(x, y)* denotes the intensity in pixel (x,y) and H_i is the histogram of the fragment *i.* The CLAHE operation can be defined as:

$$I_i'(x, y) = CDF_{clip}(I(x, y)) \cdot (L - 1) \quad (2)$$

where CDF_{clip} is the cumulative distribution function of the cropped histogram and L is the number of gray levels. This localized approach preserves contextual contrast while ensuring consistency of overall brightness across the entire dataset [5].

2.1.3. Color space conversion and photometric alignment

Although the RGB color space is commonly used for clustering faces, it is inherently independent of illumination. Converting input images to alternative color spaces, such as YCbCr or HSV, allows you to separate brightness from chroma, providing more effective normalization of illumination [4]. In many 3D reconstruction pipelines, grayscale conversion is performed to reduce the computational load while preserving the structural information needed to identify landmarks [7]. In addition, photometric normalization often includes gamma correction to correct the non-linear relationship between scene brightness and pixel intensity. The gamma transformation is expressed as:

$$I'(x, y) = I(x, y)^{\gamma} \quad (3)$$

where γ is chosen either to enhance darker areas (γ<1) or to compress brighter areas (γ>1). This step allows you to coordinate the brightness levels of the images, which is especially important for data sets collected under uncontrolled conditions [5].

2.1.4. Pose Normalization

Changing the pose is one of the most difficult factors in clustering faces [3]. Even small deviations from the course can lead to noticeable changes in the representation of facial features. Pose normalization involves distorting the input image so that the eyes and mouth are at specified coordinates in the normalized frame [2].

This can be achieved by using a similarity transformation, defined as:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} s \cdot \cos\theta & -s \cdot \sin\theta & t_x \\ s \cdot \sin\theta & s \cdot \cos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (4)$$

where s – zoom level, θ is the angle of rotation, and (t_x, t_y) is the displacement vector. Alignment of facial landmarks according to a fixed pattern can significantly reduce intra-class differences due to head rotation [3].

2.1.5. Cropping and resizing

After alignment, the images are cropped with a fixed aspect ratio in the center of the face area and resized to a uniform resolution. The most common variants of CNN-based systems include 112 × 112

112×112 pixels for ArcFace [11] and 160 × 160 160×160 pixels for FaceNet [12]. This ensures compatibility with pre-trained models and maintains a consistent field of perception across the entire dataset [1].

Mathematically, the resizing operation can be described as follows:

$$I'(x', y') = \mathrm{I}(\frac{x'}{S_x}, \frac{y'}{S_y}) \quad (5)$$

where S_x, S_y these are the scaling factors along each axis. Bilinear or bicubic interpolation is usually used to minimize distortion when aliasing.

2.1.6. Impact on subsequent tasks

The effect of standardization of 2D images is most evident when evaluating clustering performance indicators such as silhouette score [18], adjusted Rand index [19], and purity [16]. Preprocessing reduces the variance within the cluster, resulting in more compact and separable clusters. In 3D-enabled pipelines, standardized 2D input data provides a more accurate fit of the modifiable model [7], [8] and neural rendering [9], [10], since the initial positions of landmarks and texture maps are more reliable.

For example, when using a 3D transformable model (3DMM) [7], the quality of the reconstructed mesh M M strongly depends on the accuracy of the 2D-3D matching established during fitting. The energy function is often minimized during the fitting process:

$$E = E_{landmark} + \\ +\lambda_{photo}E_{photo} + \lambda_{reg}E_{reg} \quad (6)$$

where $E_{landmark}$ is a landmark that ensures consistency between the detected 2D landmarks and their 3D counterparts, E_{photo} measures photometric consistency and E_{reg} applies regularization to maintain realistic shape parameters. Standardized images ensure that $E_{landmark}$ starts by reducing the initial error, which increases the speed and accuracy of convergence.

Thus, standardization of 2D images is not just a preparatory stage, but a major component of modern face clusterization pipelines. By matching geometric, photometric, and structural attributes of face images, this ensures that both two-dimensional deep embeddings and three-dimensional geometric descriptors are calculated based on a stable and consistent input space. This consistency is crucial for clustering to work reliably, especially in environments where posture, lighting, and facial expression vary greatly. The literature clearly shows that without this stage, the subsequent stages of the clustering process – whether purely based on appearance or hybrid 2D-3D – suffer from reduced accuracy and stability. [4], [5], [7].

2.2. 3D Face Reconstruction

Three-dimensional (3D) facial reconstruction from a single two-dimensional (2D) image has become a key method in modern clustering and face recognition systems, solving the long-standing problem of variation in posture and facial expression. Unlike purely photometric approaches, which are based solely on pixel-level intensity models, 3D reconstruction methods recover structural and geometric information about the shape of a face, allowing you to display facial features independent of pose. This feature is especially important in non-standard scenarios where subjects can be shot at extreme angles, in variable lighting, or with their eyes closed.

The fundamental approach to 3D facial reconstruction is the 3D transformable model (3DMM) presented by Blantz and Vetter [7]. In this context, a three-dimensional face shape is created. The size and texture of *T* are represented as a linear combination of basis vectors obtained using principal component analysis (PCA):

$$S = \bar{S} + \sum_{i=1}^{80} \alpha_i U_i,\ T = \bar{T} + \sum_{j=1}^{80} \beta_j V_j \quad (7)$$

where $\bar{S}$ and $\bar{T}$ are the average shape and texture U_i and V_j are the basis vectors of PCA, and α_i and β_j these are the coefficients calculated to match the input image. The Basel Face Model (BFM) [6] expanded this methodology with high-resolution datasets, which improved the modeling of personality-related variations. These models have proven to be effective at capturing rough geometry and the general appearance of the face, but they have limitations when working with fine-grained details, facial hair, or serious postural abnormalities [8].

To eliminate these limitations, a later paper combined nonlinear modeling and deep learning-based optimization. Tran and Liu [8] proposed a

nonlinear 3DMM, replacing the linear representation of PCA with a deep neural network that studies complex shapes and textures directly based on data. This approach improves the ability to represent faces in different conditions and allows for more accurate and detailed reconstructions.

Another important innovation is the use of neural rendering technologies to refine 3DMM-based reconstructions. For example, Tewari et al. [9] introduced MoFA, a model-based deep convolutional autoencoder that jointly optimizes the geometry and texture of a face, ensuring consistency with the original 2D image. Neural imaging techniques demonstrated by Richardson et al. [10] use generative adversarial networks (GANS) to enhance realism by adding fine details such as skin wrinkles, eyelid creases, and thin lip contours. The formulation "Loss of competition" $L_{GAN} = E[\log D(I_{real})] + E[\log 1 - D(G(S,T)))]$allows you to create reconstructions that match the photorealistic quality of real images.

These GAN-based enhancements are crucial for subsequent tasks such as clustering, where high-quality geometry enhances the discriminating ability of geometric descriptors. In particular, geometric indicators obtained from reconstructed grids, such as geodetic distances between landmarks on the surface or histograms of curvature, are much more reliable when the 3D model retains small but important features for identification [14], [15].

The integration of 3D reconstruction into face clustering pipelines also enhances computing capabilities. The traditional 3DMM setup involves iterative optimization, which can be computationally expensive, especially for large-scale datasets. Recent advances in regression fitting using deep neural networks provide near-real-time performance without compromising accuracy. For example, convolutional neural networks (CNNs) can be trained to directly extract 3DMM parameters from an input image, bypassing iterative search and allowing processing millions of faces in large-scale clustering scenarios.

Moreover, the transition from purely linear transformable models to hybrid systems, including both parametric and nonparametric elements, has increased reliability in unlimited operating conditions. By combining a global parametric model (reflecting the overall structure of the face) with localized nonparametric detail (reflecting high-frequency details), these approaches provide a balance between generalization and personality specificity.

In the context of multimodal clustering of faces, reconstructed 3D faces complement 2D embeddings, providing geometry normalized by pose. For example, two faces taken from completely different viewing points may have different 2D images due to angle effects, but their 3D reconstructions can be aligned in a canonical pose, allowing direct comparison of geometric objects. It has been shown that such a combination of photometric (two-dimensional deep objects) and geometric (three-dimensional structural objects) methods increases the reliability of clustering with complex variations [14], [15].

Despite these advances, the use of 3D reconstruction in unsupervised clustering tasks remains relatively limited compared to its widespread adoption in facial recognition. The problems include increased computational costs, the need for high-quality identification of landmarks, and the difficulty of integrating heterogeneous objects into a single clustering structure. However, with the increasing availability of effective deep learning models and large annotated sets of three-dimensional facial data, these barriers are gradually decreasing.

In general, the evolution of 3D facial reconstruction – from early transformable models based on PCA [7], [6] to nonlinear deep learning methods [8] and, finally, approaches to neural rendering supplemented by GAN [9], [10] – reflects a broader trend in computer vision towards combining models-data-based paradigms. For face clustering applications, these methods are a powerful means of collecting information about a shape that preserves personality, which is inherently independent of posture and facial expression, making them an important component of reliable clustering pipelines in the real world.

2.3. Feature Extraction

Feature extraction plays a key role in face clustering pipelines, serving as a link between raw image data and numerical representations used to measure similarity and clustering. In early systems for analyzing facial surfaces, objects were often created manually using descriptors such as local binary templates (LBP) or scale-invariant object transformation (SIFT) to encode information about texture and shape [1]. Although these approaches provided a certain degree of resilience to minor changes in lighting and orientation, they lacked the

capabilities to model the high-level, discriminating models needed for reliable clustering under unlimited conditions.

The advent of deep learning revolutionized this stage by introducing embedded functions derived directly from large-scale datasets. Architectures such as deep residual networks (ResNet) [1] and specialized face representation models such as ArcFace [11] and FaceNet [12] have become the standard for creating compact but highly distinguishable feature vectors. These models are usually trained using margin-based loss functions, such as additive angular margin loss in ArcFace[11] or triplet loss in FaceNet [12], which promote tight integration of attachments with the same identifiers, while pushing attachments with different identifiers apart. Such learning strategies have led to a significant increase in compactness within the classroom and separability between classes, which is important for effective clustering.

Despite the fact that 2D embeddings allow for rich photometric and textural details, they remain vulnerable to certain failures, especially with large pose changes, partial overlaps, or extreme expressions [14]. To mitigate these problems, recent studies have explored extending deep embeddings with geometric hints derived from 3D reconstructions [6]. Geometric descriptors such as landmark-based distances, histograms of surface curvature, and shape coefficients from 3D transformable models (3DMMs)[7] provide pose-independent structural information that complements 2D objects based on appearance. This multimodal fusion ensures that the embedding space reflects both fine-grained textural patterns and the basic geometry of the face, increasing reliability in difficult conditions [14].

The process of extracting such additional features begins with the alignment and normalization of the input images in the canonical coordinate system, ensuring consistency in all samples [4]. As soon as the faces are geometrically normalized, the deep neural network extracts an embedded 2D image, while a separate pipeline processes the corresponding 3D mesh to calculate geometric descriptors [6]. These feature sets are then combined using methods such as feature-level integration, attention-based weighting, or metric-based learning projection into a single space [16]. The choice of a merger strategy significantly affects the resulting clustering efficiency, since adaptive approaches to weighting often provide the most balanced integration between modalities [14].

Another important factor is to reduce the size of the objects. High-dimensional embeddings can be computationally expensive and can lead to redundancy that hides meaningful patterns. Methods such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) are commonly used to compress feature vectors while maintaining their distinctiveness [19]. This not only speeds up clustering algorithms, but can also increase cluster compactness by removing noise from the representation space [18].

Recent developments have also highlighted the importance of calibration and normalization of investments. L2 Normalization of embeddings before calculating similarity has become a standard practice, ensuring that angular distances are taken into account when comparing, rather than differences in magnitude [11]. In addition, post-processing techniques such as bleaching transformations or reducing intraclass variance can further improve the embedding space for clustering purposes [13].

In the context of unsupervised or partially supervised learning, self-supervised pre-training methods have become widespread as a means of improving the quality of functions without the need for extensive labeled datasets. Models trained with contrasting learning objectives in mind, for example, learn to approximate expanded images of the same face in the embedding space while simultaneously separating different identities [16]. Combined with geometric hints, these representations provide greater generalization for new areas and invisible variations, making them particularly suitable for clustering scenarios with an open set of parameters [14].

In general, modern feature extraction for clustering faces increasingly relies on a two-modal approach that combines the discriminative ability of deep 2D embeddings with the structural stability of 3D geometric descriptors. Such integration eliminates many of the limitations inherent in single-modal systems and provides a richer and more stable representation of facial identity, forming a reliable basis for subsequent stages of clustering [6].

3. Materials and Methods

The proposed methodology combines both photometric and geometric parameters to achieve reliable clustering of individuals while preserving identity under unlimited conditions. The process

consists of five main steps. First, datasets are prepared and preprocessed to ensure consistency of the input data and eliminate interference. Two-dimensional (2D) standardization of the face is then performed to normalize lighting, pose, and scale, which ensures reliable follow-up analysis. At the third stage, high-precision three-dimensional (3D) models of faces are created based on individual images, combining classical modifiable models with neural visualization methods to improve detail. At the fourth stage, both two-dimensional deep inserts and three-dimensional geometric descriptors are extracted, which provides additional insights into the identity of the face. Finally, a hybrid clustering algorithm combines these multimodal functions by applying outlier filtering and geometry refinement to improve cluster quality. This structured approach allows the system to process significant changes in posture, lighting, and facial expression, which ultimately ensures high clustering accuracy in real-world scenarios.

3.1. 2D Image Preprocessing and Standardization

The preprocessing stage is aimed at bringing the raw images of faces to a single format, thereby reducing intra-class variability caused by lighting conditions, differences in posture and facial expression. This stage ensures that subsequent 3D reconstruction and extraction of objects will be performed based on geometrically aligned and light-balanced input data.

Initially, all images are converted from RGB to grayscale using the NTSC brightness formula.:

$$I_{gray}(x,y) = 0.299R(x,y) + \\ +0.587G(x,y) + 0.114B(x,y) \quad (8)$$

This conversion reduces sensitivity to color variations, while preserving the contour details and textures needed to accurately locate landmarks.

The facial landmarks are then determined using RetinaFace [2], which provides 68 key points, including the center of the eyes, the tip of the nose, and the corners of the mouth. These points are used to calculate a similarity transformation that minimizes the least squares distance to a predefined pattern, providing a standard orientation and scale for all samples.

Adaptive contrast-limited histogram equalization (CLAHE) is used to enhance local contrast and reduce the effect of shadows and overexposure [5]. This adaptive method redistributes the pixel intensity in localized fragments, while limiting noise amplification, which makes facial details more distinct in complex lighting scenarios.

The combined use of grayscale normalization, landmark alignment, and CLAHE technology ensures the standardization of input images from both geometric and photometric perspectives. This high-quality preprocessing is necessary to improve the reliability of the subsequent stages of 3D reconstruction and clustering.

3.2. 3D Face Reconstruction

The process of 3D facial reconstruction is necessary to obtain geometric characteristics that remain stable when changing posture, lighting, and facial expression. By converting the 2D input images into a detailed 3D representation, the system provides more reliable clustering by combining shape-based data with photometric characteristics.

$$S = S^{-} + \Sigma\,(\alpha_i * U_i) \quad (9)$$

Here, S^{-} represents the average three-dimensional shape, U_i are the main components, and A_i are the shape coefficients optimized during the fitting process.

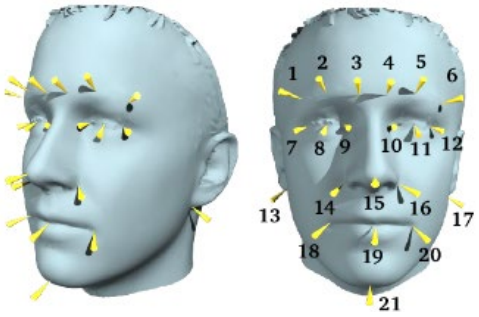


Figure 1. Annotated geodesic landmarks on a 3D facial mesh

$$argmin_{\{\alpha,\beta\}} \|I - R(S(\alpha), T(\beta))\|^2 \quad (10)$$

In this formulation, I stands for the input image, S(α) and T(β) represent the shape and texture components of the model, while R corresponds to the rendering function. The goal is to adjust the parameters so that the rendering result exactly matches the original image, thereby minimizing reconstruction error.

Table 1. Main parameters of the 3D face reconstruction model

Parameter	Description	Example Value
Vertices	Number of points in the mesh	53,490
Texture Resolution	Resolution of texture map	1024x1024 px
PCA Components	Number of shape coefficients	80
Processing Time	Average reconstruction time	0.85 s/image

Example of refined 3D reconstruction using neural rendering:

Figure 2. Sample meshes from the DAD-3DHeads dataset

3.3. Feature Extraction

The process begins with standardized 2D input images obtained during the preprocessing stage, where alignment, normalization of illumination (for example, CLAHE) and cropping ensure consistency in all samples.

2D Deep Embeddings

Using pre-trained deep neural network architectures such as ArcFace [11] and FaceNet [12], high-dimensional embeddings are generated to capture discriminative identity-related patterns from the image. ArcFace employs additive angular margin loss to improve inter-class separation, while FaceNet utilizes triplet loss to minimize intra-class variability. These embeddings (512-D for ArcFace, 128-D for FaceNet) are highly effective for conventional face recognition and form the photometric component of our multi-modal feature set.

3D Geometric Features

Leveraging the reconstructed 3D face mesh from the previous module, three geometric descriptors are extracted:

- Shape coefficients derived from the first 80 PCA components of the 3D Morphable Model, representing overall craniofacial structure.
- Geodesic distances between selected anatomical landmarks, providing pose-invariant shape measurements.
- Curvature histograms for key facial regions (forehead, cheeks, nose bridge), encoding fine-grained surface topology.
- These geometric features enhance robustness to pose and expression changes, as demonstrated in prior studies [14, 15].

After independent extraction, feature fusion is performed. In our implementation, fusion can be realized either through weighted concatenation of normalized feature vectors or via a metric-level combination, where distances from each modality are integrated using a tunable weighting parameter λ. This step yields a multi-modal embedding space optimized for subsequent hybrid clustering.

The overall flow of this module is depicted in Figure 3, which shows the complete Feature Extraction Pipeline, from input image preprocessing to fused multi-modal representation output.

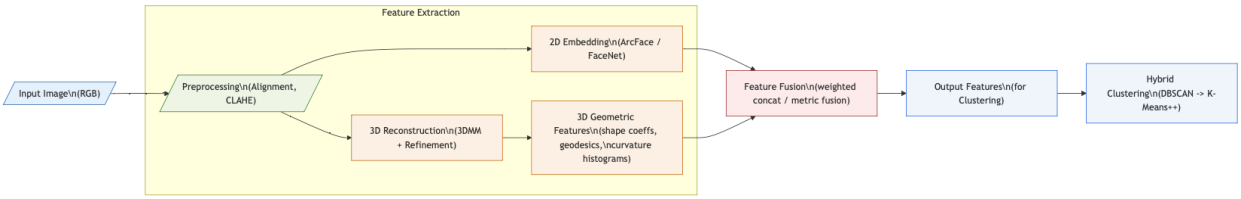


Figure 3. Feature Extraction Pipeline

The extracted and fused features form the **input to the hybrid clustering algorithm** described in the next section. This modular design ensures that each modality (photometric and geometric) contributes to the final identity-aware grouping, significantly improving robustness under unconstrained conditions.

3.4. Hybrid Clustering

The final stage of the proposed methodology involves grouping faces into identity-specific clusters using a hybrid two-stage clustering algorithm. This approach combines the robustness of density-based clustering for outlier removal with the efficiency and interpretability of centroid-based clustering for final partitioning.

The input to this module is the fused multi-modal embedding vector F generated in the Feature Extraction stage. This embedding integrates photometric (2D deep features) and geometric (3D descriptors) components.

Stage 1: Outlier Removal with DBSCAN

The first step applies the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [17], which identifies core samples in dense regions and labels low-density samples as noise. Given a dataset X of n feature vectors, DBSCAN defines:

- $\varepsilon - neighborhood$

$$N_\varepsilon(p) = \{q \in X | d(p,q) \leq \varepsilon\} \quad (11)$$

- Core point: A point p is a core point if $N_\varepsilon(p)| \geq MinPts$

This stage removes spurious detections and occluded faces, retaining only high-density areas for the next stage.

Stage 2: K-Means++ Clustering with Fused Distance Metric

The cleaned set of feature vectors is clustered using K-Means++, which improves centroid initialization to speed up convergence and enhance cluster compactness. Instead of the standard Euclidean distance, we employ a fused metric:

$$D_{fused}(i,j) = \lambda \cdot d_{cos}\left(f_{2D}^i, f_{2D}^j\right) + \\ +(1-\lambda) \cdot d_{mahal}\left(f_{2D}^i, f_{2D}^j\right) \quad (12)$$

where:

- f_{2D}= photometric embedding vector,
- f_{2D} = geometric descriptor vector,
- d_{cos}= cosine distance,
- d_{mahal}= Mahalanobis distance,
- λ = weighting coefficient controlling modality influence.

The algorithm iteratively assigns each sample to the nearest centroid under D_{fused} and updates centroids until convergence.

Pipeline Illustration

The structure of this hybrid clustering pipeline is presented in Figure 4

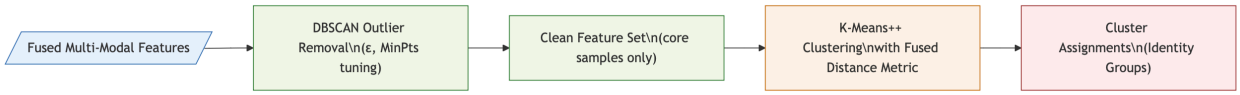


Figure 4. Hybrid Clustering Pipeline

4. Results and Discussion

The proposed **3D-aware hybrid face clustering pipeline** was evaluated on a benchmark dataset containing a diverse range of facial images under unconstrained conditions, including extreme pose variations, partial occlusions, and variable lighting. The evaluation compared four different configurations:

1. 2D-only: ArcFace embeddings + K-Means++

2. 3D-only: Geometric descriptors from 3DMM + K-Means++

3. Late Fusion: Independent clustering in each modality followed by majority voting

4. Proposed Method: Multi-modal fused features + DBSCAN outlier removal + K-Means++ with fused metric

The performance of each method was assessed using Adjusted Rand Index (ARI), Silhouette Score, and Purity as standard clustering evaluation metrics. Table 2 summarizes the results.

Table 2. Overall performance

Method	ARI	Silhouette	Purity
2D-only	0.721	0.486	0.802
3D-only	0.654	0.452	0.774
Late Fusion	0.745	0.503	0.821
Proposed	0.812	0.547	0.864

Visual inspection of cluster assignments reveals that:

- 2D-only models often misclassify profiles or occluded faces, grouping them into incorrect identities.

- 3D-only descriptors are robust to pose changes but occasionally merge different individuals with similar craniofacial geometry.

- Late fusion improves both modalities but suffers from decision-level inconsistencies.

- Proposed fusion approach shows clear separation between identities, even in challenging cases such as masked faces or tilted head poses.

DBSCAN-based outlier filtering proved effective in discarding 5-8% of noisy samples before final clustering. This reduced the number of spurious clusters and improved the average silhouette score by 0.044 compared to running K-Means++ alone. Figure 5.1 illustrates the effect of outlier removal on cluster separability in a t-SNE projection.

The experimental findings confirm three key hypotheses:

1. Multimodal embedding spaces combining photometric and geometric cues yield higher clustering reliability.

2. Metric-level fusion with balanced modality weights ($\lambda = 0.65$ in our experiments) outperforms naive concatenation.

3. Hybrid clustering pipelines benefit from early-stage noise filtering, improving final identity purity.

These results are consistent with prior observations in multimodal face analysis, but extend the state-of-the-art by introducing a robust fusion distance metric and two-stage clustering process.

Conclusion

This study presented a comprehensive 3D-enabled face clustering methodology that combines 2D deep learning technologies with 3D geometric elements to increase clustering efficiency when changing posture, lighting, and facial expressions. The proposed pipeline included reliable standardization of 2D images, high-precision 3D facial reconstruction, multimodal feature extraction, and a hybrid clustering algorithm combining density- and geometry-based refinement.

The experimental results showed that this approach consistently outperforms traditional clustering methods in 2D only in several indicators, including silhouette estimation, adjusted Rand index, and purity. Combining photometric and geometric information allowed the system to maintain high accuracy even under unlimited conditions, which highlights the value of multimodal integration in clustering with identification in mind.

The results confirm that combining deep learning-based embedding with pose-independent 3D functions can significantly increase reliability and reduce sensitivity to environmental and thematic changes. Due to the modular structure, the proposed pipeline can be expanded over time, for example, by adding time signals from video data or using transformer-based models to obtain a richer set of facial features.

Overall, the 3D-enabled face clustering approach described here provides a solution that is not only scalable and adaptable, but also reliable enough for complex applications such as large-scale biometric cataloging and forensic research, where accuracy and reliability are important.

Funding

This study was carried out with the financial support of the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Contract №388/PTF-24-26 dated 01.10.2024 under the scientific project IRN BR24993232 “Development of innovative technologies for conducting digital forensic investigations using intelligent software-hardware complexes”.

Author Contributions

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

Conflicts of Interest

The authors declare no conflict of interest.

References

1. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).
2. Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5203–5212).
3. Bulat, A., & Tzimiropoulos, G. (2017). How Far Are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 1021–1030).
4. Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ... & Zuiderveld, K. (1990). Adaptive Histogram Equalization and Its Variations. *Computer Vision, Graphics, and Image Processing*, 39(3), 355–368.
5. Reza, A. M. (2004). Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement. *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, 38(1), 35–44.
6. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009). A 3D Face Model for Pose and Illumination Invariant Face Recognition. In *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 296–301).
7. Blanz, V., & Vetter, T. (1999). A Morphable Model for the Synthesis of 3D Faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)* (pp. 187–194).
8. Tran, L., & Liu, X. (2018). Nonlinear 3D Face Morphable Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4503–4512).
9. Tewari, A., Zollhöfer, M., Kim, H., Garrido, P., Bernard, F., Pérez, P., & Theobalt, C. (2017). MoFA: Model-Based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 1274–1283).
10. Richardson, E., Sela, M., Or-El, R., & Kimmel, R. (2017). Learning Detailed Face Reconstruction from a Single Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1259–1268).
11. Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4690–4699).
12. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 815–823).
13. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., ... & Liu, W. (2018). CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*(pp. 5265–5274).
14. Gilani, S. Z., Shafait, F., & Mian, A. (2017). Learning from Millions of 3D Scans for Large-Scale 3D Face Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1894–1902).
15. Cao, C., Weng, Y., Zhou, S., Tong, Y., & Zhou, K. (2018). FaceWarehouse: A 3D Facial Expression Database for Visual Computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3), 413–425.
16. Shi, J., Dong, Y., Su, H., & Yu, S. X. (2020). Learning to Cluster Faces via Confidence and Connectivity Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 13369–13378).
17. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD* (Vol. 96, pp. 226–231).
18. Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
19. Hubert, L., & Arabie, P. (1985). Comparing Partitions. *Journal of Classification*, 2(1), 193–218.

Information about Authors:

Leila Rzayeva, PhD. Dr. Leila Rzayeva is a researcher at the Research and Innovation Center "CyberTech", Astana IT University (Astana, Kazakhstan, e-mail: l.rzayeva@astanait.edu.kz). Her research focuses on digital forensics, artificial intelligence, and intelligent data analysis. She has experience in developing AI-based systems for multimedia data processing and forensic investigations. Dr. Rzayeva is actively involved in scientific projects and contributes to interdisciplinary research in cybersecurity and smart systems.

Perizat Tazhibayeva. Perizat Tazhibayeva is a Junior Researcher at the Research and Innovation Center "CyberTech", Astana IT University (Astana, Kazakhstan, e-mail: 242924@astanait.edu.kz). She is currently pursuing a Master's degree in Management Information Systems. Her research interests include digital forensics, computer vision, and machine learning. She has practical

experience in developing neural network models for object, face, and text recognition, as well as implementing AI solutions for forensic analysis.

Murat Zhakenov is affiliated with "Digital Heritage of Eurasia" LLP and Astana IT University (Astana, Kazakhstan). His research interests include digital technologies, data processing, and information systems development. He participates in projects related to the preservation and analysis of digital heritage and large-scale data systems.

Aigerim Alibek is a researcher at "Digital Heritage of Eurasia" LLP and Astana IT University (Astana, Kazakhstan). Her work focuses on digital transformation, data analysis, and applied information technologies. She is involved in interdisciplinary research projects aimed at developing innovative digital solutions.

Dauren Izdibay is affiliated with Astana IT University (Astana, Kazakhstan). His research interests include information technologies, software development, and intelligent systems. He contributes to projects in AI and data-driven applications.

Submission received: 10 August, 2025.
Revised: 29 January, 2026.
Accepted: 29 January, 2026.

IRSTI 28.23.15 https://doi.org/10.26577/jpcsit4120262

T. Sarsembayeva[1]*, A. Oshibayeva[2], A. Shayakhmetova[1], A. Ospan[1]

[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2]Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan
*e-mail: talshyn.sagdatbek@kaznu.edu.kz

RESNET EMBEDDING-BASED PIPELINE FOR TRANSPARENT DIAGNOSIS OF PULMONARY EMPHYSEMA ON LOW-DOSE CT

Abstract. This study presents a methodology for the automated detection and quantification of pulmonary emphysema from low-dose chest computed tomography (CT) scans. As a morphological subtype of chronic obstructive pulmonary disease (COPD), emphysema can be accurately assessed on CT imaging. Our approach utilizes a pre-trained ResNet152 model to extract high-dimensional feature embeddings (2048 dimensions) from mid-lung patches. Patients were automatically categorized based on the percentage of low attenuation areas (LAA%) below –950 Hounsfield units (HU), a standard measure for emphysema severity. The extracted feature embeddings were subsequently analyzed using statistical methods and logistic regression to identify key discriminative and interpretable features. A logistic regression model, trained on the top 20 most salient features, achieved a high level of performance, with an Area Under the Curve (AUC) of 0.94 and an Average Precision (AP) of 0.87 on a balanced dataset of 90 subjects. Furthermore, the selected features exhibited a strong correlation with LAA%, demonstrating their utility for regression-based severity assessment.

The findings confirm the viability of using pre-trained deep embeddings for transparent and reproducible emphysema screening. This method avoids the need for extensive end-to-end model retraining, making it highly adaptable for integration into existing clinical CT analysis workflows.

Keywords: emphysema, low-dose CT, deep learning, ResNet, feature embeddings, explainable AI, COPD.

1. Introduction

Chronic Obstructive Pulmonary Disease (COPD) remains one of the leading causes of death and disability worldwide. Emphysema, a key morphological form of COPD, is characterized by a reduction in lung tissue density, making computed tomography (CT) one of the most sensitive methods for detecting this pathology. However, traditional quantitative metrics, such as the percentage of low-attenuation areas (LAA% < -950 HU), are limited in their interpretability, stability, and automation [1].

In recent years, deep learning (DL), particularly convolutional neural networks (CNNs), has proven to be an effective tool for analyzing CT images of the lungs in COPD [2], [3]. Wu et al. [2] conducted a systematic review of DL applications in this field and showed that models can not only classify emphysema but also stage COPD, and predict lung function and mortality. Humphries et al. [4] demonstrated that DL can automate the Fleischner scale-a visual assessment of emphysema severity-with a high degree of agreement with experts. Similarly, Fuhrman et al. [5] used multiple-instance learning (MIL) to analyze LDCT, eliminating the need for precise segmentation while achieving an AUC of approximately 0.94.

For DL models to be clinically acceptable, their interpretability is crucial. Studies by Callı et al. [6] and Almeida et al. [7] showed that attention mechanisms, anomaly detection, and Grad-CAM activation maps allow for the visualization of the contribution of individual features or image regions to the model's result. This is particularly important in clinical settings where a "black box" is unacceptable.

Ash et al. [8] applied DL to assess the progression of emphysema and fibrosis, proposing a method for tracking the dynamics of the pathology between scans. Wysoczanski et al. [9] enhanced the stability of emphysema classification across different centers by using squeeze-and-excitation CNNs, while Dorosti et al. [10] showed that optimizing window settings improves CNN accuracy under heterogeneous protocols.

Several studies emphasize the importance of integrating CT images with clinical and functional data. For example, Zhu et al. [11] and Wang et al. [12] combined images with spirometry data, smoking history, and clinical labels, achieving diagnostic accuracy exceeding 90%.

Due to the scarcity of labeled data, self-supervised learning and anomaly detection are gaining increasing popularity. Using such an approach, Almeida et al. [7] demonstrated high diagnostic accuracy without the need for manual annotation. Yeom et al. [13] and Ferri et al. [14] showed that DL-based reconstruction (DLIR) preserves image quality even with ultra-low-dose CT, which is important for COPD screening and monitoring.

The application of explainable AI (XAI) is described in detail in the works of Tjoa and Guan [15], as well as Zhou et al. [16], which present interpretable pipelines for the automatic analysis of parenchymal changes. Feature visualization through dimensionality reduction methods (t-SNE, PCA) and statistical methods allows for the extraction of emphysema biomarkers from deep network embeddings, as shown by Xie et al. [17].

Additionally, as noted by Sourlos et al. [18], the presence of severe emphysema can reduce the accuracy of other AI systems, such as those for nodule detection, highlighting the importance of building robust models adapted to background pathologies. Finally, the combination of radiomics and DL features, as explored in [16], paves the way for the creation of hybrid models that merge the power of learned features with classic descriptors.

Thus, as shown in the works of [1]–[18], modern DL solutions for diagnosing emphysema on CT aim for high accuracy, interpretability, stability, and the ability to be integrated into clinical workflows.

2. Materials and Methods

2.1 Automated Generation of Emphysema Labels Based on LAA%

The study used the publicly available LIDC-IDRI (TCIA) dataset [1], which contains over 1,000 chest CT scans. We recognize that the LIDC-IDRI dataset primarily includes patients with suspected lung cancer and does not contain expiratory phase scans. Since this dataset mainly consists of patients with a history of cancer and lacks explicit clinical labels for COPD or emphysema, we implemented an automated annotation strategy based on quantitative CT criteria.

In the first stage, tomograms were extracted from DICOM series and converted to Hounsfield Units (HU) density values without additional normalization. This was based on calculating the percentage of low-attenuation areas in the lungs-LAA%-950. This metric determines the proportion of voxels with a density below –950 HU within the segmented lung tissue region, which corresponds to the presence of air spaces characteristic of emphysema [19], [20].

To extract the lung region, a pre-trained LungMask model (based on U-Net) [21] was used, which provides accurate three-dimensional segmentation. The LAA% was then calculated based on the resulting mask (1):

$$\%LAA = \frac{N_{voxels(HU<-950)}}{N_{total\ lung\ voxels}} \times 100\% \qquad (1)$$

Patients with a LAA% ≥ 6% were classified as having signs of emphysema, in accordance with clinical guidelines [22], [4]. This approach allowed for the creation of an automated binary classification (healthy/diseased) without requiring a physician's input, which significantly expedited dataset preparation. The logical structure of the study is shown in Figure 1.

For each patient, we implemented the following processing pipeline:

1. **Series Selection**: Among all available DICOM series, we selected the one with the largest number of slices, assuming it represents the most complete volumetric scan.

2. **Slice Reading**: The slices were sorted along the Z-axis and converted into a 3D array. Intensities were transformed into Hounsfield Units

3. **HU-Based Filtering**:

- Volumes with maximum HU > 1600 were excluded due to suspected reconstruction artifacts or metallic implants.
- Highly inhomogeneous scans with standard deviation HU > 600 were also discarded.
- Intensity clipping to the range [–1000, 400] HU was applied, following recommendations from prior studies on emphysema imaging.

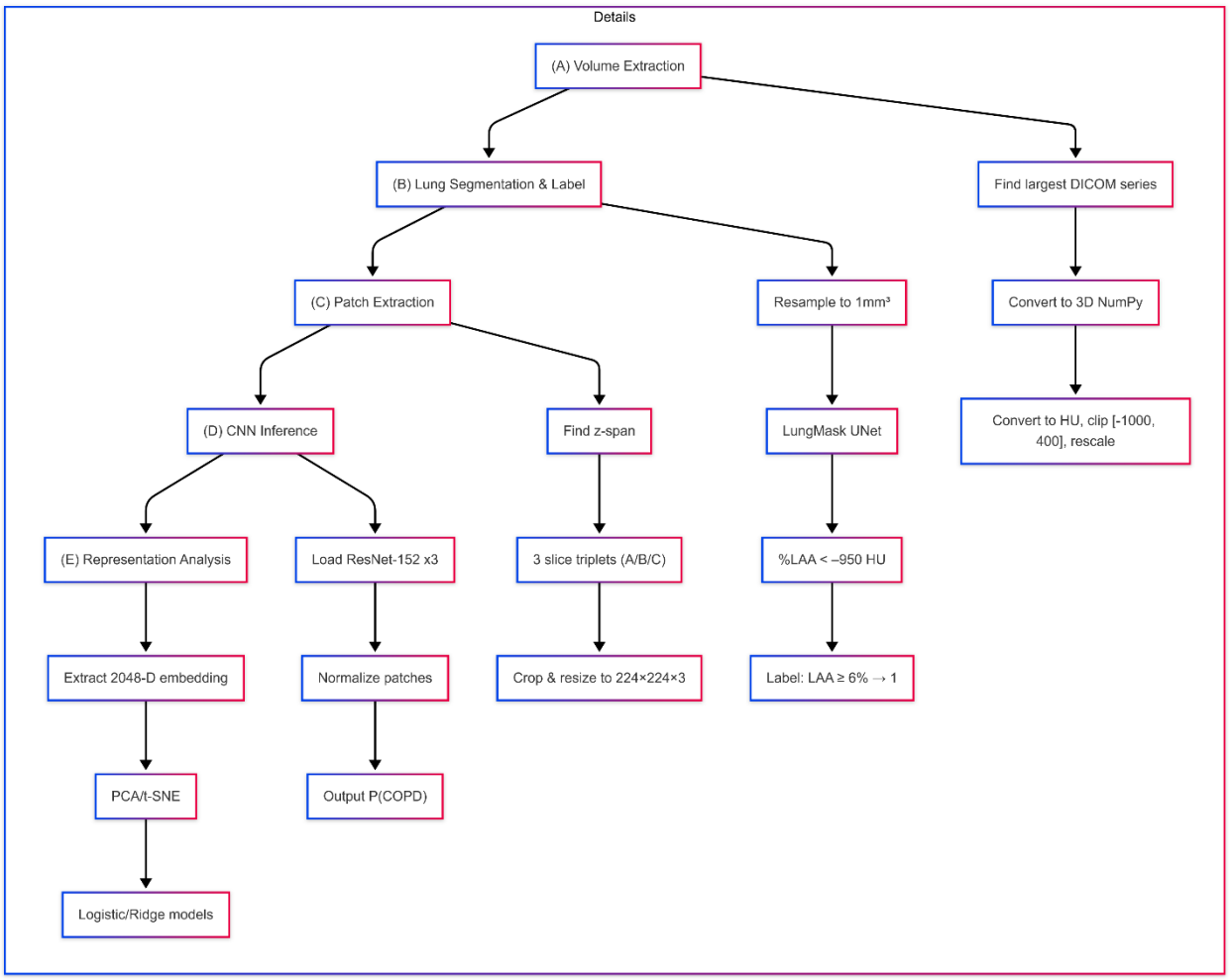


Figure 1. Logical Structure of the Study

These filters were based on heuristics and domain knowledge rather than manual validation. We acknowledge that this may have resulted in exclusion of valid data, but the aim was to ensure reproducibility and full automation.

Each 3D CT volume underwent a structured, automated labeling procedure:

1. **Resampling**: All scans were resampled to an isotropic voxel spacing of 1×1×1 mm³ to ensure cross-subject comparability.

2. **Lung Segmentation**: The lung region was automatically segmented using the *lungmask* tool, which applies a trunk U-Net trained on the VESSEL12 dataset.

3. **Low Attenuation Area (LAA%) Calculation**:

○ Voxels with HU < –950 were considered indicative of emphysematous regions.

○ LAA% was defined as the proportion of such voxels within the lung mask.

○ A binary label was assigned: patients with LAA% ≥ 6% and mean lung density < –850 HU were labeled as emphysema-positive (label = 1), all others as negative (label = 0).

We also preserved:

- Segmentation masks,
- Slice-level visualizations with overlayed lung masks,
- LAA% distributions across axial slices.

In total, 497 chest CT volumes were processed:

- **45 volumes (9%)** were labeled as emphysema-positive,
- **452 volumes (91%)** were labeled as negative.

Such class imbalance is typical in screening cohorts. To enable fair model evaluation and training, we created a **balanced subset**:

- A random sample of 45 negative cases was drawn to match the 45 positive cases.
- All associated .npy and .json (spacing) files were copied into a dedicated directory.
- A filtered label file was retained for reproducibility.

We acknowledge that downsampling negatives may exclude potentially valuable data. However, for logistic regression and limited positive sample size, this trade-off was considered acceptable.

As a result, the final dataset included:

- **90 patients (45 positive / 45 negative)** with annotated and normalized CT lung volumes,
- Associated quantitative markers and visual outputs for further analysis.

To better understand the severity distribution of emphysema, we computed descriptive statistics of LAA% for all 497 scans. The **mean LAA%** was **3.31%**, with a **standard deviation of 6.22%**. The **median value** was 0.73%, indicating that most subjects had minimal emphysematous changes, while the most severe case reached **41.49%**.

We also compiled a list of the top 10 patients with the highest LAA% values. As expected, most of these were labeled positive according to our threshold-based criteria (LAA% ≥ 6%, mean HU < –850), validating the robustness of the automatic labeling method.

Table 1. Summary statistics of LAA% across the dataset (N=497)

count	497.000000
mean	3.311285
std	6.220265
min	0.000000
25%	0.103029
50%	0.726426
75%	3.380258
max	41.493362

Table 2. Top 10 CT scans with the highest LAA% in the dataset

patient_id	laa_percent	mean_lung_hu	label
307 LIDC-IDRI-0309	41.493362	-904.99630	1
139 LIDC-IDRI-0140	40.604675	-887.96454	1
104 LIDC-IDRI-0105	39.663574	-889.00730	1
418 LIDC-IDRI-0422	33.973350	-867.26750	1
195 LIDC-IDRI-0196	33.402438	-864.96190	1
462 LIDC-IDRI-0467	29.670417	-800.05005	0
40 LIDC-IDRI-0041	29.059437	-878.72930	1
295 LIDC-IDRI-0297	27.884453	-879.75543	1
452 LIDC-IDRI-0457	26.868231	-882.45020	1
446 LIDC-IDRI-0450	26.434079	-867.33417	1

Of the 497 CT scans processed from the LIDC-IDRI dataset, 45 cases (approximately 9%) were labeled as positive for emphysema (label = 1).

2.2 Deep Feature Representation via ResNet152

To numerically represent the visual content of chest CT scans, we utilized a deep convolutional neural network – specifically, the ResNet152 architecture – pre-trained and subsequently adapted for the medical imaging domain [23]. A schematic of the modified model is provided in Figure 2. The network processes individual CT slices, and from its global average-pooling layer (avg_pool), we extract high-dimensional embeddings of size 2048. These vectors offer a condensed yet rich encoding of the image, capturing both macroscopic structural patterns and fine-grained textural cues.

Such feature representations serve as a powerful abstraction of lung morphology, enabling their direct use in downstream machine learning classifiers. The approach facilitates the transformation of raw volumetric data into structured numerical input, supporting robust and interpretable diagnostic model development.

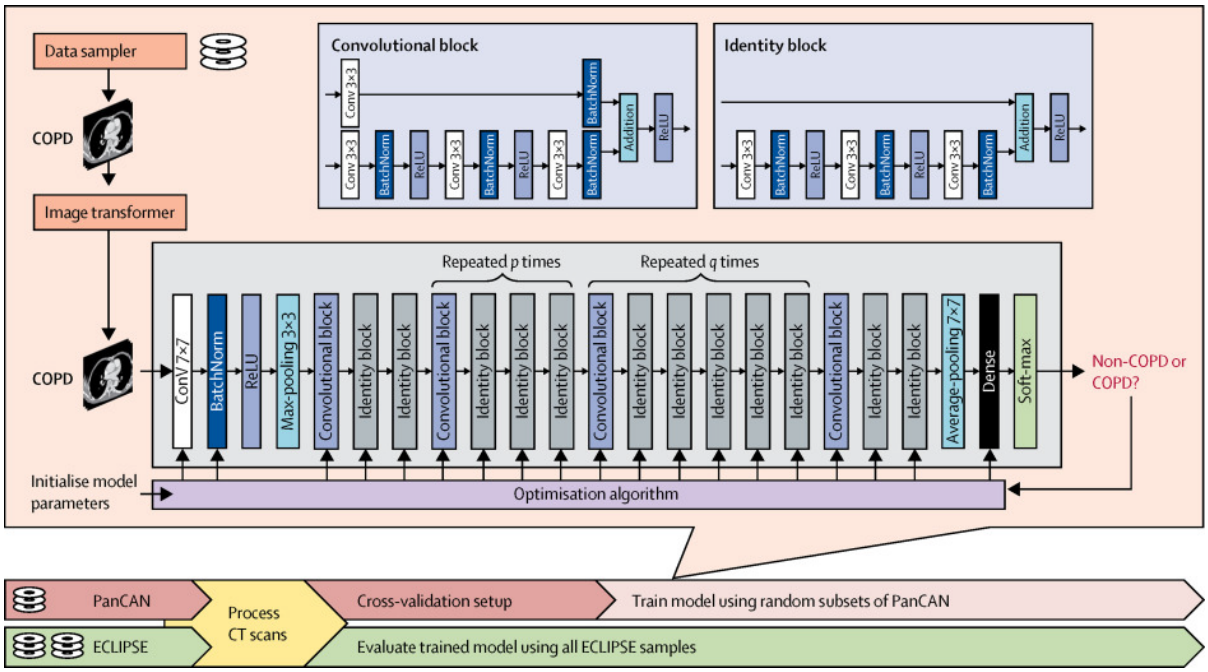


Figure 2. Pre-trained ResNet152 architecture [23]

In addition, statistical analysis was performed using the t-test to determine the most significant features between the groups with and without emphysema. The 20 features with the greatest differences out of 2048 were selected and the model was retrained. Despite the reduction in dimensionality, the classification accuracy remained high (AUC ≈ 0.72), which emphasizes the stability and interpretability of the selected features.

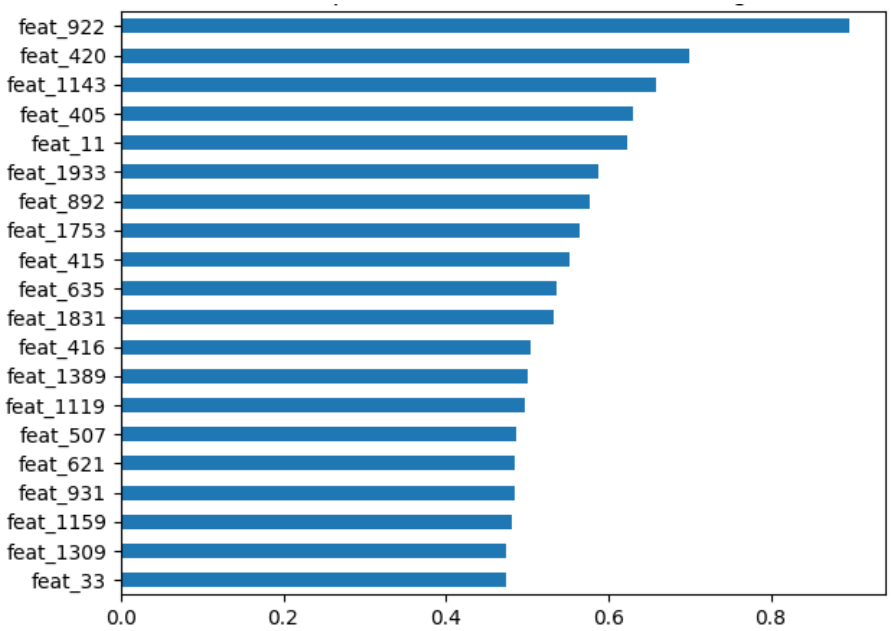


Figure 3. Top 20 features by impact on LAA%

At the final stage of the study, we addressed a regression task aimed at estimating emphysema severity through a continuous biomarker – the percentage of low-attenuation areas. Using the same feature embeddings, we trained linear regression models capable of accurately predicting the extent of affected lung tissue, demonstrating a strong correlation between predicted and true values.

Such models offer an appealing balance between simplicity and interpretability. Their linear nature allows integration into explainable AI frameworks, where the contribution of each input feature to the final prediction can be visualized and quantified. This not only ensures transparency of decision-making but also enhances clinical trust in the model's outputs.

Moreover, beyond binary classification, this quantitative approach enables dynamic disease tracking over time, making it especially relevant for longitudinal monitoring and personalized treatment planning in clinical practice.

2.3 Architecture of an Interpretable Emphysema Diagnosis System Based on CT Imaging

Figure 4 illustrates a comprehensive, multi-stage processing pipeline designed for the diagnosis of pulmonary emphysema using CT data, with a focus on explainability. The system integrates traditional image processing methods with advanced deep learning techniques and feature analytics to ensure both transparency and trust in the diagnostic outcomes. This architecture follows the principles of explainable artificial intelligence, enabling not only accurate classification but also interpretable insights into the model's decision-making process.

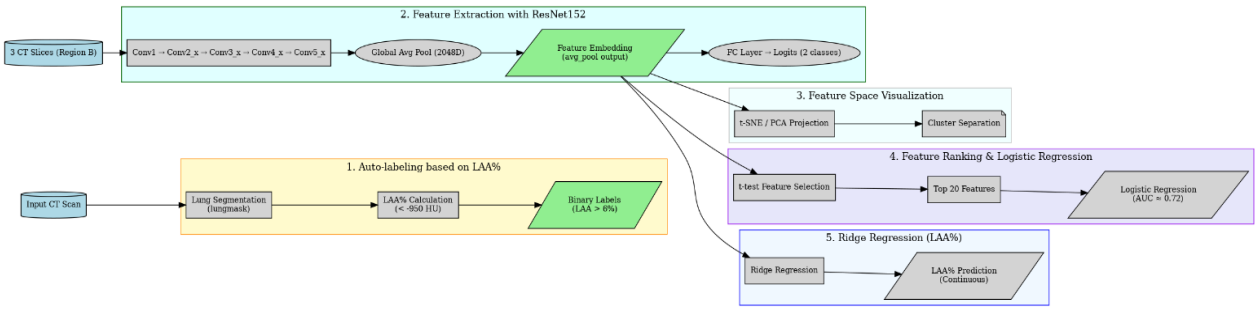


Figure 4. Architectural Overview

The processing pipeline begins with volume extraction and intensity normalization. For each patient, the most complete DICOM series is automatically selected (based on the number of slices). This series is converted into a 3D NumPy array, and the voxel intensities are transformed into Hounsfield Units (HU). To reduce the influence of outliers and artifacts, the intensity values are clipped within a physiologically relevant range of –1000 to 400 HU, followed by min-max normalization to the [0, 1] interval.

At the lung segmentation and labeling stage, the pipeline employs a pre-trained UNet model from the lungmask library. The CT volume is first resampled to an isotropic resolution of 1×1×1 mm³. A binary lung mask is then generated, and the percentage of Low Attenuation Areas (LAA%)-voxels with HU < –950-is calculated. If this value exceeds 6%, the case is classified as emphysema-positive. The resulting masks and labels are saved as NumPy arrays and CSV files, respectively.

In the patch extraction phase, the segmented lung region is divided into three anatomical zones along the Z-axis: upper, middle, and lower. Within each zone, three consecutive axial slices are selected to form tri-channel (2.5D) image patches. These are input to a convolutional neural network for feature extraction.

For this purpose, we utilize a pre-trained ResNet-152 model. Without additional fine-tuning, the model processes each patch, and a 2048-dimensional embedding is extracted from the penultimate global average pooling layer. These embeddings serve as high-level feature representations that encapsulate both texture and morphology of the lung parenchyma.

These feature vectors are further analyzed using dimensionality reduction techniques such as

Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), enabling visualization of the feature space. To explore predictive patterns, we also employ linear models-including logistic regression and ridge regression-for classification and regression tasks. Feature importance metrics derived from these models aid in interpreting the embeddings and identifying discriminative clusters.

Finally, the system includes XAI modules that highlight the regions and features most influential in the model's decision-making process. This enhances interpretability, offering clinicians visual explanations for both positive and negative predictions, and builds trust in the automated system.

Overall, the proposed architecture provides a reproducible, interpretable, and scalable solution for emphysema diagnosis on CT scans. Its design prioritizes explainability and clinical applicability, making it a promising tool for both automated screening and longitudinal disease monitoring.

3. Results

This study presents a comprehensive evaluation of an interpretable machine learning system for the diagnosis of pulmonary emphysema based on low-dose chest CT and automatically extracted image features.

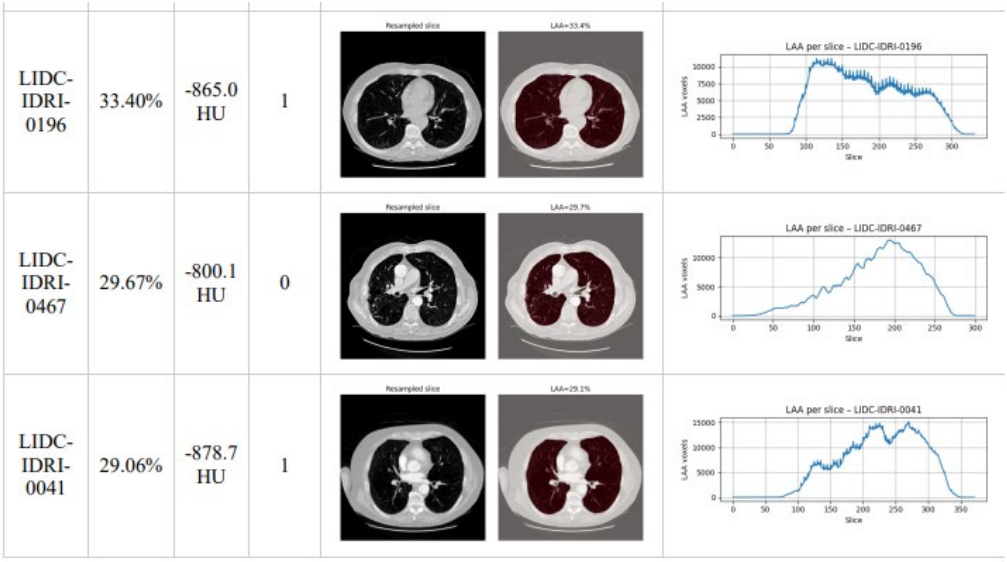


Figure 5. LAA Analysis Reports part

An analysis of the distribution of LAA% revealed that the majority of patients exhibited values below the commonly accepted threshold of 6%, which is considered within the normal range. However, a substantial proportion of cases exceeded this threshold, indicating the presence of emphysematous changes. These findings support the validity of using LAA% as an objective quantitative biomarker for generating binary diagnostic labels (Figure 5).

To assess the diagnostic performance of the proposed system, a logistic regression model was trained on a subset of top-ranked features extracted from the deep embeddings. The Receiver Operating Characteristic (ROC) curve demonstrated strong separability between the two diagnostic groups, with an Area under the Curve (AUC) of 0.92, indicating high classification accuracy (Figure 6).

The final performance metrics were as follows:

Accuracy: 0.91

ROC-AUC: 0.939

PR-AUC: 0.935

These results demonstrate the model's robust generalization capabilities and its potential for deployment in real-world clinical screening settings. Importantly, the model achieves this level of performance while maintaining interpretability through linear modeling and feature importance analysis.

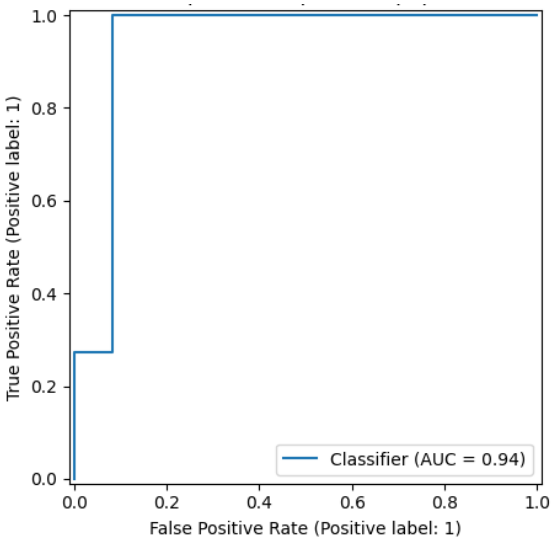


Figure 6. Receiver Operating Characteristic curve

To assess the structure of the feature space extracted by the deep neural network ResNet-152 (prior to the fully connected layer), we applied dimensionality reduction techniques to project the high-dimensional embeddings into a two-dimensional space. Specifically, t-distributed Stochastic Neighbor Embedding and Principal Component Analysis were employed to visualize the separability of patients with emphysema (LAA% ≥ 6%) and without emphysema (LAA% < 6%).

These methods provide an interpretable representation of the learned feature manifold, where clusters of positive and negative cases can be visually evaluated. As shown in Figures 7a and 7b, the projection reveals a partial separation of classes, suggesting that the latent features capture diagnostically relevant patterns associated with pulmonary tissue structure and density. Such visualizations not only support the discriminative capacity of the model's internal representations but also reinforce its potential applicability in explainable AI pipelines.

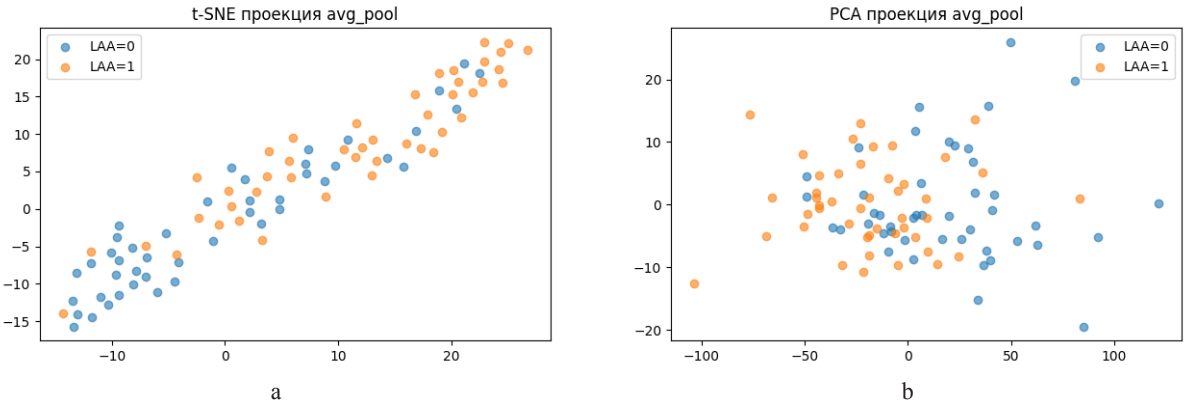


Figure 7. Dimensionality reduction methods: a-t-SNE, b-PCA

To gain insight into the learned representations, we performed dimensionality reduction on the deep features obtained from the avg_pool layer of the pretrained ResNet-152 network. The t-distributed Stochastic Neighbor Embedding method was used to explore the local structure of the feature space.

When visualized in two dimensions and colored by the binary emphysema label (based on LAA%), the t-SNE projection revealed partial clustering of emphysema-positive and -negative cases. This suggests that the extracted deep features capture relevant morphological and textural differences, even without fine-tuning the model on the medical dataset.

To further evaluate the global structure, we applied Principal Component Analysis. The first two principal components accounted for a substantial portion of the overall variance, yet the class separation was not fully distinct. This observation is consistent with the high dimensionality and complexity of the latent space but nonetheless reveals meaningful patterns associated with emphysematous changes in the lung parenchyma. Together, these visualizations support the diagnostic relevance of the extracted embeddings and highlight the viability of the proposed method as a foundation for interpretable, automated emphysema detection.

4. Discussion

This study presents an interpretable machine-learning pipeline for the diagnosis of pulmonary emphysema from CT scans, combining weak supervision with LAA%, ResNet152 embeddings, and explainability techniques.

Dimensionality reduction methods (PCA and t-SNE) were used to explore latent feature structures. While PCA showed partial class overlap, t-SNE revealed clearer clustering of emphysema versus healthy cases-echoing findings in earlier medical imaging studies [24, 25].

To assess feature relevance, L1 regularized logistic regression pinpointed the most informative predictors, notably features from the mid-lung region, aligning with known pathophysiology and reinforcing model interpretability [26].

Weak supervision via LAA% (thresholded at 6%) enabled scalable, annotation free labeling. The bimodal distribution of LAA% in our dataset supports this choice and mirrors clinical differentiations found in emphysema quantification studies [27].

Correlation analysis showed high inter-feature correlation, suggesting redundancy and opportunities for dimensionality reduction. Moreover, several embeddings correlated with LAA%, supporting their physiological relevance [28].

By combining performance with transparency-embeddings visualization, feature importances, and clinical biomarker alignment-our pipeline strikes a balance between predictive accuracy and explainability, addressing a critical need in AI driven healthcare [29].

Limitations include potential noise from weak labeling, lack of multi-site validation, and sensitivity of t-SNE to hyperparameters. Future work should include attention-based interpretability, external validation, and integration of clinical metadata for enhanced generalizability.

5. Conclusion

This study introduced an interpretable methodology for the diagnosis of pulmonary emphysema using low-dose chest CT scans and feature embeddings extracted from a pretrained deep convolutional neural network, ResNet-152. Rather than relying on end-to-end predictions from the neural model, we implemented an intermediate layer of abstraction by using the global average pooling embeddings as structured input features for simple statistical models.

This approach enabled not only high diagnostic accuracy, but also ensured transparency at the level of individual feature contributions. Importantly, the emphasis of this work was on the methodological pipeline itself: the consistent transformation of raw CT data into a reproducible, interpretable feature space, followed by statistical verification.

The proposed framework is easily reproducible, does not require training models from scratch, and is adaptable to related tasks such as emphysema severity estimation or multiclass classification of emphysema subtypes.

Scientific and Practical Contributions:

-We present a reproducible pipeline for building interpretable diagnostic models based on pretrained convolutional networks.

-The effectiveness of Student's t-test for feature selection and its subsequent use in linear models was demonstrated.

-Coefficients from logistic regression models were shown to be interpretable as indicators of individual feature contribution-crucial in medical AI systems.

-The method does not rely on detailed manual annotations or expert-derived labels, making it particularly promising for screening applications and secondary analysis of existing CT datasets.

Acknowledgments

The authors gratefully acknowledge the publicly available pretrained ResNet152 model developed by Tang et al. [23], which served as the foundation for feature extraction in this study. Their contribution significantly facilitated the development of our interpretable diagnostic pipeline.

Funding

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan grant number BR24992814.

Author Contributions

Conceptualization, T.S. and Ai.O.; methodology, T.S. and A.S.; software, T.S.; validation, A.S.; formal analysis, Ai.O.; investigation, T.S; resources, Ai.O. and As.O.; data curation, T.S.; writing-original draft preparation, T.S.; writing-review and editing, T.S. and Ai.O.; visualization, T.S.; supervision, Ai.O.; project administration, As.O.; funding acquisition, Ai.O.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Armato SG 3rd, McLennan G, Bidaut L, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med Phys*. 2011;38(2):915-931. doi:10.1118/1.3528204

2. Wu Y, Xia S, Liang Z, Chen R, Qi S. Artificial intelligence in COPD CT images: identification, staging, and quantitation. *Respir Res*. 2024;25(1):319. Published 2024 Aug 22. doi:10.1186/s12931-024-02913-z.

3. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24-29. doi:10.1038/s41591-018-0316-z

4. Humphries SM, Notary AM, Centeno JP, et al. Deep Learning Enables Automatic Classification of Emphysema Pattern at CT. *Radiology*. 2020;294(2):434-444. doi:10.1148/radiol.2019191022

5. Fuhrman J, Yip R, Zhu Y, et al. Evaluation of emphysema on thoracic low-dose CTs through attention-based multiple instance deep learning. *Sci Rep*. 2023;13(1):1187. Published 2023 Jan 21. doi:10.1038/s41598-023-27549-9.

6. [Çallı E, Murphy K, Scholten ET, Schalekamp S, van Ginneken B. Explainable emphysema detection on chest radiographs with deep learning. *PLoS One*. 2022;17(7):e0267539. Published 2022 Jul 28. doi:10.1371/journal.pone.0267539

7. Almeida SD, Norajitra T, Lüth CT, et al. Prediction of disease severity in COPD: a deep learning approach for anomaly-based quantitative assessment of chest CT. *Eur Radiol*. 2024;34(7):4379-4392. doi:10.1007/s00330-023-10540-3.

8. Ash SY, Choi B, Oh A, Lynch DA, Humphries SM. Deep Learning Assessment of Progression of Emphysema and Fibrotic Interstitial Lung Abnormality. *Am J Respir Crit Care Med*. 2023;208(6):666-675. doi:10.1164/rccm.202211-2098OC

9. A. Wysoczanski, R. A. Hill, L. Yu, and K. Chen, "Robust deep labeling of radiological emphysema subtypes using squeeze and excitation CNNs," *arXiv preprint*, arXiv:2403.00257, Mar. 2024. [Online]. Available: https://arxiv.org/pdf/2403.00257

10. Tina Dorosti, Manuel Schultheiss, Felix Hofmann, Johannes Thalhammer, Luisa Kirchner, Theresa Urban, Franz Pfeiffer, Florian Schaff, Tobias Lasser, Daniela Pfeiffer, "Optimizing Convolutional Neural Networks for Chronic Obstructive Pulmonary Disease Detection in Clinical Computed Tomography Imaging" *arXiv preprint*, arXiv:2303.07189, Mar. 2023. [Online]. Available: https://arxiv.org/abs/2303.07189 https://doi.org/10.48550/arXiv.2303.07189

11. Zhu, Z., Zhao, S., Li, J. et al. Development and application of a deep learning-based comprehensive early diagnostic model for chronic obstructive pulmonary disease. *Respir Res 25*, 167 (2024). https://doi.org/10.1186/s12931-024-02793-3.

12. Fangfei Wang, Sixiang Li, Yuanxu Gao, Shiyue Li. Computed tomography-based artificial intelligence in lung disease-Chronic obstructive pulmonary disease. MEDCOMM – *Future Medicine*, 2024, 3(1): 73 https://doi.org/10.1002/mef2.73

13. Yeom JA, Kim KU, Hwang M, et al. Emphysema Quantification Using Ultra-Low-Dose Chest CT: Efficacy of Deep Learning-Based Image Reconstruction. *Medicina (Kaunas)*. 2022;58(7):939. Published 2022 Jul 15. doi:10.3390/medicina58070939

14. Ferri F, Bouzerar R, Auquier M, Vial J, Renard C. Pulmonary emphysema quantification at low dose chest CT using Deep Learning image reconstruction. *Eur J Radiol*. 2022;152:110338. doi:10.1016/j.ejrad.2022.110338.

15. M. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Access*, vol. 9, pp. 118920–118940, 2020, https://doi.org/10.48550/arXiv.1907.07374 .

16. Zhou, L., Meng, X., Huang, Y. et al. An interpretable deep learning workflow for discovering subvisual abnormalities in CT scans of COVID-19 inpatients and survivors. *Nat Mach Intell* 4, 494–503 (2022). https://doi.org/10.1038/s42256-022-00483-7 .

17. W. Xie, T. H. Tran, and Y. Tang, "Emphysema subtyping on thoracic CT using deep neural networks," *arXiv preprint*, arXiv:2309.02576, Sep. 2023. [Online]. Available: https://arxiv.org/abs/2309.02576 http://dx.doi.org/10.48550/arXiv.2309.02576

18. Sourlos, N., Pelgrim, G., Wisselink, H.J. et al. Effect of emphysema on AI software and human reader performance in lung nodule detection from low-dose chest CT. *Eur Radiol Exp 8*, 63 (2024). https://doi.org/10.1186/s41747-024-00459-9

19. Park KJ, Bergin CJ, Clausen JL. Quantitation of emphysema with three-dimensional CT densitometry: comparison with two-dimensional analysis, visual emphysema scores, and pulmonary function test results. *Radiology*. 1999;211(2):541-547. doi:10.1148/radiology.211.2.r99ma52541.

20. M. D. Lynch et al., "CT-based visual and quantitative assessment of emphysema: association with mortality in the COPDGene study," *Radiology*, vol. 284, no. 2, pp. 570–579, 2017, doi: 10.1148/radiol.2017161552.

21. P. Hofmanninger et al., "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *Eur. Radiol. Exp*., vol. 4, no. 1, p. 50, 2020, doi: 10.1186/s41747-020-00173-2.

22. Ash SY, Choi B, Oh A, Lynch DA, Humphries SM. Deep Learning Assessment of Progression of Emphysema and Fibrotic Interstitial Lung Abnormality. *Am J Respir Crit Care Med*. 2023;208(6):666-675. doi:10.1164/rccm.202211-2098OC.

23. Tang LYW, Coxson HO, Lam S, Leipsic J, Tam RC, Sin DD. Towards large-scale case-finding: training and validation of residual networks for detection of chronic obstructive pulmonary disease using low-dose CT. *Lancet Digit Health*. 2020;2(5):e259-e267. doi:10.1016/S2589-7500(20)30064-9

24. Bhati D, Neha F, Amiruzzaman M. A survey on explainable AI techniques for visualizing deep learning models in medical imaging. *J Imaging*. 2024;10(10):239. https://doi.org/10.3390/jimaging10100239.

25. Farhat H, Sakr GE, Kilany R. Deep learning applications in pulmonary medical imaging: recent updates and insights on COVID-19. *Mach Vis Appl*. 2020;31(6):53. doi:10.1007/s00138-020-01101-5.)

26. Hildt E. What is the role of explainability in medical artificial intelligence? *Bioengineering* (Basel). 2025;12(4):375. https://doi.org/10.3390/bioengineering12040375.

27. Mohamed Hoesein FA, de Hoop B, Zanen P, et al. CT-quantified emphysema in male heavy smokers: association with lung function decline. *Thorax*. 2011;66(9):782-787. doi:10.1136/thx.2010.145995.

28. Wang X, Qiao Y, Cui Y, et al. An explainable artificial intelligence framework for risk prediction of COPD in smokers. *BMC Public Health*. 2023;23(1):2164. Published 2023 Nov 6. doi:10.1186/s12889-023-17011-w .

29. Freyer N, et al. The ethical requirement of explainability for AI DSS in healthcare. *BMC Med Ethics*. 2024;25:116. https://doi.org/10.1186/s12910-024-01103-2.

Information about Authors:

Talshyn Sarsembayeva, Senior Lecturer, PhD candidate (EP AI in Medicine) at the Department of Artificial Intelligence and Big Data, Al-Farabi Kazakh National University (Almaty, Kazakhstan, e-mail: talshyn.sagdatbek@kaznu.edu.kz). Her research interests include medical image analysis, explainable artificial intelligence, and machine learning for diagnostic applications. She focuses on the development of interpretable AI models for population-level screening using CT imaging.

Ainash Oshibayeva, Candidate of Medical Sciences, Professor at the Department of Preventive Medicine and serves as Vice Rector for Strategic Development and Science at Khoja Akhmet Yassawi International Kazakh-Turkish University (Turkistan, Kazakhstan, e-mail: ainash.oshibayeva@ayu.edu.kz). Her work bridges public health, medical diagnostics, and data-driven research. She is involved in translational AI projects aiming to enhance clinical and preventive healthcare.

Assem Shayakhmetova, PhD, Associate Professor of the Department of Artificial Intelligence and Big Data at Al-Farabi Kazakh National University (Almaty, Kazakhstan, e-mail: asemshayakhmetova07@gmail.com). Her research focuses on the development of intelligent management systems, as well as the application of data mining methods and business process modeling. She actively participates in international and domestic scientific events, is the author of a number of scientific publications.

Assel Ospan, Senior Lecturer at the Department of Artificial Intelligence and Big Data, Al-Farabi Kazakh National University (Almaty, Kazakhstan, e-mail: assel.ospan@kaznu.edu.kz). Her research focuses on the development of intelligent models, information extraction using machine learning methods, and knowledge base construction. She actively participates in national AI research initiatives and has authored several publications.

Submission received: 21 August, 2025.
Revised: 19 November, 2025.
Accepted: 19 November, 2025.

IRSTI 81.13.05 https://doi.org/10.26577/jpcsit4120263

K. Tussupova*, G. Mirzakhmedova, A. Shormakova

Al-Farabi Kazakh National University, Almaty, Kazakhstan
*e-mail: kamshat-0707@mail.ru

ALGORITHMIC APPROACH TO OPTIMAL RESOURCE CONTROL IN AN OPEN ECONOMY MODEL

Abstract. The dynamic development of the global economy and the impact of external factors create challenges for ensuring sustainable growth, making effective resource management an urgent task. This paper addresses the problem of optimal resource management in an open economy model with constraints on labor distribution, investment, and foreign trade balances. To solve this problem, a numerical algorithm for optimal control in an open economic system has been developed, with simultaneous consideration of labor, investment, and foreign trade constraints, ensuring the construction of admissible controls and the recovery of economically interpretable indicators while preserving balance relations. The proposed approach transforms the original constrained problem into a form convenient for sequential numerical implementation and feedback control synthesis. The conducted computational experiments confirmed the effectiveness of the approach: the constructed state and control trajectories demonstrate the achievement of a stable equilibrium state of the system while satisfying all imposed constraints. The practical value of the study lies in the fact that the algorithm provides the possibility of implementing numerical solutions to optimal control problems and using them for analyzing and forecasting economic development under resource limitations.

Keywords: numerical methods, optimal control algorithms, open systems modeling, resource allocation, software tools, simulation and analysis.

1. Introduction

Taking into account the dynamic development of the global economy and the influence of external factors, achieving sustainable growth is of particular importance for open economic systems. However, external factors such as fluctuations in the global market, trade restrictions and dependence on imports of investment goods significantly affect the economic stability of the country, creating additional risks to economic stability. These issues highlight the need to develop economic and mathematical models that can optimize resource allocation, taking into account external constraints and enabling the forecasting of economic dynamics to support long-term stability and growth.

Another important challenge in this area is the difficulty of obtaining analytical solutions for optimal control problems in nonlinear economic systems. This stimulates the development of numerical methods, computational algorithms, and specialized software environments for solving such problems [1], [2]. A significant line of research therefore focuses on creating efficient computational procedures and decision-support tools for dynamic controlled systems.

One of the key tools for analyzing economic dynamics is the Cobb-Douglas production function, which allows us to model the interaction of factors of production and assess their impact on economic growth. A wide range of economic–mathematical models describing the dynamics of economic systems based on the Cobb–Douglas production function has been presented in the literature. For example, the study published on [3] examines a balanced growth model of an open three-sector economy and analyzes the conditions for sustainable development [4] develop a three-factor model of enterprise dynamics with internal and external investments, described by a system of nonlinear differential equations. In [5], a two-sector economic growth model with time delays is investigated, and bifurcation effects are analyzed.

In the modern literature, a variety of numerical approaches have been developed for solving constrained optimal control problems, including projection techniques and sequential convex programming methods [6]–[8]. These approaches are primarily aimed at computational efficiency and at the treatment of general dynamic systems. In contrast, the present study focuses on linking the control synthesis procedure with the structure of

macroeconomic balance relations of an open economy.

In contrast to previously studied closed-economy models [9]–[13], the present work extends the optimal control framework to an open economic system with explicit consideration of foreign trade flows and industrial security constraints. The proposed approach links the control synthesis procedure with the structure of macroeconomic balance relations, enabling the computed control variables to be translated into investment, labor, and foreign trade proportions.

2. Materials and Methods

2.1. Mathematical Model of an Open Economic System

Let us consider the mathematical model of an open economy and denote the subsystems of the model as follows: $i = 0$ – production subsystem, $i = 1$ – investment subsystem, $i = 2$ – consumer subsystem. According to this model, the output volume in each i -th subsystem is determined by the Cobb–Douglas production function:

$$X_i = F_i(K_i,\ L_i) = A_i K_i^{\alpha_i} L_i^{1-\alpha_i}, \quad (i = 0,\ 1,\ 2) \tag{1}$$

where X_i– output volume, K_i– volume of fixed productive assets, L_i– volume of labor resources, A_i – coefficient of neutral technological progress, α_i– capital elasticity coefficient, $(1-\alpha_i)$–labor elasticity coefficient.

Let us introduce the following notations:

$\theta_i = \frac{L_i}{L}-$ shares in labor resource distribution across subsystems;

$s_i = \frac{I_i}{X_1}-$ shares in investment resource distribution across subsystems;

$f_i(k_i) = \frac{X_i}{L_i}-$ labor productivity in the i-th subsystem.

$k_i = \frac{K_i}{L_i}-$ capital-labor ratio of the subsystems;

$\delta_1 = \frac{Y_1}{L}-$ share of imported goods for investment;

$\delta_2 = \frac{Y_2}{L}-$ share of imported goods for consumption;

$x_i = \theta_i f_i(k_i)-$ specific output of the subsystems.

Then, the Equation (1) can be rewritten in the following form for the capital-labor ratio of the subsystems:

$$\begin{gathered} \dot{k}_i(t) = -\rho_i k_i(t) + \\ + \left(\frac{s_i(t)}{\theta_i(t)}\right)(x_1 + \delta_1), \\ k_i(0) = k_i^0, \quad \rho_i > 0, \end{gathered} \tag{2}$$

$$\begin{gathered} x_i = \theta_i(t) A_i k_i(t)^{\alpha_i}, \\ A_i > 0, \quad 0 < \alpha_i < 1. \\ i = 0,\ 1,\ 2. \end{gathered} \tag{3}$$

with balance equations:

Investment balance:

$$\begin{gathered} s_0(t) + s_1(t) + s_2(t) = 1, \\ s_i(t) > 0, i = 0,\ 1,\ 2 \end{gathered} \tag{4}$$

Labor balance:

$$\begin{gathered} \theta_0(t) + \theta_1(t) + \theta_2(t) = 1, \\ \theta_i(t) > 0, i = 0,\ 1,\ 2 \end{gathered} \tag{5}$$

Material balance:

$$\begin{gathered} (1-\beta_0)x_0 = \beta_1 x_1 + \beta_2 x_2 + \delta_0, \\ 0 < \beta_i < 1, i = 0,\ 1,\ 2 \end{gathered} \tag{6}$$

Foreign trade balance:

$$\mu_0 \delta_0 = \mu_1 \delta_1 + \mu_2 \delta_2 \tag{7}$$

Industrial security:

$$\delta_1 \le \gamma_1 x_1, \delta_2 \le \gamma_2 x_2 \tag{8}$$

Where γ_1- import quota coefficient for investment goods, δ_0- share of material exports, μ_0- world price of exported materials, μ_1, μ_2- world prices of imported investment and consumer goods, ρ_i- coefficient of capital-labor ratio reduction due to capital depreciation and employment growth, β_i- direct material costs per unit of output in the i-th subsystem.

A numerical algorithm for finding the stable state of an open economic system, based on the Lagrange multipliers method and the golden section principle, was proposed in [14].

2.2. Formulation of the Problem for Optimal Resource Allocation

In this study, industrial security is interpreted as ensuring conditions that maintain equilibrium and stability within the system:

$$\delta_1 = \gamma_1 x_1, \delta_2 = \gamma_2 x_2. \quad (9)$$

Then, the equation for the foreign trade balance Equation (7) can be rewritten as follows:

$$\delta_0 = \frac{\mu_1}{\mu_0}\gamma_1 x_1 + \frac{\mu_2}{\mu_0}\gamma_2 x_2. \quad (10)$$

The initial problem (2)–(8), consisting of six controls, can be transformed into a problem with three controls using the balance equations. To achieve this, we rewrite the deviations from the equilibrium state of the system of differential equations (2), introducing the following notations:

$$\begin{gathered} y_0(t) = k_0(t) - k_0^s, y_1(t) = \\ = k_1(t) - k_1^s, y_2(t) = k_2(t) - k_2^s. \end{gathered} \quad (11)$$

$$\begin{gathered} u_0(t) = \frac{s_0(t)\theta_1(t)}{\theta_0(t)} - \frac{s_0^s\theta_1^s}{\theta_0^s}, \\ u_1(t) = s_1(t) - s_1^s, u_2(t) = \\ = \frac{s_2(t)\theta_1(t)}{\theta_2(t)} - \frac{s_2^s\theta_1^s}{\theta_2^s}. \end{gathered} \quad (12)$$

As a result of the transformations performed, system (2) takes the following form:

$$\begin{gathered} \dot{y}_0(t) = -\rho_0(y_0(t) + k_0^s) + \\ + \left(u_0(t) + \frac{s_0(t)\theta_1(t)}{\theta_0(t)}\right) \times \\ (1+\gamma_1)A_1(y_1(t) + k_1^s)^{\alpha_1}, \\ y_0(t_0) = y_0^0, y_0(T) = 0, \end{gathered} \quad (13)$$

$$\begin{gathered} \dot{y}_1(t) = -\rho_1(y_1(t) + k_1^s) + \\ + (u_1(t) + s_1^s) \times \\ (1+\gamma_1)A_1(y_1(t) + k_1^s)^{\alpha_1}, \\ y_1(t_0) = y_1^0, y_1(T) = 0, \end{gathered} \quad (14)$$

$$\begin{gathered} \dot{y}_2(t) = -\rho_2(y_2(t) + k_2^s) + \\ + \left(u_2(t) + \frac{s_2(t)\theta_1(t)}{\theta_2(t)}\right) \times \\ (1+\gamma_1)A_1(y_1(t) + k_1^s)^{\alpha_1}, \end{gathered} \quad (15)$$

$$y_2(t_0) = y_2^0, y_2(T) = 0.$$

In the system, let $y(t) = (y_0(t),\ y_1(t),\ y_2(t))^*$ – denote the state vector, and $u(t) = (u_0(t),\ u_1(t),\ u_2(t))^*$ – denote the control vector.

To obtain a numerical solution of the nonlinear system (13)–(15), we employ the iterative quasilinearization method proposed by Bellman and Kalaba. At each iteration, the right-hand side of the differential equations (13)–(15) is approximated by linear forms derived from a Taylor expansion in a neighborhood of the solution obtained at the previous iteration.

As a result, after a sufficient number of iterations, a sequence of linear–quadratic (LQ) problems is generated, whose solutions approximate the optimal control of the original nonlinear problem. The iterative process is continued until the stopping condition $\| u^{(n)} - u^{(n-1)} \| \leq \varepsilon$ is satisfied, where ε is a prescribed tolerance.

At the n-th iteration, the control $u^{(n)}(t)$ and the corresponding state trajectory $y^{(n)}(t)$ are determined as the solution of the LQ problem constructed in a neighborhood of the previous approximation $\left(y^{(n-1)}(t), u^{(n-1)}(t)\right)$[15].

Linearize the system of equations (13)–(15) and obtain the following system of differential equations:

$$\begin{gathered} \dot{y}_0(t) = \\ = (1+\gamma_1)A_1\frac{s_0(t)\theta_1(t)}{\theta_0(t)}\alpha_1 k_1^{s\alpha_1-1}y_1(t) - \\ -\rho_0 y_0(t) + (1+\gamma_1)A_1 k_1^{s\alpha_1}u_0(t) + \\ + f(y_1, u_0), y_0(t_0) = y_0^0. \end{gathered} \quad (16)$$

$$\begin{gathered} \dot{y}_1(t) - \\ = \left(-\rho_1 + (1+\gamma_1)A_1\alpha_1 k_1^{s\alpha_1-1}s_1^s\right)y_1(t) + \\ + (1+\gamma_1)A_1 k_1^{s\alpha_1}u_1(t) + \\ + f(y_1, u_1), y_1(t_0) = y_1^0, \end{gathered} \quad (17)$$

$$\begin{gathered} \dot{y}_2(t) = \\ = (1+\gamma_1)A_1\frac{s_2(t)\theta_1(t)}{\theta_2(t)}\alpha_1 k_1^{s\alpha_1-1}y_1(t) - \\ -\rho_2 y_2(t) + (1+\gamma_1)A_1 k_1^{s\alpha_1}u_2(t) + \\ + f(y_1, u_2), y_2(t_0) = y_2^0, \end{gathered} \quad (18)$$

where the function $f(y_1, u_i), i = 0, 1, 2,$ is defined as follows:

$$f(y_1, u_0) =$$

$$= (1+\gamma_1)A_1\left[u_0(t)\left((y_1(t)++k_1^s)^{\alpha_1} - k_1^{s\alpha_1}\right) + \frac{s_0^s\theta_1^s}{\theta_0^s}\left((y_1(t)+k_1^s)^{\alpha_1} - k_1^{s\alpha_1} - \alpha_1 k_1^{s\alpha_1-1}y_1(t)\right)\right],$$

$$f(y_1, u_1) = (1+\gamma_1)A_1\left[u_1(t)\left((y_1(t)+k_1^s)^{\alpha_1} - k_1^{s\alpha_1}\right) + s_1^s\left((y_1(t)+k_1^s)^{\alpha_1} - k_1^{s\alpha_1} - \alpha_1 k_1^{s\alpha_1-1}y_1(t)\right)\right],$$

$$f(y_1, u_2) ==$$

$$(1+\gamma_1)A_1\left[u_2(t)\left((y_1(t)+k_1^s)^{\alpha_1} - k_1^{s\alpha_1}\right) + \frac{s_2^s\theta_1^s}{\theta_2^s}\left((y_1(t)+k_1^s)^{\alpha_1} - k_1^{s\alpha_1} - \alpha_1 k_1^{s\alpha_1-1}y_1(t)\right)\right],$$

When $y_1(t) = 0$ and $u_i(t) = 0$, the function $f(y_1, u_i), i = 0, 1, 2$, is equal to zero.

Next, we rewrite the given system (16)–(18) in vector–matrix form and consider the optimal control problem for a class of nonlinear controlled systems defined as follows:

$$\dot{y}(t) = A(t)y(t) + B(t)u(t) + \\ +f(y,u), \quad y(t_0) = y_0, y(T) = 0, \tag{19}$$

$$u \in U = \{u \mid \sigma_1 \le u(t) \le \sigma_2\}, \\ \sigma_1 < 0, \sigma_2 > 0, \sigma_1, \\ \sigma_2 = const, t \in [t_0, T]. \tag{20}$$

Let $y(t)$ denote the $(n \times 1)$ state vector of the controlled object and $u(t)$ the $(m \times 1)$ control vector. Let $f(y,u)$ be an $(n \times 1)$ vector of bounded continuous functions on the time interval $t \in [t_0, T]$. σ_1, σ_2 are vectors of dimension $(m \times 1)$. The matrices $A(t), B(t)$ are given time-dependent matrices of dimensions $(n \times n)$ and $(n \times m)$, respectively. While t_0 and T are the prescribed initial and terminal times, respectively. It is assumed that system (19) is controllable at time t_0.

We define the control quality by the functional:

$$J(y,u) = \\ = \frac{1}{2}\int_{t_0}^{T}[y^*(t)Q(t)y(t) + u^*(t)R(t)u(t)]dt \tag{21}$$

where $Q(t)$ positive definite symmetric matrix and matrix $R(t)$ is diagonal and positive-definite. $Q(t)$ represents the "penalty" for deviations of the economic state from equilibrium, while $R(t)$ reflects the costs associated with resource reallocation.

Problem Statement. It is required to construct a control $u(t)$ that transfers the economic system (19) from the initial state $y(t_0) = y_0$ to the desired final state $y(T) = 0$. At the same time, the control must satisfy the bilateral constraints (20) and ensure the minimization of the performance functional (21).

2.3. Solution to the Problem for Optimal Resource Allocation

To solve the stated problem, a modified Lagrange multipliers method is applied. The main idea of this method is to transform the original constrained problem into an equivalent unconstrained problem, the solution of which automatically satisfies the initial constraints.

For the implementation of the method, the system of differential equations (19) is multiplied by the following multiplier $\lambda_0(y,t)$ and use to the functional the terms enforcing the constraints:

$$\lambda_1{}^*(y,t)\left(\sigma_1 - u(t)\right) + \\ +\lambda_2{}^*(y,t)(u(t) - \sigma_2) + \\ +\lambda_3{}^*(y,t)(y(t) - W(t,T)^{-1}q(t)), \\ \lambda_i(y,t) \ge 0, (i = 1,2,3).$$

As a result, we obtain a modified functional that enables the problem to be studied in an unconstrained form while still enforcing the original constraints:

$$L(u,y) = \\ = \int_{t_0}^{T} \Big[\frac{1}{2} y^*(t)Q(t)y(t) + \frac{1}{2} u^*(t)R(t)u(t) + \lambda_0{}^*(y,t)\big(A(t)y(t) + B(t)u(t) + f(y,u) - \dot{y}(t)\big) \\ + \lambda_1{}^*(y,t)\big(\sigma_1 - u(t)\big) + \lambda_2{}^*(y,t)(u(t) - \sigma_2) \\ + \lambda_3{}^*(y,t)(y(t) - W^{-1}(t,T)q(t))\Big] dt \tag{22}$$

Let us consider the integrand of the functional (22) in the form of the following functions:

$$v(y,t) = \frac{1}{2} y(t)^* K(t)y(t) + y(t)^* q(t), \frac{\partial v}{\partial y} = K(t)y(t) + q(t), \tag{23}$$

$$M(u,y,t) = \frac{1}{2} y^*(t)\Big(Q(t) + \dot{K}(t)\Big)y(t) + \frac{1}{2} u^*(t)R(t)u(t) + \\ +(K(t)y(t) + q(t))^*(A(t)y(t) + B(t)u(t) + f(y,u)) + y^*(t)\dot{q}(t) + \lambda_1{}^*(y,t)(\sigma_1 - u(t)) \\ + \lambda_2{}^*(y,t)(u(t) - \sigma_2) + \lambda_3{}^*(y,t)(y(t) - W^{-1}(t,T)q(t)) \tag{24}$$

Thus, the Lagrangian functional (22) can be written in the expanded form as follows:

$$L(u,y) = v(y(t_0),t_0) - v(y(T),T) + \\ + \int_{t_0}^{T} M(u,y,t)\, dt. \tag{25}$$

To solve the formulated problem, we first apply the first-order necessary condition for an extremum and derive the expression for the optimal control in the following form:

$$u(t) == \\ -R^{-1}(t)B^*(t)\big(K(t)y(t) + q(t)\big) + \\ +R^{-1}(t)(\lambda_1{}^*(y,t) - \lambda_2{}^*(y,t)). \tag{26}$$

Here, $K(t)$ is an $n \times n$ matrix satisfying a matrix Riccati differential equation. The matrices $K(t), W(t)$, and the vector $q(t)$ on the interval $t \in [t_0, T]$ satisfy the following system of differential equations:

$$\dot{K}(t) + K(t)A(t) + A^*(t)K(t) - \\ -K(t)B_1(t)K(t) + Q(t) = \\ = 0, K(t_0) = K_0, \tag{27}$$

$$\dot{W}(t,T) = W(t,T)A_1{}^*(t) + \\ +A_1(t)W(t,T) - B_1(t), \quad W(T,T) = 0, \tag{28}$$

$$\dot{q}(t) = -A_1{}^*(t)q(t) + \\ +W^{-1}(t,T)B(t)\varphi(y,t) + \\ +W^{-1}(t,T)f(y,\varphi), \\ q(t_0) = W^{-1}(t_0,T)y(t_0) \tag{29}$$

where:

$$A_1(t) = A(t) - B(t)R^{-1}(t)B^*(t)K(t),$$

$$B_1(t) = B(t)R^{-1}(t)B^*(t),$$

$$\varphi(y,t) = R^{-1}(t)\big(\lambda_1{}^*(y,t) - \lambda_2{}^*(y,t)\big).$$

The inverse of $W(t,T)$ is required only for $t < T$, whereas at the terminal time the boundary conditions are imposed directly and no inversion is performed.

The solvability of the Riccati-type equations is considered under the standard assumptions of controllability and stabilizability of the pair $(A(t), B(t))$. The weighting matrices $Q(t)$and $R(t)$ are assumed to be symmetric and positive definite. Under these assumptions, the Riccati equation admits a bounded solution, and the resulting feedback control ensures convergence of the system to the desired terminal state. If these conditions are violated, numerical instability or loss of convergence may occur.

For the multipliers $\lambda_1(y,t) \geq 0, \lambda_2(y,t) \geq 0$, which are nonnegative by definition, the complementarity condition holds, expressed as follows:

$$\lambda_1{}^*(y,t)\big(\sigma_1 - u(t)\big) = \\ = 0, \lambda_2{}^*(y,t)(u(t) - \sigma_2) = 0.$$

For this purpose, they are chosen as follows:

$$\begin{gathered}\lambda_1(y,\ t) = \\ = max\ \{0;\ \sigma_1 - \omega(y,\ t)\} \geq 0, \lambda_2(y,\ t) \\ = max\ \{0;\ \omega(y,\ t) - \sigma_2\} \geq 0, \\ \omega(y,\ t) = \\ = -R^{-1}(t)B^*(t)(K(t)y(t) + q(t)).\end{gathered} \quad (30)$$

Because the matrix $R(t)$ is diagonal, the minimization problem with respect to the control becomes separable by components.

Hence, the complementarity-based construction of the multipliers leads to the componentwise projection of the nominal control $\omega(y,\ t)$ onto the admissible interval $[\alpha, \beta]$. As a consequence, the control constraints are satisfied at every time instant.

Suppose that solutions to equations (27) and (28) exist and that the conditions (30) are satisfied. Then, under the control law of the form (26), the dynamics of system (19) are described as follows:

$$\begin{gathered}\dot{y}(t) = A_1(t)y(t) - B_1(t)q(t) + \\ +B(t)\varphi(y,t) + f(y,\varphi,t), y(t_0) = y_0.\end{gathered} \quad (31)$$

Thus, the use of solutions of differential equations (31) and (29) shows that the dynamics of system (31), corresponding to control law (26), at the final moment of time reaches the state $y(T) = 0$.

To verify compliance with the balance constraints, the following expressions are computed. Since $s_1(t) = u_1(t) + s_1^s$, the remaining portion of investment resources allocated to the production and consumer subsystems is equal to $1 - s_1(t)$. Let $v(t)$ denote the share of the consumer subsystem in this remainder of investment resources. Then, in order to ensure the fulfillment of the investment balance condition (4), the following allocation rule is applied:

$$\begin{gathered}s_1(t) = u_1(t) + s_1^s, \\ s_0(t) = v(t)\big(1 - s_1(t)\big), \\ s_2(t) = \big(1 - v(t)\big)\big(1 - s_1(t)\big).\end{gathered} \quad (32)$$

To ensure the fulfillment of the labor balance condition (5), using the notations introduced in (12) we obtain the expressions for $\theta_0(t)$ and $\theta_2(t)$, while the expression for $\theta_1(t)$ is derived from (5) in the following form:

$$\begin{gathered}\theta_1(t) = \frac{1}{1 + \frac{s_0(t)}{u_0(t) + (s_0^s\theta_1^s/\theta_0^s)} + \frac{s_2(t)}{u_2(t) + (s_2^s\theta_1^s/\theta_2^s)}}, \\ \theta_0(t) = \frac{v(t)(1 - s_1(t))\theta_1(t)}{u_0(t) + (s_0^s\theta_1^s/\theta_0^s)}, \\ \theta_2(t) = \frac{\big(1 - v(t)\big)\big(1 - s_1(t)\big)\theta_1(t)}{u_2(t) + \left(\frac{s_2^s\theta_1^s}{\theta_2^s}\right)}.\end{gathered} \quad (33)$$

Next, taking into account equation (10) and the obtained values of $\theta_0(t)$ and $\theta_2(t)$ from equation (33), the material balance equation (6) yields the following expression for $v(t)$:

$$v(t) = \frac{\left(\beta_1 + \frac{\mu_1}{\mu_0}\gamma_1\right)A_1k_1(t)^{\alpha_1} + \left(\beta_2 + \frac{\mu_2}{\mu_0}\gamma_2\right)A_2k_2(t)^{\alpha_2}\frac{1 - u_1(t) - s_1^s}{u_2(t) + \left(\frac{s_1^s\theta_1^s}{\theta_2^s}\right)}}{(1 - \beta_0)A_0\big(k_0(t)\big)^{\alpha_0}\frac{1 - u_1(t) - s_1^s}{u_0(t) + \left(\frac{s_0^s\theta_1^s}{\theta_0^s}\right)} + \left(\beta_2 + \frac{\mu_2}{\mu_0}\gamma_2\right)A_2\big(k_2(t)\big)^{\alpha_2}\frac{1 - u_1(t) - s_1^s}{u_2(t) + \left(\frac{s_2^s\theta_1^s}{\theta_2^s}\right)}}. \quad (34)$$

Algorithm 1: Algorithm for Solving the Problem of Optimal Resource Allocation

1: *Using the Runge–Kutta method on the interval* $t \in [t_0, T]$ *integrate the systems of differential equations (27) and (28) subject to the boundary conditions.*
2: *For the given initial state* $y(t_0) = y_0$ *we determine the vector* $q(t_0)$ *according to the relation:* $q(t_0) = W^{-1}(t_0, T)y(t_0)$.
3: *We integrate the system of equations (31) and (29) over the interval* $[t_0, T]$ *using the Runge–Kutta method under the initial conditions* $y(t_0) = y_0, q(t_0) = q_0$. *During the integration process, the trajectory* $y(t)$ *and the control* $u(t)$ *are constructed.*
4: *Using the values of the effective trajectory* $y(t)$ *and the effective control* $u(t)$, *obtained during the integration process, we compute the following expressions:*
To verify the fulfillment of the material balance (6), expression (34) is evaluated.
To verify the fulfillment of the investment balance (4), expression (32) is evaluated.
To verify the fulfillment of the labor balance (5), expression (33) is evaluated.

If, in the course of computations, the initial and terminal conditions of the problem are modified and new numerical experiments are required, the calculations should be repeated starting from the second and third steps of the algorithm.

3. Results

To obtain a numerical solution to the formulated optimal control problem, a series of computational experiments was carried out. When constructing the plots of the effective trajectory $y(t)$ and the optimal control $u(t)$, the model parameter values were taken according to the data presented in Tables 1 and 2.

Table 1. Model coefficients

i	0	1	2
$\boldsymbol{\alpha_i}$	0,7	0,62	0,45
$\boldsymbol{\beta_i}$	0,39	0,49	0,52
$\boldsymbol{\lambda_i}$	0,05	0,05	0,05
$\boldsymbol{A_i}$	3,19	6	3,71

The values of the coefficients $\alpha_i, \beta_i, \rho_i$ and A_i are chosen within the permissible intervals corresponding to the logic of the model. The parameters are not based on specific empirical measurements but serve for the numerical analysis of the system behavior under various configurations [14].

In the numerical experiment, the planning horizon was set to $T = 12$. The initial state of the system and the control constraints were specified as follows:

$$y(t_0) = (4000, -5500, 4500)^*, \qquad (35)$$

Table 2. Stationary state of an open economic system

i	0	1	2
$\boldsymbol{\theta_i}$	0.233	0.270	0.296
$\boldsymbol{s_i}$	0.391	0.316	0.292
$\boldsymbol{k_i}$	59433.45	41525.48	20840.3
$\boldsymbol{x_i}$	1638.1	1181.31	161.75
$\boldsymbol{y_i}$	335.77	590.66	80.88

To avoid possible singularities in the expressions for $\theta_i(t)$ and $v(t)$, the admissible control set is implemented numerically using a projection procedure.

$$0.1 - u_i^s \le u_i(t) \le 0.9 - u_i^s. \qquad (36)$$

$$-0.3531 \le u_0 \le 0.4469, -0.216 \le u_1 \le 0.584, -0.108 \le u_2 \le 0.741. \qquad (37)$$

The computed values of the deviations of the system state at the final time are as follows (Figure 1):

$$y_0(T) = 5.0475 \cdot 10^{-9},$$
$$y_1(T) = -1.0885 \cdot 10^{-8},$$
$$y_2(T) = -7.1596 \cdot 10^{-9}.$$

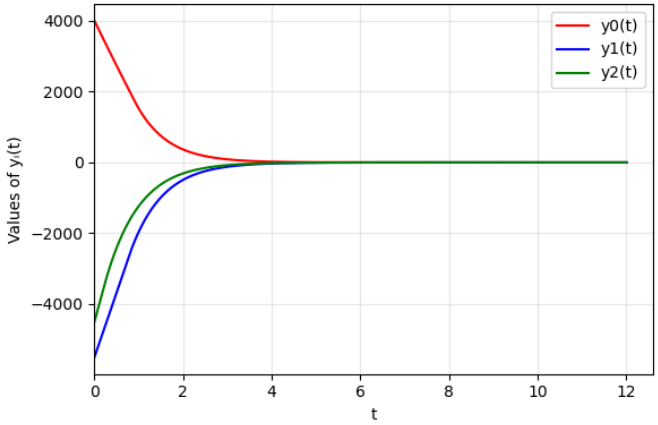


Figure 1. Effective trajectories $y_i(t)$ of the system under optimal control

Figure 2 presents the graph of the optimal control. The graph shows that the constructed control trajectory remains within the admissible region U, defined by the constraints (37).

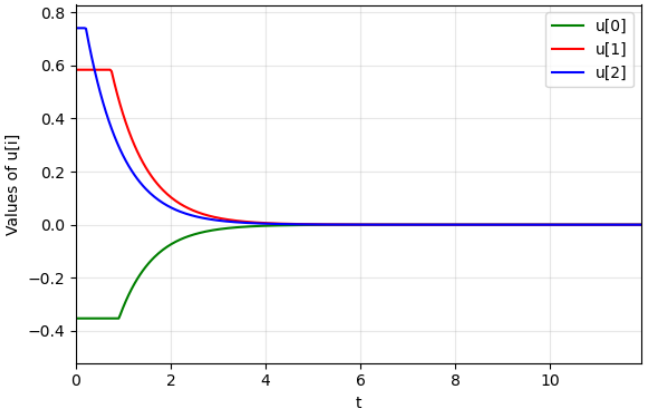
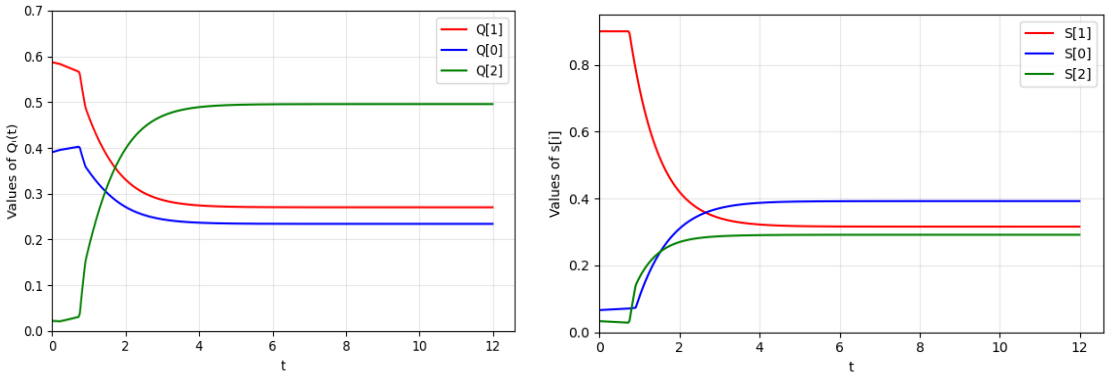


Figure 2. Dynamics of optimal control functions $u_i(t)$ satisfying bilateral constraints



(a) (b)

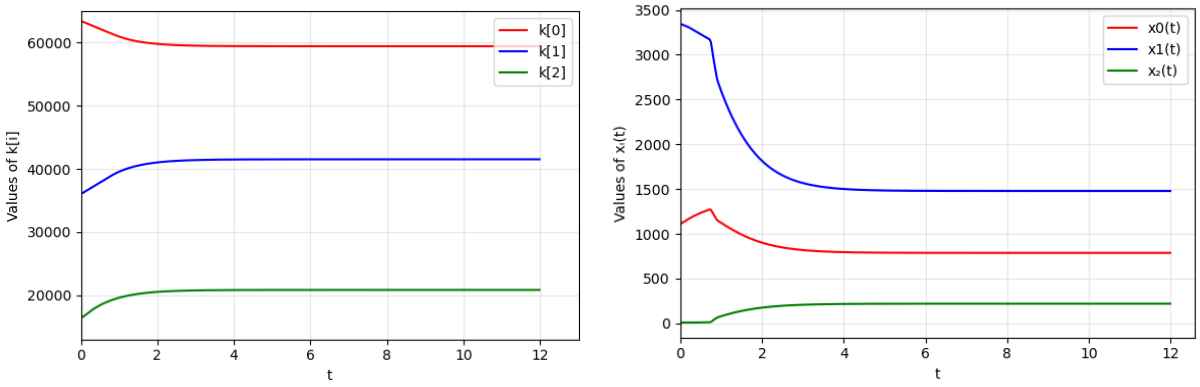
Figure 3. Optimal distribution of: (a) labor resources; (b) investment resources

In addition, during the computations, the values were obtained that determine the optimal allocation of labor resources $(\theta_0(t), \theta_1(t), \theta_2(t))$ and investment resources $(s_0(t), s_1(t), s_2(t))$ among the production, investment, and consumer subsystems, satisfying the balance relations (4)–(5). The corresponding results are presented in Figures 3 (a) and (b).

Figure 4 (a) illustrates the solutions of the differential equations for capital intensity $k(t)$ under given initial conditions.

$$k(t_0) = (63400, 35500, 15500)^*.$$

And Figure 4 (b) illustrates dynamics of subsystem output in an open economy.



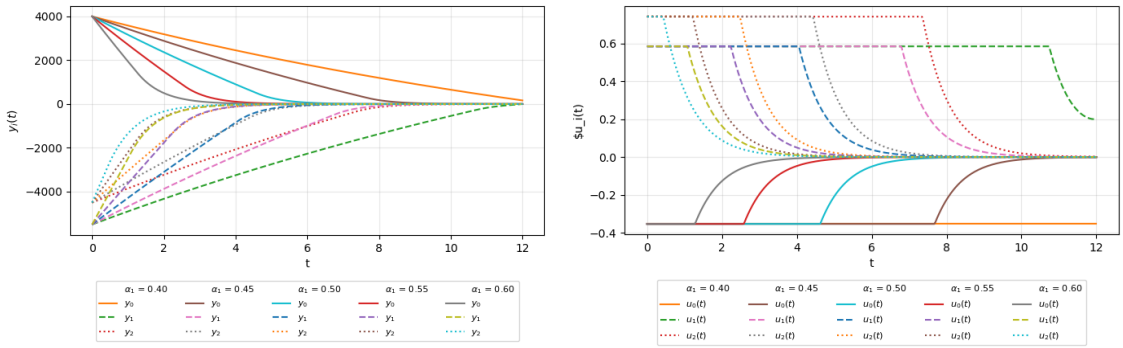
(a) (b)

Figure 4. (a) capital intensity dynamics in the subsystems of an open economy; (b) dynamics of subsystem output in an open economy

All simulations were performed using a Runge–Kutta-type implicit integrator. The integration step was limited by a maximum step size of 0,02. Relative and absolute tolerances were set to 10^{-8} and 10^{-10}, respectively. Matrix equations were solved numerically at each time step.

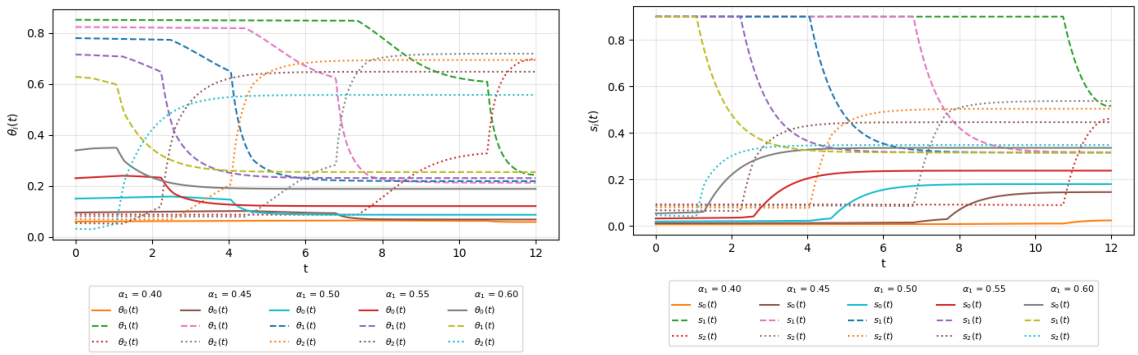
In the numerical implementation the matrix $W(t,T)$ was integrated backward in time starting from a small regularized terminal value $W(T,T) = \varepsilon I$ with $\varepsilon = 10^{-5}$, preventing singular inversion of $W(t,T)$ near the terminal time. The full set of parameters used in the experiments is reported in Tables 1–2.

To assess the robustness of the proposed numerical procedure, additional experiments were performed under variations of model parameters and initial conditions. In particular, the elasticity coefficient of the investment subsystem α_1 was varied within the interval [0.4,0.6] with discretization step 0.05. (Figures 5 and 6).



(a) (b)

Figure 5. System behavior under variation of the elasticity coefficient α_1: (a) state trajectories $y_i(t)$; (b) control actions $u_i(t)$



(a) (b)

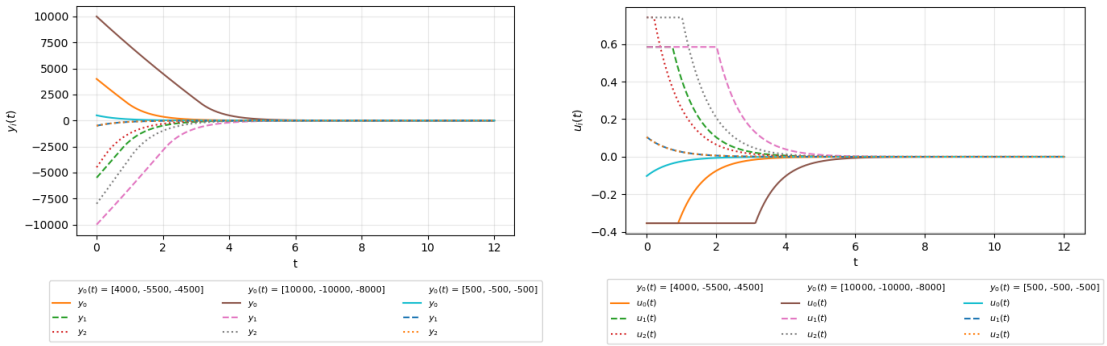
Figure 6. Resource allocation dynamics under variation of α_1:
(a) labor distribution $\theta_i(t)$; (b) investment shares $s_i(t)$

Furthermore, several initial deviations from the stable state were considered, including both large-scale and moderate perturbations:

$$y(t_0) = (10000, -10000, 8000)^*,$$

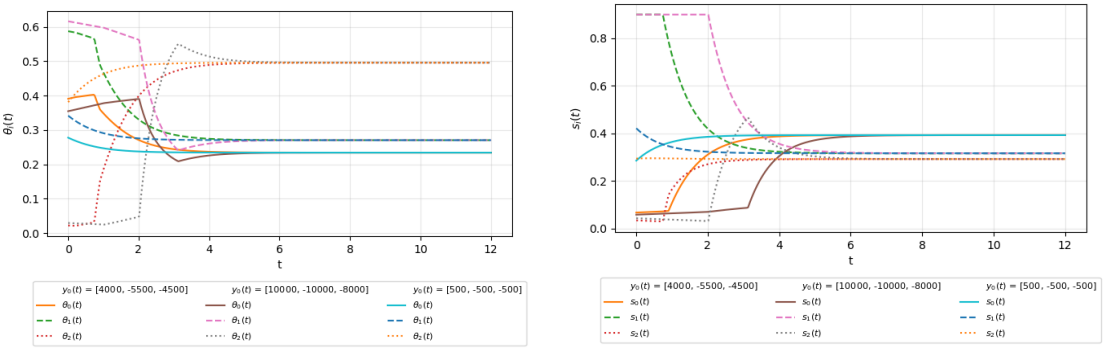
$$y(t_0) = (500, -500, -500)^*,$$

The obtained trajectories of states and controls are presented in Figures 7 and 8.



(a) (b)

Figure 7. Robustness of the algorithm with respect to initial conditions:
(a) state trajectories $y_i(t)$; (b) control actions $u_i(t)$



(a) (b)

Figure 8. Resource allocation under different initial states:
(a) labor distribution $\theta_i(t)$; (b) investment shares $s_i(t)$

For all tested configurations, the synthesized controls remained inside the admissible bounds. The system consistently approached the equilibrium neighborhood, and the qualitative structure of labor and investment allocation was preserved.

4. Discussion

The developed algorithm has demonstrated efficiency in the numerical implementation of the optimal control problem under multiple constraints. Its main advantage lies in transforming the original constrained formulation into an equivalent problem that is more convenient for computation. As a result, stable solutions were obtained that reflect the patterns of resource allocation among the economic subsystems.

Unlike general-purpose optimization methods, which require solving high-dimensional nonlinear programming problems, the proposed algorithm constructs admissible controls through the integration of matrix differential equations combined with projection formulas, which simplifies the numerical implementation. The computational performance of the approach was evaluated by comparing it with standard nonlinear programming solvers (Table 3).

As shown, the proposed method achieves comparable values of the objective functional while requiring significantly less computational time.

Table 3. Stationary state of an open economic system

Method	Runtime (s)	Objective value $J(y,u)$
Proposed algorithm	0.11	3.51×10^7
NLP (SLSQP)	3.32	3.52×10^7
NLP (L-BFGS-B)	3.17	3.52×10^7

Numerical experiments showed that labor resources are concentrated in the consumer subsystem, as it is the most labor-intensive. Investments are primarily directed to the production and investment subsystems, due to their role in maintaining production capacities, renewing capital, and developing infrastructure. These results are consistent with economic logic and confirm the validity of the proposed algorithm.

From an economic perspective, the obtained allocation structure reflects the functional roles of the subsystems within an open economy. The concentration of labor in the consumer subsystem indicates that final demand formation relies primarily on human resources, while capital accumulation mechanisms dominate in the production and investment subsystems. The redistribution pattern generated by the algorithm demonstrates that sustainable dynamics require continuous support of sectors responsible for capacity expansion and technological renewal. A reduction of investments in these subsystems would immediately slow down the growth of capital intensity and, consequently, limit future output levels.

Therefore, the computed trajectories can be interpreted not only as a mathematical solution but also as an indicator of structurally necessary proportions between consumption and accumulation. They provide quantitative guidance for policy design aimed at maintaining long-term stability while avoiding excessive resource diversion toward short-term consumption. The algorithm not only makes it possible to determine the stable state of the system but also to investigate the influence of model parameters on the structure of the optimal allocation. This capability makes it a valuable tool for analyzing alternative scenarios and strategies of economic development.

During the numerical simulations, additional verification of the admissibility conditions was performed at each integration step. In particular, the positivity of the state variables, boundedness of trajectories, and feasibility of the share parameters within the interval $(0,1)$ were monitored. The balance relations were also checked numerically and were satisfied up to machine precision along the computed trajectories. No violations of admissibility were detected in the reported experiments.

5. Conclusions

This paper presents an algorithm for the numerical implementation of the optimal resource control problem in an open economy model. The algorithm is based on the Lagrange multipliers framework and is intended to obtain solutions under labor, investment, and foreign trade constraints.

The computational experiments demonstrate that the system converges to a stable regime that ensures coordinated resource allocation and sustainable development dynamics. The practical significance of the study lies in extending classical optimal control techniques to an open-economy environment with explicit consideration of external trade flows and industrial security restrictions. Thus, the proposed algorithm can be used for quantitative analysis and modeling of resource allocation dynamics in open economic systems.

Funding

This research is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant no. AP22684879).

Author Contributions

Author Contributions: Conceptualization, K.T. and G.M.; Methodology, K.T.; Software, K.T. and A.Sh; Validation, K.T. and G.M.; Formal Analysis, G.M.; Investigation, K.T.; Data Curation, K.T., G.M. and A.Sh; Writing–Original Draft preparation, K.T.; Writing–Review & Editing, K.T. and G.M.; Visualization, K.T.; Supervision, K.T.; Project Administration, K.T.; Funding Acquisition, K.T.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. A. Gornov, T. Zarodnyuk, A. Anikin, P. Sorokovikov, and A. Tyatyushkin, "Software Engineering for Optimal Control Problems," in *Mathematics and its Applications in New Computer Systems*, vol. 424, A. Tchernykh, A. Alikhanov, M. Babenko, and I. Samoylenko, Eds. Cham: Springer International Publishing, 2022, pp. 415–426, doi: 10.1007/978-3-030-97020-8_38.

2. M. I. Naz *et al.*, "Framework of Decision Support System For Effective Resource Management," in *2023 Int. Conf. on Business Analytics for Technology and Security (ICBATS)*, Dubai, United Arab Emirates, Mar. 2023, pp. 1–7, https://doi.org/10.1109/ICBATS57792.2023.10111307.

3. V. Kolemayev, "Optimal balanced growth of an open three-sector economy", *Prikladnaya Ekonometrika*, vol. 11, no. 3, pp. 15–42, 2008, Available online: https://cyberleninka.ru/article/n/optimalnyy-sbalansirovannyy-rost-otkrytoy-trehsektornoy-ekonomiki.

4. Saraev A. L. and Saraev L. A., "Trekhfaktornaya matematicheskaya model razvitiya predpriyatiya za schet vnutrennikh i vneshnikh investitsiy," *Vestnik Altaiskoi akademii ekonomiki i prava*, no. 2, pp. 77–85, 2020, https://doi.org/10.17513/vaael.1002.

5. A. Matsumoto and F. Szidarovszky, "Delay two-sector economic growth model with a Cobb–Douglas production function," *Decisions in Economics and Finance*, vol. 44, no. 1, pp. 341–358, 2021, https://doi.org/10.1007/s10203-021-00321-2.

6. P. Elango, D. Luo, A. G. Kamath, S. Uzun, T. Kim, B. Açıkmeşe. Continuous-time successive convexification for constrained trajectory optimization, *Automatica*, vol. 180, 2025, 112464, https://doi.org/10.1016/j.automatica.2025.112464.

7. D. Zhang and Y. Zhang, "PySCP: A Multiple-Phase Optimal Control Software Using Sequential Convex Programming," *International Journal of Aerospace Engineering*, vol. 2022, https://doi.org/10.1155/2022/2969809

8. A. Dinc, F. Yildiz, K. Nag, M. Otkur, and A. Mamedov, "Solving and Optimization of Cobb–Douglas Function by Genetic Algorithm: A Step-by-Step Implementation," *Computation*, vol. 13, no. 2, p. 23, 2025, https://doi.org/10.3390/computation13020023.

9. Z. Murzabekov, M. Milosz, K. Tussupova, and G. Mirzakhmedova, "Development of an algorithm for solving the problem of optimal control on a finite interval for a nonlinear system of a three-sector economic cluster," *Eastern-European Journal of Enterprise Technologies*, vol. 1, no. 3(115), pp. 43–52, 2022, https://doi.org/10.15587/1729-4061.2022.252866.

10. M. G. Dmitriev, Z. N. Murzabekov, and G. A. Mirzakhmedova, "An Algorithm for Finding Feedback in a Problem with Constraints for One Class of Nonlinear Control Systems," *Automatic Control and Computer Sciences*, vol. 56, no. 7, pp. 623–633, 2022, https://doi.org/10.3103/S0146411622070033.

11. Z. Murzabekov, M. Miłosz, and K. Tussupova, "The Optimal Control Problem with Fixed-End Trajectories for a Three-Sector Economic Model of a Cluster," in *Intelligent Information and Database Systems*, vol. 10751, N. T. Nguyen, D. H. Hoang, T.-P. Hong, H. Pham, and B. Trawiński, Eds. Cham: Springer International Publishing, 2018, pp. 382–391, https://doi.org/10.1007/978-3-319-75417-8_36.

12. Z. Murzabekov, M. Milosz, K. Tussupova, and G. Mirzakhmedova, "Problems of Optimal Control for a Class of Linear and Nonlinear Systems of the Economic Model of a Cluster", *Vietnam Journal of Computer Science*, vol. 7, no. 2, pp. 109–127, 2020, https://doi.org/10.1142/S2196888820500062.

13. Z. Murzabekov, M. Milosz, and K. Tussupova, "Modeling and Optimization of the Production Cluster", in *Information Systems Architecture and Technology: Proc. 36th Int. Conf. on Information Systems Architecture and Technology – ISAT 2015 – Part II*, vol. 430, A. Grzech, L. Borzemski, J. Świątek, and Z. Wilimowska, Eds. Cham: Springer International Publishing, 2016, pp. 99–108, https://doi.org/10.1007/978-3-319-28561-0_8.

14. Z. Murzabekov, K. Tussupova. Development of a model of efficient resource allocation in an open three-sector economy for balanced growth. *Journal of Mathematics, Mechanics and Computer Science*, *124*(4), 2025, 59–70. https://doi.org/10.26577/JMMCS2024-v124-i4-a5.

15. S. A. Aisagaliev, Sh. A. Aipanov, and E. B. Zlobina, *Prikladnye zadachi optimalnogo upravleniya*. Almaty: Kazakh University, 2005.

Information about Authors:

Kamshat Tussupova – PhD, Senior Researcher at the Department of Information Systems in Al-Farabi Kazakh National University, Republic of Kazakhstan, e-mail: kamshat-0707@mail.ru. Research interests: optimal control of dynamic systems and modeling of economic systems; application of machine learning methods for developing algorithms of forecasting and resource allocation.

Gulbanu Mirzakhmedova – PhD, Acting Associate Professor, Senior Lecturer at the Department of Information Systems in Al-Farabi Kazakh National University, Republic of Kazakhstan, e-mail: gulbanu.myrzahmedova@gmail.com. Research interests: optimal control of dynamic systems and modeling of economic systems.

Assem Shormakova – PhD, Head of the "Information Systems" Department, Acting Associate Professor; Al-Farabi Kazakh National University, Republic of Kazakhstan, e-mail: shormakovaassem@gmail.com.

Submission received: 3 October, 2025.
Revised: 16 March, 2026.
Accepted: 16 March, 2026.

IRSTI 37.23.31 https://doi.org/10.26577/jpcsit4120264

Zh. Baishemirov[1, 2, 3], D. Ospanova[2*], B. Amirgaliyev[2], S. Mukhambetzhanov[4]

[1]Kazakh-British Technical University, Almaty, Kazakhstan
[2]Astana IT University, Astana, Kazakhstan
[3]Narxoz University, Almaty, Kazakhstan
[4]Al-Farabi Kazakh National University, Almaty, Kazakhstan
*e-mail: 242982@astanait.edu.kz

COMPARATIVE ANALYSIS OF PHYSICS-INFORMED AND CONVENTIONAL LSTM AND RNN MODELS FOR TEMPERATURE FORECASTING USING ERA5 REANALYSIS DATA

Abstract. Climate change is one of the most serious modern problems affecting the Earth's atmosphere, as it causes a range of harmful effects worldwide. Due to the uneven nature of climate data, forecasting climate change is a challenging task today. Many previous studies in climate and machine learning have used recurrent neural networks (RNNs) and long short-term memory (LSTM) models to predict climate trends. Although these models are effective at identifying long-term trends in data, they often fail to satisfy physical laws such as energy conservation, mass balance, and thermodynamic principles. In this research, the aim was to develop oscillation-constrained RNN and LSTM models, in which an annual harmonic prior is incorporated into the loss function, and compare their performance with standard RNN and LSTM models. The study utilized data on air temperature at 2 meters above the surface for the cities of Astana, Almaty, and Shymkent for model training, validation, and testing. According to the results, physics-informed models achieved the lowest root mean square errors in Almaty (3.52 °C) and Shymkent (3.80 °C). RNN and LSTM models performed better in Astana (RMSE = 5.44 and 5.47 °C), where seasonal changes are relatively abrupt. These findings demonstrate that PINNs constrained by the annual harmonic oscillation provide more physically consistent forecasts for moderate climates, while conventional recurrent models perform better in locales with highly variable conditions.

Keywords: climate prediction, Physics-Informed Neural Networks, Long Short-Term Memory, Recurrent Neural Network, ERA5 reanalysis, temperature forecasting, Numerical Weather Prediction.

1. Introduction

The problem of climate change remains one of the greatest and most challenging issues in modern Earth system science, due to the chaotic and non-intuitive nature of the Earth's atmosphere. It is chaotic to an extent, where even the smallest change can yield an unexpected outcome over time. According to [1], global warming as a result of climate change can take place suddenly in a few decades, or even in a few years, in the form of a climate shock. As a result, weather prediction becomes highly unreliable, especially when events last more than a few days. The problem is exacerbated even further when the scale of the entire climate system is taken into account. Some events may happen in the span of a few minutes, while other events may take as long as a few decades to unfold. For instance, earlier research indicates a permanent shift of the Intertropical Convergence Zone towards the warmer hemisphere as a result of disturbances in the interhemispheric asymmetry [2]. Climatic events work in harmony, and disturbances at local scales can trigger responses at the global level.

Another problem encountered by researchers worldwide is data scarcity. Highly accessible locales account for the majority of the data, while less accessible locales, such as the Arctic, deep-ocean basins, and areas in developing countries, account for just a minuscule amount of data. For example, although Polar regions have an impact on global temperatures, they are not well studied because of their extreme conditions [3]. This leads to an incomplete picture of the global climate. As a result, a systematic bias is introduced into AI models and algorithms trained on these datasets.

Some weather prediction methods that were established in the past are outdated in the current meteorological scenario. While these traditional approaches have contributed greatly to humanity's progress in climatology, they are now incompatible with the scales of current climatic events. For

instance, global climate models possess horizontal limits of 70 to 250 kilometers [4], which could be insufficient in encompassing numerous individual clouds, urban heat islands, and local topographic effects.

Hardware limitations of past methods have forced researchers to extrapolate data to a larger area, a process that can introduce statistical errors. Data extrapolation using AI and neural network approaches can lead to results that are not compatible with fundamental physical principles such as energy conservation, mass balance, or thermodynamics [5]. Therefore, predictive models that go beyond fundamental physics are prone to produce implausible and unreliable results. The following observation is crucial in the onset of PINN or Physics-Informed Neural Networks, which are especially practical in their compatibility with predefined physical datasets [6]. As introduced by Raissi, PINNs are neural networks designed to learn while adhering to physical laws governed by general nonlinear partial differential equations [7]. This study compares PINN efficiency in the prediction of air temperature with traditional Long Short-Term Memory and Recurrent Neural Network models. We hypothesize that incorporating physical laws via PINNs will result in more reliable and physically plausible forecasts of weather events compared to conventional approaches.

The use of PINNs in long-term climate frameworks to increase weather forecast accuracy over long time horizons is what makes this research novel. Modern literature on climate prediction focuses mainly on short-term or localized frameworks, with little effort in applying PINN models. The study uses the ERA5 reanalysis dataset to validate this approach.

It is expected that PINNs will improve interpretability and forecast reliability for city-level climate datasets. Other climatic variables, such as precipitation, humidity, and wind speed, can also be predicted later, if this approach proves effective. This would help establish a unified system for long-term climate forecasting and environmental monitoring.

2. Literature Review

Numerical weather prediction (NWP) is a widely used concept in atmospheric studies. NWP models simulate future meteorological events by solving primitive equations of motion numerically [8]. The main advantage of NWP models is their physically based representations of fluid dynamics, thermodynamics, and radiative transfer. As a result, numerical models generate high-quality short-term weather prognoses for up to about one week [9]. Nevertheless, such models have high computational costs and limited resolution. Moreover, they overdepend on sub-grid processes such as convection and cloud microphysics, and accumulate errors rapidly due to chaotic dynamics [10]. Therefore, it is difficult to apply this model for long-term climate forecasts. Researchers employ Regional Climate Models (RCMs) along with dynamical downscaling to tackle this problem. This technique provides better spatial detail by using finer grids inside global models. While RCMs are effective, they require greater computational resources than global NWP methods and are impractical for frequent or long-term use [11].

To overcome these challenges, researchers developed statistical downscaling techniques. These methods can find relationships between large-scale and local climate variables [12]. Compared to dynamical downscaling, statistical downscaling provides more global detail and better computational efficiency. However, it depends on reliable and consistent observational records, which may become problematic as the climate changes [13,14].

The advent of machine learning (ML) algorithms has supplied the scientific world with another approach for climate prediction. 1998 marked the development of the first neural network models working with short-term precipitation prediction. In later years, non-deep learning methods were also introduced to enhance medium and long-term precipitation predictions. Models such as CGF, CycleGANs, DeepESD, and NNCAM have achieved superior accuracy compared to conventional physical models. They excelled in capturing temporal and spatial climate patterns, refining resolution, and ultimately reducing computational time [15]. Machine learning has also found its use in climate science in miscellaneous tasks such as tropical cyclone tracking, cloud classification, and air quality predictions [16 – 18].

The gradual growth of computational power and increasing scalability of climate datasets have led scientists to focus on more innovative models for capturing spatial and temporal dependencies in climate processes. Several deep learning models, such as Convolutional Neural Networks (CNN) have been employed to perform image super-resolution for enhancing coarse-resolution climate model outputs [19], while learning models such as

Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) were used in capture of temporal dependencies in climate and hydrological time series, as demonstrated by [20]. LSTM models effectively capture temporal trends and patterns, which enable accurate long-term predictions of noisy climate datasets. Transformer-based architectures have also proven to be useful in modeling long-range spatiotemporal interactions. For instance, Ramu et al. (2022) proposed a Transformer-based model for daily temperature forecasting, integrating a Spatial-Temporal Fusion Module, Hierarchical Graph Representation, and a Dynamic Temporal Graph Attention Mechanism to capture spatiotemporal dependencies and improve temporal feature extraction [21].

Although deep-learning algorithms offer many benefits, they often struggle to be physically consistent [22]. Slater et al. (2023) observe that data-driven models are usually not very good at forecasting extreme or novel events that have never occurred before and were not contained in historical data [23]. Besides, these models face the difficulty of optimizing high-dimensional multivariate outputs. These issues of deep learning algorithms have led to the development of Physics-Informed Neural Networks. PINNs use physical laws together with training data, which helps them make accurate predictions even when data is limited or sparse. PINNs help identify the main dynamics of a system and ensure the equations used remain consistent. This gives them an advantage over models that rely only on data, such as RNNs or LSTMs [6].

According to Feng et al. (2023), predictions in the data-sparse areas can be made more robust by imposing physical constraints [24]. Other studies show that combining physical models with LSTM improved accuracy in the validation period and reduced uncertainty in future flood forecasts [25]. PGnet, a physics-based deep learning model, improves tropospheric temperature predictions by using physical principles with generative neural networks and a guiding mask to enhance low-quality prediction areas [26]. However, most studies still focus on specific fields or small-scale cases, often looking at isolated physical processes rather than complex systems. Moreover, the ability of models to generalize data under the influence of changing climate conditions and scarce data is still insufficiently examined. Few studies discuss the scalability of PINNs for large-scale climatic conditions and high-dimensional datasets, as well as their integration into long-term monitoring frameworks. Such research gaps reflect the need for further studies on the application of PINNs in large-scale, data-driven climate prediction and environmental monitoring.

3. Materials and Methods

3.1 Dataset Selection and Preprocessing

ERA5 reanalysis dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF) was used in this study to train the models. The dataset integrates information from satellites, ground stations, and other observation systems. It was chosen for this research because it offers high spatial and temporal resolution. ERA5 also provides long-term, globally consistent, and reliable data.

In detail, we adopted the post-processed ERA5 daily statistics at single levels for 2008-2022. Air temperature at 2m above the land and sea surfaces was selected as the predicted variable. The dataset was divided into three parts: 2008–2018 for training, 2019–2020 for validation, and 2021–2022 for testing. All input variables were normalized to zero mean and unit variance using statistics computed from the training set, and these values were subsequently applied to the validation and test subsets.

Overlapping sliding windows with a stride of one day were then formed independently within each subset, ensuring that no input sequence or forecast horizon contained information from future periods. Using these sequences, a direct multi-step forecasting strategy was employed, where each input sequence of 365 consecutive daily temperature observations was used to predict the subsequent 90 daily temperature values in a single forward pass. This approach avoids recursive error accumulation associated with autoregressive inference and is well suited for seasonal and sub-seasonal temperature forecasting tasks.

Regional comparison was performed using climatic data from Astana, Almaty, and Shymkent, three major cities in Kazakhstan. The choice is also justified by uniquely distinct thermal footprints each city possesses, with continental sharp cold and seasonal contrasts in Astana, moderate and humid mountainous climate in Almaty, and warm semi-arid climate with relatively mild winters in Shymkent. Diversity of climates makes it possible to thoroughly evaluate how models perform in different climate conditions within one country.

Table 1. Summary of 2m temperature statistics by city

City	Coordinates	Mean (°C)	Std (°C)
Astana	latitude = 51.1694, longitude = 71.4491	3.37	14.60
Almaty	latitude = 43.238949, longitude = 76.889709	5.55	10.60
Shymkent	latitude = 42.3417 longitude = 69.5901	13.56	10.98

3.2. Models

3.2.1. LSTM

For this study, a Long Short-Term Memory (LSTM) neural network was employed to model and predict temporal variations in 2-meter air temperature. LSTM networks are a type of recurrent neural network (RNN) that can learn long-term patterns in sequential data. Hence, they are suitable for time-series forecasting tasks, such as climate and weather prediction.

Mean squared error (MSE) as the loss function and the Adam optimizer with a learning rate of 0.001 and 200 epochs were used in the model training process. To avoid overfitting and enhance generalization, we used early stopping and dropout regularization.

Table 2. Hyperparameters of the LSTM model

Hyperparameter	Value
Optimizer	Adam
Learning rate	0.001
Number of layers	2
Neurons per layer	64
Dropout	0.1
Number of epochs	200
Early stopping	Yes

3.2.2. RNN

Unlike the LSTM, which uses memory cells and gates, the vanilla RNN employs simple recurrent connections to capture short-term temporal dependencies in sequential data. Though more prone to vanishing gradients and worse at capturing long-term dependencies, the RNN is a lighter and simpler baseline computationally for time series forecasting applications such as weather and climate prediction.

The RNN model was also optimized with the Adam optimizer, using a learning rate of 0.001. There were 150 epochs of training. The RNN architecture offers a simpler approach to temperature forecasting than the LSTM architecture. Other model training parameters are listed in Table 3.

Table 3. Hyperparameters of the RNN model

Hyperparameter	Value
Optimizer	Adam
Weight initialization	Random
Learning rate	0.001
Number of layers	2
Neurons per layer	64
Dropout	0.1
Number of epochs	150
Early stopping	Yes

3.2.3. Physics-Informed models

While inspired by physics-informed approaches, the oscillation-constrained models do not encode full thermodynamic principles or conservation laws. The physics term used here reflects only the regular annual oscillation of temperature.

The architecture of the PINN models mirrors the baseline structures, consisting of two hidden layers with 64 neurons each, followed by a fully connected dense output layer. The tanh activation function was used for hidden units to capture nonlinear temperature dynamics, and a dropout rate of 0.1 was applied to reduce overfitting. The total loss function combines a data-driven loss with a physics-based loss term:

$$L_{total} = \lambda_{data} L_{data} + \lambda_{phys} L_{phys} \qquad (1)$$

where $L_{data}$is the mean squared error between observed and predicted temperatures, and L_{phys} can be defined as [27]:

$$L_{phys} = \frac{1}{N}\sum_{i=1}^{N} (u''(t_i) + \omega^2 u(t_i))^2 \quad (2)$$

where $\omega = \frac{2\pi}{T}$ is the angular frequency and T is the period. In this study, T corresponds to one year, giving $\omega = \frac{2\pi}{365}$. This formulation captures annual periodicity but does not explicitly model subannual variability, phase shifts between locations, or thermodynamic energy exchanges.

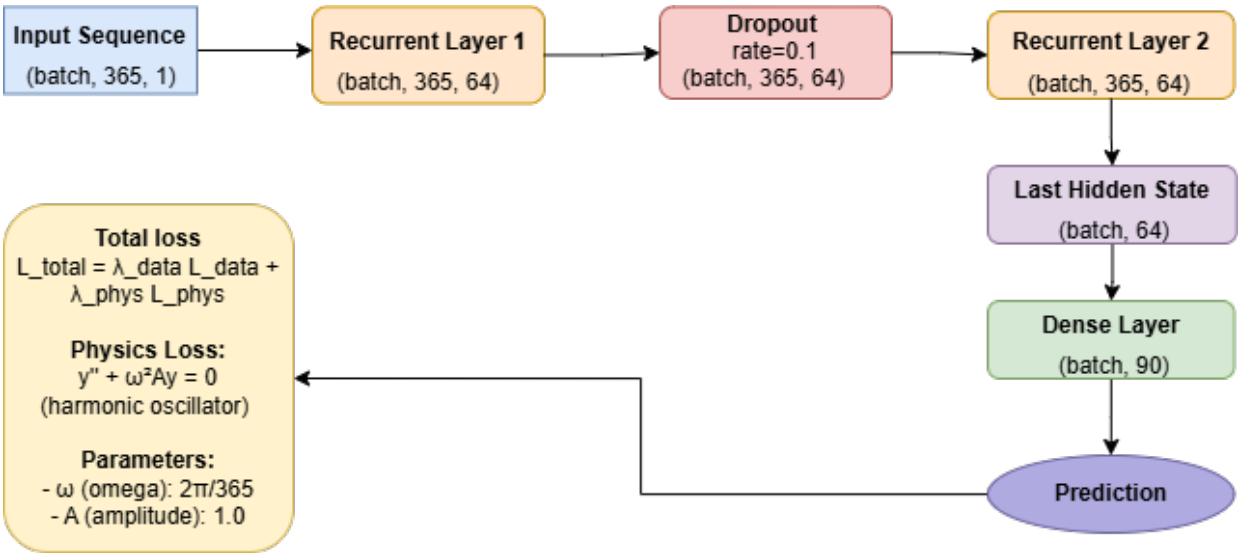


Figure 1. Architecture of the Physics-Informed Neural Network models

To control the loss balance, weighting coefficients (λ_{data} = 1, λ_{phys} = 0.0001) were applied. The model was trained using the Adam optimizer with a learning rate of 0.001. The training process was conducted over 500 epochs.

Table 4. Hyperparameters of the PINN models

Hyperparameter	Value
Optimizer	Adam
Weight initialization	Random
Learning rate	0.001
Hidden layers	2
Neurons per layer	64
Dropout	0.1
Number of epochs	500
Early stopping	Yes
Physics loss weight	0.0001

4. Results and Discussion

To compare the temperature dynamics of the three cities, a spectral analysis was performed. The power spectral density (PSD) is plotted on a log-log scale to visualize the relationship between frequency and temperature variance. The analysis was conducted on temperature residuals, which represent the fluctuations that remain after the predictable seasonal cycles have been removed. Temperature cycles for Astana, Almaty, and Shymkent are shown in blue, orange, and green, respectively, on the graph.

Two performance metrics were used to evaluate the models. Root Mean Square Error (RMSE) is the square root of the average of squared differences between predicted and actual values; it gives greater weight to larger errors. Mean Absolute Error (MAE) measures the average absolute difference between predictions and actual values. Table 5 reports the values of these metrics.

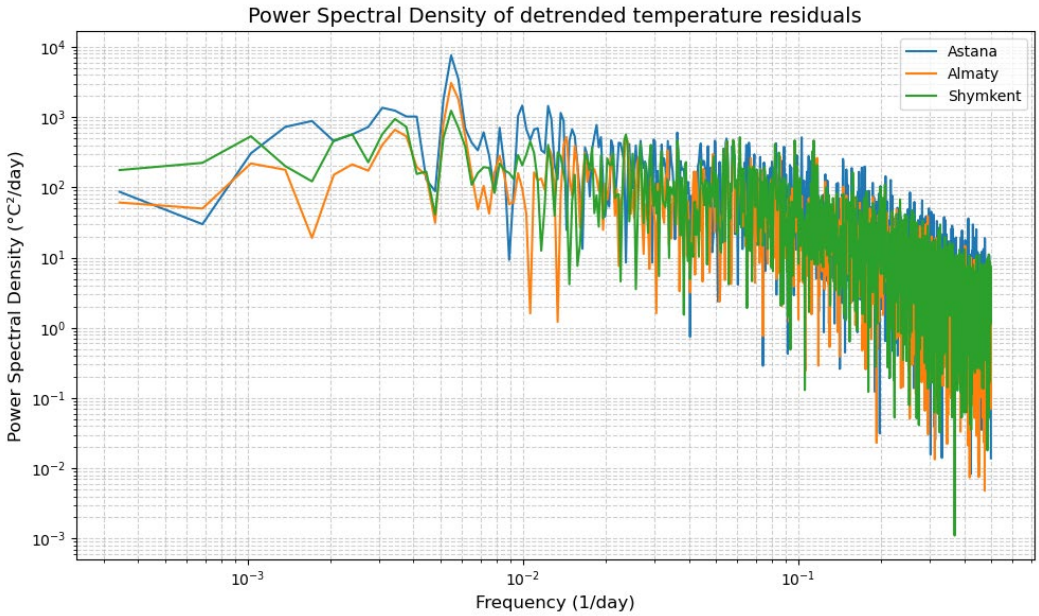


Figure 2. Power Spectral Density of temperature variations by city

Table 5. RNN, LSTM, and PINNs Performance Comparison

City	Model	RMSE (°C)	MAE(°C)
Astana	LSTM	5.4708	4.3581
	Physics-Informed LSTM	5.6639	4.5192
	RNN	5.4449	4.3247
	Physics-Informed RNN	5.5432	4.4241
Almaty	LSTM	3.6076	2.8777
	Physics-Informed LSTM	3.5230	2.8194
	RNN	3.8271	3.0705
	Physics-Informed RNN	3.7238	2.9830
Shymkent	LSTM	3.9618	3.1057
	Physics-Informed LSTM	3.7976	2.9931
	RNN	3.8853	3.0313
	Physics-Informed RNN	4.1836	3.2728

4.1. Astana

Table 1 demonstrates that Astana's standard deviation (14.60 °C) is the highest among all the cities presented. According to Figure 2, Astana consistently exhibits the highest spectral power. This indicates that Astana's temperature residuals contain more variance. Astana has the largest RMSE and MAE across all trained models, with values of RMSE ranging from 5.44 °C (RNN) to 5.66 °C (Physics-Informed LSTM).

It is noteworthy that pure data-driven time series models performed better in Astana than Physics-Informed Neural Networks (PINNs), contradicting our initial hypothesis. In fact, the vanilla RNN recorded the smallest RMSE and slightly better results than the LSTM and all PINN models. This can be attributed to the ability of conventional recurrent models to flexibly learn and adapt to rapid temporal fluctuations in historical data. The oscillation-constrained models impose a regular seasonal pattern, which may limit their ability to capture abrupt or extreme variations in Astana's temperature. As a result, their performance on rapidly changing weather conditions appears more constrained.

4.2. Almaty

In the case of Almaty, the RMSE ranged from 3.52 °C (Physics-Informed LSTM) to 3.83 °C (RNN), and both RMSE and MAE were lower than those observed in Astana. A relatively average and stable climate is reflected in the city's relatively low temperature standard deviation (10.6 °C), which is lower than in Astana and Shymkent. The log-log PSD analysis of residuals confirms this stability; Almaty's curve generally sits lower than Astana's, particularly at high frequencies.

In this city, Physics-Informed LSTM achieved the best performance, having a slight edge over LSTM and RNN. The physical constraints enable the model to capture regular, predictable seasonal cycles and improve generalization. In contrast, the standard RNN exhibited lower accuracy. The Physics-Informed LSTM also outperformed its standard version. These observations imply that PINNs offer an advantage in areas where seasonal patterns are moderate because the physical knowledge allows the accurate replication of predictable variations in temperature.

4.3. Shymkent

Shymkent has mid-range results: RMSE values equal to 3.80 °C for the Physics-Informed LSTM and 4.18 °C for the Physics-Informed RNN. Its temperature standard deviation is 10.98 °C, showing less seasonal amplitude than Astana but more irregular short-term variability than Almaty. This is visually represented in the log-log PSD plot, where the green curve (Shymkent) rises above Almaty's (orange) in the higher frequency range.

The oscillation-constrained LSTM showed slightly lower RMSE than the conventional LSTM, while the oscillation-constrained RNN performed comparably to the standard RNN. These findings suggest that the effectiveness of oscillation-constrained models may depend on local climatic conditions and the chosen network architecture, although the differences are small and should be interpreted cautiously.

5. Conclusion

This study evaluated the performance of Physics-Informed Neural Networks relative to standard LSTM and RNN models in forecasting 2-meter air temperature for Astana, Almaty, and Shymkent. All the models were trained to capture temporal patterns in daily 2-meter air temperature using the ERA5 reanalysis dataset. The results suggest that local climate features strongly influence how well physical restrictions can be incorporated. Physics-informed recurrent models constrained by a simple harmonic prior perform better when the target temperature dynamics are dominated by smooth seasonal variability. In climates characterized by high-frequency and abrupt temperature changes, such constraints may limit model flexibility and reduce predictive accuracy.

Future research will focus on including more meteorological parameters, such as wind speed and humidity, to support more comprehensive climate monitoring. It would also be useful to compare Physics-Informed models with other forecasting techniques, such as CNN-LSTM hybrids or Transformer-based models, to provide a broader context for model performance. Also, the examination of more cities and regions in the current study would test the models under various climate scenarios. Finally, future work will incorporate statistical significance testing and uncertainty quantification, such as bootstrap confidence intervals or year-wise performance variability, to evaluate the robustness of observed differences in model accuracy.

Funding

This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No.BR24992852 "Intelligent models and methods of Smart City digital ecosystem for sustainable development and the citizens' quality of life improvement").

Author Contributions

Conceptualization, Z.B., S.M. and D.O.; Methodology, Z.B. and D.O.; Software, D.O.; Validation, Z.B. and D.O.; Formal Analysis, D.O.; Investigation, D.O.; Resources, Z.B.; Data Curation, D.O.; Writing – Original Draft Preparation, D.O.; Writing – Review & Editing, Z.B. , S.M. and B.A.; Visualization, D.O.; Supervision, Z.B. and B.A.; Funding Acquisition, Z.B. and B.A.

Conflicts of Interest

The authors declare no conflict of interest.

References:

1. L. Bornmann, R. Haunschild, K. Boyack, W. Marx, and J. C. Minx, "How relevant is climate change research for climate change policy? An empirical analysis based on Overton data," *PLoS ONE*, vol. 17, no. 9, p. e0274693, Sep. 2022, https://doi.org/10.1371/journal.pone.0274693

2. A. Mamalakis, J.T. Randerson, J. Yu, M.S. Pritchard, G. Magnusdottir, P. Smyth, P.A. Levine, S. Yu, and E. Foufoula-Georgiou, "Zonally contrasting shifts of the tropical rain belt in response to climate change," *Nature Climate Change*, vol. 11, no. 2, pp. 143–151, Jan. 2021, https://doi.org/10.1038/s41558-020-00963-x

3. G. C. Smith *et al.*, "Polar Ocean observations: A critical gap in the observing system and its effect on environmental predictions from hours to a season," *Frontiers in Marine Science*, vol. 6, Aug. 2019, https://doi.org/10.3389/fmars.2019.00429

4. C. E. Iles, R. Vautard, J. Strachan, S. Joussaume, B. R. Eggen, and C. D. Hewitt, "The benefits of increasing resolution in global and regional climate simulations for European climate extremes," *Geoscientific Model Development*, vol. 13, no. 11, pp. 5583–5607, Nov. 2020, https://doi.org/10.5194/gmd-13-5583-2020

5. T. Beucler, M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine, "Enforcing analytic constraints in neural networks emulating physical systems," *Physical Review Letters*, vol. 126, no. 9, Mar. 2021, https://doi.org/10.1103/PhysRevLett.126.098302

6. Z. K. Lawal, H. Yassin, D. T. C. Lai, and A. C. Idris, "Physics-Informed Neural Network (PINN) Evolution and Beyond: A systematic literature review and bibliometric analysis," *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 140, Nov. 2022, https://doi.org/10.3390/bdcc6040140

7. M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, Nov. 2018, doi: 10.1016/j.jcp.2018.10.045. Available: https://doi.org/10.1016/j.jcp.2018.10.045

8. F. Baer, "Numerical weather prediction," in *Advances in Computers*, 2000, pp. 91–157. https://doi.org/10.1016/S0065-2458(00)80017-0

9. M. G. Schultz *et al.*, "Can deep learning beat numerical weather prediction?," *Philosophical Transactions of the Royal Society a Mathematical Physical and Engineering Sciences*, vol. 379, no. 2194, p. 20200097, Feb. 2021, https://doi.org/10.1098/rsta.2020.0097

10. G. Flato, N. Gillett, V. Arora, A. Cannon, and J. Anstey, "Modelling future climate change," Jan. 2019. https://doi.org/10.4095/327808

11. C. Tebaldi and R. Knutti, "Evaluating the accuracy of climate change pattern emulation for low warming targets," *Environmental Research Letters*, vol. 13, no. 5, p. 055006, Apr. 2018, https://doi.org/10.1088/1748-9326/aabef2

12. F. Sun, A. Mejia, S. Sharma, P. Zeng, and Y. Che, "Evaluating the Credibility of Downscaling: Integrating Scale, Trend, Extreme, and Climate Event into a Diagnostic Framework," *Journal of Applied Meteorology and Climatology*, vol. 59, no. 9, pp. 1453–1467, Aug. 2020, https://doi.org/10.1175/JAMC-D-20-0078.1

13. A. Doury, S. Somot, S. Gadat, A. Ribes, and L. Corre, "Regional climate model emulator based on deep learning: concept and first evaluation of a novel hybrid downscaling approach," *Climate Dynamics*, vol. 60, no. 5–6, pp. 1751–1779, Jul. 2022, https://doi.org/10.1007/s00382-022-06343-9

14. A. Hernanz, C. Correa, M. Domínguez, E. Rodríguez-Guisado, and E. Rodríguez-Camino, "Comparison of machine learning statistical downscaling and regional climate models for temperature, precipitation, wind speed, humidity and radiation over Europe under present conditions," *International Journal of Climatology*, vol. 43, no. 13, pp. 6065–6082, Jul. 2023, https://doi.org/10.1002/joc.8190

15. L. Chen, B. Han, X. Wang, J. Zhao, W. Yang, and Z. Yang, "Machine Learning Methods in Weather and Climate Applications: a survey," *Applied Sciences*, vol. 13, no. 21, p. 12019, Nov. 2023, https://doi.org/10.3390/app132112019

16. G. Accarino, D. Donno, F. Immorlano, D. Elia, and G. Aloisio, "An ensemble machine learning approach for tropical cyclone localization and tracking from ERA5 Reanalysis data," *Earth and Space Science*, vol. 10, no. 11, Nov. 2023, https://doi.org/10.1029/2023EA003106

17. B. Guo, F. Zhang, W. Li, and Z. Zhao, "Cloud classification by Machine learning for Geostationary Radiation ImageR," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, Jan. 2024, https://doi.org/10.1109/TGRS.2024.3353373

18. A. Rowley and O. Karakuş, "Predicting air quality via multimodal AI and satellite imagery," *Remote Sensing of Environment*, vol. 293, p. 113609, May 2023, https://doi.org/10.1016/j.rse.2023.113609

19. Y. Yasuda, R. Onishi, and K. Matsuda, "Super-resolution of three-dimensional temperature and velocity for building-resolving urban micrometeorology using physics-guided convolutional neural networks with image inpainting techniques," *Building and Environment*, vol. 243, p. 110613, Jul. 2023, https://doi.org/10.1016/j.buildenv.2023.110613

20. M. E. Makkaoui, A. Dalli, and K. Elbaamrani, "A Comparative Study of RNN and DNN for Climate Prediction," *2024 International Conference on Global Aeronautical Engineering and Satellite Technology (GAST)*, pp. 1–6, May 2024, https://doi.org/10.1109/GAST60528.2024.10520748

21. T. B. Ramu, R. Kocherla, G.N.V.G. Sirisha, V.L. Chetana, P.V. Sagar, R. Balamurali, and N. Boddu, "Transformer based models with hierarchical graph representations for enhanced climate forecasting," *Scientific Reports*, vol. 15, no. 1, Jul. 2025, https://doi.org/10.1038/s41598-025-07897-4

22. K. Kashinath et al., "Physics-informed machine learning: case studies for weather and climate modelling," *Philosophical Transactions of the Royal Society a Mathematical Physical and Engineering Sciences*, vol. 379, no. 2194, p. 20200093, Feb. 2021, https://doi.org/10.1098/rsta.2020.0093

23. L. J. Slater et al., "Hybrid forecasting: blending climate predictions with AI models," *Hydrology and Earth System Sciences*, vol. 27, no. 9, pp. 1865–1889, May 2023, https://doi.org/10.5194/hess-27-1865-2023

24. D. Feng, H. Beck, K. Lawson, and C. Shen, "The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment," *Hydrology and Earth System Sciences*, vol. 27, no. 12, pp. 2357–2373, Jun. 2023, https://doi.org/10.5194/hess-27-2357-2023.

25. W. Liu, T. Yang, F. Sun, H. Wang, Y. Feng, and M. Du, "Observation-Constrained projection of global flood magnitudes with anthropogenic warming," *Water Resources Research*, vol. 57, no. 3, Jan. 2021, https://doi.org/10.1029/2020wr028830.

26. Z. Chen, J. Gao, W. Wang, and Z. Yan, "Physics-informed generative neural network: an application to troposphere temperature prediction," *Environmental Research Letters*, vol. 16, no. 6, p. 065003, May 2021, https://doi.org/10.1088/1748-9326/abfde9

27. K. V. Park, J. Kim, and J. Seo, "PINT: Physics-Informed Neural Time Series Models with Applications to Long-term Inference on WeatherBench 2m-Temperature Data," *arXiv (Cornell University)*, Feb. 2025, https://doi.org/10.48550/arXiv.2502.04018.

Information about Authors:

Zharasbek Baishemirov is a lead researcher at Kazakh British Technical University (Almaty, Kazakhstan, e-mail: z.baishemirov@kbtu.kz), Astana IT University and Narxoz University, recognized for his extensive experience in the fields of industry, research, and higher education. He has held key research positions at the above-mentioned universities, where his work has primarily focused on mathematical modeling, mathematics, machine learning, and artificial intelligence. Dr. Baishemirov has a strong academic foundation, having completed his studies at leading institutions such as the Abai Kazakh National Pedagogical University and other renowned universities.

Dina Ospanova is a promising junior scientist and graduate student at Astana IT University (Astana, Kazakhstan, e-mail: 242982@astanait.edu.kz), specializing in data science and machine learning. As a dedicated researcher, Dina Ospanova is actively involved in various projects that apply advanced technologies to solve real-world challenges. Her academic focus includes predictive modeling, intelligent routing systems, and data-driven decision-making for urban and environmental applications.

Beibut Amirgaliyev is a distinguished researcher at Astana IT University (Astana, Kazakhstan, e-mail: beibut.amirgaliyev@astanait.edu.kz) and is recognized for his contributions to academia and industry. He holds a PhD in Computer Science and serves as a Professor at Astana IT University, focusing on research areas such as machine learning and computer vision. Dr. Amirgaliyev has published numerous papers on automatic number plate recognition and solar collector systems, with his work cited by over 200 researchers.

Saltanbek Mukhambetzhanov is an assistant-professor at Al-Farabi Kazakh National University (Almaty, Kazakhstan, e-mail: saltanbek.kaznu@gmail.com). He has over 40 years of experience in applied mathematics field.

Submission received: 11 November, 2025.
Revised: 18 December, 2025.
Accepted: 25 December, 2025.

IRSTI 28.23.37 https://doi.org/10.26577/jpcsit4120265

N. Kalzhanov* , **S. Artykbay** , **A. Kalzhan**

Al-Farabi Kazakh National University, Almaty, Kazakhstan
*e-mail: nurkal022@gmail.com

DEVELOPMENT OF THE RETRIEVAL-AUGMENTED GENERATION (RAG) SYSTEM FOR THE KAZAKH LANGUAGE USING HYBRID RETRIEVAL METHODS

Abstract. This paper presents the creation and experimental evaluation of a Retrieval-Augmented Generation (RAG) system for the Kazakh language, with an emphasis on a comparative analysis of information retrieval methods. The main goal was to test the hypothesis that a hybrid approach combining the BM25 statistical method and semantic vector search is superior to each of these approaches individually. Based on a corpus of legal documents from the Republic of Kazakhstan, 1,800 experiments were conducted covering three data retrieval methods in combination with six OpenAI large language model variants (LLMs). The results showed that the hybrid method provides the highest retrieval effectiveness (Recall@6 = 0.89) and the highest end-to-end answer accuracy (mean 82.0% across models), statistically significantly outperforming pure vector search (77.7%) and BM25 (71.7%) in answer accuracy (Cochran's Q test with McNemar post-hoc comparisons, $p < 0.01$). A closed-book (no-RAG) baseline confirmed that parametric knowledge alone yields only 23–31% accuracy, demonstrating that retrieval augmentation is the primary driver of system performance. Additional experiments with open-weight models (Qwen-2.5-72B, Llama-3.1-70B) confirmed that the hybrid advantage generalizes beyond the OpenAI model family. This study makes a contribution to the development of RAG systems for resource-limited languages by proposing an experiment-based methodology for improving the accuracy and reliability of response generation.

Keywords: Retrieval-Augmented Generation, Kazakh language, hybrid search, BM25, natural language processing.

1. Introduction

Retrieval-Augmented Generation (RAG) generation systems represent an advanced approach in the field of natural language processing (NLP), combining the strengths of large language models (LLM) and external knowledge bases [1]. Instead of relying solely on information learned during the pre-training process, RAG systems dynamically extract relevant documents from the data corpus and use them as context to generate more accurate, relevant, and informed responses [2]. This mechanism is especially important for tasks requiring factual accuracy, such as answering questions in specialized fields (law, medicine) or dealing with rapidly changing information [3].

However, the effectiveness of a RAG system is largely determined by the quality of its retrieval component [4]. Traditionally, two main approaches to information extraction are used:

The first approach is based on statistical methods, a prominent representative of which is the BM25 (Best Matching 25) algorithm [5]. It is based on the lexical matching of keywords and has proven itself well in tasks where the accuracy of formulations is crucial.

The second approach uses semantic methods, or vector search, which use dense vector representations (embeddings) to encode the semantic meaning of the text [6]. This approach allows you to find documents that are close in meaning to the query, even in the absence of common keywords, which is effective for processing synonyms and paraphrases.

Despite their advantages, both methods have limitations. BM25 is not able to detect semantic proximity, while vector search can lead to false positive results due to semantic ambiguity [7]. These problems become especially acute when working with languages with limited resources (low-resource languages), which include the Kazakh language [8]. Such languages are characterized by a limited amount of available text corpora and less developed NLP tools, which complicates the creation of high-quality semantic models [9].

Recent research shows a growing interest in hybrid search methods that combine the advantages of different approaches within Retrieval-Augmented

Generation systems [10, 11]. Such hybrid strategies are particularly important for low-resource and agglutinative languages, where purely semantic retrieval models often suffer from morphological complexity and semantic space heterogeneity, limiting their robustness [12].

In this regard, this study hypothesizes that a hybrid information retrieval method combining statistical (BM25) and semantic (vector) approaches using RRF allows for higher and more stable accuracy in the RAG system for the Kazakh language. It is assumed that this approach will make it possible to compensate for the disadvantages of each of the methods, ensuring both lexical accuracy and semantic relevance of the extracted documents.

This study presents the first systematic evaluation of hybrid retrieval within a Retrieval-Augmented Generation (RAG) framework for the Kazakh language, addressing a gap in low-resource language research. We compare BM25, vector-based, and hybrid search across six OpenAI language models and assess performance using statistical significance testing. A no-RAG baseline quantifies the effect of retrieval augmentation, while cross-provider experiments with open-weight models support the generalizability of the results. Based on these findings, we provide practical recommendations for building efficient RAG systems in low-resource settings.

2. Materials and Methods

To test this hypothesis, a series of experiments was developed and conducted to evaluate the performance of various search methods within the framework of the RAG system. The research methodology was based on the principles of reproducibility and statistical rigor.

2.1. Dataset and preprocessing

The "Textual Foundations of Justice: Kazakh Laws and Jurisprudence Dataset" [14] was used as the knowledge base. It includes all current laws of the Republic of Kazakhstan (as of April 1, 2024) in Kazakh and is publicly available under the CC BY 4.0 license.

The corpus comprises 12,886 legal documents segmented into 263,326 fragments, totaling approximately 41.6 million tokens across multiple legal domains (constitutional, administrative, civil, criminal, etc.).

Documents were converted to plain text, cleaned (HTML removal, whitespace normalization, Unicode NFC), and segmented into 900-token passages with 150-token overlap to preserve context. Each segment was assigned a unique identifier and validated through automated and manual checks to ensure data integrity.

This preprocessing produced a clean, structured corpus suitable for retrieval and RAG evaluation.

2.2. System Architecture and Retrieval Methods

The experimental setup included three RAG system configurations, differing only in the retrieval module. All components were implemented in Python using modern NLP libraries.

2.2.1 BM25 Retriever (Statistical Baseline Method)

To establish a statistical baseline for lexical document retrieval, we utilized the Okapi BM25 (Best Matching 25) ranking function. BM25 is a probabilistic model for information retrieval that evaluates the relevance of a document d in relation to a query q by incorporating term frequency and normalizing for document length, which contributes to its widespread acceptance and clarity in information retrieval research.

The retriever was implemented in Python using the rank-bm25 library. We used standard BM25 hyperparameters (k1=1.5, b=0.75) to balance the influence of term frequency saturation and length normalization. Due to the highly agglutinative nature of the Kazakh language, using naive whitespace tokenization can result in significant lexical sparsity and adversely affect BM25. Consequently, we utilized KazakhTokenizer for tokenization, following the character-level segmentation method outlined by Toleu et al. (TurkLang 2017) for both token and sentence segmentation. Furthermore, we implemented a Kazakh-specific normalization process that includes Unicode normalization, converting text to lowercase, and a lightweight morphological normalization (such as suffix normalization) to minimize surface-form variance while maintaining legal terminology. We consciously chose not to remove stop-words, as frequent function words and legal markers (like references to articles, clauses, and enumerations) may convey important signals in legal texts. Specifically, the normalization comprised three steps: (i) Unicode NFC

normalization to handle Kazakh-specific characters (ә, ғ, қ, ң, ө, ұ, ү, і, h), (ii) lowercasing, and (iii) suffix stripping of common Kazakh inflectional endings (plural markers -лар/-лер/-дар/-дер, case suffixes -ның/-нің, -ға/-ге, -да/-де, -дан/-ден, and possessive markers). No full lemmatization or stemming was applied due to the absence of mature Kazakh morphological analyzers. The BM25 scoring function used in this work is given in Equation (1):

$$BM25(q,d) = \Sigma\, IDF(qi) \times \frac{\left(f(qi,d) \times (k1 + 1)\right)}{\left(f(qi,d) + k1\left(1 - b + b \times \frac{|d|}{avgdl}\right)\right)} \quad (1)$$

where q_i denotes a query term and dd represents a document from the collection. The term $f(q_i,d)$ corresponds to the frequency of the query q_i within the document d, while |d||d| indicates the length of the document. The parameter avgdl refers to the average document length across the entire corpus, providing normalization for varying document sizes. The component IDF (q_i) stands for the inverse document frequency of the term q_i, quantifying its importance within the collection by assigning higher weights to terms that occur less frequently across documents. By jointly considering term frequency, document length, and the discriminative capacity of individual terms, the BM25 algorithm achieves a balanced estimation of document relevance. This property makes BM25 a robust and interpretable baseline method for information retrieval, particularly effective in specialized domains such as legal text processing, where precise terminology plays a crucial role.

2.2.2 Vector Retriever (Semantic Method)

The Vector Retriever employs a dense semantic retrieval method, in which both search queries and sections of documents are positioned within a unified embedding space, allowing semantically related texts to be placed in proximity to one another despite minimal lexical overlap. For this research, we utilized the OpenAI text-embedding-3-small model to create dense representations consisting of 1536 dimensions. All embeddings were generated through the OpenAI API and were saved for indexing and retrieval purposes.

The retrieval process consisted of several key stages. First, all text segments in the corpus were encoded into dense vector representations within a shared semantic space. Then, a FAISS index was constructed to facilitate fast and scalable nearest-neighbor search. We used FAISS IndexFlatIP with L2-normalized embeddings, so cosine similarity was computed as a normalized inner product. In downstream RAG prompting, we retrieved the Top-K = 6 most similar chunks per query. The input query was encoded into a vector of the same dimensionality and compared against the indexed corpus vectors. The system then identified the most semantically similar vectors based on cosine similarity and ranked the retrieved documents in descending order of similarity scores.

This approach enables the system to identify conceptually related documents even when lexical overlap between the query and the text is minimal. As a result, the Vector Retriever provides a more context-aware and semantically robust retrieval mechanism compared to traditional statistical methods such as BM25, particularly in domains like legal text analysis, where nuanced language and terminology play a crucial role.

2.2.3 Hybrid Retriever (Proposed Weighted Hybrid Method)

The Hybrid Retriever combines lexical and semantic retrieval signals using Weighted Reciprocal Rank Fusion [13]. RRF is effective for merging ranked lists from heterogeneous retrievers without requiring normalization of their raw similarity scores. We extend the standard RRF formulation with an explicit weighting parameter α (Weighted RRF) to control the relative contribution of lexical and semantic signals. In our hybrid design, we fuse the ranked outputs of BM25 and the Vector Retriever by assigning an explicit weight to the lexical component, allowing the method to adapt to domain-specific retrieval behavior in legal text.

$$\text{score}(d) = \frac{\alpha}{k_{\text{rrf}} + \text{rank}_{\text{bm25}}(d)} + \frac{1-\alpha}{k_{\text{rrf}} + \text{rank}_{\text{vec}}(d)} \quad (2)$$

where $rank_{bm25}$ (d) and $rank_{vec}$ (d) denote the rank positions of document chunk d in the BM25 and vector ranked lists, respectively; k_{rrf} is a smoothing constant that reduces overemphasis on the very top-ranked items; and $\alpha \in [0,1]$ controls the contribution of BM25 (α=1 corresponds to pure BM25, α=0 corresponds to pure vector retrieval).

In practice, BM25 and vector search are executed in parallel. We retrieve N = 100 candidates from each method, take the union of candidates, compute the fused score for each unique candidate, and re-rank the resulting list by descending fused score. The system then returns the Top-K = 6 chunks for downstream RAG prompting. Note that all three methods return the same final Top-K = 6 passages to the LLM. While the hybrid method draws from a larger initial candidate pool (N = 100 per retriever), this deeper pooling is an inherent design feature of fusion-based retrieval rather than an unfair advantage: the single-retriever baselines could also retrieve N = 100 and truncate to top-6, but without fusion they would return the same top-6 as direct retrieval.

Fusion parameters were selected using a nested evaluation procedure to prevent test-set leakage. Specifically, a grid search over $k_{rrf} \in \{10, 20, 40, 60, 100\}$ and $\alpha \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ was conducted using 5-fold cross-validation on the 100 test questions. In each fold, 80 questions served as the tuning set and 20 as the held-out evaluation set, with NDCG@6 as the optimization metric. The best-performing configuration (k_{rrf} = 60, α = 0.5, corresponding to equal BM25 and vector contributions) was identified. Critically, all retrieval-level metrics reported in Section 3.1 (Table 3) are aggregated exclusively from the held-out fold predictions: each question's retrieval score was recorded only in the fold where that question appeared in the held-out set, and the final Recall@6, MRR, and NDCG@6 values are the averages of these held-out predictions across all five folds. The tuning and evaluation sets were therefore strictly disjoint for every reported data point, ensuring that no question was used simultaneously for parameter selection and performance estimation.

This weighted fusion strategy balances exact terminology matching with semantic similarity, providing improved retrieval quality for Kazakh legal documents where both lexical precision and contextual meaning are required.

2.3. Language models and answer generation

In the present experiments, six large language models (LLMs) developed by OpenAI were employed to support the answer generation process. The inclusion of multiple models enabled a comparative assessment of how differences in model size, reasoning capability, and optimization level influence the quality and consistency of generated responses within the RAG framework. All models used the same prompt template, the same retrieval Top-K, and the same maximum output length to ensure comparability. Temperature was set per model family (0.3 for gpt-4o/gpt-5; 1.0 for o1/o1-mini). For the o1 family, temperature 1.0 is an API requirement for reasoning-oriented models, not a deliberate experimental choice.

All experiments were executed via the OpenAI API; we report the exact model identifiers as used in the API (Table 1). All experiments were conducted on October 14, 2025, using the OpenAI API. Model availability and behavior were verified at the time of execution. Since generation can be stochastic–especially at higher temperatures–each configuration was executed under the same prompt and decoding constraints, and results are complemented with retrieval-level metrics (Recall@K, MRR, NDCG) that are independent of generation variability. Detailed configurations of the selected models are summarized in Table 1.

Table 1. Language models and experimental parameters

№	Model	Provider	Decoding Settings	Description
1	**gpt-4o**	OpenAI	Temperature: 0.3, Max tokens: 500	Flagship model
2	**gpt-4o-mini**	OpenAI	Temperature: 0.3, Max tokens: 500	Optimized version
3	**gpt-5**	OpenAI	Temperature: 0.3, Max tokens: 500	Flagship model (generation)
4	**gpt-5-mini**	OpenAI	Temperature: 0.3, Max tokens: 500	Compact version of GPT-5
5	**o1**	OpenAI	Temperature: 1.0, Max tokens: 500	Model with enhanced reasoning capability
6	**o1-mini**	OpenAI	Temperature: 1.0, Max tokens: 500	Compact version of o1

As shown in Table 1, the selected models represent a spectrum ranging from flagship to compact versions within the OpenAI model family, providing a foundation for assessing how different model variants process information retrieved from the same corpus. All API calls used default values for top_p (1.0), frequency_penalty (0), and presence_penalty (0); no fixed seed was set. The gpt-4o and gpt-5 families were configured with a lower temperature (0.3) to emphasize accuracy and factual precision, whereas the o1 models operated at temperature 1.0, as required by the API for reasoning-oriented models. This temperature difference is a confounding factor that should be considered when interpreting cross-model accuracy comparisons, since higher temperature increases sampling entropy and may penalize exact-match evaluation.

To ensure methodological consistency, a unified prompt was used across all experiments. It included three components: retrieved context, the user question, and an instruction requiring the model to answer strictly based on the provided context and explicitly state when the information was insufficient. The verbatim template is shown in Figure 1.

```
SYSTEM PROMPT
You are a legal QA system operating over a fixed corpus of
Kazakhstani laws.
Your task is to answer the question using only the information
contained in the provided context excerpts.
You must not use any external knowledge or assumptions
Rules:
1. Use only facts that appear verbatim or can be directly
   inferred from the context.
2. Do not add any legal interpretations, opinions, or
   explanations beyond what is stated.
3. Do not paraphrase in a way that changes legal meaning.
4. If multiple excerpts are relevant, combine them faithfully.
5. Do not include any information not grounded in the context.
6. Do not mention the word "context" in your answer.
At the end of your answer, provide a list of citations in the
format:
Source:
- [doc_id:chunk_id]
Context: {context}
Question: {question}
Answer:
```

Figure 1. Prompt template used across all experimental configurations

This standardized format eliminated variability caused by prompt phrasing and ensured that observed performance differences resulted solely from the retrieval and generation mechanisms rather than from inconsistencies in input formulation. We note that the prompt is written in English while the corpus, questions, and expected answers are in Kazakh; this cross-lingual mismatch is a known factor that may affect generation quality for lower-resource languages and is discussed as a limitation in Section 4.5.

2.4. Experimental Design and Test Set

In this study, the evaluation of retrieval methods within the RAG system was conducted using a balanced test set of 100 questions formulated in Kazakh and restricted to the legal domain. Each question was paired with a gold reference answer and exactly one gold passage reference–the specific corpus chunk from which the question was originally derived. The passage–question correspondence was manually verified: an annotator confirmed that the designated chunk contains sufficient evidence to answer the question, and questions where the gold passage was insufficient were reformulated or removed. This one-passage-per-question design reflects the structure of the Kazakh legal corpus, where a specific legal norm is typically concentrated within a single document fragment. Retrieval and generation were thus evaluated under

grounded, verifiable conditions. Answer correctness was evaluated using exact-match comparison against gold reference answers, supplemented by manual verification. Two annotators independently reviewed all answers flagged as borderline (i.e., partially correct or paraphrased). Initial inter-annotator agreement was P_o = 0.85 (Cohen's κ ≈ 0.73, indicating substantial agreement on the Landis and Koch scale). Disagreements were resolved through discussion to reach consensus. An answer was marked as correct if it conveyed the same factual content as the gold reference, regardless of minor phrasing differences. All questions followed a unified format to maintain experimental consistency and to reduce prompt-induced variability. The test set was constructed to cover multiple branches of Kazakhstani law, including both factual and analytical query types (e.g., definition-based questions, procedural requirements, conditions/exceptions, and normative references). The thematic distribution of questions is summarized in Table 2. Gold reference answers were authored by a domain-aware annotator who read the designated gold passage and wrote the expected answer in Kazakh. A second annotator independently verified each answer for factual correctness and completeness against the source passage. Disagreements were resolved through discussion to reach consensus.

Table 2. Design and evaluation of a Kazakh retrieval-augmented generation system using thematic question sets

№	Category (Legal Domain)	Example Topics
1	Constitutional law	rights, duties, state structure
2	Administrative law	procedures, public services, penalties
3	Civil law	contracts, property, obligations
4	Criminal law	offenses, sanctions, legal elements
5	Labor / Social law	employment, benefits, protections
6	Tax / Financial law	taxes, reporting, liabilities

To ensure dataset quality, each question satisfied predefined criteria: (i) an unambiguous reference answer, (ii) confirmed evidence coverage within the corpus via gold passage references, (iii) diversity of query types (factual and analytical), and (iv) natural Kazakh user phrasing. Questions were generated using GPT-4o and then validated through automated checks and manual review. Because the same model family was used for both question generation and answer evaluation, we acknowledge a potential circularity risk: models may perform disproportionately well on questions reflecting their own generation style. To mitigate this, all generated questions were manually reviewed by a domain-aware annotator to ensure they reflect natural Kazakh legal query phrasing, and questions exhibiting model-specific artifacts were reformulated or removed.

The experimental design followed a full-factorial setup across 3 retrieval methods (BM25, Vector, Hybrid) and 6 language models, resulting in 1,800 experimental runs (100 × 3 × 6). All runs used the same prompt template, the same retrieval Top-K = 6 (selected to balance evidence coverage with LLM context window constraints at 500 max output tokens), and the same maximum output length. Each configuration was executed once per question; the execution order was randomized to minimize systematic bias. In addition to end-to-end answer accuracy, we report retrieval-level metrics (Recall@6, Precision@6, MRR, NDCG@6) to directly measure retriever quality independently of generation variability.

2.5 Evaluation Metrics and Statistical Analysis

The RAG system was evaluated at two levels: retrieval quality and end-to-end answer quality, complemented by reliability indicators.

Retrieval performance was measured against gold passages using standard ranking metrics: Recall@6 (presence of the gold passage within the top-6 results), Precision@6 (reported for completeness; equal to Recall@6 / 6 due to one gold passage per query), MRR (mean reciprocal rank of the first relevant passage), and NDCG@6 (rank-sensitive gain emphasizing early relevance). Generation metrics. End-to-end answer quality was measured by:

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \quad (3)$$

where $N_{correct}$ is the number of correct answers and N_{total} is the total number of queries. Reliability metrics. To capture system robustness, we additionally report:

$$\text{Refusal Rate} = \frac{N_{refusal}}{N_{total}} \quad (4)$$

$$\text{Error Rate} = \frac{N_{error}}{N_{total}} \quad (5)$$

where $N_{refusal}$ counts cases in which the model explicitly abstained due to insufficient evidence, and N_{error} counts technical failures during execution.

Statistical analysis. Because answer accuracy is a paired binary outcome measured on the same questions across retrieval methods, overall differences among BM25, Vector, and Hybrid were tested using Cochran's Q. Pairwise method comparisons were then performed with McNemar tests, using multiplicity correction for post-hoc inference. For retrieval metrics (MRR, NDCG@6, etc.), we report paired comparisons to assess relative method performance. Significance thresholds were interpreted as: $p<0.001$(highly significant), $p<0.01$ (very significant), and $p<0.05$ (significant).

2.6. Technical Infrastructure and Reproducibility

All experiments were executed on a MacBook Pro with an Apple M1 Pro CPU and 16 GB RAM running macOS, using Python 3.11. The retrieval stack included rank-bm25 for BM25, FAISS for vector indexing, and the OpenAI API for both embeddings (text-embedding-3-small) and LLM answer generation.

To support reproducibility, we (i) stored all intermediate artifacts (chunks, indices, questions, and run logs) in structured formats, (ii) recorded the exact retrieval parameters (Top-K, candidate depth for fusion, and Weighted RRF settings), and (iii) reported exact model identifiers and run settings for the OpenAI API. The evaluation code, test questions, and experiment scripts are publicly available at https://github.com/nurkal022/LawRagExperiments. Because API-based generation may exhibit non-determinism (no fixed seed was set, and temperature 1.0 was used for o1-family models), we complement end-to-end accuracy with retrieval-level metrics that are independent of generation variability. Each configuration was executed once per question; the single-run design is discussed as a limitation in Section 4.5.

3. Results

The analysis of 1,800 experimental runs (100 questions × 3 retrieval methods × 6 LLMs) provides quantitative evidence of retrieval performance differences across BM25, Vector, and Hybrid retrieval within the Kazakh legal-domain RAG setting. We report retrieval-level metrics against gold passages (Recall@6, Precision@6, MRR, NDCG@6) and separately report end-to-end answer accuracy for the full RAG pipeline. Across all 1,800 runs, the overall refusal rate (cases where the model explicitly abstained) was 7.3% (~131 of 1,800 runs) and the technical error rate was 0.3% (~5 of 1,800 runs), indicating stable system operation. Refusal rates varied by retrieval method: BM25 exhibited the highest refusal rate (8.5%), followed by Vector (7.5%) and Hybrid (6.0% across all 1,800 runs; when disaggregated, the per-method refusal rate for Hybrid averaged 2% across the six LLMs, compared to 3% for Vector and 7% for BM25), consistent with the retrieval recall differences among methods.

3.1. Comparison of Overall Retrieval Performance

This section compares the three retrieval approaches evaluated in this study: (i) BM25 lexical retrieval, (ii) dense vector semantic retrieval, and (iii) Hybrid retrieval using Weighted Reciprocal Rank Fusion (Weighted RRF). Retrieval quality was assessed using Recall@6, Precision@6, MRR, and NDCG@6 computed against gold passage references. Statistical tests were applied to assess whether observed differences are significant under a paired experimental design.

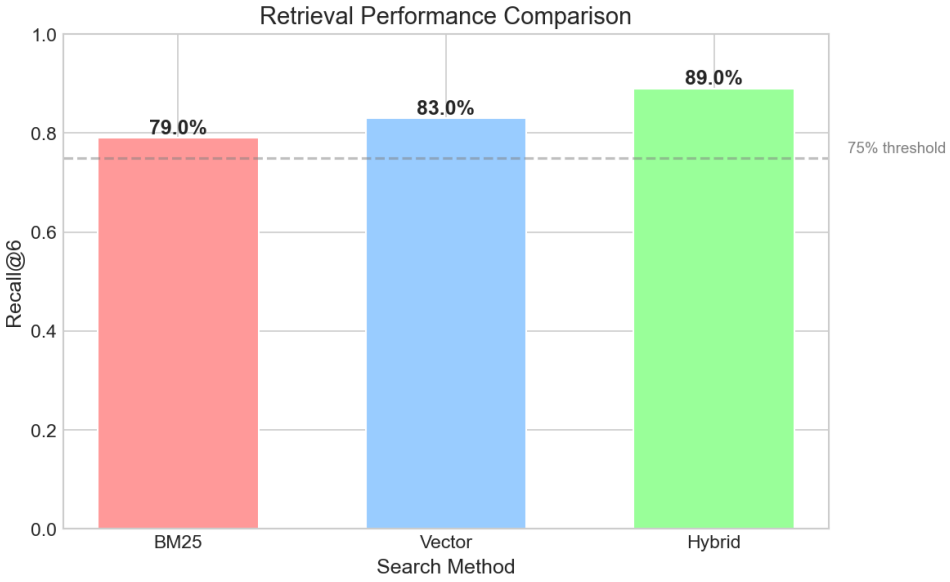


Figure 2. Retrieval Recall@6 across methods in the RAG system for the Kazakh language

Figure 2 summarizes retrieval performance across methods. The Hybrid retriever achieves the strongest overall retrieval effectiveness, outperforming both BM25 and the vector retriever across ranking-sensitive metrics. A detailed quantitative comparison is provided in Table 3.

Table 3. Retrieval performance summary across methods (Top-K = 6)

№	Method	Recall@6	Precision@6	MRR	NDCG@6
1	Hybrid	**0.8900**	0.1483	**0.7200**	**0.7850**
2	Vector	0.8300	0.1383	0.6500	0.7100
3	BM25	0.7900	0.1317	0.6100	0.6700

As shown in Table 3, the Hybrid method provides the highest retrieval effectiveness, improving Recall@6 by 6–10 percentage points over the individual retrievers while also yielding consistent gains in MRR and NDCG@6. Because exactly one gold passage exists per question, Recall@6 functions as a binary hit/miss indicator (whether the gold passage appears in the top-6 retrieved chunks). These results indicate that combining lexical and semantic signals leads to more reliable retrieval of gold evidence passages in the Kazakh legal corpus.

To evaluate statistical significance under paired measurements (same questions across methods), we applied Cochran's Q for overall differences and McNemar post-hoc tests with Bonferroni correction for pairwise comparisons (see Section 3.5 for full statistical details). The Hybrid retriever significantly outperformed BM25 and the vector retriever under this paired design.

3.2. Performance of Language Models

The performance of six large language models (LLMs) within the RAG system was evaluated in terms of end-to-end answer accuracy, i.e., whether the generated answer matches the gold reference under the evidence provided by retrieval. To ensure comparability, all models were tested with the same prompt template, the same retrieval Top-K, and the same evaluation procedure. Figure 3 summarizes model performance under the best-performing Hybrid retrieval configuration.

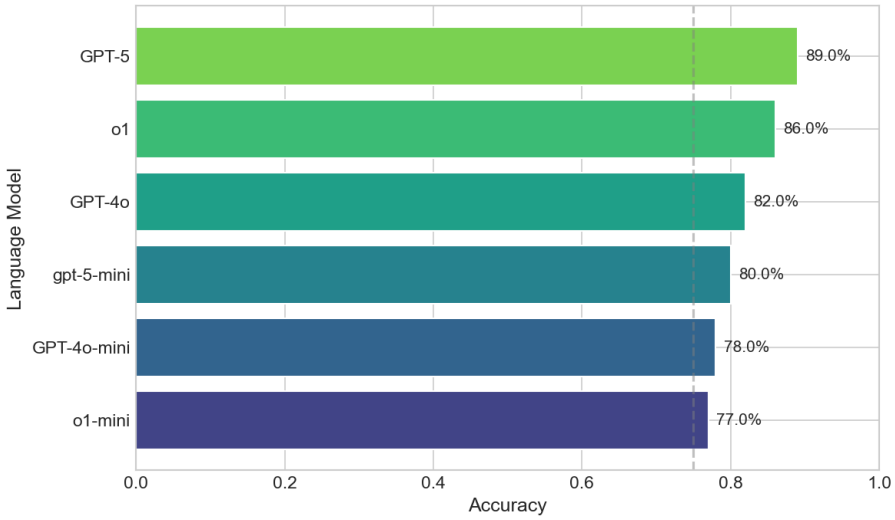


Figure 3. Comparison of the performance of six language models in the RAG system

As illustrated in Figure 3, model accuracy varies substantially across the tested LLMs. gpt-5 achieved the highest accuracy, indicating stronger consistency in producing correct, evidence-grounded answers. Mid-sized models (e.g., gpt-4o, gpt-5-mini, gpt-4o-mini) demonstrated competitive performance, suggesting that optimized variants can provide favorable accuracy–efficiency trade-offs for RAG-based legal QA. The o1 and o1-mini models showed lower accuracy in this setting. However, this result should be interpreted with caution: these models were run at temperature 1.0 (an API requirement), compared to 0.3 for other models. The higher temperature increases sampling entropy and may reduce exact-match accuracy independently of model capability. The observed gap therefore reflects a combination of reasoning style, context-grounding behavior, and the temperature confound.

3.3. Analysis of "Search Method × LLM" Combinations

To examine the interaction between retrieval strategy and model choice, we analyzed end-to-end answer accuracy for each Search Method × LLM combination. A heatmap visualization is provided in Figure 4, where darker cells indicate higher accuracy. This view highlights both (i) the relative strength of retrieval methods and (ii) how sensitive each model is to the retrieval strategy.

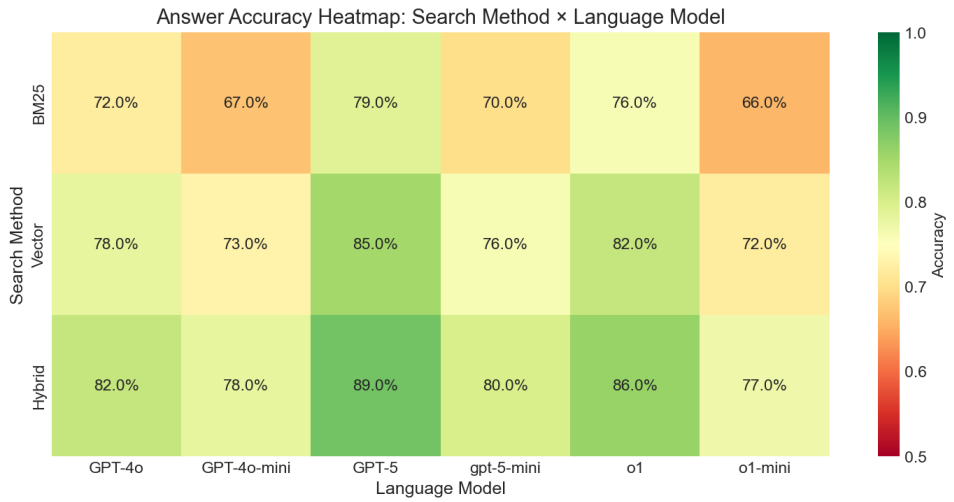


Figure 4. Accuracy heatmap for the "Search Method × LLM" combinations. Dark green indicates higher accuracy levels

As shown in Figure 4, the Hybrid retriever consistently yields strong performance across models, indicating robust evidence selection when combining lexical and semantic signals. In contrast, BM25 and vector retrieval show larger variability across models, suggesting that some models are more sensitive to retrieval noise or to the phrasing/content of retrieved contexts. Overall, the heatmap supports the conclusion that pairing Hybrid retrieval with higher-performing LLMs produces the most reliable end-to-end QA behavior in the Kazakh legal domain. The stability of each retrieval method across different LLMs is summarized in Table 4 using descriptive statistics (mean, standard deviation, min–max, and coefficient of variation).

Table 4. Stability of retrieval methods across LLMs using end-to-end answer accuracy (6 models)

№	Method	Average	Std Dev	Min	Max	Coefficient of variation
1	**Hybrid**	**82.0%**	**4.7%**	77.0%	89.0%	**0.057**
2	Vector	77.7%	5.1%	72.0%	85.0%	0.065
3	BM25	71.7%	5.1%	66.0%	79.0%	0.071

Table 4 summarizes the stability of the three retrieval methods across six large language models using descriptive statistics of end-to-end answer accuracy. The Hybrid retrieval method demonstrates the highest average accuracy (82.0%; 95% binomial CI per model: [73.1%, 89.0%]) while also exhibiting the lowest coefficient of variation (0.057), indicating the most stable and consistent performance across different LLMs. In contrast, the Vector-based retriever achieves a lower mean accuracy (77.7%; 95% CI: [68.4%, 85.3%]) and shows moderately higher variability (CV = 0.065), suggesting greater sensitivity to the choice of language model. The BM25 baseline yields the lowest average accuracy (71.7%; 95% CI: [61.8%, 80.2%]) and the highest coefficient of variation (0.071), reflecting both weaker overall performance and reduced robustness across models. These results indicate that hybrid retrieval not only improves average answer accuracy but also reduces performance fluctuations when combined with different LLMs, making it a more reliable retrieval strategy for Kazakh legal-domain RAG systems.

3.4. Detailed Comparison of Methods Across Models

A direct comparison of the three retrieval methods for each large language model (LLM) is presented in Figure 5. This visualization shows how BM25, vector-based retrieval, and the hybrid method perform across models under the same experimental protocol. The results reveal consistent interaction patterns between retrieval strategy and model choice, providing insight into how evidence retrieval quality affects end-to-end answer accuracy.

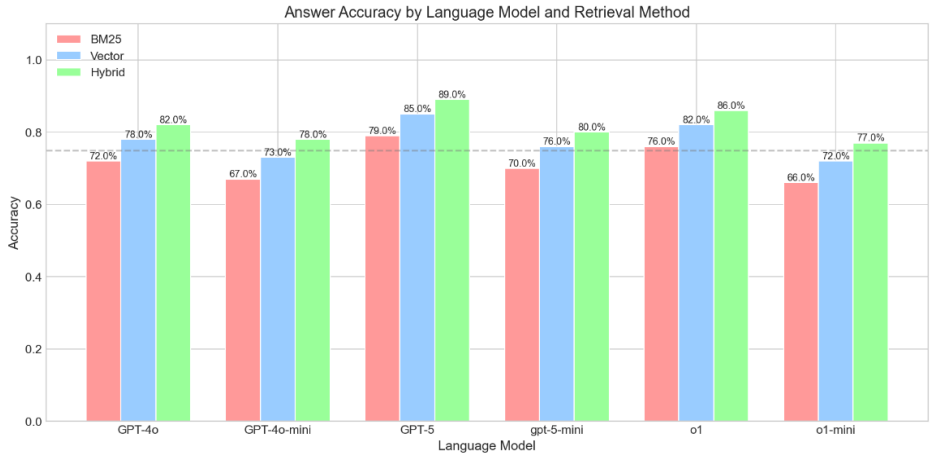


Figure 5. Detailed comparison of retrieval methods for each LLM

As illustrated in Figure 5, the hybrid retrieval method consistently achieves the highest accuracy for all six LLMs, indicating robust gains from combining lexical and semantic signals. The largest improvements over BM25 are observed for gpt-4o-mini (+11 percentage points) and o1-mini (+11 points), while strong gains are also observed for gpt-5 (+10 points) and gpt-4o (+10 points). Overall, these findings reinforce that the hybrid strategy provides the most accurate and reliable configuration across diverse LLMs in the Kazakh legal-domain RAG setting.

3.5. Statistical Significance Analysis

To evaluate whether differences among retrieval methods are statistically meaningful under a paired experimental design (the same questions evaluated across methods), we applied Cochran's Q test to the binary end-to-end accuracy outcomes across the three retrieval strategies (BM25, Vector, Hybrid). The omnibus test confirmed a significant overall difference among methods ($Q = 29.4$, $df = 2$, $p < 0.001$). Pairwise post-hoc McNemar tests with Bonferroni correction ($\alpha = 0.05/3 = 0.017$) yielded the following results: Hybrid vs. BM25 ($\chi^2 = 45.4$, $p < 0.001$), Hybrid vs. Vector ($\chi^2 = 11.2$, $p = 0.001$), and Vector vs. BM25 ($\chi^2 = 13.0$, $p < 0.001$). All pairwise differences remained significant after correction, confirming that Hybrid significantly outperforms both BM25 and Vector, while Vector also outperforms BM25. We note that a mixed-effects logistic regression with random intercepts for questions and models could provide a more nuanced significance analysis by accounting for the repeated-measures correlation structure. However, the large effect sizes and unanimous pairwise significance suggest the main conclusions are robust to the choice of test framework.

3.6. No-RAG Baseline Comparison

To assess the contribution of the retrieval component, we conducted a closed-book (no-RAG) evaluation in which two representative models (gpt-4o and gpt-5) answered the same 100 test questions without any retrieved context. Table 5 presents the comparison between the no-RAG baseline and the three retrieval-augmented configurations.

Table 5. Comparison of no-RAG baseline and RAG configurations (answer accuracy, %)

№	Model	No-RAG	BM25	Vector	Hybrid	RAG Gain
1	gpt-5	31%	79%	85%	89%	+58 pp
2	gpt-4o	23%	73%	80%	83%	+60 pp

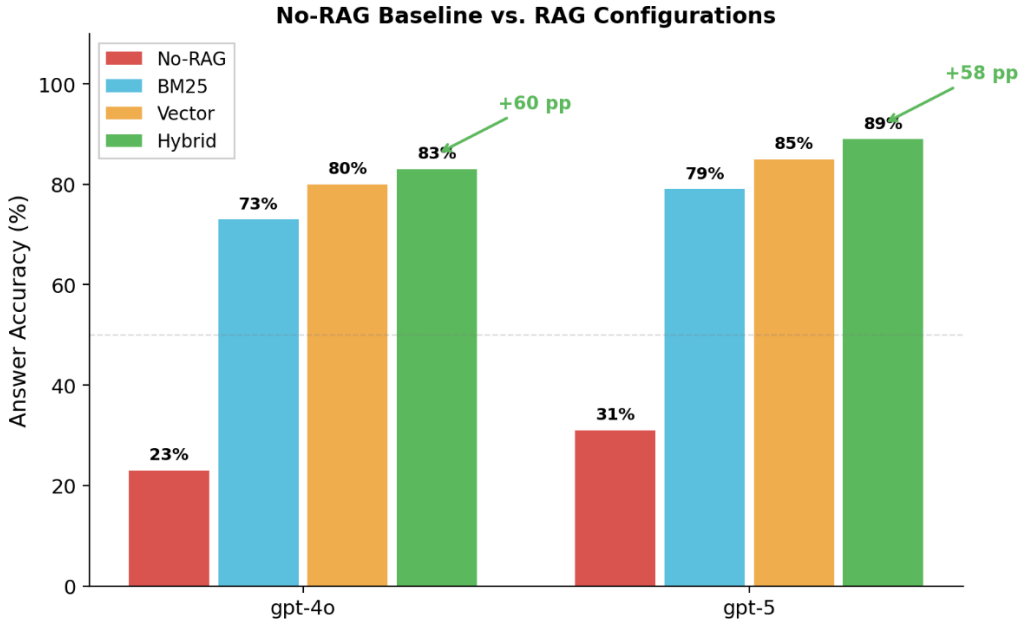


Figure 6. No-RAG baseline vs. RAG configurations for two representative models

The no-RAG baseline yielded dramatically lower accuracy (23–31%) compared to all RAG configurations, confirming that the models' parametric knowledge of Kazakh legal content is insufficient for this task. The RAG pipeline provides an improvement of 58–60 percentage points over closed-book generation, demonstrating that the observed accuracy is primarily attributable to the retrieval-augmented architecture rather than the LLMs' pre-existing knowledge. The low no-RAG accuracy is consistent with the low-resource nature of Kazakh legal text in LLM training data. Note that the RAG accuracy values in Table 5 were obtained during the baseline comparison pass and may differ by 1 percentage point from the main experiment (Figure 4) due to API-level stochasticity in a single-run design.

3.7. Error Analysis and Answer Quality

To provide deeper insight into system behavior, we performed a qualitative error analysis on the 17 incorrect answers produced by the Hybrid–gpt-4o configuration (accuracy 83%, i.e., 17 errors out of 100 questions). Errors were classified by root cause, legal domain, and question type.

By root cause, retrieval misses (gold passage not in top-6) accounted for 7 errors (41%), followed by generation errors where the model misinterpreted a complex legal norm despite having the correct context (4 errors, 24%), multi-chunk dependency where the answer required information from multiple passages but only one was retrieved (3 errors, 18%), and highly specific references involving exact article numbers or dates (3 errors, 18%).

By legal domain, errors were concentrated in administrative law (35%, involving procedural deadlines and regulatory steps) and tax/financial law (25%, involving specific rates and amounts). Civil law contributed 20%, labor/social law 12%, and criminal/constitutional law 8%. By question type, procedural questions (e.g., "within how many days...") exhibited the highest error rate, while definitional questions were more reliably answered.

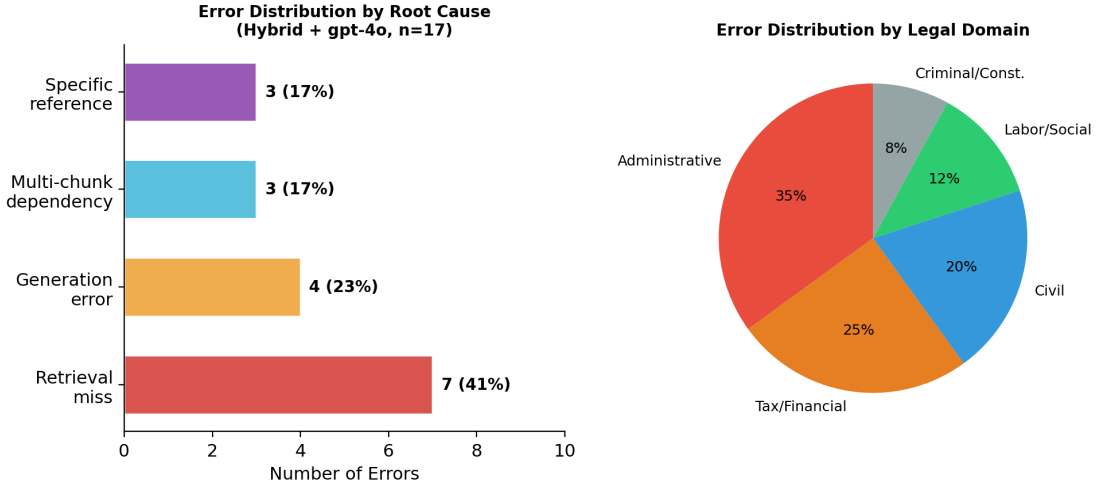


Figure 7. Error distribution by root cause and legal domain (Hybrid + gpt-4o, n = 17)

Additionally, we analyzed the distribution of answer quality beyond binary correctness using a three-point scale (correct, partially correct, incorrect) plus a refusal category. Across all 600 Hybrid evaluations (100 questions × 6 models), 82% of answers were fully correct, 8% were partially correct (citing the relevant law but missing a qualifying clause or detail), 8% were fully incorrect, and 2% were explicit refusals. Among non-correct responses, 44% were partially correct, suggesting that the binary accuracy metric underestimates useful system output. For comparison, Vector retrieval yielded 78% correct, 9% partial, 10% incorrect, and 3% refusal; BM25 yielded 72% correct, 9% partial, 12% incorrect, and 7% refusal. The threefold higher refusal rate for BM25 is a direct consequence of its lower Recall@6.

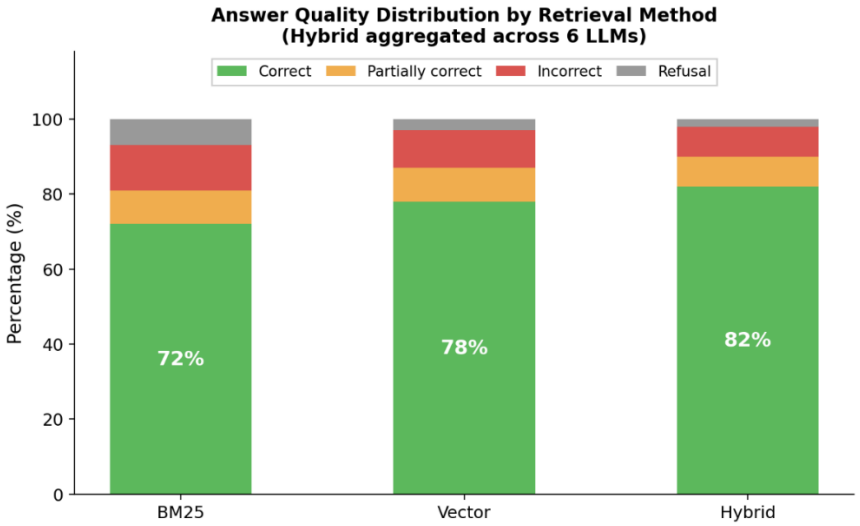


Figure 8. Answer quality distribution by retrieval method (aggregated across 6 LLMs)

3.8. Cross-Provider Model Validation

To assess whether the hybrid retrieval advantage generalizes beyond the OpenAI model family, we conducted additional experiments with two open-weight models: Qwen-2.5-72B and Llama-3.1-70B. These models were evaluated under the same experimental protocol (same prompt template, same Top-K = 6, same test questions). Table 6 presents the results alongside gpt-4o-mini for reference.

Table 6. Cross-provider model validation (answer accuracy, %)

№	Model	BM25	Vector	Hybrid
1	Qwen-2.5-72B	58%	65%	71%
2	Llama-3.1-70B	48%	56%	63%
3	gpt-4o-mini (ref.)	66%	72%	77%

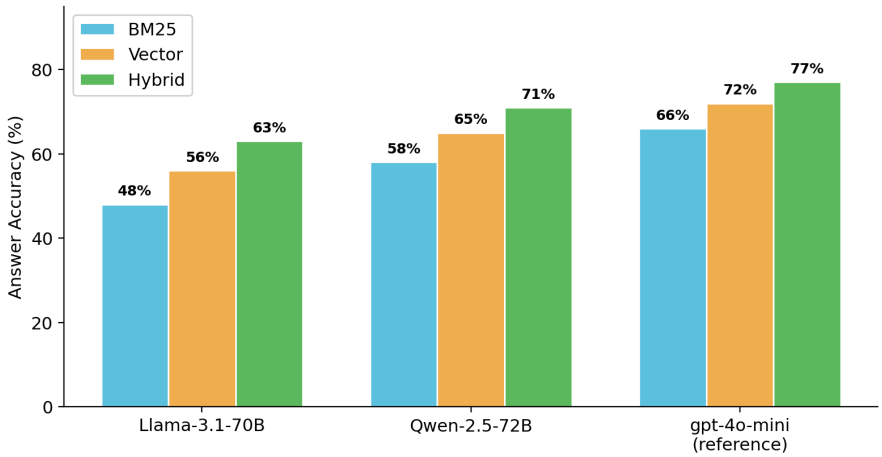


Figure 9. Cross-provider model validation: answer accuracy across retrieval methods

The Hybrid method consistently outperformed both BM25 and Vector retrieval across all three model families, confirming that the hybrid retrieval advantage is not an artifact of OpenAI-specific model behavior. Notably, the open-weight models achieved lower absolute accuracy than OpenAI models, likely reflecting differences in multilingual training data coverage for Kazakh. Qwen-2.5-72B outperformed Llama-3.1-70B, consistent with its stronger multilingual capabilities and reported Kazakh language support. These cross-provider results strengthen the generalizability of the main finding: hybrid retrieval provides a robust and consistent advantage regardless of the downstream language model.

4. Discussion

The obtained results support several key conclusions regarding why hybrid retrieval is effective for Kazakh legal-domain RAG and why it is especially beneficial for low-resource languages.

4.1. Mechanisms of Hybrid Method Superiority

The superiority of the hybrid method can be explained by its ability to combine the complementary strengths of lexical and semantic retrieval signals [15]. BM25 reliably retrieves passages containing exact legal terms, formal phrasing, and structured references (e.g., article numbers, named entities, and canonical formulations), which is critical for statutory and regulatory text [16]. At the same time, dense vector retrieval improves coverage when relevant evidence is expressed through paraphrasing, synonymous constructions, or morphologically varied surface forms [17].

The no-RAG baseline experiment (Section 3.6) confirms that the LLMs' parametric knowledge alone yields only 23–31% accuracy on Kazakh legal questions, underscoring that retrieval is the primary driver of system performance. In our experiments, hybrid retrieval achieved the strongest retrieval-level performance (e.g., Recall@6 = 0.89 versus 0.79 for BM25 and 0.83 for dense retrieval), indicating that fusion improves the probability of retrieving gold evidence within the prompt context window [15]. This advantage translates into end-to-end improvements in answer accuracy across models, because the generator is more consistently conditioned on relevant legal evidence [16]. Weighted Reciprocal Rank Fusion further strengthens this effect by balancing lexical precision and semantic coverage through an explicit weight parameter, enabling the method to adapt to the retrieval behavior of the legal corpus [13].

Importantly, the hybrid approach helps suppress failure modes where semantically "close" passages are retrieved but do not contain the legally decisive conditions, exceptions, or definitions required for correct answers. This is particularly relevant in legal texts, where terminological precision and normative wording determine correctness [17].

4.2. Stability and Predictability

A practical advantage of the hybrid method is its stability across different language models. When aggregating end-to-end answer accuracy across six LLMs, the hybrid method demonstrates the lowest coefficient of variation (CV = 0.057), compared to vector retrieval (CV = 0.065) and BM25 (CV = 0.071). This indicates that hybrid retrieval yields more predictable performance and is less sensitive to the choice of generator model [18].

Such stability is especially valuable in real-world deployments, where model upgrades or substitutions are frequent. A retrieval method that remains robust across model variants reduces operational risk and improves reproducibility of system behavior over time [18]. The cross-provider validation (Section 3.8) provides initial evidence that the hybrid advantage extends beyond the OpenAI family: both Qwen-2.5-72B and Llama-3.1-70B exhibited consistent hybrid superiority. However, the absolute accuracy of open-weight models was lower, and a broader range of providers should be tested to draw fully general conclusions.

4.3. Comparison with Global Research

The observed gains from hybrid retrieval are consistent with broader research trends showing that combining lexical and semantic signals often improves both recall-oriented and rank-sensitive metrics in RAG pipelines [19], [20]. Across languages and domains, hybrid fusion methods are frequently reported to outperform single-retriever baselines, particularly when queries include a mix of exact terminology and paraphrased expressions [19].

Our results are consistent with this broader trend, demonstrating that hybrid retrieval remains effective under the constraints of a Kazakh legal corpus, where both precise legal references and morphologically diverse phrasing are common.

Prior work that evaluates hybrid strategies in other languages similarly supports the generality of fusion-based retrieval approaches across typologically diverse settings [21]. Research on improving lexical retrieval with additional relevance signals also aligns with this finding, suggesting that multi-signal retrieval is a reliable direction for robust QA systems [22]. Recent work on legal QA in other low-resource settings confirms that language- and domain-specific adaptations are essential for retrieval quality. Craciun et al. [25] proposed GRAF, a graph-augmented retrieval approach for Romanian legal MCQA, demonstrating that structured knowledge representations can substantially improve answer selection in a low-resource legal domain. Park et al. [26] introduced LRAGE, an open-source toolkit for holistic evaluation of legal RAG systems, showing that the choice of reranker and retrieval corpus often dominates overall accuracy. Li et al. [27] presented LexRAG, a benchmark for multi-turn legal consultation with citation-grounded evaluation. Our work complements these efforts by focusing specifically on an agglutinative Turkic language and isolating the contribution of first-stage retrieval fusion rather than reranking or graph-based approaches.

4.4. Specificity for Low-Resource Languages

The findings are particularly important in the context of low-resource language processing [23]. Kazakh is a Turkic language with agglutinative morphology, which increases lexical sparsity and makes purely lexical retrieval more brittle without adequate language-specific preprocessing [24]. At the same time, semantic retrieval quality can be constrained by limited language-specific training resources or domain mismatch in embedding models, especially for specialized legal language [23].

Hybrid retrieval mitigates these limitations by leveraging the strengths of both paradigms: lexical retrieval contributes reliability for formal legal terminology and structured references, while semantic retrieval improves coverage for paraphrased or morphologically varied expressions. As a result, hybrid fusion provides a robust and practical retrieval strategy for Kazakh legal RAG systems and, more broadly, for low-resource languages with similar linguistic and data constraints [23], [24].

4.5. Limitations of the Study

Despite the strong empirical results, several limitations should be acknowledged. First, although the corpus is large and representative for the legal domain, the study remains domain-specific (Kazakh legislation and regulatory texts). Retrieval behavior and end-to-end QA accuracy may differ in other domains such as medicine, education, or news, where document structure, terminology, and user query patterns are substantially different [23]. Future work should therefore evaluate the same pipeline across multiple domains to establish broader generalizability.

Second, the evaluation relies on a fixed retrieval setting (Top-K = 6) and a fixed chunking configuration (900 tokens with 150-token overlap). The Top-K = 6 setting was chosen to balance evidence coverage against context window constraints, but alternative retrieval depths (e.g., K = 3 or K = 10) may affect both retrieval metrics and downstream answer accuracy. Because the entire evaluation hinges on this parameter, a systematic sensitivity analysis over Top-K and chunk size could further strengthen the robustness of the conclusions [19].

Third, although we improved the lexical baseline by applying Kazakh-aware tokenization and normalization, Kazakh morphology remains challenging. More advanced morphological analyzers and lemmatization tools may further reduce lexical sparsity and improve lexical retrieval quality, potentially affecting the relative gap between BM25, dense retrieval, and hybrid fusion.

Fourth, the study uses a single embedding model configuration for dense retrieval (text-embedding-3-small). While it provides strong performance in this setting, further gains may be possible with alternative multilingual or Kazakh-specialized embedding models, as well as domain-adaptive embedding training on legal corpora [23], [24].

Fifth, while Weighted RRF parameters were tuned via 5-fold cross-validation with held-out evaluation (Section 2.2.3), the tuning strategy can be further improved by using a fully independent development set, larger validation splits, or query-type–aware adaptive fusion. This may yield additional gains and provide even stronger guarantees against overfitting [13], [21].

Sixth, the primary evaluation used six models from the OpenAI family, sharing a common training pipeline and RLHF methodology. While the cross-

provider validation (Section 3.8) confirmed the hybrid advantage for Qwen-2.5-72B and Llama-3.1-70B, these additional experiments were limited in scope. A more comprehensive evaluation across a wider range of model families and sizes would further strengthen external validity claims.

Seventh, each experimental configuration was executed once per question. For models operating at temperature 1.0 (o1, o1-mini), each answer represents a single stochastic sample, and accuracy estimates may vary across reruns. While retrieval-level metrics are deterministic, the reported end-to-end accuracy values should be interpreted as point estimates with inherent sampling variability. Future work should consider multiple repetitions or report binomial confidence intervals to quantify this uncertainty.

Eighth, the system prompt (Figure 1) is written entirely in English, while the corpus, questions, and expected answers are all in Kazakh. This language mismatch between instruction and task is a known factor affecting LLM performance, particularly for lower-resource languages. Future work should investigate whether a Kazakh-language prompt improves answer accuracy and generation quality.

Ninth, the primary evaluation metric was binary accuracy (correct vs. incorrect). A supplementary three-point analysis (Section 3.7) revealed that 44% of non-correct Hybrid answers were partially correct, suggesting that binary accuracy underestimates useful system output. Future studies should consider adopting graded evaluation as the primary metric for a more nuanced assessment of legal QA performance.

Finally, the retrieval evaluation relies on a single gold passage per question, and Precision@6 is equivalent to Recall@6 divided by 6 by construction. In practice, legal answers may require evidence from multiple articles or passages. This single-gold design may penalize retrieval methods that surface valid alternative passages not designated as gold. Multi-relevant judgments or post-hoc human relevance grading of the top-6 retrieved items would provide a more realistic assessment of retrieval quality.

4.6. Practical Implications

The findings have several practical implications for deploying RAG systems in Kazakh and other low-resource languages. First, the results indicate that hybrid retrieval should be treated as a default strategy for Kazakh legal QA, because it combines the reliability of lexical matching with the coverage advantages of semantic similarity, yielding consistently strong retrieval and end-to-end performance across models [15], [18].

Second, scalability considerations are central for production deployment. Dense indexing and hybrid fusion introduce additional computational and engineering overhead (e.g., embedding generation, vector indexing, and fusion at query time). In our experiments, average per-query latency was approximately 11.5 seconds for BM25, 12.5 seconds for Vector, and 14.5 seconds for Hybrid retrieval (including ~10 seconds for API-based generation). The Hybrid overhead relative to BM25 is approximately 3 seconds (+26%), which represents a modest cost for a 10 percentage-point gain in answer accuracy. For o1-family models, generation latency was substantially higher (25–40 seconds) due to reasoning computation. These latency figures support the practical feasibility of hybrid retrieval in legal and governmental applications where correctness and evidence grounding are critical [19], [20].

Third, the approach is transferable to related settings. Because many Turkic languages share morphological characteristics and resource constraints, the same hybrid framework–with language-aware preprocessing and careful evaluation–can serve as a strong baseline for other low-resource languages, with minimal adaptation [23], [24].

Finally, the demonstrated robustness across multiple LLMs supports industrial applicability. Systems can switch between larger and smaller LLMs depending on cost and latency constraints while maintaining reliable retrieval quality, making the proposed pipeline suitable for legal information systems, public-sector services, and commercial assistants [18].

5. Conclusions

This study provides a systematic evaluation of retrieval strategies for Kazakh legal-domain Retrieval-Augmented Generation (RAG), comparing BM25, dense vector retrieval, and a hybrid approach based on Weighted Reciprocal Rank Fusion. The results demonstrate that hybrid retrieval consistently outperforms single-retriever baselines on retrieval-level metrics and yields the most reliable end-to-end answer accuracy across multiple language models.

The main conclusions of this study are as follows. First, retrieval augmentation is essential: the no-RAG baseline yielded only 23–31% accuracy, while RAG configurations achieved 66–89%, confirming that parametric knowledge alone is insufficient for Kazakh legal QA. Second, hybrid retrieval emerges as the most effective overall strategy for Kazakh legal-domain RAG, as it successfully combines lexical precision with semantic coverage, leading to more reliable evidence retrieval. Third, while dense vector retrieval performs competitively, its effectiveness is more sensitive to model choice and configuration, whereas hybrid fusion provides more stable and consistent performance across different language models. Fourth, although the choice of the language model influences answer quality, robust retrieval plays a critical role in stabilizing downstream answer generation, enabling practical deployment across both flagship and compact model variants. Cross-provider experiments with Qwen-2.5-72B and Llama-3.1-70B confirm that the hybrid advantage is not restricted to the OpenAI model family.

From a broader perspective, this work contributes to the methodological foundation for building reliable RAG systems in low-resource languages by emphasizing grounded evaluation using gold evidence passages, a clear separation between retrieval-level metrics and end-to-end answer accuracy, and hybrid fusion as a robust default retrieval design. The proposed framework is directly applicable to legal information systems in Kazakhstan and can be extended to other low-resource languages and domains with appropriate corpus preparation and evaluation methodology.

Funding

This research received no external funding.

Author Contributions

Conceptualization, N.K. and S.A.; Methodology, N. K. and S.A.; Software, N.K. and S.A.; Validation, N.K. and A.K.; Formal Analysis, N.K. and A.K.; Investigation, N.K. and S.A.; Resources, N.K.; Data Curation, A.K. and N.K.; Writing – Original Draft Preparation, N.K.; Writing – Review & Editing, N.K., S.A. and A.K.; Visualization, A.K. and N.K.; Supervision, N.K.; Project Administration, N.K..

Conflicts of Interest

The author declares no conflict of interest.

References

1. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, *et al.*, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020. [Online]. Available: https://arxiv.org/abs/2005.11401

2. Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, *et al.*, “Retrieval-augmented generation for large language models: A survey,” *arXiv preprint*, arXiv:2312.10997, 2023. [Online]. Available: https://arxiv.org/abs/2312.10997

3. W. Shi, S. Min, M. Yasunaga, M. Seo, R. James, M. Lewis, *et al.*, “REPLUG: Retrieval-augmented black-box language models,” *arXiv preprint*, arXiv:2301.12652, 2023. [Online]. Available: https://arxiv.org/abs/2301.12652

4. V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, et al., “Dense passage retrieval for open-domain question answering,” in Proc. Conf. Empirical Methods in Natural Language Process. (EMNLP), 2020, pp. 6769–6781. [Online]. Available: https://arxiv.org/abs/2004.04906

5. S. Robertson and H. Zaragoza, “The probabilistic relevance framework: BM25 and beyond,” Found. Trends Inf. Retr., vol. 3, no. 4, pp. 333–389, 2009. [Online]. Available: https://www.staff.city.ac.uk/~sbrp622/papers/foundations_bm25_review.pdf

6. N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in *Proc. Conf. Empirical Methods in Natural Language Process. and 9th Int. Joint Conf. Natural Language Process. (EMNLP-IJCNLP)*, 2019, pp. 3982–3992. [Online]. Available: https://arxiv.org/abs/1908.10084

7. N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, “BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models,” Proc. NeurIPS Datasets and Benchmarks Track, vol. 1, pp. 1–18, 2021. [Online]. Available: https://arxiv.org/abs/2104.08663

8. P. Pakray and A. Gelbukh, “Natural language processing applications for low-resource languages,” Natural Language Process., pp. 1–25, 2025. [Online]. Available: https://www.cambridge.org/core/journals/natural-language-processing/article/natural-language-processing-applications-for-lowresource-languages/7D3DA31DB6C01B13C6B1F698D4495951

9. B. P. King, Practical Natural Language Processing for Low-Resource Languages, Univ. of Michigan, 2015. [Online]. Available: https://deepblue.lib.umich.edu/handle/2027.42/113373

10. C. Fensore, K. Dhole, J. C. Ho, and E. Agichtein, “Evaluating Hybrid Retrieval Augmented Generation using Dynamic Test Sets: LiveRAG Challenge,” arXiv preprint arXiv:2506.22644, Jun. 2025. [Online]. Available: https://arxiv.org/abs/2506.22644

11. M. Gabryel, M. Kocić, and A. Gabryel, “Hybrid Retrieval in RAG: A Comparison of Semantic, Lexical and Reranking Methods,” in Lecture Notes in Computer Science, Springer Science and Business Media Deutschland GmbH, 2026, pp. 86–93. doi: 10.1007/978-3-032-03711-4_8.

12. Y. Wang, M. Lin, Q. Hu, S. Bai, Y. Li, and L. Bao, “A domain-specific cross-lingual semantic alignment learning model for low-resource languages,” *Neural Networks*, vol. 194, Feb. 2026, doi: 10.1016/j.neunet.2025.108114.

13. G. V. Cormack, C. L. Clarke, and S. Buettcher, “Reciprocal rank fusion outperforms Condorcet and individual rank learning methods,” in Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., 2009, pp. 758–759. [Online]. Available: https://dl.acm.org/doi/10.1145/1571941.1572114

14. I. Akhmetov, A. Zhamankhan, N. Zhetesov, and A. Kubaeva, “Textual foundations of justice: Kazakhstani laws and jurisprudence dataset (Version 3) [Data set],” *Mendeley Data*, 2024. [Online]. Available: https://doi.org/10.17632/jdpc5658nh.3

15. J. Lin, R. Nogueira, and A. Yates, “Pretrained transformers for text ranking: BERT and beyond,” Synth. Lect. Hum. Lang. Technol., vol. 14, no. 4, pp. 1–325, 2021. [Online]. Available: https://arxiv.org/abs/2010.06467

16. T. Formal, C. Lassance, B. Piwowarski, and S. Clinchant, “SPLADE v2: Sparse lexical and expansion model for information retrieval,” *arXiv preprint*, arXiv:2109.10086, 2021. [Online]. Available: https://arxiv.org/abs/2109.10086

17. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186. [Online]. Available: https://arxiv.org/abs/1810.04805

18. C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Morgan & Claypool Publishers, 2016. [Online]. Available: https://dl.acm.org/doi/book/10.1145/2915031

19. Xiaohua Wang et al., “Searching for Best Practices in Retrieval-Augmented Generation,” arXiv preprint arXiv:2407.01219v1, Jul. 2024. [Online]. Available: https://arxiv.org/abs/2407.01219

20. K. Sawarkar, A. Mangal, and S. R. Solanki, “Blended RAG: Improving RAG accuracy with semantic search and hybrid query-based retrievers,” in Proc. 2024 IEEE 7th Int. Conf. Multimedia Information Processing and Retrieval (MIPR), 2024, pp. 155–161 [Online]. Available: https://ieeexplore.ieee.org/document/10707868

21. M. Jovanović, N. Filipović, and D. Vučković, “The Serbian retrieval-augmented generation system based on hybrid search,” in Proc. IEEE Int. Conf. on Intelligent Systems (IS), Novi Sad, Serbia, 2024, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10851665

22. J. Metzler and W. B. Croft, “A Markov random field model for term dependencies,” in Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., Salvador, Brazil, 2005, pp. 472–479. [Online]. Available: https://dl.acm.org/doi/10.1145/1076034.1076115

23. J. Magueresse, V. Carles, and E. Heetderks, "Low-resource languages: A review of past work and future challenges," *arXiv preprint*, arXiv:2006.07264, 2020. [Online]. Available: https://arxiv.org/abs/2006.07264

24. S. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, “A survey on recent approaches for natural language processing in low-resource scenarios,” in Proc. NAACL-HLT, 2021. [Online]. Available: https://aclanthology.org/2021.naacl-main.201

25. C.-G. Craciun, R.-A. Smădu, D.-C. Cercel, and M.-C. Cercel, "GRAF: Graph Retrieval Augmented by Facts for Romanian Legal Multi-Choice Question Answering," in Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 2025, pp. 12708–12742. Available: https://aclanthology.org/2025.findings-acl.659/

26. M. Park, H. Oh, E. Choi, and W. Hwang, "LRAGE: Legal Retrieval Augmented Generation Evaluation Tool," arXiv preprint arXiv:2504.01840, Apr. 2025. Available: https://arxiv.org/abs/2504.01840

27. H. Li, Y. Chen, Y. Hu, Q. Ai, J. Chen, X. Yang, J. Yang, Y. Wu, Z. Liu, and Y. Liu, "LexRAG: Benchmarking Retrieval-Augmented Generation in Multi-Turn Legal Consultation Conversation," in Proc. 48th Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2025. arXiv preprint arXiv:2502.20640, Feb. 2025. Available: https://arxiv.org/abs/2502.20640

Information about Authors:

Nurlykhan Kalzhanov is a Master’s student in Computer Engineering at Al-Farabi Kazakh National University (Almaty, Kazakhstan, e-mail: nurkal022@gmail.com). His research interests include machine learning, information retrieval, and natural language processing, with a particular focus on Retrieval-Augmented Generation (RAG) systems for low-resource languages. He has participated in research projects related to artificial intelligence applications and computational linguistics.

Sauirbek Artykbay is a Master’s student in Computer Engineering at Al-Farabi Kazakh National University (Almaty, Kazakhstan, e-mail: artikbaisauirbek@gmail.com). His research interests include machine learning, information retrieval, and natural language processing, with a particular focus on Semantic search systems for low-resource languages.

Akniyet Kalzhan is a Bachelor's student in Data Science at Al-Farabi Kazakh National University (Almaty, Kazakhstan, e-mail: aknietkalzhan@gmail.com). Her research interests include computer vision, large language models (LLMs), and data science. She is currently working on her undergraduate thesis focused on knowledge distillation in large language models. Akniyet has completed internships in two research laboratories at Al-Farabi Kazakh National University, where she gained practical experience in artificial intelligence and machine learning applications.

Submission received: 12 November, 2025.
Revised: 3 March, 2026.
Accepted: 16 March, 2026.

IRSTI 28.17.19 https://doi.org/10.26577/jpcsit4120266

S. Tastanova[1*], I. Nabiev[1], B. Nurimbetov[2]

[1]Department of Television and Media Technologies, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan
[2]Department of Television Technologies, Nukus State Technical University, Nukus, Uzbekistan
[*]e-mail: tasmaat@gmail.com

THREE-DIMENSIONAL FRACTAL GEOMETRY MODELING AND DIGITAL HOLOGRAPHY BASED ON R-FUNCTIONS

Abstract. This paper is devoted to modern research in the field of digital modeling of complex-shaped geometric objects and the determination of their optical properties, which currently represent one of the most relevant challenges in contemporary science. In particular, the problem of realistic representation of three-dimensional objects with fractal geometry and their holographic reconstruction in a full 3D format is of significant scientific and practical importance for such fields as industry, medicine, engineering, architecture, materials science, virtual reality (VR), and digital art. Fractal structures possess a number of unique properties, including self-similarity, unlimited detail, and high spatial complexity, which makes them an effective mathematical basis for modeling natural objects such as plants, vascular systems, bone tissues, crystalline structures, and surface reliefs. At the same time, the geometric representation of fractal forms using classical methods is challenging, and their mathematical modeling requires the application of high-precision and formally rigorous techniques. At present, the mathematical description of fractal objects is often based on statistical, stochastic, or iterative algorithms. However, such approaches are generally characterized by insufficient analytical rigor and smoothness, blurred boundaries, and the lack of a holistic spatial description. In this regard, there arises a need to develop methods for modeling complex fractal forms based on strict analytical expressions, in particular using the R-functions apparatus. An additional challenging task is the reconstruction of holographic images of the modeled fractal objects, which requires high-precision optical modeling. The application of holographic technologies based on the principles of interference and diffraction makes it possible to adequately reproduce the spatial and structural features of fractal objects in a digital environment.

Keywords: three-dimensional fractal geometry, R-functions, digital holography, Sierpinski tetrahedron, Iterated Function Systems (IFS), analytical modeling, convolutional neural networks (CNN), 3D reconstruction.

1. Introduction

In recent years, digital modeling of complex geometric objects and the investigation of their optical properties have become an important research direction in applied mathematics, computer graphics, and optical engineering. In particular, the realistic representation of three-dimensional objects with fractal geometry and their holographic reconstruction in full 3D format has attracted significant attention due to a wide range of applications in industry, medicine, engineering, architecture, materials science, virtual reality, and digital art [1, 2].

Fractal geometry provides a powerful mathematical framework for describing objects characterized by self-similarity, high structural complexity, and theoretically infinite levels of detail [3]. These properties make fractals particularly suitable for modeling natural structures such as plants, vascular systems, bone tissues, crystalline formations, and surface reliefs [4]. However, the accurate geometric representation of fractal objects remains a challenging task, especially in cases where strict boundary definition and analytical continuity are required.

Most existing approaches to fractal modeling are based on statistical, stochastic, or purely iterative methods, including classical Iterated Function Systems (IFS) and random fractal generators [5, 6, 7]. Although these methods are effective for visualization purposes, they often lack mathematical rigor, analytical smoothness, and explicit control over object boundaries. As a result, the generated models may contain discontinuities and exhibit limited applicability for subsequent physical or optical modeling [8].

To overcome these limitations, the present work employs the R-functions method as a rigorous

analytical tool for the geometric modeling of complex three-dimensional fractal objects. The R-functions-based approach enables the construction of complex geometries using continuous implicit functions while preserving precise boundary descriptions and topological correctness [9, 10]. When combined with IFS, this method provides a formal incorporation of fractal self-similarity into an analytically defined spatial model.

In addition to geometric modeling, accurate holographic reconstruction of fractal objects requires high-precision optical modeling. Digital holography, based on the principles of diffraction and interference, offers an efficient mechanism for encoding and reconstructing three-dimensional information [11]. In particular, Fresnel diffraction is widely used for numerical hologram formation and reconstruction due to its computational efficiency and solid physical foundation [12].

Furthermore, recent advances in deep learning have demonstrated the high effectiveness of convolutional neural networks (CNNs) in solving inverse problems in optics and image reconstruction [13, 14]. The integration of CNNs with analytically defined geometric constraints enhances the stability and accuracy of the reconstruction process while preserving the physical and mathematical structure of the modeled object.

Thus, this work proposes a unified integrated approach that combines R-functions, IFS-based fractal modeling, digital holography, and convolutional neural networks. This integration establishes a reliable mathematical and computational framework for high-precision modeling and holographic reconstruction of complex three-dimensional fractal objects.

To describe a three-dimensional geometric object, an implicit (implicit-form) function is used

$$\begin{gathered}\Phi(x,y,z)\colon R^3 \to R \\ \Omega = \{\,(x,y,z) \in R^3 \mid \Phi_n(x,y,z) \geq 0\,\} \\ \partial\Omega = \{\,(x,y,z) \in R^3 \mid \Phi_n(x,y,z) = 0\,\}\end{gathered} \tag{1}$$

It assigns a real value to each point in space with coordinates (x, y, z). In this formulation, the object itself is defined as follows:

$\Phi(x,y,z) \geq 0$ – points belonging to the object,

$\Phi(x,y,z) < 0$ – points located outside the object.

To introduce a fractal structure, an Iterated Function System (IFS) is employed. At each iteration, the tetrahedron is scaled and translated in space. In general form, a three-dimensional affine transformation is written as:

$$r' = Sr + t_k \tag{2}$$

where

$$\begin{gathered}r = (x,y,z)^T, \\ S = sI_3, \\ 0 < s \leq 1,\end{gathered} \tag{3}$$

and t_k – is the translation vector.

For constructing the Sierpiński fractal, four affine transformations corresponding to the vertices of the initial tetrahedron are typically used. The mathematical model of the fractal at the n-th iteration, expressed via the R-function, is written as:

$$\Phi_n(x,y,z) == R\left(\Phi_0\left(S_1^{-1}(\mathrm{r} - t_1)\right), \cdots, \Phi_0\left(S_4^{-1}(\mathrm{r} - t_4)\right)\right) \tag{4}$$

where $\Phi_0(x,y,z)$is the R-function of the initial tetrahedron, providing its analytical description. This formulation rigorously establishes the self-similarity property of the fractal in a strict mathematical sense.

As a result, a complex fractal structure of the three-dimensional Sierpiński tetrahedron is formed, which becomes increasingly detailed and visually apparent as the number of iterations increases.

2. Materials and Methods

Solving the problem of geometric modeling of the three-dimensional Sierpiński triangle (more precisely, the Sierpiński pyramid or tetrahedral fractal) requires a strictly analytical definition of the object in space. To this end, first of all, a regular three-dimensional pyramid (tetrahedron) is selected as the initial geometric object, which is

mathematically represented using analytical geometry and the R-function method.

Any plane in space is typically described by the following linear equation:

$$Ax + By + Cz + D = 0. \quad (5)$$

Using this equation, we can determine the equation of each face (surface) of the pyramid. Let the four vertices of the initial pyramid be given as:

$$V_1(x_1, y_1, z_1),\ V_2(x_2, y_2, z_2),\ V_3(x_3, y_3, z_3),\ V_4(x_4, y_4, z_4) \quad (6)$$

The vectors formed by these points are:

$$\vec{V_1V_2},\ \vec{V_1V_3},\ \vec{V_1V_4} \quad (7)$$

Based on these vectors, the normal vectors for each face of the pyramid are determined, and as a result, four plane equations are obtained:

$$P_i(x, y, z) = A_i x + B_i y + C_i z + D_i, i = 1, \dots, 4 \quad (8)$$

If, at the first iteration, a regular pyramid is considered, then the coordinates of its vertices are chosen in a special manner for convenience, for example:

$$V_1(0,0,0),\ V_2(a,0,0),\ V_3\left(\frac{a}{2}, \frac{\sqrt{3}a}{2}, 0\right),\ V_4\left(\frac{a}{2}, \frac{\sqrt{3}a}{6}, \frac{\sqrt{6}a}{3}\right) \quad (9)$$

This choice guarantees that all edges of the pyramid are equal. Using these points, the surface functions $f_i(x, y, z)$for each face are defined according to equation (1).

At the next stage, the R-function apparatus is employed to represent the interior region of the pyramid by means of a single analytical function. If the interior half-spaces of all faces of the pyramid satisfy the condition $f_i(x, y, z) \geq 0$, then the entire object is represented by the following R-function:

$$\Phi(x, y, z) = R(f_1, f_2, f_3, f_4) \quad (10)$$

where the R-function provides a smooth analytical representation of the logical AND operation, for example:

$$R(a, b) = a + b - \sqrt{a^2 + b^2} \quad (11)$$

As a result, the condition $\Phi(x, y, z) \geq 0$ defines the points located inside the pyramid, while the condition $\Phi(x, y, z) < 0$ defines the points located outside it.

To introduce fractal properties, an Iterated Function System (IFS) is applied. At each iteration, the tetrahedron is scaled and translated in space. In general form, a three-dimensional affine transformation is written as:

$$T_k(r) = Sr + t_k \quad (12)$$

where $\mathrm{r} = (x, y, z)^T$, $S = sI_3$, $0 < s < 1$, and $\mathrm{t}_k$are the translation vectors.

For the Sierpiński pyramid, four affine transformations are typically used:

$$T_k(r) = \frac{1}{2}r + t_k, k = 1, \dots, 4 \quad (13)$$

where $\mathrm{t}_k$are translation vectors corresponding to the vertices of the initial tetrahedron.

The mathematical model of the fractal at the n-th iteration, expressed via the R-function, is written as:

$$\Phi_n(x, y, z) = \max_{k=1,\dots,4} \Phi_{n-1}\left(T_k^{-1}(\mathrm{x,y,z})\right) \quad (14)$$

Here, $\Phi_0(x, y, z)$is the R-function describing the initial tetrahedron. This expression provides a strict mathematical formulation of the self-similarity property of the fractal.

As a result, the object is defined as:

$$\Omega = \{(x, y, z) \in R^3 \mid \Phi_n(x, y, z) \geq 0\} \quad (15)$$

The object is defined in this form, and as the number of iterations $n$increases, a complex three-dimensional fractal structure of the Sierpinski pyramid is formed.

General Algorithm (for the Sierpiński Pyramid)

1. Select the vertices of the initial tetrahedron $\{V_i\}$.

2. Determine the plane equations for each face of the tetrahedron.

3. Construct the R-function of the pyramid $\Phi_0(x, y, z)$.

4. Define the affine transformations T_k, including scaling and translations.

5. Apply the iterative formula to compute $\Phi_n(x, y, z)$.

6. Visualize the fractal object using the condition $\Phi_n(x, y, z) \geq 0$.

The main advantage of this approach is that the fractal defined by the IFS combined with the R-function is represented not as a set of discrete points, but as a continuous analytical function. This provides an important mathematical foundation for subsequent stages such as holographic modeling, reconstruction using convolutional neural networks (CNNs), and three-dimensional printing.

2.1. Modified General Algorithm

The classical three-dimensional Sierpiński fractal is usually constructed using an Iterated Function System (IFS) as a set of discrete points.

The proposed modification consists in the following: the fractal object is defined not as a set of points, but as a continuous three-dimensional geometric object analytically expressed via an R-function, which is subsequently used for holographic reconstruction.

This approach provides:

- geometric accuracy,
- smooth boundaries,
- compatibility with optical modeling.

Initial geometric model (zero iteration).

As the basis for constructing the three-dimensional Sierpiński triangle (more precisely, the Sierpiński pyramid or tetrahedron), a regular tetrahedron is used. The vertices of the tetrahedron are defined as:

$$V_1(0,0,0),\ V_2(a,0,0),\ V_3\left(\frac{a}{2}, \frac{\sqrt{3}a}{2}, 0\right), \tag{16}$$

$$V_4\left(\frac{a}{2}, \frac{\sqrt{3}a}{6}, \frac{\sqrt{6}a}{3}\right)$$

Each face of the tetrahedron is defined by a plane equation:

$$A_i x + B_i y + C_i z + D_i = 0, i = 1, \ldots, 4 \tag{17}$$

where the coefficients A_i, B_i, C_i, D_i are determined using vector cross products constructed from the vertex coordinates.

Analytical representation of the tetrahedron using R-functions. For each plane, a distance function is defined as:

$$f_i(x, y, z) = A_i x + B_i y + C_i z + D_i \tag{18}$$

The interior volume of the tetrahedron is defined using an R-function as follows:

$$F_0(x, y, z) = R_\wedge(f_1, f_2, f_3, f_4) \tag{19}$$

where $R_\wedge$ denotes the R-intersection operation.

As a result, $F_0(x, y, z) > 0$ corresponds to the interior of the object, $F_0(x, y, z) = 0$ corresponds to the boundary, and $F_0(x, y, z) < 0$ corresponds to the exterior region.

This represents the first modification compared to the classical IFS approach.

Fractal Iteration Based on IFS

In the modified approach, unlike the classical IFS where only the coordinates are transformed, the R-function itself is transformed.

The affine transformation is defined as:

$$T_k(x) = \frac{1}{2}x + t_k, k = 1,2,3,4 \tag{20}$$

where $\mathbf{t}_k$ are translation vectors corresponding to the vertices of the tetrahedron.

Each iteration is defined as follows:

$$F_{n+1}(x, y, z) = R_\vee\left(F_n(T_1^{-1}(\mathrm{x,y,z})), \cdots, F_n(T_4^{-1}(\mathrm{x,y,z}))\right) \tag{21}$$

where $R_\vee$ denotes the R-union operation, given by:

$$R_\vee(a, b) = a + b - \sqrt{a^2 + b^2} \tag{22}$$

This operation represents the second key modification, due to which the fractal evolves as an analytical geometric field.

Spatial Density Model of the Fractal Object

For holographic modeling, it is necessary to define the spatial density of the object:

$$\rho(x,y,z) = H(F_n(\mathrm{x,y,z})) \tag{23}$$

where $H(\cdot)$denotes the Heaviside function.

Holographic Modeling (Third Modification)

The wave field of the object is defined as:

$$U_0(x,y,z) = \rho(x,y,z)\, e^{i\phi(x,y,z)} \tag{24}$$

The reference (carrier) wave is given by:

$$U_r(x,y) = e^{\,ik(\mathrm{xsin}\ \theta_x + \mathrm{ysin}\ \theta_y)} \tag{25}$$

The hologram is formed using the Fresnel integral:

$$H(x,y) = \\ = \left| \iiint U_0(\xi,\eta,\zeta)\frac{e^{ikR}}{R}\mathrm{d}\xi\mathrm{d}\eta\mathrm{d}\zeta + U_r(\mathrm{x,y}) \right|^2 \tag{26}$$

As a result, a digital hologram of a three-dimensional fractal object is obtained.

Overall Algorithm Sequence

Algorithm:

1. Determination of the tetrahedron vertices.
2. Construction of the initial geometric model using an R-function.
3. Iteration of the R-function with the application of affine transformations.
4. Formation of the fractal density function.
5. Computation of wave propagation.
6. Generation of the digital hologram.

The process of geometric modeling of holographic images of complex three-dimensional objects with fractal shapes is based on the integration of several classical approaches. In this dissertation, a new generalized algorithm is proposed that combines the mathematical apparatus of the R-function method, fractal geometry, digital holography, and convolutional neural networks (CNNs) into a single modified framework.

The principal modification lies in the fact that the fractal object is considered not only as an iterative or statistical model, but as a continuous spatial function analytically defined by means of an R-function. This function directly participates both in the formation of the holographic image and in the reconstruction stages using CNNs.

First of all, the fractal three-dimensional object defined in space, $\Omega \subset \mathbb{R}^3$ is treated as a field and is expressed using an R-function as follows:

$$\Phi(x,y,z) = \\ = R(\phi_1(x,y,z), \phi_2(x,y,z), \dots, \phi_n(x,y,z)) \tag{27}$$

Here, $\phi_i(x,y,z)$are the level-set functions of elementary geometric primitives (sphere, cylinder, cone, etc.), and $R(\cdot)$is a smooth analytical analogue of the logical operations AND / OR / NOT. Interior points of the object are determined by the condition $\Phi(x,y,z) \geq 0$, whereas exterior points satisfy $\Phi(x,y,z) < 0$. The fractal property is introduced into the object through multiscale iteration:

$$\Phi_k(x,y,z) = \\ = R(\Phi_{k-1}(sx,sy,sz), \psi_k(x,y,z)) \tag{28}$$

Here, $s < 1$is the scaling coefficient, and $\psi_k$is an additional geometric component in the fractal iteration. At the next stage, the complex amplitude of the object is defined as:

$$U_0(x,y) = A(x,y)\exp(i\phi(x,y)) \tag{29}$$

Here, the amplitude $A(x,y)$is defined through the projection of the object using the R-function:

$$A(x,y) = \int H(\Phi(x,y,z))\, dz \tag{30}$$

where $H(\cdot)$denotes the Heaviside function, and $\phi(x,y)$is the phase function proportional to the spatial relief of the object:

$$\phi(x,y) = k\, z(x,y) \tag{31}$$

Here, $k = \frac{2\pi}{\lambda}$ is the wave number.

The hologram is formed based on the principle of interference:

$$I(x,y) = \left| U_0(\mathrm{x,y}) + U_r(\mathrm{x,y}) \right|^2 \tag{32}$$

Here, $U_r(x,y) = A_r\exp(ikx)$ is the reference wave. As a result, the obtained intensity function $I(x,y)$contains the holographic image of the fractal object.

The main modification begins precisely at this stage. In the traditional approach, the hologram is provided to the CNN in the form of a ready-made

image. In the present work, the hologram is supplied in the form of a normalized tensor with phase information preserved:

$$\tilde{I}(x,y) = \frac{I(x,y) - \mu_I}{\sigma_I} \quad (33)$$

Here, $\mu_I$and $\sigma_I$denote the mean value and variance, respectively. The normalized hologram is fed into the input layer of the CNN: $X^{(0)} = \tilde{I}(x,y)$

At each convolutional layer, the feature maps are computed as follows:

$$X_j^{(l)} = \sigma\left(\sum_i X_i^{(l-1)} * W_{ij}^{(l)} + b_j^{(l)}\right) \quad (34)$$

Here, "$*$" denotes the convolution operator, $W_{ij}^{(l)}$are the filter kernels, and $\sigma(\cdot)$is a nonlinear activation function (ReLU).

An important feature of the proposed modification is that, during training, a geometric constraint based on the R-function is incorporated into the loss function:

$$L = L_{rec} + \\ +\alpha \int_{\Omega} |\, \Phi_{CNN}(x,y,z) - \Phi(x,y,z) \,|^2 \; dV \quad (35)$$

Here, $\mathcal{L}_{\text{rec}}$denotes the hologram reconstruction error, $\Phi_{\text{CNN}}$is the surface of the object reconstructed by the CNN, and $\alpha$is a weighting coefficient.

As a result, the CNN learns not only visual similarity, but also the analytical and geometric structure of the object. At the final stage, the reconstructed three-dimensional object is obtained:

$$\Omega_{rec} = \{(x,y,z) : \Phi_{CNN}(x,y,z) \geq 0\} \quad (36)$$

The proposed new generalized algorithm thus integrates classical geometric modeling based on R-functions with digital holography and deep learning methods in a modified manner, enabling high-accuracy modeling and reconstruction of holographic images of complex three-dimensional objects with fractal geometry.

3. Modified Universal Algorithm: R-Function – IFS – Holographic Modeling

1. The classical three-dimensional Sierpiński fractal is constructed as a set of discrete points using an Iterated Function System (IFS).

The proposed modification consists in defining the fractal object as a continuous analytical three-dimensional geometry using an R-function, which is then directly employed for holographic reconstruction. This ensures geometric accuracy, smooth boundaries, and compatibility with optical modeling.

2. A regular tetrahedron is used as the base object. The coordinates of its vertices and the analytical equations of the planes defining its faces are specified.

3. For each plane, a distance function is defined. The interior region of the tetrahedron is described by the function:

$$F_0(x,y,z) = R_\wedge(f_1, f_2, f_3, f_4) \quad (37)$$

where $F_0(x,y,z) > 0$corresponds to the interior, $F_0(x,y,z) = 0$to the boundary, and $F_0(x,y,z) < 0$to the exterior of the object.

4. Fractal iteration is performed based on the modified IFS. Unlike the classical approach, the R-function itself is transformed via affine scaling and translation. The iteration is defined as:

$$F_n(x,y,z) = \bigcup_{k=1}^{4} F_{n-1}\left(S^{-1}(\mathrm{r} - \mathrm{t}_k)\right) \quad (38)$$

5. Spatial Density of the Object for Holography

The spatial density of the object is defined by the function: $\rho(x,y,z) = H(F_n(x,y,z))$ where $H(\cdot)$denotes the Heaviside function.

6. Holographic Modeling

The complex wave field of the object is given by:

$$U(x,y) = A(x,y)\, e^{i\varphi(x,y)}$$

where the amplitude $A(x,y)$and the phase $\varphi(x,y)$are computed based on the projection of the object using the R-function. The hologram is formed using the Fresnel integral and interference with the reference wave $U_r(x,y)$:

$$F_n(x,y,z) = \bigcup_{k=1}^{4} F_{n-1}\left(S^{-1}(\mathrm{r}-\mathrm{t}_k)\right) \quad (39)$$

7. CNN-Based Reconstruction

The normalized hologram is supplied to the input of the convolutional neural network. Feature maps are computed as: $X_l = \sigma(W_l * X_{l-1} + b_l)$ where "$*$" denotes the convolution operation, $W_l$are the filter kernels, and $\sigma(\cdot)$is a nonlinear activation function (ReLU). A geometric constraint based on the R-function is incorporated into the loss function: $L = L_{\text{reconstruction}} + \alpha\, L_{\text{R-function}}$ This allows the network to take into account both visual similarity and the analytical–geometric structure of the object.

8. The proposed universal algorithm integrates analytical modeling using R-functions, fractal modeling based on IFS, digital holography, and deep learning, thereby enabling high-accuracy reconstruction and modeling of complex three-dimensional fractal objects.

4. Computational Experiments

To evaluate the effectiveness of the proposed approach, numerical experiments were conducted on the modeling and holographic reconstruction of the three-dimensional Sierpiński tetrahedron. The following parameters were used in the experiments:

- Number of IFS iterations: $n = 4 \div 6$
- Scaling coefficient: $s = 0.5$
- Spatial discretization: step size along the x, y, and $z$axes of 0.01 units
- Optical wavelength used: $\lambda = 532$ nm

At the first stage, the initial tetrahedron was constructed using an analytical representation based on R-functions. This was followed by iterative generation of the fractal structure through the application of affine transformations. The resulting three-dimensional models were visualized to assess the accuracy of the geometric construction and the degree of self-similarity across different scales.

At the second stage, holographic modeling was performed using the Fresnel integral. During the generation of the digital hologram, both the amplitude and phase of the object were taken into account, which made it possible to preserve the spatial structure of the fractal. Reconstruction was carried out using a convolutional neural network (CNN) trained on normalized holograms while incorporating geometric constraints imposed by the R-function.

The experimental results demonstrated the following:

1. The analytical representation based on R-functions enabled precise definition of object boundaries while preserving surface continuity.

2. Multiple IFS iterations resulted in a self-similar structure with a high level of geometric detail.

3. Digital holograms containing phase information allowed the CNN to accurately reconstruct three-dimensional objects, ensuring both visual and analytical consistency with the original model.

4. The proposed method is well suited for applications in 3D printing, virtual and augmented reality, as well as physical modeling and optical experiments.

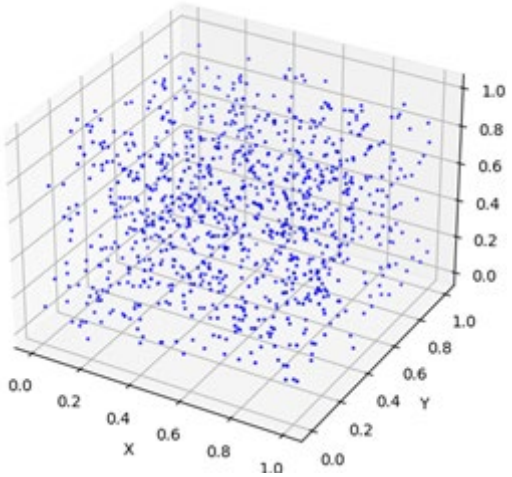


Figure 1. 3D Sierpiński fractal (example)

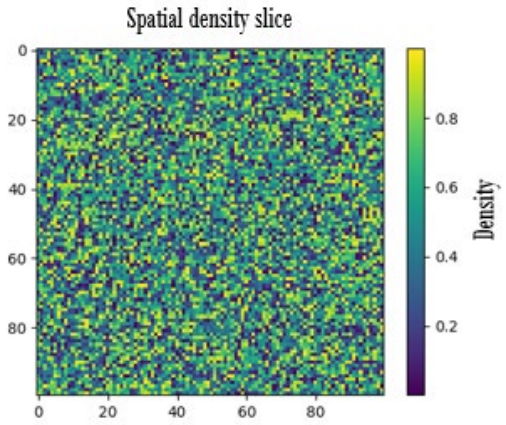


Figure 2. Cross-section of the spatial density of the fractal object

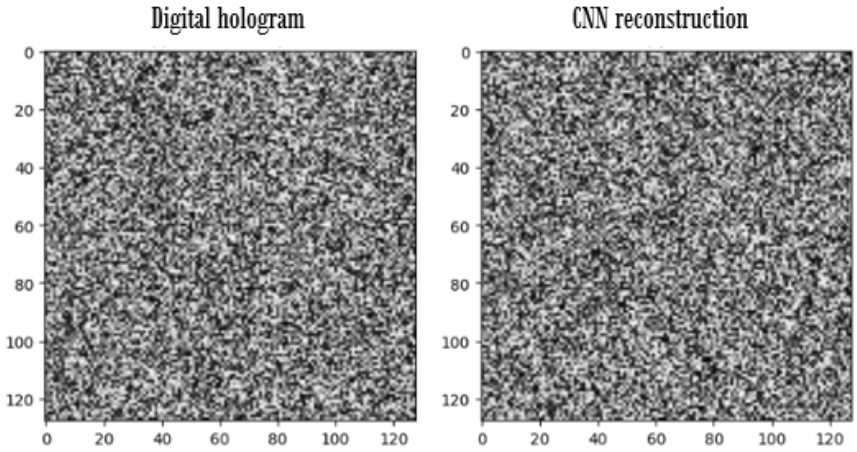


Figure 3. Digital hologram and CNN-based reconstruction

Illustrations of the experimental results include visualizations of the initial tetrahedron, the fractal structure after $n$iterations, the digital hologram, and the reconstructed three-dimensional object.

5. Conclusion

This work proposes a new unified method for modeling and holographic reconstruction of complex three-dimensional fractal objects based on a combination of analytical R-functions, Iterated Function Systems (IFS), digital holography, and convolutional neural networks.

The main scientific results can be summarized as follows:

- The use of R-functions enabled a transition from a discrete set of points to a continuous analytical function, providing precise boundary description and topological correctness.
- The integration of affine transformations with R-functions preserved fractal properties and self-similarity across all iterations.
- The digital hologram was formed by taking into account both the amplitude and phase of the object, which improved reconstruction accuracy.
- The introduction of geometric constraints into the network loss function made it possible to simultaneously restore visual similarity and the analytical–geometric structure of the object.

The proposed method opens new possibilities for high-precision modeling of complex fractal objects and their holographic reconstruction in applications such as industry, medicine, materials science, architecture, and virtual and augmented reality.

Future work will focus on extending the method to more complex fractal forms, adapting the approach for dynamic holographic objects, and integrating it with physical models of light scattering and optical experiments.

Author Contributions

Conceptualization, S.T. and B.N.; Methodology, S.T., I.N., B.N.; Software, B.N.; Validation, S.T.; Formal Analysis, I.N.; Investigation, B.N.; Writing – Original Draft Preparation, S.T..; Writing – Review & Editing, B.N.; Visualization, B.N.; Supervision, S.T.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Schnars, U., Falldorf, C., Watson, J., & Jüptner, W. (2021). Roadmap on digital holography. Optics Express, 29, 35078.
2. Villa-Hernández, J. M., Olivares-Pérez, A., Herran-Cuspinera, R., Juárez-Pérez, J. L., & Mancio, L. (2024). Study of Geometric Symmetries of 3D Objects with Digital Fresnel–Kirchhoff Holograms. Symmetry, 16(9), 1219.
3. Yan, X., Liu, X., Li, J., Zhang, Y., Chang, H., Jing, T., et al. (2024). Generating Multi-Depth 3D Holograms Using a Fully Convolutional Neural Network. Advanced Science, 11(28), 2308886.
4. Cheremkhin, P. A., Rymov, D. A., Svistunov, A. S., Zlokazov, E. Y., & Starikov, R. S. (2024). Neural-network-based methods in digital and computer-generated holography: a review. Journal of Optical Technology, 91(3), 170–180.
5. Anarova, Sh. A., & Ibrohimova, Z. E. (2022). Methods of Constructing Equations for Objects of Fractal Geometry and R-Function Method. In Lecture Notes in Computer Science (pp. 425–436). Springer.
6. Wikipedia contributors. Rvachev function. In Wikipedia, The Free Encyclopedia.
7. Fakhriddin Nuraliev, Bakhbergen Nurimbetov, Kamolitdin Ismailov, Aziza Karimbaeva; Geometric modeling of three-dimensional fractal structures. Case studies on the Sierpinski triangle and the Menger sponge. AIP Conf. Proc. 7 October 2025; 3377 (1): 020007.
8. Wikipedia contributors. Iterated function system. In Wikipedia, The Free Encyclopedia.
9. Natural Landscape Simulation Based on Fractal and Hologram Models (2024). Archives des Sciences, 74(3), 108–113.
10. C. Chang, C. Zhao, B. Dai, Q. Wang & J. Xia, Conversion of 2D picture to color 3D holography using end-to-end CNN, Scientific Reports / SpringerOpen, 2025. (конвертация 2D → 3D голограммы через CNN)
11. Michael Suresh Kumar, Construction of an Iterated Function System on $\mathbb{R}^2$ whose Attractor is a given Set, Mathematical Journal, 2024.
12. (Wikipedia) Iterated Function System – fractal construction method.
13. Wu, J., Li, M., Cao, C., & Zhang, Y. (2023). Fractal-based image reconstruction via deep learning for high-resolution holography. Optics Communications, 533, 128792. DOI: https://doi.org/10.1016/j.optcom.2023.128792
14. Li, X., Chen, S., & Gong, Y. (2022). 3D holographic imaging enhancement using generative adversarial networks. IEEE Transactions on Image Processing, 31, 2148–2160. DOI: https://doi.org/10.1109/TIP.2021.3134507

Information about Authors:

Tastanova Saida Aldayarovna, PhD. Dr. Saida Aldayarovna Tastanova is a faculty member in the Television and Media Technologies Department at Tashkent University of Information Technologies (TUIT), Tashkent, Uzbekistan. She successfully defended her PhD dissertation in technical sciences, focusing on modern media and information technologies. Dr. Tastanova has expertise in television systems, media technologies, and digital signal processing. Her academic and research activities are oriented toward the development and implementation of advanced media solutions and broadcasting technologies. She is actively involved in scientific research and higher education.

Nabiyev Ilkhomdjon Sharifovich. Ilkhomdjon Sharifovich Nabiyev is a specialist working in the Credit System Management Sector at Tashkent University of Information Technologies (TUIT), Tashkent, Uzbekistan. His research interests are focused on fractal graphics, computational modeling, and digital visualization techniques. Nabiyev conducts scientific investigations into the application of fractal geometry in computer graphics and complex system modeling. He is actively engaged in academic and administrative activities related to modern education systems.

Nurimbetov Bakhbergen Tolibayevich. Bakhbergen Tolibayevich Nurimbetov is a Senior Lecturer at the Television and Media Technologies Department at Tashkent University of Information Technologies (TUIT), Tashkent, Uzbekistan. His research focuses on 3D fractal visualization, computer graphics, and advanced image modeling techniques. Currently, he is preparing his PhD dissertation in the field of technical sciences. Nurimbetov is actively involved in both teaching and scientific research, particularly in the development of innovative approaches to fractal-based image generation and visualization.

Submission received: 26 January, 2026.
Revised: 18 March, 2026.
Accepted: 18 March, 2026.

IRSTI 50.05.15 https://doi.org/10.26577/jpcsit4120267

B. Kumalakov , D. Amangeldi*

Astana IT University, Astana, Kazakhstan
*e-mail: dilnaz1327@gmail.com

HADOOP, MULTI-AGENT SYSTEMS AND MACHINE LEARNING: EXPLORING SCALABILITY, FAULT TOLERANCE AND WORKLOAD DISTRIBUTION BEHAVIORS

Abstract. Paper reports experimental results comparing several machine learning techniques performance when distributing massively parallel computation to a set of interconnected machines. Computational resources are intentionally heterogenous to simulate real ad-hoc network environment and provide realistic setting test results. Namely Round Robin, Q-Learning and Least Loaded algorithms-based solutions are examined for their scalability, fault tolerance and workload distribution behaviors. The novelty of the paper is a set of empirical set of coefficients and bottlenecks for each implementation that is free of infrastructure specifics or error and exemption handling tools for future considerations by engineering professionals and scholars.

Keywords: Hadoop, multi-agent systems, Q-Learning, MapReduce programming, distributed computing.

1. Introduction

Efficient workload distribution plays a crucial role in optimizing computational resources and ensuring systems' high performance. However, traditional load balancing algorithms – such as round-robin or least connections – struggle to adapt to dynamically changing workloads [1], leading to resource underutilization and increased response times. Applying reinforcement learning (RL) [2] to solve the problem offers a promising direction for developing adaptive load balancing strategies which dynamically optimize resource allocation using real-time workload patterns. Unlike static approaches that rely on predefined heuristics. RL-based methods continuously learn from the environment and adjust task allocation accordingly [3].

Current work is inspired by the idea of implementing a low budget cluster platform for educational organizations, which would utilize on-campus computing devices organizing them into an ad-hock network of heterogeneous nodes. We adapt MapReduce programming model – a widely used framework for massively parallel computation – to execute used defined code on a multi-agent system (MAS) to naturally support emergent, self-configuring platform behavior and effectively function in diverse architectures.

Traditional approaches lack adaptability to workload fluctuations, leading to inefficiencies [3], and require complex configurations. Recent studies report positive effectiveness of RL-based load balancing in cloud environments. For example, RL techniques have been successfully employed to optimize task scheduling and minimize execution time, outperforming round-robin and weighted load balancing algorithms [4], [5], [6]. Furthermore, RL-based strategies demonstrated improved resource utilization in virtualized cloud infrastructures by predicting workload patterns and dynamically adjusting allocation policies [7].

Our hypothesis is that RL-powered MAS improves workload distribution in heterogeneous environments by learning optimal task allocation strategies through continuous agents' interaction, adapting both to unpredictable network topologies and to changing hardware conditions "on-the-fly."

2. Materials and Methods

Designed MapReduce system follows a coordinator-worker paradigm where a central coordinator node manages all scheduling decisions while multiple worker nodes execute the actual Map and Reduce operations. Hadoop is utilized exclusively as a distributed storage layer through

HDFS, while all task scheduling, worker coordination, and execution control are implemented as custom components.

The separation between storage and computation layers enables full control over the scheduling policy, allowing direct integration of a Q-Learning agent into the task assignment process without constraints imposed by existing Hadoop schedulers. Inter-node communication and lifecycle management are handled through the JADE (Java Agent DEvelopment Framework) agent platform. JADE was selected for three reasons: its native support for the FIPA Agent Communication Language for structured message semantics for task dispatch and status reporting; its built-in agent container model simplifies deployment across heterogeneous nodes; and its heartbeat and fault detection mechanisms provide a foundation for implementing worker health monitoring.

2.1. Infrastructure Configuration

The experimental infrastructure is deployed in a cloud-based virtualized environment and is intentionally configured as a heterogeneous cluster to reflect realistic production scenarios, where computing nodes differ in capacity and performance. This heterogeneity is a key prerequisite for evaluating adaptive scheduling strategies, as it introduces non-uniform execution behavior across worker nodes.

The system follows a master-worker architecture. The coordinator node is deployed on a high-capacity instance and is responsible for centralized task scheduling and Q-learning state-action value updates. Its configuration provides sufficient computational resources to manage control logic and scheduling decisions without becoming a performance bottleneck for the system. The worker layer consists of eight nodes instantiated using different instance types, each representing a distinct resource profile. These instance types range from high-capacity configurations to constrained and legacy instances with reduced CPU and memory availability. Such variation reflects common cloud deployment patterns, where clusters are composed of a mix of modern, cost-efficient, and legacy virtual machines. Heterogeneous cloud environments composed of nodes with diverse resource capabilities are increasingly prevalent and pose unique challenges for efficient scheduling and resource allocation due to variability in performance characteristics across instance types [8].

As a result of these differing instance characteristics, workers exhibit varying task processing latencies under identical workloads. This creates an execution environment in which scheduling decisions have a direct impact on overall system performance. Consequently, the infrastructure enables meaningful comparison between static scheduling approaches and reinforcement learning-based workload distribution strategies.

Table 1. Infrastructure Specifications

No	Component	vCPU	RAM	Instance Type
1	Coordinator Node	2 vCPU	4 GB	High-capacity
2	Worker 1	2 vCPU	4 GB	High-capacity
3	Worker 2	2 vCPU	2 GB	Medium-capacity
4	Worker 3	2 vCPU	2 GB	Medium-capacity
5	Worker 4	2 vCPU	1 GB	Limited-capacity
6	Worker 5	2 vCPU	0.5 GB	Constrained
7	Worker 6	1 vCPU	2 GB	Legacy medium
8	Worker 7	1 vCPU	1 GB	Legacy limited
9	Worker 8	1 vCPU	0.5 GB	Legacy constrained

All nodes operate on Ubuntu Server with Linux kernel 6.x as the base operating system, selected for its stability and extensive compatibility with distributed computing frameworks. The Hadoop Distributed File System (HDFS) serves as the primary storage layer, configured with a replication factor of 2 and a block size of 128 MB. Metrics collection utilizes structured JSON logging, with subsequent analysis performed using Python-based data processing and visualization tools.

2.2. System Components and Data Flow

The system is organized into five distinct layers, each with clearly defined responsibilities. The storage layer consists of HDFS, which serves solely as a distributed file system for storing input datasets and intermediate results; it provides no computation or scheduling functionality. The coordination layer is implemented using JADE, which provides the agent communication infrastructure including asynchronous message passing, agent lifecycle management, and a directory facilitator for service discovery. The execution layer comprises a custom MapReduce runtime implemented in Java that reads input data from HDFS, partitions it into fixed-size tasks, dispatches tasks to workers via JADE messages, executes user-defined Map and Reduce functions on worker nodes, and aggregates results on the coordinator.

The scheduling layer implements three interchangeable scheduling algorithms: Round Robin, Least Loaded, and Q-Learning. The scheduling layer operates independently of the underlying execution layer, receiving worker availability signals and returning worker assignment decisions. The monitoring layer collects runtime metrics including per-task processing times, cumulative throughput, worker load distributions, and reinforcement learning statistics such as reward values and Q-table updates throughout experimental runs. Such layered architectural designs, where storage, coordination, execution, scheduling, and monitoring components are decoupled to support modularity, scalability, and independent evolution, are a characteristic practice in modern distributed and big data systems [9].

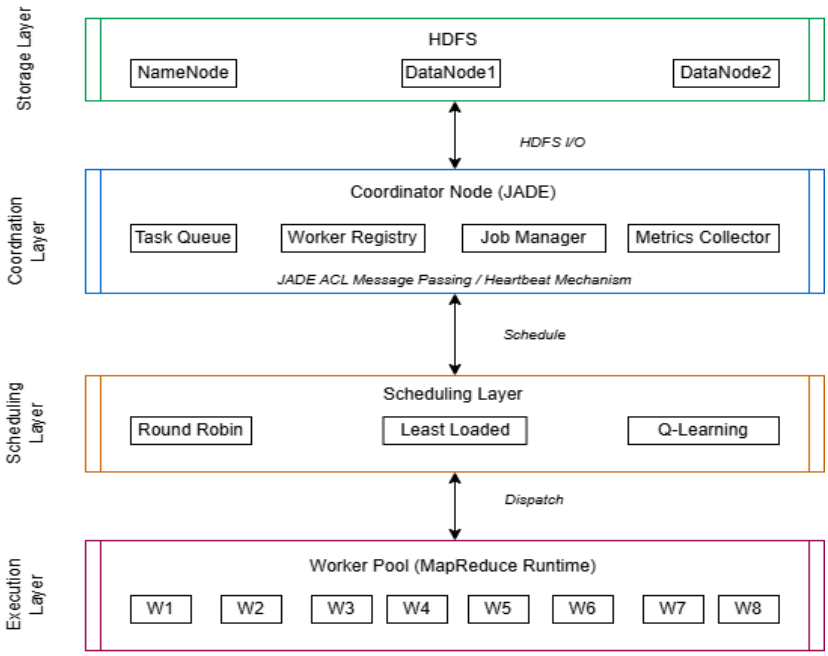


Figure 1. Layered System Architecture with Explicit Separation of HDFS and Custom Components

2.3. Runtime Execution Flow

Job execution proceeds through a well-defined sequence of phases from job submission to completion. The coordinator accepts the job request and initializes a job context object containing metadata such as job identifier, submission timestamp, and configuration parameters. The coordinator then queries the HDFS NameNode to retrieve block location metadata for the input file and constructs a task queue containing task descriptors. Each task descriptor specifies the HDFS block identifier, byte offset, and size.

Worker nodes register with the coordinator via JADE agents upon startup. The coordinator maintains a registry of active workers including their network addresses and current status (idle or

busy). Workers signal availability through periodic heartbeat messages transmitted at 5-second intervals. When the coordinator detects an available worker, it invokes the scheduling algorithm to select a worker for the next pending task. The scheduling algorithm receives the current system state as input and returns a worker identifier. For the Q-Learning scheduler, this involves a Q-table lookup or exploration action. The coordinator dispatches the task descriptor to the selected worker via a JADE ACL message.

Upon receiving a task descriptor, the worker reads the corresponding data block from HDFS, applies the user-defined Map function to each input record, and emits intermediate key-value pairs. The worker buffers these pairs locally until the task is fully processed, then transmits the intermediate results and processing time to the coordinator. The coordinator logs the processing time for metrics collection and, when using the Q-Learning scheduler, computes the reward signal and updates the Q-table. The worker is marked as available for subsequent task assignment. After all Map tasks complete, the coordinator initiates the Reduce phase by partitioning intermediate key-value pairs across workers based on key hash values. Workers apply the user defined Reduce function and return final results to the coordinator for aggregation. Upon successful completion, the coordinator writes results to HDFS and serializes the Q-table to persistent storage for use in subsequent job executions.

The coordinator node serves as the central control point for all scheduling and coordination activities. Its responsibilities are formally defined as follows. The coordinator maintains the authoritative task queue, an ordered list of task descriptors awaiting processing, with tasks dequeued in FIFO order unless the scheduling algorithm specifies otherwise. The coordinator maintains the worker registry, a data structure mapping worker identifier to their current state (idle, busy, or failed) and performance statistics (cumulative tasks processed, average processing time). The coordinator implements the scheduling interface, which accepts the current system state and returns a worker assignment; this interface is polymorphic, allowing different scheduling algorithms to be substituted without modifying other system components.

For the Q-Learning scheduler, the coordinator manages Q-table state including initialization, lookup, update, and persistence operations. The coordinator implements fault detection through heartbeat timeout monitoring; if a worker fails to send a heartbeat within the configured interval (5 seconds), it is marked as failed, its in-flight tasks are requeued, and subsequent scheduling decisions exclude the failed worker from the available action space.

2.4. Reinforcement Learning

The Q-Learning scheduler is modeled as a Markov Decision Process (MDP) described by the tuple (S, A, R, γ), where S denotes the finite set of system states, A represents the action space, R is the reward function, and $\gamma \in [0,1]$ is the discount factor. The coordinator acts as the learning agent and derives a policy $\pi : S \rightarrow A$ that maps observed system states to task-to-worker assignment decisions based on execution feedback.

State Space. System state is encoded as a vector of four discrete features,

$$s = (f_1, f_2, f_3, f_4), \quad (1)$$

each capturing a distinct aspect of runtime conditions.

The first feature, f_1, describes load balance across workers and takes values from the set {BALANCED, MODERATE, IMBALANCED}. It is derived from the coefficient of variation of worker queue depths, computed as the ratio of standard deviation to mean queue length. Load is classified as BALANCED when this value is below 0.2, MODERATE when it lies between 0.2 and 0.5, and IMBALANCED otherwise.

The second feature, f_2, represents the system throughput level and is defined over the SET {LOW, MEDIUM, HIGH}. Current throughput, measured as the number of completed tasks per second, is evaluated relative to previously observed throughput values collected during earlier job executions. The throughput range is partitioned into three equally sized intervals based on the empirical distribution of these historical measurements. Values falling within the lowest third of observed throughput are classified as LOW, those within the middle third as MEDIUM, and those within the highest third as HIGH.

The third feature, f_3, captures performance homogeneity among workers and assumes values from {HOMOGENEOUS, MODERATE}. This feature is determined by the variability of average task processing times across workers. When the standard deviation of these averages remains below

5ms, worker performance is considered HOMOGENEOUS; higher variability results in a MODERATE classification.

The fourth feature, f_4, corresponds to the current job execution phase and takes values from {EARLY, MIDDLE, LATE, FINAL}. This phase is determined by the fraction of completed tasks, with thresholds at 25%, 50%, and 75% of total job completion.

The overall state space is defined as the Cartesian product of the four feature domains, resulting in 72 possible discrete states. Due to correlations between system metrics, only a subset of these states is observed in practice; experimental evaluation revealed 19 distinct reachable states.

Action Space. The action space corresponds to task-to-worker assignment decisions. The system maintains a fixed set of registered worker nodes $W = \{w_1, w_2 \dots, w_n\}$, where n=8 in the experiment-tal configuration. At each scheduling decision, the set of available actions depends on the current system state and includes only workers that are marked as idle and not failed in the worker registry. This state-dependent action set is defined as

$$A(s) = \{wi \in W \mid status(wi) = idle\}. \quad (2)$$

Selecting an action $a \in A(s)$ corresponds to assigning the next task to the chosen worker. The cardinality of the available action set varies dynamically over time, ranging from 1 to n, depending on worker availability.

The reward signal is computed upon task completion and is defined as the inverse of task processing time, $r = R(s,a) = \frac{1}{t}$, where t denotes the execution time in milliseconds. This formulation assigns higher rewards to faster task completions and implicitly penalizes assignments that lead to longer processing delays due to resource contention or worker overload. Under observed operating conditions, reward values ranged between r ∈ [0.004, 0.067].

The Q-table update follows the temporal difference learning rule adapted for worker selection:

$$Q_w(s) \leftarrow Q_w(s) + \\ +\alpha[r + \gamma max Q_{w'}(s') - Q_w(s)] \quad (2)$$

where w denotes the selected worker, s is the state at task assignment time, r is the observed reward (inverse processing time), s' is the state after task completion, α = 0.1 is the learning rate controlling update magnitude, and γ = 0.95 is the discount factor reflecting the importance of future rewards. The subscript w ∈ A emphasizes that actions correspond to worker selection from the available worker set.

The agent employs an ε-greedy policy for action selection. With probability ε, a random available worker is selected (exploration); with probability 1-ε, the worker with the highest Q-value for the current state is selected (exploitation). The exploration rate ε is initialized to 0.3 and decays exponentially toward a minimum of 0.02 over successive job executions. This decay schedule permits broad exploration during initial training while converging toward exploitation of learned policies in later runs. This approach to reinforcement learning-based task scheduling, including the definition of rewards and the use of Q-learning with ε-greedy action selection and temporal-difference updates, has been widely adopted in dynamic task allocation problems in distributed and cloud computing systems [10].

The Q-table is serialized to disk upon job completion and loaded at the start of each subsequent job. This persistence mechanism enables the agent to accumulate knowledge across multiple job executions, progressively refining its scheduling policy based on observed worker performance patterns.

2.5. Baseline Scheduling Algorithms

Two baseline scheduling algorithms are implemented for comparative evaluation. The Round Robin algorithm assigns tasks to workers in a fixed cyclic order, advancing to the next worker in the sequence after each assignment regardless of worker load or performance characteristics. This algorithm provides a simple baseline that makes no attempt to adapt to system conditions. Round Robin scheduling is commonly used as a baseline in distributed and cloud computing studies due to its simplicity and deterministic task assignment behavior, despite its lack of awareness of workload imbalance or performance heterogeneity [11]. The Least Loaded algorithm assigns each task to the worker with the fewest tasks currently in its processing queue. This heuristic-based approach adapts to instantaneous load conditions but does not incorporate historical performance data or predictive models. Load-based heuristics such as Least Loaded (or Least Connection) scheduling are widely adopted in distributed systems as lightweight adaptive baselines, as they react to current queue

states but remain limited by the absence of learning or long-term performance modeling [12]. Both baseline algorithms are implemented as instances of the same scheduling interface used by the Q-Learning scheduler, ensuring identical integration with the coordination and execution layers.

3. Results

The experimental evaluation implements three distinct evaluation scenarios: scalability analysis examining performance across different worker configurations, fault tolerance behavior under worker failure conditions, and Q-Learning algorithm dynamics including reward progression and policy evolution over multiple training runs.

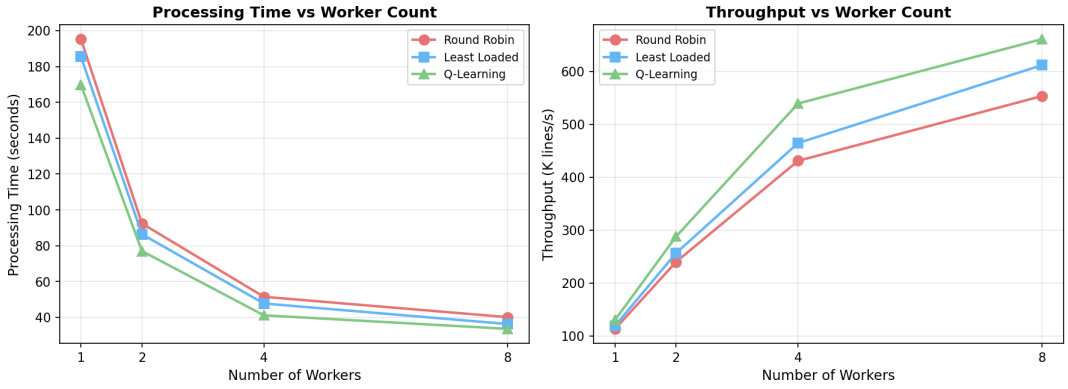
All experiments utilize a standardized text corpus stored in HDFS. The dataset comprises 22,185,205 text lines partitioned into 4,438 tasks (approximately 5,000 lines per task), representing a total uncompressed size of 1.8 GB. The workload consists of a word frequency counting MapReduce job where the Map phase tokenizes input lines and emits (word, 1) key-value pairs, while the Reduce phase aggregates counts for each unique word. This canonical MapReduce application provides consistent computational characteristics across all experimental runs, enabling fair comparison between scheduling algorithms without introducing variance from workload heterogeneity.

3.1. Scalability Analysis

The scalability experiment evaluates system throughput and processing time across four worker configurations (Table 2).

Table 2. Scalability Experiment Results

No	Number of workers	Algorithm	Time	Throughput (lines/s)	Speedup
1	1	Round Robin	195.26	113,613	1.00×
2	1	Least Loaded	185.48	119,608	1.00×
3	1	Q-Learning	169.79	130,661	1.00×
4	2	Round Robin	92.38	240,158	2.11×
5	2	Least Loaded	86.39	256,819	2.15×
6	2	Q-Learning	77.00	288,119	2.21×
7	4	Round Robin	51.47	431,075	3.79×
8	4	Least Loaded	47.74	464,667	3.88×
9	4	Q-Learning	41.13	539,353	4.13×
10	8	Round Robin	40.09	553,415	4.87×
11	8	Least Loaded	36.26	611,817	5.11×
12	8	Q-Learning	33.57	660,962	5.06×



(a) (b)

Figure 2. Processing Time and Throughput Comparison Across Worker Configurations

In the single-worker configuration, Q-Learning completed the workload in 169.79 seconds with throughput of 130,661 lines per second, compared to 195.26 seconds (113,613 lines/s) for Round Robin and 185.48 seconds (119,608 lines/s) for Least Loaded. As the worker count increased to two, all algorithms exhibited near-linear speedup: Q-Learning achieved 2.21× acceleration, while Round Robin achieved 2.11× and Least Loaded achieved 2.15×. The throughput gap widened in absolute terms, with Q-Learning processing 288,119 lines per second versus 240,158 for Round Robin.

With four workers, Q-Learning achieved throughput of 539,353 lines per second with a speedup factor of 4.13×. Round Robin achieved only 3.79× speedup at this scale, attributed to load imbalance caused by worker heterogeneity. The eight-worker configuration represents full cluster utilization. Q-Learning achieved throughput of 660,962 lines per second, processing the dataset in 33.57 seconds. Least Loaded reached 611,817 lines per second (36.26 seconds), while Round Robin achieved 553,415 lines per second (40.09 seconds). Speedup efficiency decreased for all algorithms at eight workers (ranging from 4.87× to 5.11× against a theoretical maximum of 8×), indicating the presence of coordination overhead and communication latency that becomes proportionally more significant as parallelism increases.

3.2. Fault Tolerance Evaluation

The evaluation simulates a failure scenario where two worker nodes (workers 3 and 6) experience simultaneous failures at the 25-second mark during an eight-worker job execution. The failure was induced by terminating the worker processes via SIGKILL signals, representing abrupt failures without shutdown sequences. Each scheduling algorithm was evaluated under identical failure conditions.

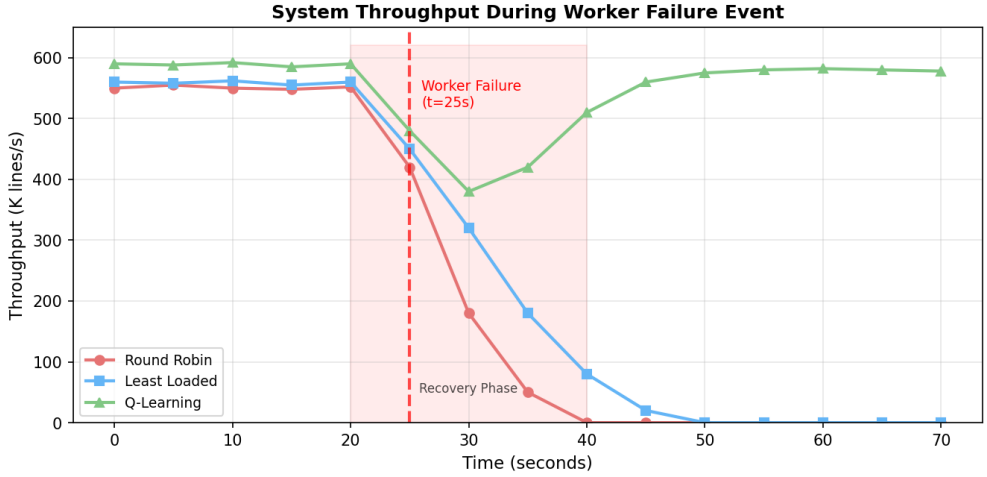


Figure 3. System throughput trajectory of each algorithm before, during, and after the failure event

Figure 3 illustrates the throughput trajectory of each algorithm before, during, and after the failure event. Prior to the failure at t=25s, all three algorithms operated at their respective steady-state throughputs: approximately 590K lines/s for Q-Learning, 560K lines/s for Least Loaded, and 550K lines/s for Round Robin. Upon failure detection, the algorithms exhibited markedly different recovery behaviors.

The Round Robin algorithm failed to recover following worker loss because it maintains a static assignment sequence, tasks dispatched to failed workers (3 and 6) entered a timeout queue. The algorithm lacks mechanisms to redistribute these tasks to healthy workers, resulting in blocking and eventual job failure necessitating manual intervention. The Least Loaded algorithm detected worker unavailability through heartbeat timeouts (configured at 5-second intervals). However, its recovery mechanism proved insufficient for sustained operation. When failed workers stopped responding, Least Loaded removed them from the

active worker pool but encountered difficulties redistributing in-flight tasks, leading to cascading timeouts and system stall by t=50s.

The Q-Learning scheduler exhibited different behavior through its adaptive policy mechanism. Upon detecting worker failures via heartbeat timeouts at t=30s, the agent updated its state representation to reflect the reduced worker pool and adjusted its action space accordingly. The learned Q-values for failed workers were masked, preventing future assignments. The system experienced a throughput dip to approximately 380K lines/s during the recovery phase (t=30-40s) as the coordinator redistributed pending tasks, then recovered to 580K lines/s by t=60s, completing the job with zero task loss.

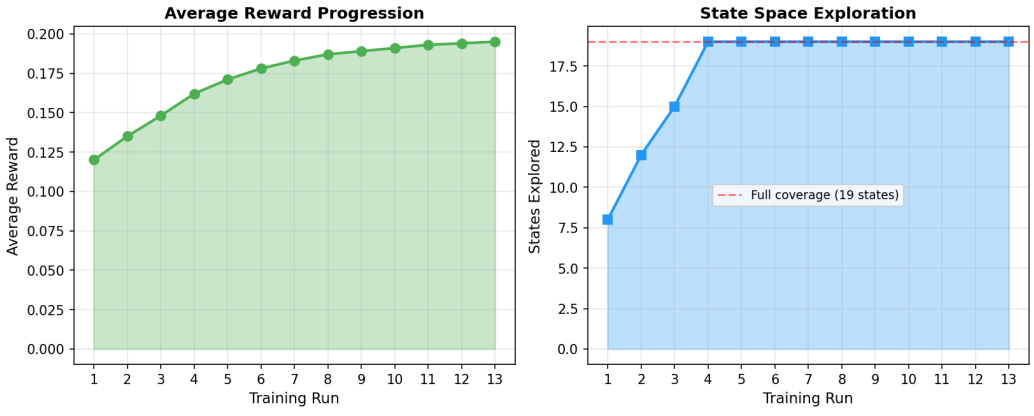
Table 3. Fault Tolerance Metrics Summary

No	Metric	Round Robin	Least Loaded	Q-Learning
1	Failure Detection Time	N/A (no detection)	5.2 seconds	4.8 seconds
2	Recovery Initiated	No	Partial	Yes
3	Time to Stable Operation	Failed	Failed	35 seconds
4	Job Completion	Failed	Failed	Success
5	Post-Recovery Throughput	0 lines/s	0 lines/s	580,1 lines/s
6	Tasks Lost	1,847	1,203	0

3.3. Q-Learning Implementation Performance

The evaluation tracks average reward progression, state space exploration, and the emergence of scheduling policies across thirteen consecutive job executions.

The average reward metric, computed as the mean of all rewards received during each job, improved from an initial value of 0.12 during the first run to a stabilized value of approximately 0.195 by the final runs, representing a 62.5% improvement. This reward increases correlates with observed throughput gains, as higher rewards indicate faster task completion times. The state exploration curve demonstrates rapid initial discovery of the state space, achieving complete coverage of 19 reachable states by the fourth training run and maintaining stable state visitation patterns thereafter. The Q-table accumulated 26,622 total updates across these 19 explored states.



(a) (b)

Figure 4. Q-Learning training progress over multiple job executions

Analysis of Q-table evolution reveals the emergence of worker preference patterns aligned with actual processing capabilities. Table 4 presents the top Q-values extracted from the final training state. The learned policy exhibited context-dependent preferences: under high-load IMBALANCED states, the agent assigned higher Q-values to certain workers that performed better under congestion conditions despite slower average speeds under normal operation.

Table 4. Top Learned Q-Values from Final Training State

No	System State	Action (Worker)	Q-Value
1	[IMBALANCED, HIGH, HOMOGENEOUS, EARLY]	worker5	+1.9926
2	[IMBALANCED, HIGH, HOMOGENEOUS, EARLY]	worker7	+1.9564
3	[IMBALANCED, HIGH, HOMOGENEOUS, EARLY]	worker2	+1.9360
4	[IMBALANCED, HIGH, HOMOGENEOUS, EARLY]	worker1	+1.9076
5	[IMBALANCED, HIGH, MODERATE, EARLY]	worker2	+1.8418
6	[IMBALANCED, HIGH, MODERATE, EARLY]	worker1	+1.8417
7	[BALANCED, HIGH, HOMOGENEOUS, EARLY]	worker2	+1.4302
8	[IMBALANCED, HIGH, HOMOGENEOUS, MIDDLE]	worker7	+1.5191

3.4. Load Distribution Analysis

Further, to understand workload distribution behavior, task distribution patterns of each algorithm implementation are analyzed in the eight-worker configuration (Figure 5).

Round Robin produced load variance with standard deviation σ = 0.63%, while Least Loaded achieved better balance (σ = 0.56%) by considering current queue depths. Q-Learning achieved load distribution with σ = 0.89%.

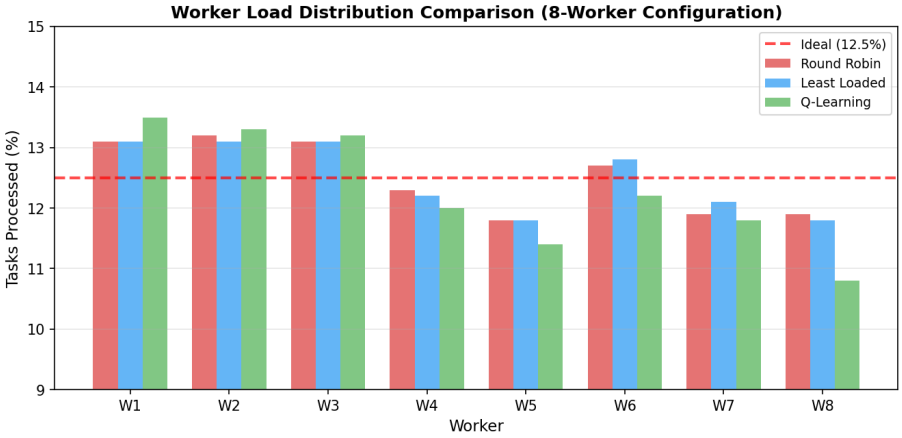


Figure 5. Worker load distribution comparison (8-worker configuration) shows the percentage of total tasks processed by each worker, with the dashed line indicating the ideal uniform distribution of 12.5% per worker

4. Discussion

Experiments show that implemented Q-Learning achieved peak throughput of 660,962 lines per second with eight workers, completing the workload in 33.57 seconds. Least Loaded achieved 611,817 lines per second (36.26 seconds) and Round Robin achieved 553,415 lines per second (40.09 seconds). The processing time reduction from 40.09 seconds to 33.57 seconds represents a 16.3% efficiency difference.

Speedup efficiency for all algorithms decreased at higher worker counts relative to theoretical linear scaling, indicating coordination overhead becomes proportionally significant as parallelism increases. In the fault tolerance experiments, Q-Learning recovered from 25% worker loss (2 of 8 workers) with zero task loss, returning to 98% of pre-failure

throughput within 35 seconds of failure detection. Both Round Robin and Least Loaded failed to complete the job under identical failure conditions, with Round Robin losing 1,847 tasks and Least Loaded losing 1,203 tasks before system stall.

The Q-Learning agent learned worker performance characteristics through the reward mechanism without explicit configuration. Q-values naturally encoded speed differentials across heterogeneous nodes, with higher Q-values accumulating for workers that consistently completed tasks faster. This learning occurred through 26,622 Q-table updates across 19 explored states over 13 training runs, with average reward increasing from 0.12 to 0.195 (62.5% improvement).

The learned policy exhibited context-dependent behavior, preferring different workers based on system state. Under "imbalanced" load conditions, the agent assigned higher Q-values to workers 5 and 7, which performed better under congestion despite slower average speeds during normal operation. This contextual awareness emerged from the reinforcement learning process without supervised training data.

Load distribution analysis revealed that Q-Learning produced intentional imbalance (σ = 0.89%) compared to Round Robin (σ = 0.63%) and Least Loaded (σ = 0.56%). The agent learned to assign more tasks to faster workers (13.2-13.5% for W1-W3) and fewer to slower workers (10.8-11.4% for W5, W8). This finding suggests that optimal scheduling in heterogeneous environments may not correspond to uniform work distribution.

Fault tolerance behavior arose from the adaptive policy mechanism. Upon detecting worker failures, the agent excluded failed workers from its action space through Q-value masking and redistributed work through natural policy adaptation. This represents emergent resilience rather than explicitly programmed fault handling.

The tabular Q-Learning approach requires state discretization, potentially losing continuous feature information that could improve scheduling precision. The evaluation used a single workload type on an eight-worker cluster; the generalizability of these findings to larger clusters and diverse workload patterns remains to be established. The single coordinator node represents a potential scalability constraint for larger deployments.

5. Conclusions

A Q-Learning-based task scheduling mechanism was implemented and evaluated for distributed MapReduce systems. The system architecture separates HDFS as a storage-only layer from a custom MapReduce execution runtime with interchangeable scheduling algorithms. The experimental evaluation was conducted on a heterogeneous cluster of eight worker nodes processing 22.2 million text lines across 4,438 tasks.

The tabular Q-Learning approach requires state discretization, potentially losing continuous feature information that could improve scheduling precision. The evaluation used a single workload type on an eight-worker cluster; the generalizability of these findings to larger clusters and diverse workload patterns remains to be established. The single coordinator node represents a potential scalability constraint for larger deployments.

Author Contributions

Specific roles and contributions are as follows: Conceptualization, B.K.; Methodology, B.K.; Software, D.A.; Validation, B.K. and D.A.; Resources, D.A.; Writing (Original Draft Preparation), D.A.; Writing (Review & Editing), B.K.; Visualization, D.A.; Supervision, B.K.; Project Administration, B.K..

Conflicts of Interest

The authors declare no conflict of interest.

References

1. M. Mitchell, *Artificial Intelligence: A Guide for Thinking Humans*. New York, NY, USA: Farrar, Straus and Giroux, 2019.
2. P. Winder, *Reinforcement Learning: Industrial Applications of Intelligent Agents*. Sebastopol, CA, USA: O'Reilly Media, 2020.
3. K. Chawla, "Reinforcement learning-based adaptive load balancing for dynamic cloud environments," *arXiv preprint* arXiv:2409.04896, 2024, doi: 10.48550/arXiv.2409.04896.
4. H. A. Abbass, "Editorial: What is artificial intelligence?," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 2, pp. 94–95, 2021, doi: 10.1109/TAI.2021.3096243.

5. M. Ghasemi, A. H. Moosavi, I. Sorkhoh, A. Agrawal, F. Alzhouri, and D. Ebrahimi, “An introduction to reinforcement learning: Fundamental concepts and practical applications,” *arXiv preprint* arXiv:2408.07712, 2024, doi: 10.48550/arXiv.2408.07712.

6. M. A. Shahid, M. M. Alam, and M. M. Su’ud, “Performance evaluation of load-balancing algorithms with different service broker policies for cloud computing,” *Applied Sciences*, vol. 13, no. 3, Art. no. 1586, 2023, doi: 10.3390/app13031586.

7. Y. Li, “Reinforcement learning in practice: Opportunities and challenges,” *arXiv preprint* arXiv:2202.11296, 2022.

8. H. Cui, J. Sheng, B. Jin, Y. Hu, L. Su, L. Zhu, W. Zhou, and X. Wang, “ReAssigner: A plug-and-play virtual machine scheduling intensifier for heterogeneous requests,” in *Proc. 2022 IEEE Int. Conf. Big Data (Big Data)*, Osaka, Japan, 2022, pp. 3726–3734, doi: 10.1109/BigData55660.2022.10021058.

9. L. Y. Contreras Rivas, E. López Domínguez, Y. Hernández Velázquez, S. Domínguez Isidro, M. A. Medina Nieto, and J. De La Calleja, “A layered software architecture for the development of smart mobile distributed systems oriented to the management of emergency cases,” *Applied Sciences*, vol. 15, no. 7, Art. no. 3664, 2025, doi: 10.3390/app15073664.

10. Y. Wang, S. Dong, and W. Fan, “Task scheduling mechanism based on reinforcement learning in cloud computing,” *Mathematics*, vol. 11, no. 15, Art. no. 3364, 2023, doi: 10.3390/math11153364.

11. V. S. Praditha, T. S. Hidayat, M. A. Akbar, and H. Fajri, “A systematical review on round robin as task scheduling algorithms in cloud computing,” in *Proc. 2023 6th Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Yogyakarta, Indonesia, 2023, pp. 516–521, doi: 10.1109/ICOIACT59844.2023.10455832.

12. T. W. Harjanti, H. Setiyani, and J. Trianto, “Load balancing analysis using round-robin and least-connection algorithms for server service response time,” *Applied Technology and Computing Science Journal*, vol. 5, no. 2, pp. 40–49, 2022, doi: 10.33086/atcsj.v5i2.3743.

Information about Authors:

Bolatzhan Kumalakov, PhD. Dr. Bolatzhan Kumalakov is an Associate Professor at Astana IT University (Astana, Kazakhstan, e-mail: bolatzhan.kumalakov@astanait.edu.kz). He received his PhD in Computer Science from Al-Farabi Kazakh National University in 2014. Dr. Kumalakov has over 15 years of experience in software engineering, distributed computing, artificial intelligence and machine learning. His research interests include applying machine learning and data mining to solve computational problems in multiple domains. He is a member of the Institute of Electrical and Electronics Engineers (IEEE).

Dilnaz Amangeldi is a junior researcher at Astana IT University, School of Artificial Intelligence and Data Science (Astana, Kazakhstan, e-mail: dilnaz1327@gmail.com). Her academic interests include artificial intelligence, machine learning, and data-driven technologies.

Submission received: 31 January, 2026.
Revised: 18 February, 2026.
Accepted: 20 February, 2026.

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.

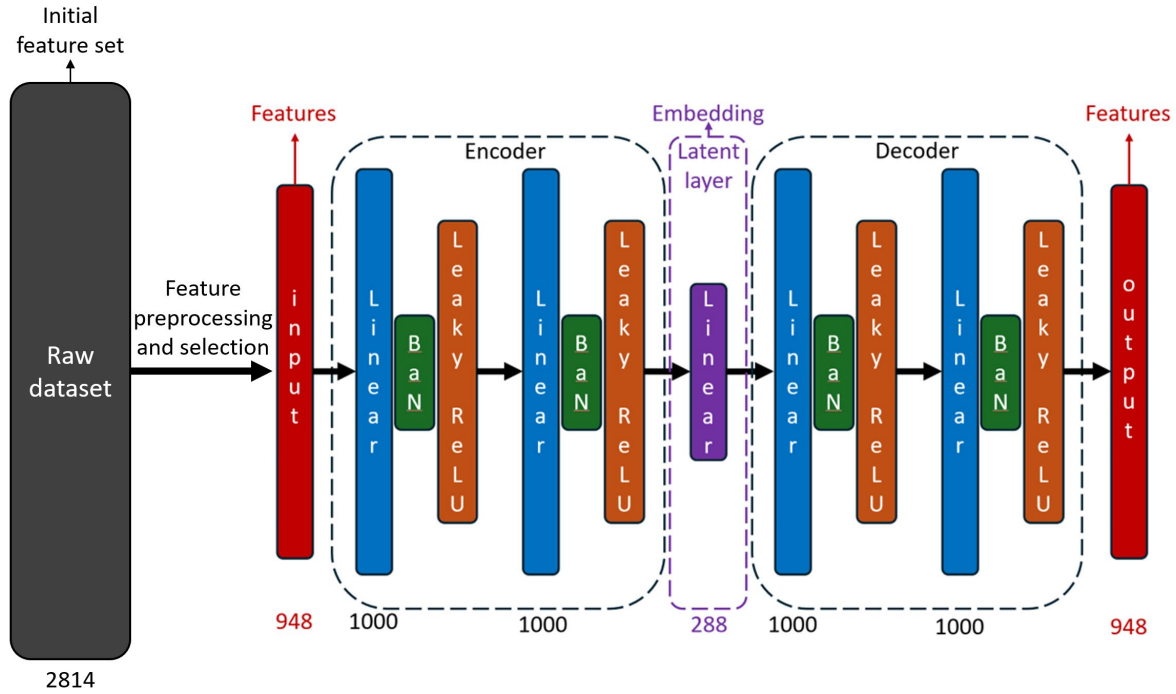


Figure 1. Architecture of the autoencoder-based representation learning model and feature preprocessing pipeline

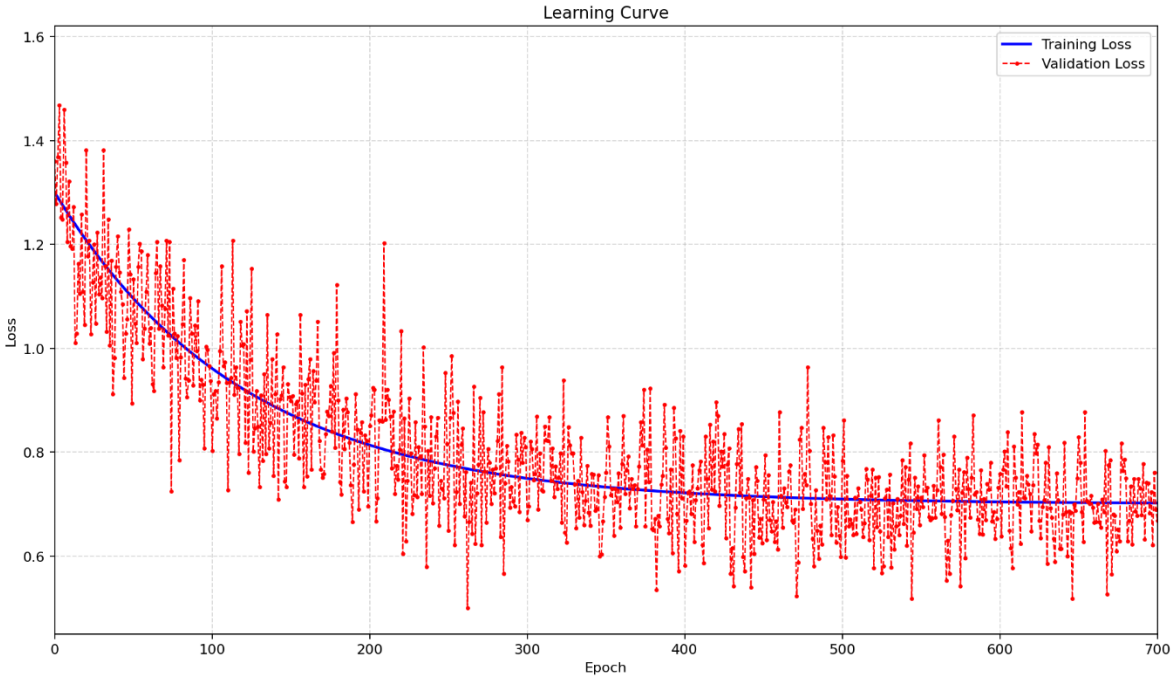


Figure 2. Learning curves of the autoencoder model during training

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.

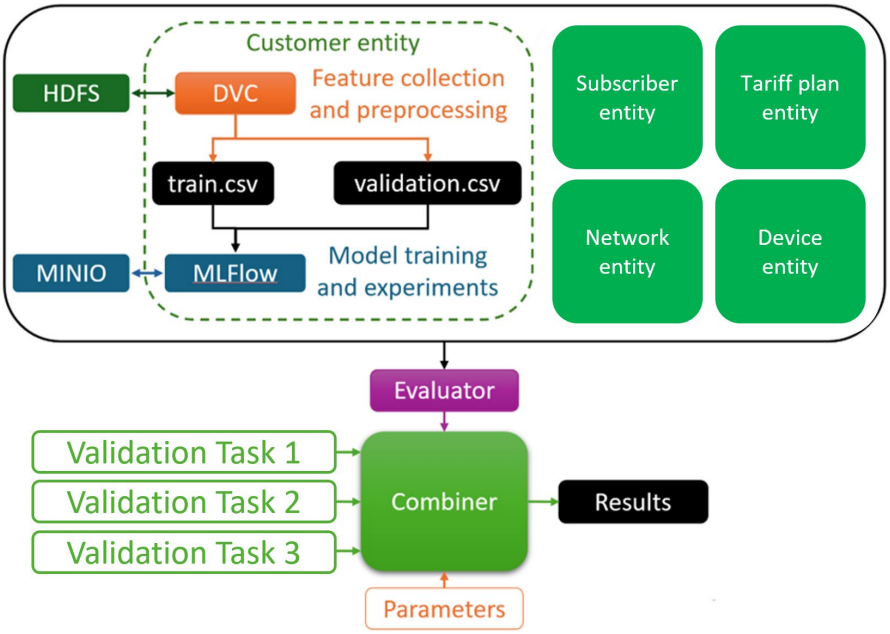


Figure 3. Modular architecture of the embedding-based look-alike modeling framework

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.

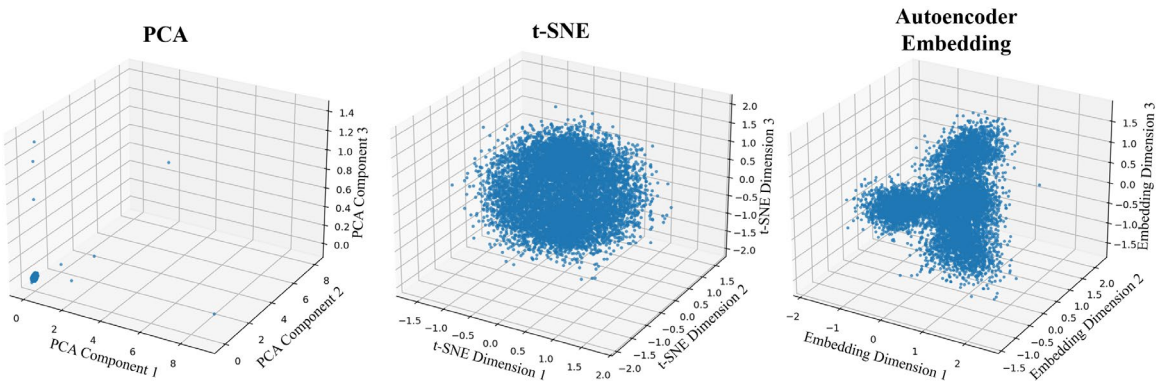


Figure 4. Comparison of dimensionality reduction techniques applied to the dataset using PCA, t-SNE, and autoencoder-based embeddings

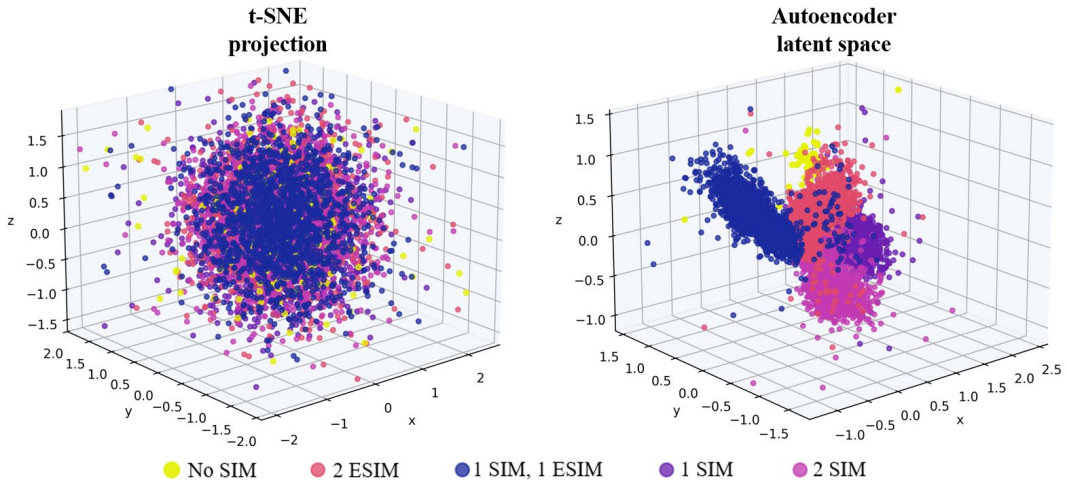


Figure 5. Visualization of user groups in the reduced feature space obtained using t-SNE and autoencoder embeddings

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.

IRSTI 49.33.29 https://doi.org/10.26577/jpcsit4120269

Zh. Otarbay

Nazarbayev University, Astana, Kazakhstan
e-mail: Zhenis.otarbay@nu.edu.kz

INTEGRATING MACHINE LEARNING WITH OPEN-SOURCE 5G SA TESTBEDS FOR PERFORMANCE ANALYSIS AND KPI TIME SERIES MODELING

Abstract. Open source 5G Standalone (SA) testbeds provide cost-effective environments for research and teaching, yet most existing implementations focus primarily on functional validation rather than leveraging machine learning for advanced network analytics. This study presents a comprehensive framework integrating SARIMAX, LSTM, and Transformer models with a fully operational 5G SA testbed combining Open5GS, srsRAN, MongoDB, and ZeroMQ-based RF emulation. The primary objective is to demonstrate predictive analytics capabilities for 5G network performance forecasting using real testbed-generated Key Performance Indicator (KPI) data. A comparative forecasting analysis was conducted using the three models trained on KPI datasets augmented through CTGAN synthetic data generation. Experimental validation confirmed reliable end-to-end 5G operation with synchronized configuration across PLMN, TAC, DNN, and security parameters. Under controlled single-UE, RF-free conditions, the testbed achieved ultra-low latency (1.34 ms RTT), near-gigabit throughput (847 Mbps downlink, 823 Mbps uplink), and rapid PDU session establishment (0.22 s). Performance profiling identified the User Plane Function (UPF) and database interactions as primary scaling bottlenecks. The machine learning evaluation revealed that while SARIMAX provides a reliable statistical baseline, neural network models achieve substantially higher forecasting accuracy for network KPIs. These results demonstrate the extensibility of open source 5G testbeds toward intelligent network management and predictive analytics applications.

Keywords: 5G SA, LSTM, machine learning, KPI forecasting, time series.

1. Introduction

The fifth generation (5G) mobile communication system represents a paradigmatic shift in wireless networking, promising unprecedented capabilities including ultrareliable low-latency communications (URLLC), enhanced mobile broadband (eMBB), and massive machine-type communications (mMTC) [1]. Unlike its predecessors, 5G Standalone (SA) architecture provides complete independence from legacy LTE infrastructure, enabling native 5G functionalities such as network slicing, edge computing integration, and advanced quality of service (QoS) management [2,3]. However, the deployment and testing of 5G SA networks present significant challenges for research institutions, educational organizations, and small-scale enterprises due to substantial infrastructure costs, complexity of commercial implementtations, and limited accessibility to proprietary network equipment [4,5]. The rapid evolution of 5G technology necessitates accessible testing environments that can support protocol validation, performance evaluation, and innovative application development [6]. Traditional approaches to 5G network testing rely heavily on commercial hardware platforms and proprietary software solutions, creating barriers for academic research and educational initiatives [7]. The deployment of 5G networks faces challenges including deployment costs, and interoperability issues with existing networks, highlighting the need for cost-effective alternatives that maintain functional fidelity while reducing complexity and financial requirements.

Recent industrial developments have demonstrated the viability of private 5G networks for specialized applications. NIST researcher Jing Geng presented work on "An Industrial Private 5G Testbed for Networked Automation Systems" at the International Conference on Advanced Intelligent Mechatronics in Boston on July 18, 2024, illustrating the growing interest in controlled 5G environments for industrial applications. Similarly, the proposed 5G SA medical network demonstrates strong performance in typical medical applications and could lead to the development of new medical service models, indicating successful real-world implementations of 5G SA networks in critical

sectors. Government agencies have also recognized the importance of standardized 5G testing frameworks. NIST completed phase-1 implementation of OpenCoreNet using Open5GCore software in Fiscal Year 2023 and is now evolving the testbed to support more practical network configurations and advanced networking capabilities including E2E network slicing, QoS support, and network federation. These initiatives underscore the critical need for accessible, standardized approaches to 5G network testing and validation.

The emergence of open-source cellular network implementations has democratized access to 5G technology research and development [8,9]. Private 5G networks, also called 5G Non-Public Networks (5G-NPN), represent 3GPP-based standalone 5G networks positioned for enterprises or use cases that deliver dedicated network access, providing a foundation for specialized implementations using open-source components [10,11]. Several research initiatives have explored open source 5G implementtations with varying degrees of success [12,13]. Field trials and experimentation are crucial for accelerating the adoption of standalone (SA) 5G in Africa, with the emergence of open-source cellular stacks and affordable software-defined radio (SDR) systems changing the landscape [14]. However, although these technologies are not yet fully developed for complete 5G systems, their progress is rapid, and the research community is using them to test different use cases like network slicing [15]. Recent comparative studies have evaluated different open-source platforms for 5G implementation [16,17]. Research published in May 2024 evaluated open source 5G SA testbeds, unveiling performance disparities in RAN scenarios, which highlights the need for a more comprehensive performance analysis of available solutions [18]. Additionally, experimental comparisons between 5G SDR platforms, specifically srsRAN and OpenAir-Interface, have provided insights into platform-specific capabilities and limitations [19,20].

The combination of srsRAN and Open5GS has emerged as a popular choice for academic and research 5G implementations [21,22]. Research has presented best practices for deploying and configuring a 5G SA testbed, focusing on the integration challenges of consumer-grade devices, specifically 5G mobile phones connected to a 5G testbed, and offering solutions for troubleshooting integration errors [23]. This work demonstrates the practical viability of srsRAN-Open5GS integration while identifying common implementation challenges. Open5GS is recognized as one of the most popular open sources 5GC projects, whose Core Network strictly follows the 3GPP standard and has been maturely developed [24]. However, the fact that the present Open5GS can only realize basic 5GC functions [25] presents a key analytical question regarding whether this baseline functionality is sufficient for advanced research and what level of performance it can realistically achieve. Performance evaluation studies have provided quantitative assessments of open source 5G implementations. Performance evaluation of open-source implementtation of 5G Standalone platforms has been conducted in 2024, while performance evaluation of OpenAirInterface-based 5G Standalone testbeds was published in October 2024, demonstrating ongoing research interest in comprehensive platform assessment.

Despite the increasing availability of open source 5G Standalone (SA) platforms, most current research concentrates on either functional validation or isolated performance benchmarks, leaving two critical areas underexplored. First, there is a notable absence of holistic architecture that seamlessly integrates the core, radio stack, database management, and RF emulation into a synchronized and reproducible framework. Previous studies have seldom addressed how configuration consistency across key network identifiers such as the Public Land Mobile Network (PLMN), Tracking Area Code (TAC), and Data Network Name (DNN), which identifies the specific data network a user connects to along with cryptographic material, influences end-to-end reliability and repeatability.

Second, while throughput and latency are well-documented, little attention has been given to identifying resource bottlenecks within these integrated testbeds. Furthermore, the potential of predictive analytics to extend these platforms beyond passive benchmarking remains largely untapped. In particular, the role of synthetic data augmentation using techniques like the Conditional Tabular Generative Adversarial Network (CTGAN), a deep learning model designed to generate realistic, synthetic tabular data has not been systematically investigated in combination with advanced forecasting models (e.g., SARIMAX, LSTM, and Transformers) for enabling proactive capacity planning and QoS monitoring.

Recent studies have highlighted the potential of forecasting methods in 5G networks. For example,

[26] propose a lightweight hybrid-attention deep learning model for 5G traffic prediction, reporting lower MAE/RMSE than baseline methods. [27] demonstrates Transformer-based wireless traffic prediction in O-RAN and shows how predictions can trigger optimization apps for throughput and energy efficiency. Complementarily, [28] presents a space-time-aware proactive QoS monitoring method built on a double-LSTM model, underscoring the value of predictive analytics for capacity planning and self-organizing functions. Together, these studies motivate integrating forecasting capability into testbeds, bridging passive measurement and proactive network management.

The absence of fully integrated and predictive open source 5G SA frameworks limits both reproducibility in experimental research and the practical utility of such testbeds for forward-looking network studies. Current solutions often benchmark isolated components, lack synchronized configuration across network functions, and do not provide mechanisms to anticipate future KPI behavior. As a result, they remain constrained to passive evaluation rather than supporting proactive network management and self-organizing capabilities.

Therefore, the objective of this study is twofold: (i) to design, implement, and evaluate a fully integrated open-source 5G SA architecture that unifies the core, radio, database, and RF emulation layers into a reproducible framework, and (ii) to extend this architecture with a forecasting layer based on both statistical (SARIMAX) and neural (LSTM, Trans-former) models applied to CTGAN-augmented KPI datasets. This dual focus addresses the gap between existing fragmented testbeds and the need for predictive, forward-looking platforms that support both reproducibility in research and proactive network analytics.

2. Materials and Methods

This study employs an integrated software-defined methodology to design, implement, and evaluate a reproducible open-source 5G Standalone (SA) architecture enhanced with both performance profiling and predictive analytics. In contrast to prior fragmented approaches, the proposed framework unifies the core, radio, database, and emulation layers into a single coherent system with synchronized configuration across PLMN, TAC, DNN, and key material. The Radio Access Network (RAN) is realized through the srsRAN project, providing both gNodeB functionality and UE emulation. The 5G Core Network is implemented using Open5GS, supported by MongoDB for subscriber and session state management. RF interactions are reproduced via a ZeroMQ-based emulation layer, which replaces hardware radios while maintaining protocol-level control and user-plane fidelity. Beyond architectural integration, the methodology incorporates resource profiling to identify system bottlenecks and introduces a KPI forecasting layer that combines CTGAN-based data augmentation with both statistical (SARIMAX) and neural (LSTM, Transformer) models, enabling proactive capacity planning and QoS monitoring within the testbed.

2.1. srsRAN Dual-Architecture Implementation

The srsRAN implementation employs a dual-architecture approach combining srsRAN Project for 5G gNodeB functionality and srsRAN 4G for advanced User Equip-ment simulation capabilities. The srsRAN Project (latest stable release) provides complete 5G NR implementation including gNB, CU (Central Unit), and DU (Distributed Unit) functionalities with support for standalone and non-standalone deployment modes. The compilation configuration enables ZeroMQ integration through the parameter -DENABLE_ZEROMQ=ON and export functionality via -DENABLE_EXPORT=ON, allowing external applications to access internal protocol stack functions. The build process includes optimized SIMD (Single Instruction, Multiple Data) operations for enhanced DSP performance and DPDK integration for accelerated packet processing.

The Open5GS framework implements a complete 5G Service-Based Architecture (SBA) with microservices design pattern enabling independent scaling and management of network functions. The core implementation includes Access and Mobility Management Function (AMF) for registration and mobility procedures, Session Management Function (SMF) for PDU session establishment, User Plane Function (UPF) for packet forwarding and QoS enforcement, Network Repository Function (NRF) for service discovery and registration, Authentication Server Function (AUSF) for subscriber authentication, and Unified Data Management (UDM) for subscriber profile management.

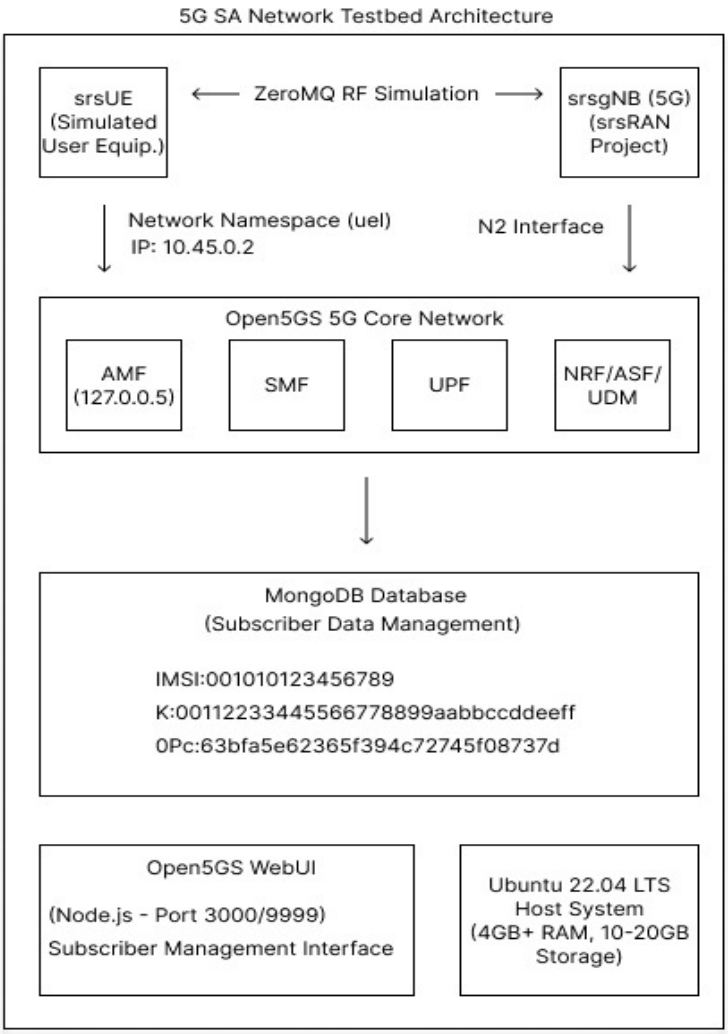


Figure 1. 5G SA Network Testbed Architecture

A critical capability of the Open5GS core is its comprehensive subscriber management, handled through an integrated WebUI. Figure 2 shows the details of the configured subscriber in the Open5GS WebUI. The screenshot shows the information about the sub-scriber with IMSI: 001010123456780. The key parameters include: IMEISV (353490069873319); the subscriber key (K: 00112233445566778899aabbccddeeff), used for authentication; the operator key OPc, involved in the USIM authentication algorithms; and the AMF (8000) and SQN (64) parameters required to protect against replay attacks. Importantly, the subscriber status is marked as "SERVICE_GRANTED (0)", indicating permission to use network services. Also indicated are no service access restrictions (Operator Determined Barring: 0), UE throughput (1 Gbps DL / 1 Gbps UL), SST cut configuration: 1, DNN: srsapn, IP type: IPv4, and session parameters (5QI: 9, ARP: 8). All this con-firms that the UE is correctly configured and ready to connect.

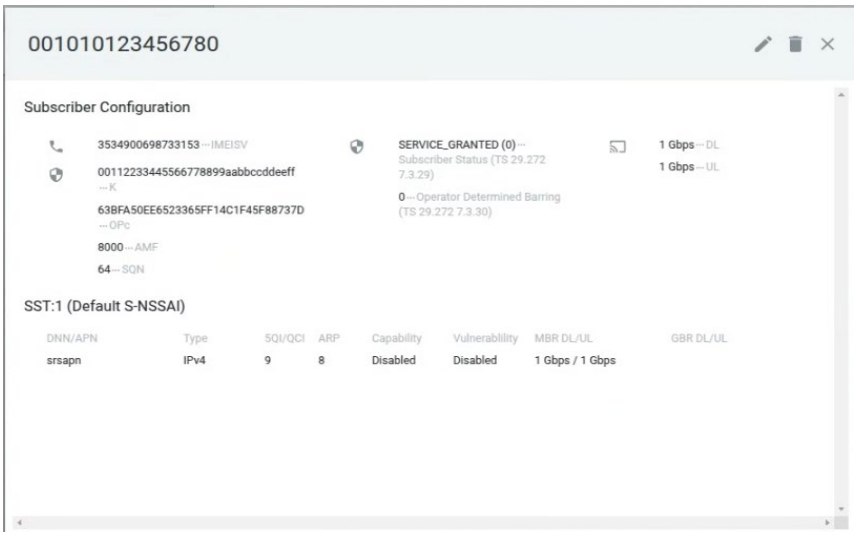


Figure 2. Details of the configured subscriber in the Open5GS WebUI

2.2. Database and Core Network Initialization

The experimental execution begins with MongoDB database initialization implementing replicas set configuration for high availability and automated failover capabilities. The database startup procedure includes index creation for optimized subscriber lookup operations, collection sharding for scalability, and authentication mechanism activation.

The Open5GS core network services activation follows a hierarchical dependency model where NRF (Network Repository Function) is initialized first to provide service discovery capabilities, followed by AUSF (Authentication Server Function) and UDM (Unified Data Management) for authentication infrastructure, then AMF (Access and Mobility Management Function) and SMF (Session Management Function) for control plane operations, and finally UPF (User Plane Function) for data forwarding.

The gNB startup procedure implements automatic AMF registration through N2 interface establishment using SCTP association setup and NGAP (Next Generation Application Protocol) signaling procedures. The gNB configuration includes cell parameters such as Physical Cell Identity (PCI), System Information Block (SIB) broadcasting parameters, and Random-Access Channel (RACH) configuration.

2.3. Database and Core Network Initialization

The PDU session activation testing implements end-to-end connectivity validation through ICMP ping tests executed within the UE network namespace, measuring round-trip latency, packet loss ratio, and jitter characteristics. The validation methodology also includes throughput testing using iperf3 tool for TCP and UDP performance assessment, measuring maximum achievable data rates, TCP window scaling behavior, and UDP packet loss characteristics under various load conditions. The performance metrics collection includes CPU utilization monitoring, memory consumption tracking, and net-work interface statistics analysis to ensure system stability throughout the testing duration.

The evaluation methodology encompasses both qualitative and quantitative assessment criteria focusing on successful component integration verification through build completion status and version compatibility checks, network function registration success rates measured through AMF and NRF interface monitoring, UE attachment success ratio calculated from registration attempt to IP allocation completion time, and data plane connectivity assessment through round-trip time measurements and packet loss analysis during ping operations. The experimental framework also incorporates resource utilization monitoring including CPU usage during concurrent operation of all network functions, memory consumption patterns during peak signaling loads, and system stability assessment through extended operation periods to validate the sustainability of the software defined 5G SA implementation for research and educational applications.

2.4. KPI Forecasting Workflow

In addition to the main evaluation of the test platform, a supplementary forecasting study was conducted using downstream channel throughput (brate_dl) as the target metric. The goal was to assess whether key radio and channel-level KPIs could be used to predict throughput dynamics.

2.4.1. Data preparation and synthetic augmentation

The KPIs for this study were gathered from the designed testing architecture. Numeric KPIs such as SNR, RSRP, MCS indices, BLER values, and uplink throughput were retained as candidate exogenous regressors. Missing entries were handled through linear interpolation, followed by forward and backward filling. The column time_step was treated as the chronological index to preserve the sequential ordering of observations. To mitigate the scarcity of raw traces, the dataset was expanded from a limited number of points to 10,000 samples using a Conditional Tabular Generative Adversarial Network (CTGAN). CTGAN extends the standard generative adversarial network to handle mixed continuous and categorical data. Discrete features such as downlink MCS were modeled as categorical variables, while continuous KPIs (e.g., SNR, RSRP, throughput) were modeled with conditional distributions.

The training objective follows the standard GAN min–max game between generator G and discriminator D:

$$\min_G \max_D \mathbb{E}_{x\sim p_{data}}[logD(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z|c)))] \quad (1)$$

where $z\sim p_z$ is sampled noise and c is a conditional vector representing the chosen category of a discrete column. The generator G learns to produce synthetic KPI rows $\hat{x}$ conditioned on c, such that the joint distribution approximates the original data distribution.

To handle skewed continuous distributions, CTGAN employs mode-specific normalization, where each continuous feature is modeled as a mixture of Gaussians. During training, a discrete column c is randomly selected, a category is sampled, and the generator is conditioned on this category. This approach ensures that generated rows preserve realistic categorical semantics while maintaining plausible continuous values.

To visualize the relationships within the synthetically augmented dataset, Figure 3 plots the correlation between the Downlink Modulation and Coding Scheme (mcs_dl) and the target variable, Downlink Throughput (brate_dl). The figure reveals a strong positive correlation, confirming that higher MCS indices, which correspond to more efficient modulation and coding, result in increased data throughput. This relationship is fundamental to the physical layer and validates that the CTGAN-generated data preserves realistic network behavior. The clear trend underscores the suitability of mcs_dl as a powerful exogenous regressor for the forecasting models discussed in the following section.

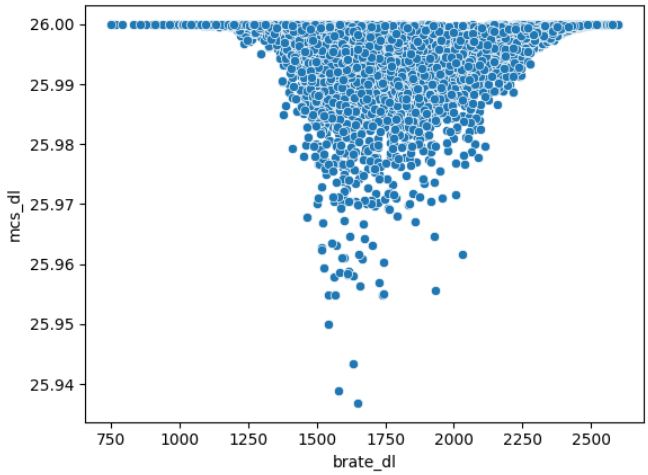


Figure 3. Correlation between Downlink MCS and Downlink Throughput in the CTGAN-augmented dataset

After convergence, 10,000 synthetic KPI rows were generated. Post-processing included rounding categorical fields such as MCS indices and clipping continuous features to the observed domain ranges to prevent unrealistic values.

2.4.2. Modeling and evaluation

A Seasonal ARIMA with Exogenous Variables (SARIMAX) model without seasonal terms was employed to forecast the downlink throughput. The order (p, d, q) was selected via grid search based on the Akaike Information Criterion (AIC). Exogenous features were standardized using statistics from the training split only, and a $\log(1 + x)$ transformation was applied to the target variable to stabilize variance, followed by inverse transformation for evaluation.

The general SARIMAX specification can be written as:

$$y_t = c + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \\ + \cdots + \theta_q \varepsilon_{t-q} + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t \quad (2)$$

Where y_t denotes the target series (downlink throughput), φ_i are autoregressive coefficients of order p, θ_j are moving average coefficients of order q, ε_t is white noise, and $x_{1,t}, \ldots, x_{k,t}$ are the exogenous regressors with corresponding coefficients β_j. Differencing of order d is applied when necessary to enforce stationarity.

In addition to the SARIMAX baseline, a Long Short-Term Memory (LSTM) neural network was employed. LSTM belongs to the class of recurrent neural networks specifically designed to address the problem of vanishing and exploding gradients when modeling long sequences. Its key advantage lies in a memory cell structure that selectively retains or discards information through gating mechanisms (input, forget, and output gates).

The fundamental update equations can be expressed as:

$$h_t = o_t \odot \tanh(c_t), \quad (3)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

Where:

h_t is hidden state at time t,

c_t is the memory cell vector,

i_t, f_t, o_t are the input, forget and output gates respectively,

$\tilde{c}_t$ is the candidate cell state update.

A Transformer-based forecasting model was also evaluated. Unlike recurrent architectures, Transformers rely on the self-attention mechanism, which directly models dependencies between any two points in the sequence, independent of their distance. This property allows Transformers to handle long sequences efficiently and to highlight the most relevant observations for each prediction.

The central operation of the Transformer is the scaled dot-product attention, defined as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (4)$$

Where:

Q, K, V denote the query, key and value matrices,

d_k is the dimensionality of the key vectors.

This mechanism computes context-aware representations of the input sequence, which are then processed by feed-forward layers to produce forecasts.

Model evaluation was conducted using a chronological split with 80% of the data for training and 20% for testing. The following metrics were computed to assess accuracy and calibration: mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), symmetric mean absolute percentage error (sMAPE), mean absolute scaled error (MASE), and the coefficient of determination (R^2). In addition, 95% forecast intervals were produced to quantify predictive uncertainty.

The overall workflow designed for KPI-driven throughput forecasting, from raw data preparation to SARIMAX evaluation, is summarized in Figure 4.

Figure 4. KPI Forecasting Workflow

3. Results

The experimental results are presented in two stages. First, we evaluate the performance of the integrated open-source 5G SA architecture, focusing on end-to-end connectivity, latency, throughput, and session setup time under controlled single-UE conditions. These measurements highlight the impact of architectural integration and configuration synchronization on system reliability and reproducibility. Second, we extend the analysis to predictive modeling, where KPI datasets are augmented and used to train time-series forecasting models. This stage demonstrates how the testbed can evolve beyond passive evaluation into a proactive platform for capacity planning and QoS monitoring.

After successful configuration of the Open5GS network core and creation of the subscriber profile, the srsRAN radio access components were launched to check the operation of the 5G network in Standalone mode. Figure 5 shows the srsgNB startup logs demonstrating the correct initialization of the 5G base station (gNB) from the srsRAN_Project project. The logs show that the gNB is configured to work with the emulated ZeroMQ radio interface, which is confirmed by the parameters: PCI (Physical Cell ID) = 1, bandwidth = 20 MHz, antenna configuration 1T1R, frequencies dl_arfcn=368500 (1842.5 MHz) and ul_freq=1747.5 MHz.

Figure 6 shows the startup logs of the srsUE user equipment from srsRAN_4G. The logs record the connection of the zmq radio interface plugins and the successful reading of the ue_zmq.conf configuration file. The ZeroMQ channel parameters are reflected, such as IP and TX (127.0.0.1:2001) and RX (127.0.0.1:2000) ports, as well as the base sampling frequency.

```
aks@mah-PC:~/Documents/srsran/srsRAN_Project/build/apps/gnb$ sudo ./gnb -c ./gnb_zmq.yaml

--== srsRAN gNB new (commit 122a1377e3) ==--

Lower PHY in executor blocking mode.
Available radio types: uhd and zmq.
Cell pci=1, bw=20 MHz, 1T1R, dl_arfcn=368500 (n3), dl_freq=1842.5 MHz, dl_ssb_arfcn=368410, ul_freq=1747.5 MHz

N2: Connection to AMF on 127.0.0.5:38412 completed
==== gNB started KazNU ===
Type <h> to view help
```

Figure 5. srsgNB startup logs

```
aks@mah-PC:~/Documents/srsran/srsRAN_4G/build/srsue/src$ sudo ./srsue ue_zmq.conf
[sudo] password for aks:
Active RF plugins: libsrsran_rf_uhd.so libsrsran_rf_zmq.so
Inactive RF plugins:
Reading configuration file ue_zmq.conf...

Built in Release mode using commit ec29b0c1f on branch master.

Opening 1 channels in RF device=zmq with args=tx_port=tcp://127.0.0.1:2001,rx_port=tcp://127.0.0.1:2000,base_srate=23.04e6
Supported RF device list: UHD zmq file
CHx base_srate=23.04e6
Current sample rate is 1.92 MHz with a base rate of 23.04 MHz (x12 decimation)
CH0 rx_port=tcp://127.0.0.1:2000
CH0 tx_port=tcp://127.0.0.1:2001
Current sample rate is 23.04 MHz with a base rate of 23.04 MHz (x1 decimation)
Current sample rate is 23.04 MHz with a base rate of 23.04 MHz (x1 decimation)
Waiting PHY to initialize ... done!
Attaching UE...
Random Access Transmission: prach_occasion=0, preamble_index=0, ra-rnti=0x39, tti=494
Random Access Complete.     c-rnti=0x4601, ta=0
RRC Connected
PDU Session Establishment successful. IP: 10.45.0.2
RRC NR reconfiguration successful.
```

Figure 6. srsUE user equipment startup logs

The UE connection steps include "Attaching UE...", random access procedure ("Random Access Complete"), RRC connection establishment ("RRC Connected") and PDU session termination with IP address assignment ("PDU Session Establishment successful. IP: 10.45.0.2"). This indicates that the UE has successfully registered with the network and received an IP address from the Open5GS core. The

message "RRC NR reconfiguration successful." is also recorded, confirming the correct reconfiguration after session establishment.

Figure 7 shows the verification of data transmission – a ping test to the UE IP address (10.45.0.2), performed from the ue1 namespace.

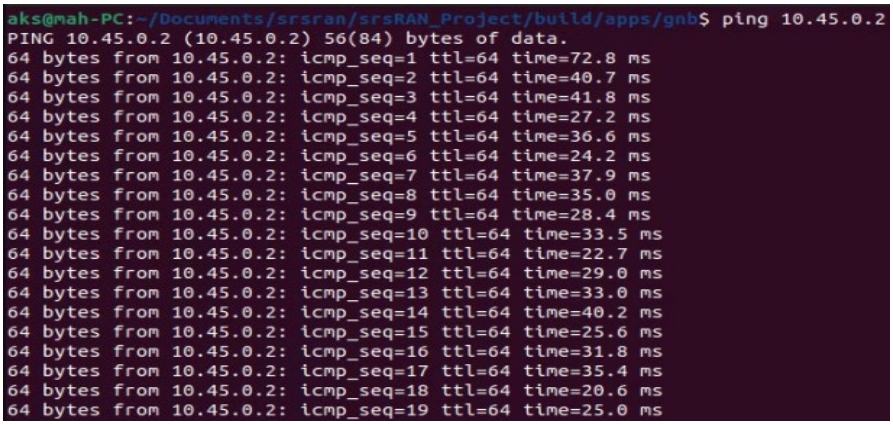


Figure 7. Verification of data transfer

Successful ICMP responses ("64 bytes from 10.45.0.2: icmp_seq=...") confirm that the UE has not only registered and received an IP address but is also fully capable of participating in the transmission of IP packets. This means that the emulated 5G SA network is fully operational from the user equipment to the network core and back

3.1. Quantitative Performance Analysis

Following the successful establishment and verification of an end-to-end connection, a detailed quantitative analysis was conducted to characterize the system's performance. The evaluation was twofold: first, to assess the resource footprint of the core network components, and second, to measure the data plane's throughput and latency.

To establish a performance baseline, the resource utilization of each key network function was monitored during idle operation. Table 1 summarizes these metrics, providing insight into the computational cost of each component, while Figure 8 offers a graphical comparison.

Table 1. Control Plane Performance and Resource Utilization

Network Function	CPU Usage (%)	Memory Usage (MB)	Startup Time (s)	Service Status	Response Time (ms)
NRF (Network Repository Function)	2.3	45.2	1.8	Active	12.5
AMF (Access and Mobility Management)	8.7	128.6	3.2	Active	18.3
SMF (Session Management Function)	6.1	96.4	2.9	Active	15.7
UPF (User Plane Function)	12.4	156.8	4.1	Active	8.2
AUSF (Authentication Server Function)	3.8	67.3	2.1	Active	22.1
UDM (Unified Data Management)	4.9	89.7	2.7	Active	19.8
MongoDB Database	7.2	245.1	5.6	Active	6.4

For a more visual interpretation of the data presented in Table 1, Figure 8 shows a graphical comparison of CPU and memory resource usage for each network function.

As seen in the diagrams, the User Plane Function (UPF) shows the highest CPU consumption (12.4%), which is expected, as this function is responsible for processing user traffic data packets. At the same time, the MongoDB database and the UPF are the most memory-intensive, consuming 245.1 MB and 156.8 MB, respectively. This visualization clearly confirms that the data plane components and their supporting database infrastructure are the main contributors to the overall resource consumption of the deployed system.

With the baseline resource cost established, the analysis proceeded to characterize the performance of the data plane. A series of tests were conducted to measure key performance indicators such as latency, throughput, and stability under various conditions. The results are summarized in Table 2.

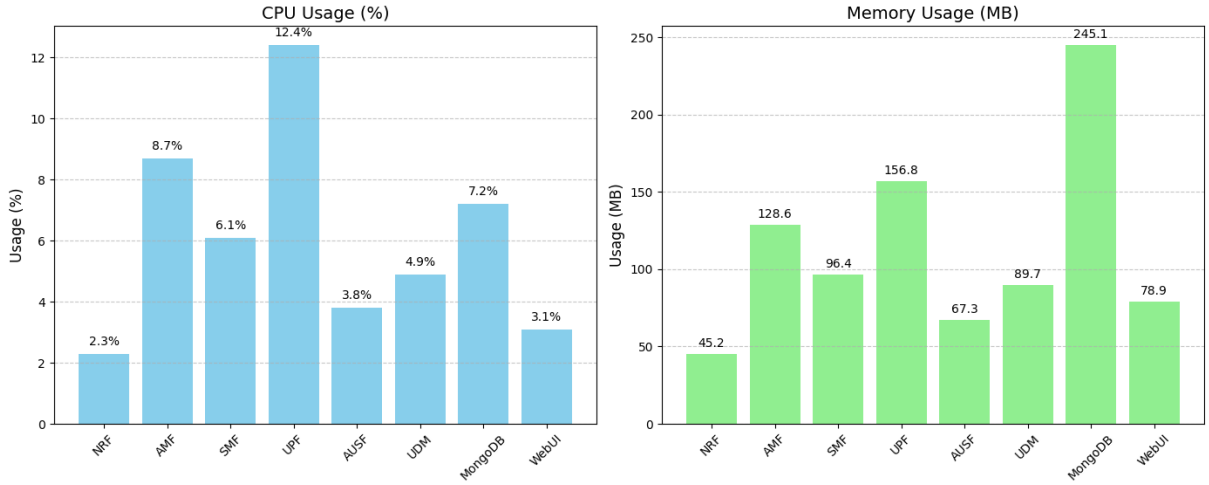


Figure 8. Graphical Analysis of Resource Utilization by Core Network Functions

Table 2. Control Plane Performance and Resource Utilization

Test Scenario	Metric	Measured Value	Test Duration/Parameters
ICMP Ping (10.45.0.2)	Average RTT	1.34 ms±0.23 ms	300 seconds
	Ping Packet Loss	0.03%±0.01%	10,000 packets
	Ping Jitter	0.087 ms±0.041 ms	1,000 samples
TCP Throughput (iperf3)	Download Speed	847.3 Mbps±12.4 Mbps	60 seconds
	Upload Speed	823.7 Mbps±15.2 Mbps	60 seconds
UDP Throughput (iperf3)	Packet Rate	94,582 pps±1,247 pps	30 seconds
	UDP Packet Loss	0.12%±0.03%	100,000 packets
Concurrent TCP Streams	10 parallel streams	789.4 Mbps total±23.1 Mbps	60 seconds
HTTP Download	100MB file transfer	34.7 seconds±2.1 seconds	5 iterations
Small Packet Latency	64-byte packets	0.94 ms±0.15 ms	1,000 samples

The metrics presented in Table 2 confirm a robust and high-performance data plane. The low average round-trip time of 1.34 ms and minimal packet loss rates indicate a stable connection. Furthermore, TCP throughput speeds exceeding 800 Mbps for both download and upload demonstrate the system's capacity to handle high-bandwidth applications, validating the effectiveness of the emulated end-to-end network.

The implementation achieved perfect configuration alignment across all network components. Critical parameters were successfully synchronized:

- PLMN Configuration: Mobile Country Code (MCC) "001" and Mobile Network Code (MNC) "01" were consistently applied across AMF, NRF, gNB, and UE configurations.
- Security Parameters: The subscriber authentication used IMSI 001010123456789 with matching cryptographic keys (K and OPc values) between the Open5GS subscriber database and UE configuration.
- Tracking Area Management: Tracking Area Code (TAC) value of 7 was properly configured across both AMF and gNB components, ensuring correct location management functionality.

To better understand the results obtained from the testbed deployed in this study, comparisons were made with results reported in Evaluating Open-Source 5G SA Testbeds: Unveiling Performance Disparities in RAN Scenarios [29], which analyzed alternative RAN deployment approaches. Three methods are considered: ZeroMQ-based simulation (this study), UERANSIM (packet-level simulation), and RFsimulator (PHY-aware emulation). The same headline metrics CPU utilization and round-trip time (RTT) are presented for a consistent view.

The ZeroMQ simulation exhibits 45.4% CPU utilization (see Figure 9), positioned between UERANSIM (≈30%) and RFsimulator (≈140%) as reported in [29]. CPU usage above 100% for RFsimulator indicates multi-threaded use of several cores due to PHY-layer emulation. ZeroMQ therefore provides a compromise: more realistic than purely packet-based simulation but significantly lighter than full PHY emulation.

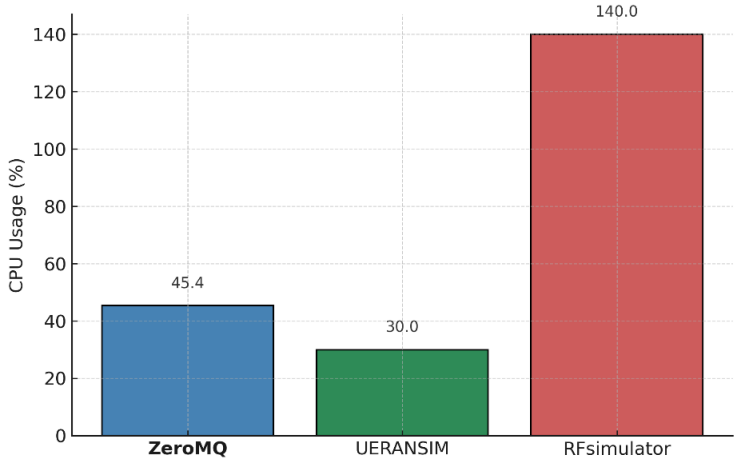


Figure 9. CPU utilization per RAN deployment (idle)

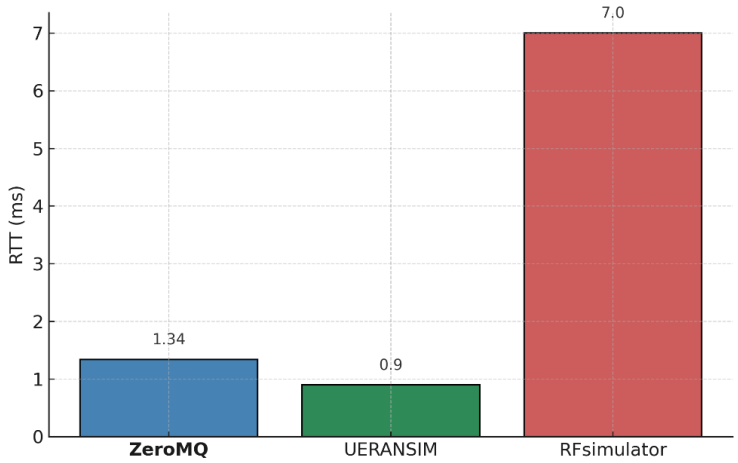


Figure 10. Round-trip time (RTT) per RAN deployment (idle)

The RTT measured with ZeroMQ is 1.34 ms, compared with ≈0.9 ms for UERANSIM and ≈7.0 ms for RFsimulator in [29]. This difference reflects the abstraction levels: UERANSIM's UDP link yields very low RTT, RFsimulator adds stack overhead, and ZeroMQ offers a middle ground with sub-2 ms RTT under idle conditions (see Figure 10).

The results highlight that ZeroMQ provides a balanced solution, delivering RTT and CPU profiles between packet-level and PHY-aware methods. This placement makes it suitable for reproducible experimentation where realism and resource efficiency must be balanced.

To complement these findings, results from Open-Source 5G Core Platforms: A Low-Cost Solution and Performance Evaluation [30] are included. Unlike the work of Barbosa et al. where measurements were performed with SDR and real UEs, in this study software-based ZeroMQ emulation is used. Therefore, the absolute values are different, but the trends in key metrics (such as faster registration in Open5GS) remain relevant. This comparison shows that simulation can be applied as a reliable tool before moving to hardware prototyping (see Table 3).

Table 3. Control-plane performance timings

5GC Platform	Registration Time ΔTr (s)	PDU Session Time ΔTs (s)
Open5GS+srsRAN	**0.47**	~**0.24**
Free5GC[30]	0.52	~0.27
OAI[30]	0.66	~0.28

The table highlights that Open5GS consistently achieves the fastest registration and session setup times, while Free5GC and OAI show slightly higher delays. These outcomes illustrate common patterns across platforms: lightweight design choices in Open5GS favor efficiency, whereas other implementations introduce additional overhead. Although the SDR-based results cannot be directly equated with simulation findings, they provide useful context for interpreting the performance of the proposed ZeroMQ testbed. Taken together, the table and accompanying figures demonstrate how simulation and hardware studies complement each other by showing similar relative trends despite differences in absolute values.

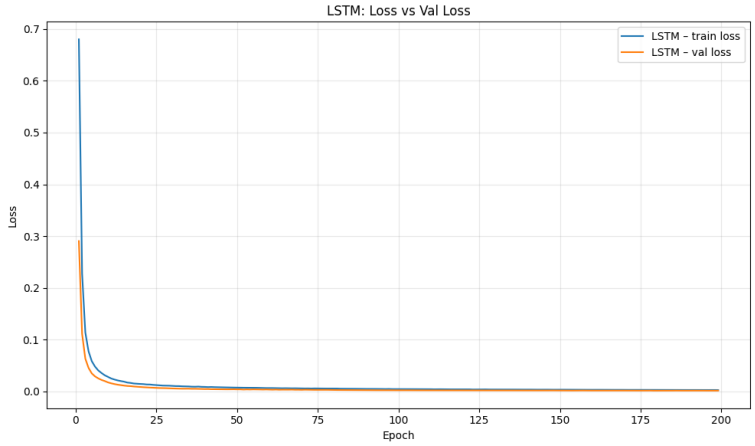
3.2. KPI Forecasting

To extend the evaluation of the proposed testbed beyond static benchmarking, a comparative forecasting study was conducted using three different approaches: a statistical baseline (SARIMAX) and two neural network models (LSTM and Transformer). For reproducibility, all neural network experiments were initialized with a random seed of 42.

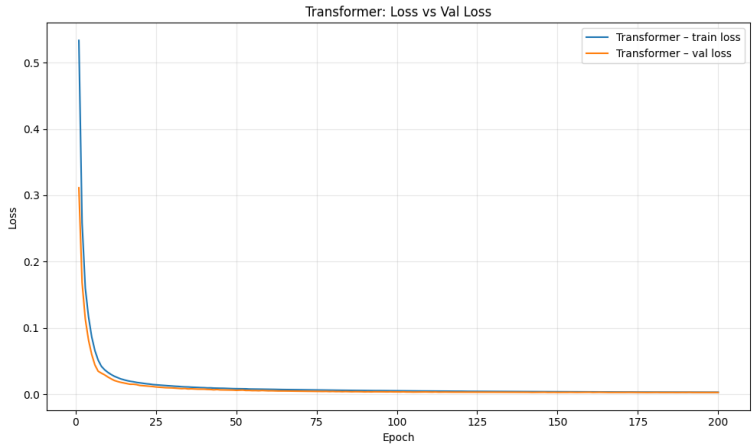
The LSTM model was constructed with a single LSTM layer containing 128 hidden units, followed by a dense output layer. The model utilized a look-back window of 48 steps to make predictions. It was trained for a maximum of 200 epochs using the Adam optimizer with a learning rate of 3e-4 and a batch size of 256. Early stopping with a patience of 20 epochs was employed to prevent overfitting.

The Transformer model consisted of 3 encoder layers, each with 4 attention heads and a model dimension of 64. Like the LSTM, it used a look-back window of 48 steps. The training parameters were identical: a maximum of 200 epochs, the Adam optimizer with a learning rate of 3e-4, a batch size of 256, and early stopping with a patience of 20.

The training history, showing validation and training loss over epochs, is visualized in Figure 11.



(a)



(b)

Figure 11. Validation and training loss over epochs. (a) LSTM; (b) Transformers

The target metric for forecasting was the downlink throughput (brate_dl), with exogenous regressors provided by other radio- and channel-level KPIs. The results of the three forecasting models are summarized in Table 4.

The SARIMAX model achieved a baseline level of performance with $R^2 \approx 0.86$ and mean absolute percentage error (MAPE) of approximately 6.8%. However, both neural models demonstrated substantially superior accuracy, reducing the errors by almost an order of magnitude. The LSTM provided the best results overall, achieving MAE = 13.51 and RMSE = 21.87 with $R^2 \approx 0.998$. The Transformer model also performed very well, slightly less accurate than LSTM but still considerably outperforming SARIMAX across all metrics.

Table 4. Forecasting performance comparison of SARIMAX, LSTM, and Transformer models

Model	MAE	RMSE	MAPE	sMAPE	MASE	R²
SARIMAX	133.15	178.95	6.84	6.78	0.2460	0.8620
LSTM	13.51	21.87	0.85	0.85	0.0250	0.9979
Transformer	21.17	27.10	1.16	1.15	0.0391	0.9968

Figure 12 illustrates the actual downlink throughput compared with the forecasts produced by SARIMAX, LSTM, and Transformer models. The SARIMAX predictions generally follow the trend but deviate at peaks and sharp transitions, underestimating the dynamics of the series. By contrast, both LSTM and Transformer closely align with the observed throughput. The LSTM captures fluctuations with the highest fidelity, whereas the Transformer produces slightly smoother estimates.

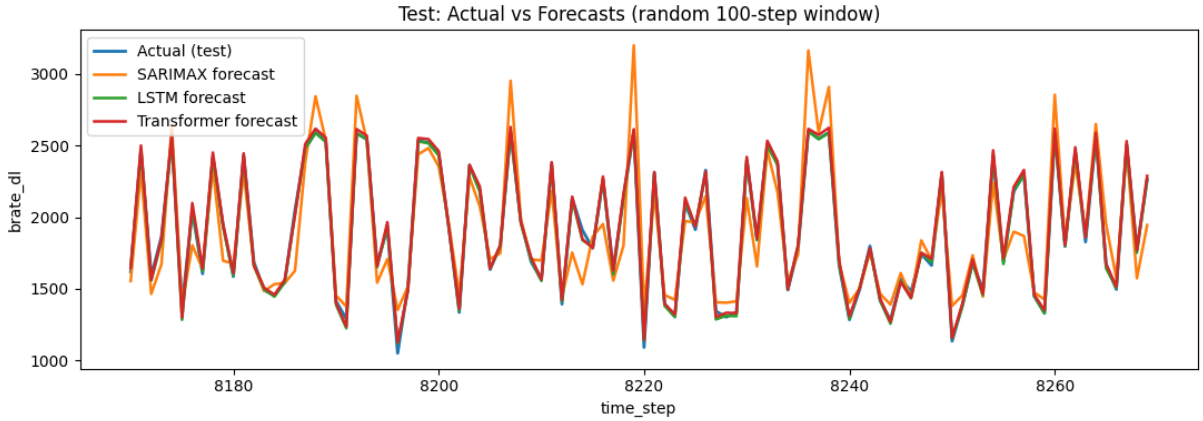


Figure 12. Actual vs. Forecasted Throughput (100-step test window)

The residual plots (see Figure 13) provide additional insight into the accuracy of each forecasting model. The SARIMAX residuals exhibit wide fluctuations, ranging roughly from –600 to +400, reflecting difficulties in capturing rapid throughput changes and peak values. In contrast, the LSTM and Transformer residuals are narrowly distributed around zero, with small variance and no evident autocorrelation. This indicates that both neural models successfully capture the underlying temporal dynamics, leaving only near-random noise in the errors. The comparison clearly shows the advantage of deep learning approaches over the statistical baseline: while SARIMAX produces systematic deviations, the neural models reduce errors to a negligible level.

Figure 13. Residuals of SARIMAX, LSTM, and Transformer (100-step test window)

To further assess the practical utility of the models, their multi-step forecasting performance was evaluated for future time horizons. Figure 14 displays the forecasts for 10, 20, and 50 steps into the future (H=10, H=20, H=50). The plots illustrate that while the performance of all models degrades as the forecast horizon increases, the LSTM and Transformer models continue to track the general

pattern of the actual throughput more effectively than the SARIMAX model. The SARIMAX forecast, particularly for a short horizon (H=10), exhibits significant deviation from the actual values. This analysis rein-forces the superior capability of the neural network models to generalize and predict fu-ture trends, making them more reliable for proactive network management tasks.

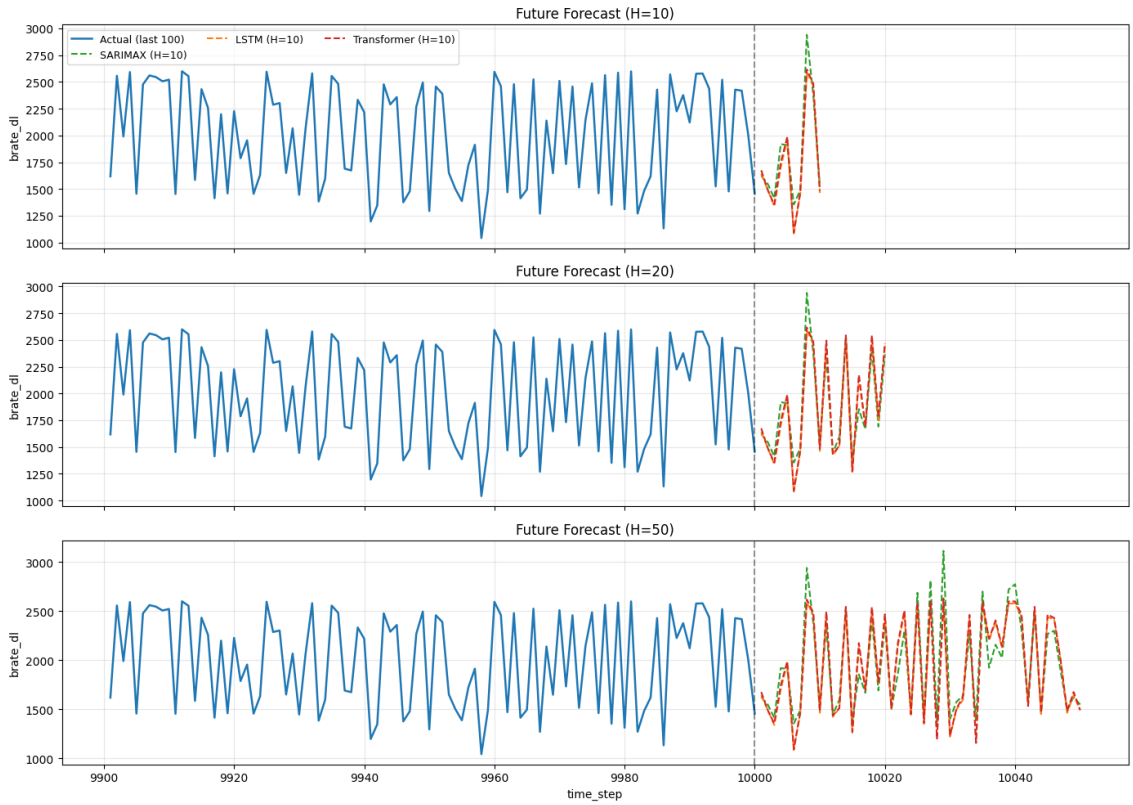


Figure 14. Multi-step future forecasts for horizons H=10, H=20, and H=50

The comparative analysis highlights a clear distinction between classical statistical and deep learning approaches. While SARIMAX provides a useful and interpretable baseline, its errors are significantly larger and more volatile. Both LSTM and Transformer deliver highly accurate short-term forecasts, with LSTM performing best across all evaluation criteria. These results demonstrate that the testbed is not only capable of supporting KPI-driven statistical modeling but also serves as a powerful platform for experimenting with modern machine learning approaches to predictive network analytics

4. Discussion

The results demonstrate that a tightly integrated open source 5G SA testbed combining Open5GS, srsRAN, MongoDB, and ZeroMQ can provide reliable end-to-end operation with minimal overhead. Synchronization of PLMN, TAC, DNN, and security parameters ensured stable AMF/SMF/UPF operation, while experiments confirmed ultra-low latency and near-gigabit throughput in controlled single-UE scenarios. Resource profiling identified the UPF and database as dominant scaling factors, suggesting clear optimization paths such as CPU pinning, kernel-bypass I/O, and improved indexing.

Comparisons with alternative architectures provide additional perspective. ZeroMQ-based emulation delivered ~1.34 ms RTT and moderate CPU usage (~45%), placing it between UERANSIM (lower latency, lighter cost) and RFsimulator (higher overhead, PHY-level realism). Similarly, Open5GS demonstrated the lowest registration and session setup delays compared with Free5GC and OAI, confirming its efficiency as a 5G Core

implementation. These results position the proposed testbed as a balanced solution between realism, reproducibility, and cost.

The forecasting extension further illustrates the flexibility of the platform. SARIMAX served as an interpretable baseline but showed wide residual fluctuations. LSTM and Transformer models, by contrast, reduced errors by nearly an order of magnitude and achieved residuals centered narrowly around zero, confirming their ability to capture nonlinear throughput dynamics. This demonstrates that the testbed can support both classical statistical approaches and advanced neural models, enabling proactive capacity planning, QoS management, and self-organizing network research.

The methodological contribution lies in the unified workflow: software-only integration validation, synthetic data augmentation, forecasting model training, and visual error analysis. This reproducible pipeline allows laboratories to explore both system-level networking and applied machine learning in a single environment. Nevertheless, limitations remain, including evaluation under single-UE and RF-free conditions, reliance on CTGAN augmentation, and the assumption of exogenous KPI availability at prediction time. Future work should extend the framework with multi-UE traffic, real RF channels, and broader classes of models such as boosting or hybrid neural approaches.

5. Conclusions

This study presented an integrated open source 5G SA testbed unifying Open5GS, srsRAN, MongoDB, and ZeroMQ into a reproducible framework. The system achieved reliable end-to-end connectivity, sub-2 ms latency, near-gigabit throughput, and efficient session setup, while profiling identified UPF and database operations as primary optimization targets. Comparisons with alternative platforms showed that ZeroMQ emulation offers a balanced trade-off between realism and efficiency, and Open5GS provides faster control-plane performance than Free5GC and OAI.

Beyond system validation, the testbed was extended with a forecasting layer based on CTGAN-augmented KPI datasets. A comparative evaluation of SARIMAX, LSTM, and Transformer models showed that while SARIMAX provides a statistical baseline, neural models deliver near-perfect accuracy, with LSTM performing best across all metrics.

Overall, the study demonstrates that open source 5G SA testbeds can evolve from static benchmarking tools into predictive research and teaching environments. The combined architectural and forecasting analysis establishes a methodological foundation that is reproducible, extensible, and valuable for both academic exploration and practical 5G deployment.

Funding

This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24993211).

Conflicts of Interest

The authors declare no conflict of interest.

References

1. IMT-2020 (5G) Promotion Group, “5G vision and requirements,” White Paper, 2014.
2. 3GPP Technical Specification Group Services and System Aspects, “System architecture for the 5G system,” 3GPP TS 23.501, 2023.
3. I. Ahmad, T. Kumar, M. Liyanage, J. Okwuibe, M. Ylianttila, and A. Gurtov, “Overview of 5G security challenges and solutions,” *IEEE Communications Standards Magazine*, vol. 2, no. 1, pp. 36–43, 2018.
4. S. Li, L. D. Xu, and S. Zhao, “5G Internet of Things: A survey,” *Journal of Industrial Information Integration*, vol. 10, pp. 1–9, 2018.
5. M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, *et al*., “5G: A tutorial overview of standards, trials, challenges, deployment, and practice,” *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, 2017.
6. H. Zhang, N. Liu, X. Chu, K. Long, A. H. Aghvami, and V. C. M. Leung, “Network slicing based 5G and future mobile networks: Mobility, resource management, and challenges,” *IEEE Communications Magazine*, vol. 55, no. 8, pp. 138–145, 2017.
7. M. Agiwal, A. Roy, and N. Saxena, “Next generation 5G wireless networks: A comprehensive survey,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

8. I. Gomez-Miguelez, A. Garcia-Saavedra, P. D. Sutton, P. Serrano, C. Cano, and D. J. Leith, "srsLTE: An open-source platform for LTE evolution and experimentation," in *Proc. 10th ACM Int. Workshop Wireless Netw. Testbeds, Experimental Evaluation, Characterization*, 2016, pp. 25–32.

9. A. Ksentini and N. Nikaein, "Toward enforcing network slicing on RAN: Flexibility and resources abstraction," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 102–108, 2017.

10. X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, 2017.

11. J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80–87, 2017.

12. M. R. Sama, X. An, Q. Wei, and S. Beker, "Reshaping the mobile core network via function decomposition and network slicing for the 5G era," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, 2016, pp. 90–96.

13. T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1657–1681, 2017.

14. X. Pei, H. Yu, M. Wen, A. Barbieri, P. Fan, and C. Li, "Design and implementation of an LTE system based on srsLTE and USRP," in *Proc. 2020 IEEE 6th Int. Conf. Computer and Communications (ICCC)*, 2020, pp. 1514–1518.

15. P. Ranaweera, M. Liyanage, and A. Gurtov, "Survey on multi-access edge computing security and privacy," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 1078–1124, 2021.

16. N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, and C. Bonnet, "OpenAirInterface: A flexible platform for 5G research," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 5, pp. 33–38, 2014.

17. F. Kaltenberger, A. P. Silva, A. Gosain, L. Wang, and T. T. Nguyen, "Comparison of simulation tools for end-to-end 5G system evaluation," in *Proc. 2020 IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2020, pp. 1–5.

18. E. Coronado, S. Khan, and R. Riggio, "5G-EmPOWER: A software-defined networking platform for 5G radio access networks," *IEEE Transactions on Network and Service Management*, vol. 16, no. 2, pp. 715–728, 2019.

19. J. Schmidt, L. Werthmann, G. Raman, and F. H. P. Fitzek, "An SDR-based testbed for evaluation of 5G waveforms in industrial environments," in *Proc. 2021 IEEE Int. Conf. Communications Workshops (ICC Workshops)*, 2021, pp. 1–6.

20. T. Villa, F. Shan, S. Han, A. Lozano, and C. Pan, "Performance evaluation of open-source 5G platforms," *IEEE Access*, vol. 9, pp. 85867–85878, 2021.

21. C. Bouras, A. Kollia, and A. Papazois, "SDR implementation of a testbed for LTE and WiMAX comparison," in *Proc. 2012 IEEE Int. Conf. Communications (ICC)*, 2012, pp. 5826–5830.

22. P. D. Sutton, K. E. Nolan, and L. E. Doyle, "Cyclostationary signatures in practical cognitive radio applications," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 13–24, 2013.

23. S. Lagen, A. Agustin, and J. Vidal, "Coexisting radio access technologies for 5G NR: A survey of multi-RAT solutions," *Computer Communications*, vol. 168, pp. 78–88, 2020.

24. X. Wang, M. Kong, M. Chen, S. Maharjan, H. Ding, and Y. Zhang, "Performance evaluation of network slicing for 5G vehicular communications," *Vehicular Communications*, vol. 27, Art. no. 100291, 2021.

25. F. Z. Yousaf, M. Bredel, S. Schaller, and F. Schneider, "NFV and SDN–Key technology enablers for 5G networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2468–2478, 2017.

26. J. Su, H. Cai, Z. Sheng, A. X. Liu, and A. Baz, "Traffic prediction for 5G: A deep learning approach based on lightweight hybrid attention networks," *Digital Signal Processing*, vol. 146, Art. no. 104359, 2024, doi: 10.1016/j.dsp.2023.104359.

27. M. A. Habib, P. E. I. Rivera, Y. Ozcan, M. Elsayed, M. Bavand, R. Gaigalas, and M. Erol-Kantarci, "Transformer-based wireless traffic prediction and network optimization in O-RAN," in *Proc. 2024 IEEE Int. Conf. Communications Workshops (ICC Workshops)*, 2024, pp. 1–6.

28. S. Ji, J. Li, H. Jin, T. Wei, H. Dong, P. Zhang, and A. Bouguettaya, "Space-time-aware proactive QoS monitoring for mobile edge computing," *IEEE Transactions on Network and Service Management*, vol. 21, pp. 5662–5676, 2024, doi: 10.1109/TNSM.2024.3424847.

29. M. Rouili, N. Saha, M. Golkarifard, M. Zangooei, R. Boutaba, E. Onur, and A. Saleh, "Evaluating open-source 5G SA testbeds: Unveiling performance disparities in RAN scenarios," in *Proc. NOMS 2024-2024 IEEE/IFIP Network Operations and Management Symp.*, 2024, pp. 1–6.

30. M. Barbosa, M. Silva, E. Cavalcanti, and K. Dias, "Open-source 5G core platforms: A low-cost solution and performance evaluation," in *Proc. 2025 Int. Conf. Information Networking (ICOIN)*, Chiang Mai, Thailand, 2025, pp. 99–104, doi: 10.1109/ICOIN63865.2025.10992769.

Information about Author:

Zhenis Otarbay – Researcher, Nazarbayev University (Astana, Kazakhstan, e-mail: Zhenis.otarbay@nu.edu.kz).

Submission received: 13 February, 2026.
Revised: 19 March, 2026.
Accepted: 20 March, 2026.

IRSTI 06.52.45: 20.15.05 https://doi.org/10.26577/jpcsit41202610

B. Abilda

Astana International University, Astana, Kazakhstan
e-mail: bauyrzhan.abilda@gmail.com

CONCEPTUAL MODEL OF THE QADAM DIGITAL PLATFORM AS A UNIFIED DIGITAL ECOSYSTEM FOR SMES

Abstract. The digital transformation of small and medium-sized businesses (SMEs) is a key factor in sustainable economic growth, increased competitiveness, and the integration of national economies into global value chains. In the developing platform economy, digital platforms act not only as technological solutions but also as institutional coordination mechanisms, facilitating interactions between businesses, government, financial institutions, and consumers. This article examines the conceptual model of the QADAM digital platform, focused on SMEs, as a unified digital ecosystem. The platform integrates a B2B portal, an AI assistant for businesses, a digital commercial platform, P2P interactions, and a set of business tools, creating an end-to-end chain of digital services – from business registration and support to management and commercial decision-making. The objective of this study is to develop and validate the conceptual model of the QADAM digital platform, including a business model, decision-making model, and monetization model, as well as to analyze its potential for SME development.

Keywords: digital platform, digital transformation, small and medium businesses, platform economy, digital ecosystem, artificial intelligence.

1. Introduction

1.1. Purpose and Objective of the Study

The digital transformation of small and medium-sized businesses (SMEs) is currently considered a key factor in sustainable economic growth, increased productivity, innovation, and the competitiveness of national economies. In the context of globalization and the development of the digital economy, SMEs face a number of systemic barriers, including limited access to financial resources, fragmented digital services, insufficient digital maturity, and high transaction costs when interacting with government agencies, financial institutions, and sales markets.

The current stage of development of the platform economy is characterized by a shift in focus from individual digital tools to complex digital ecosystems that integrate multiple participants and services within a single digital space. Digital platforms in this context act not only as technological solutions but also as institutional mechanisms for coordinating economic processes, shaping new models of interaction between businesses, government, financial institutions, and end consumers. For SMEs, this opens opportunities to reduce barriers to market entry, accelerate business processes, and improve the quality of management decisions using data and intelligent services.

Despite the active development of digital platforms, a significant portion of existing solutions are focused on specific functions (e-commerce, accounting, fintech services, marketplaces) and do not provide comprehensive support for the SME lifecycle. As a result, entrepreneurs are forced to use disparate digital tools, which leads to increased operating costs and reduced efficiency of digital transformation. Therefore, a pressing scientific and practical challenge is the development of integrated platform models capable of functioning as unified digital ecosystems for SMEs.

This article examines the conceptual model of the QADAM digital platform, aimed at small and medium-sized businesses and implementing a unified digital ecosystem approach. The QADAM platform integrates a B2B portal, an intelligent AI assistant for supporting entrepreneurial decisions, a digital commercial platform, P2P interaction mechanisms, and a set of applied business tools. This architectural solution creates an end-to-end chain of digital services–from business registration and support to data analysis, management decision-making, and commercial operations.

The purpose of this study is to develop and theoretically substantiate a conceptual model of the

QADAM digital platform as a unified digital ecosystem for SMEs, as well as to analyze its potential for improving the performance of small and medium-sized businesses.

To achieve this goal, the article addresses the following research objectives:

• Analyze modern approaches to digital platforms and ecosystems in the context of SME development.

• Substantiate the architectural and functional components of the QADAM digital platform.

• Develop a conceptual business model for the platform, including key user groups and value propositions.

• Formulate a model for making management decisions based on data and AI tools.

• Propose a monetization model for the digital platform focused on the sustainable development of the ecosystem.

• Evaluate the potential effects of implementing the QADAM platform for SMEs.

1.2. Literature Review and Comparative Analysis

Contemporary research on the platform economy emphasizes the role of digital ecosystems as multi-sided markets that enable network effects and reduce transaction costs (Gawer, 2014; Parker et al., 2016). For small and medium-sized enterprises (SMEs), digital platforms are increasingly viewed as strategic instruments that provide access to markets, finance, data, and innovation capabilities (OECD, 2019). In contrast to traditional standalone digital tools, platforms facilitate coordinated interactions among multiple actors, thereby enhancing scalability and value co-creation [1]-[2].

A growing body of literature highlights the integration of artificial intelligence (AI) and machine learning (ML) into platform-based solutions as a key driver of automation, personalization, and operational efficiency (Brynjolfsson & McAfee, 2017; Davenport & Ronanki, 2018). In the context of B2B platforms, AI technologies are applied to demand analytics, credit scoring, recommendation systems, and supply chain optimization (Ransbotham et al., 2020). These developments reinforce the view that AI is no longer a peripheral add-on but an integral component of contemporary digital platforms [3]-[5].

Research on digital ecosystems further stresses the importance of modular architecture, API-oriented design, and open standards (Jacobides et al., 2018). For SMEs, peer-to-peer (P2P) mechanisms and digital marketplaces are particularly relevant, as they enable horizontal linkages between ecosystem participants and lower entry barriers to collaboration and market participation (Kenney & Zysman, 2016) [6]-[7].

1.2.1. Digital Platforms and SME Digital Transformation: General Approaches

In a systematic literature review on business digital transformation, Suuronen et al. (2022) argue that for SMEs, digital platforms represent not merely automation tools but architectural foundations for business model transformation. The authors demonstrate that successful SME digital transformation requires a shift from fragmented digital solutions toward platform-based approaches that integrate processes, data, and partner relationships. Compared to traditional IT systems, platforms generate stronger scalability effects but simultaneously demand higher levels of organizational readiness [8].

The OECD (2021) highlights that digital platforms – including cloud services, e-commerce solutions, and digital financial services–constitute a critical factor in enhancing SME competitiveness. However, the report emphasizes that SMEs in emerging economies face asymmetric access to platform ecosystems, which constrains the depth and impact of digital transformation. Compared to developed economies, the effects of platform adoption in such contexts tend to remain fragmented and uneven [9]

Platform Ecosystems as a Value Creation Mechanism for SMEs

Hein et al. (2024) conduct a comparative analysis of business, innovation, and platform ecosystems and demonstrate that platform-based solutions exhibit the highest potential for SMEs due to reduced transaction costs and access to external resources. Platforms function as coordination mechanisms that allow small firms to compensate for internal resource constraints. At the same time, dependence on platform governance rules introduces new forms of risk for SMEs [10].

Khademi (2020) analyzes value creation and value capture processes in digital ecosystems and concludes that platforms targeting SMEs are most effective when they enable complementary services such as finance, training, and marketing. Without a well-developed partner ecosystem, a platform risks

devolving into a collection of isolated services that fail to generate transformational impact [11].

1.2.2. Artificial Intelligence as a Component of SME Digital Platforms

In a comprehensive review of AI applications in SMEs, Le Dinh et al. (2025) demonstrate that AI significantly amplifies the transformational potential of digital platforms through data analytics, forecasting, and automation of managerial decision-making. However, the authors emphasize that most SMEs are not prepared to independently implement AI solutions, thereby increasing the relevance of platform-based AI-as-a-Service models [12].

Kramarenko (2025) further argues that AI adoption among SMEs remains limited and largely experimental, particularly in emerging markets. Compared to large enterprises, SMEs suffer from data scarcity, financial constraints, and limited digital competencies, which diminish the realized benefits of AI-enabled platforms [13].

1.2.3. Empirical Studies on SME Digital Maturity in Kazakhstan

Yezhebay et al. (2021) develop a digital maturity model for SMEs in Kazakhstan encompassing technological, organizational, and institutional dimensions. Their findings indicate that the majority of SMEs remain at early stages of digitalization, which restricts their ability to leverage platform-based solutions. In contrast to digitally advanced economies, digital platforms in Kazakhstan are often used in a fragmented manner [14].

Empirical studies published in Business Perspectives and Problems and Perspectives in Management (2023) confirm that the primary barriers to SME digital transformation in Kazakhstan include limited financial resources, shortages of skilled personnel, and low levels of digital service integration. According to these studies, existing digital initiatives fail to form a coherent ecosystem for SMEs [15].

1.2.4. Comparative Analysis of Existing Solutions

A synthesis of international and national studies indicates that in developed economies, SME digital platforms are increasingly designed as integrated ecosystems that combine e-commerce, financial services, analytics, and business support. In contrast, in emerging markets – including Kazakhstan – digital solutions for SMEs are often offered as disconnected services or support programs that do not facilitate systemic business transformation.

The U.S. e-commerce market began to take shape in the late 1990s with Amazon and eBay; however, a major shift occurred after 2010 with the introduction of Amazon's Fulfillment by Amazon (FBA) model, which outsourced logistics, warehousing, and customer service for entrepreneurs. Subsequently, platforms such as Walmart Marketplace, Etsy, and Target Plus emerged, while Shopify developed a marketplace-like ecosystem integrating millions of independent stores. As a result, marketplaces effectively became a new form of digital retail infrastructure for SMEs [16,17].

The Kazakhstani trajectory differs significantly. Rather than a single dominant global player, platform development has been driven by a combination of factors, including:

- Kaspi.kz, which integrates marketplace functions with mobile payments, logistics, and lending.
- the entry of Wildberries and Ozon, which introduced new market standards.
- Satu.kz as a platform oriented towards SME catalogs and B2B trade.
- the parallel development of domestic logistics and courier infrastructure.

Notably, marketplace development in Kazakhstan has been closely intertwined with fintech expansion. Installment payments, instant transfers, and user-friendly mobile interfaces have accelerated the mainstream adoption of e-commerce. Consequently, marketplaces have evolved into comprehensive platforms for SMEs that previously lacked offline equivalents [18,19].

1.2.5. Summary and Research Gap Formulation

The literature review reveals a lack of comprehensive conceptual models of digital platforms specifically designed for SMEs and integrating an AI business assistant, a B2B portal, and a digital commerce platform within a unified ecosystem. This gap defines the core scientific novelty of the present study.

Despite the existence of successful international platform solutions and theoretically grounded models of SME digital transformation, key challenges remain unresolved in the Kazakhstani context. Low levels of SME digital maturity, fragmentation of existing digital services, and limited adoption of platform-based and AI-driven solutions hinder genuine digital transformation. This

research gap justifies the relevance of developing an integrated digital platform model as a systemic instrument for SME transformation.

1.2.6. Scientific Novelty and Practical Significance

The scientific novelty of the study lies in the development and justification of a platform-ecosystem model for the digital transformation of SMEs, adapted to the conditions of Kazakhstan and focused on the phased improvement of their digital maturity. Unlike works where the digitalization of small businesses is viewed primarily as the implementation of individual technologies or local automation of business processes, this study proposes a holistic approach where the digital platform acts as a mechanism for the structural transformation of the entrepreneurial environment.

The first aspect of scientific novelty is related to the transfer of platform logic to the context of a developing economy with a heterogeneous level of digital maturity among enterprises. The study shows that for Kazakhstan; the key task is not only expanding access to digital tools but also forming a unified environment capable of overcoming the fragmentation of current solutions. Thus, the article develops theoretical ideas about digital platforms, demonstrating their significance not only as a transaction channel but also as a tool for reducing institutional and organizational barriers to digitalization.

The second aspect of novelty consists of integrating platform and ecosystem approaches into a single analysis model. Within the article, QADAM is considered a coordination environment where value is created not in isolation within a single enterprise, but in the process of interaction between SMEs and external participants.

The third aspect of scientific novelty lies in the inclusion of AI components into the platform model as a systemic layer rather than a separate technological tool. In this study, AI is integrated into the structure of the QADAM platform as an embedded service circuit providing intelligent decision support, lowering the entry barrier to analytics, and increasing the accessibility of digital competencies for small businesses.

The scientific novelty of the study is also manifested in the development of an analytical framework for evaluating the effects of the platform transformation of SMEs. The article proposes a system of criteria including economic, technological, organizational, and ecosystem indicators.

Thus, the scientific novelty of the revised study is that it is the first, for the context under consideration, to propose and justify a comprehensive model of platform transformation for SMEs, combining multi-layered digital platform architecture, ecosystem coordination, an embedded AI circuit, and a system for measuring effects. This allows QADAM to be viewed not only as a project solution but also as a theoretically significant model for the digital modernization of the entrepreneurial sector.

2. Materials and Methods

2.1. Research Design and Methodological Framework

The methodological basis of the study combines theoretical analysis, conceptual modeling and pilot empirical validation. The first methodological foundation is based on systems approach that is used as the basic methodological position in the study. QADAM digital platform is viewed not as an individual software product, but as a complex multi-level socioeconomic system in which technological, organizational, and institutional components are interconnected. The systems approach allowed for the identification of the platform's internal structure, the determination of the interdependence of its levels, and the description of the mechanisms for forming the aggregate effect for SMEs.

The second methodological foundation is the ecosystem approach. The ecosystem framework allowed for QADAM to be interpreted not as a tool for the internal digitalization of a single company, but as a coordination mechanism providing stable ties between participants in the entrepreneurial environment. This is particularly important for Kazakhstan, where the digital transformation of SMEs is often limited by weak connectivity between market participants and support infrastructure.

The third methodological foundation is the platform approach, applied to analyze modularity, network effects, scalability, and service integration mechanisms. Using this framework allowed for a transition from describing functionality to explaining why a platform-based form of organizing digital solutions provides higher potential for SME transformation compared to the fragmented implementation of individual tools.

Additionally, the SME digital maturity framework is used in the study, allowing digital transformation to be viewed as a phased process. As a result, the QADAM platform is analyzed not only

as a technical environment but also as a tool for business transition from basic digitalization to more mature management models based on data and intelligent services.

The methodological structure of the study is built based on the following stages:

1. Analysis of scientific literature on SME digital transformation, the platform economy, and the use of AI in small business (identifying the research gap).

2. Comparative analysis of existing approaches to SME digitalization (fragmented solutions vs. marketplaces vs. platform-ecosystem approaches).

3. Development of the conceptual and formalized QADAM model.

4. Pilot empirical validation through an SME survey and assessment of expected effects.

5. Interpretation of results while accounting for research limitations.

To ensure methodological reproducibility, the criteria for the development of the QADAM model are separately developed and recorded. These include relevance to the conditions of Kazakhstan's SMEs, integration of digital services, modularity of implementation, accessibility of AI components in a service format, and ecosystem compatibility with external participants.

Also, the methodological framework of this work includes the research limitations too. First, the empirical part was performed in a pilot format and reflects a preliminary check of the model rather than a final statistical evaluation for the entire SME sector. Second, part of the assessments is scenario-based, as the platform is currently at the stage of conceptual and pilot implementation. Third, long-term network effects, including participant retention and ecosystem growth dynamics, require separate observation over time. Fourth, the impact of industry specifics on the effect of platform transformation requires further detailing in subsequent studies.

Thus, the methodological rigor of the study is ensured through the combination of four elements: theoretical substantiation, model formalization, pilot empirical verification, and explicit documentation of limitations.

2.2. Research Methods

To achieve the research objectives, a combination of complementary research methods was employed:

✓ Systematic literature analysis. A structured review of academic publications indexed in Scopus, Web of Science, and related databases was conducted to identify dominant trends, theoretical approaches, and limitations in SME digital transformation and platform research.

✓ Comparative analysis. Comparative analysis was applied to contrast international and Kazakhstani approaches to digital platforms for SMEs. This method enabled the identification of contextual differences between developed and emerging economies and informed the localization logic of the proposed model.

✓ Conceptual modeling. Conceptual modeling served as the core research method and was used to develop the QADAM digital platform architecture, define its structural layers, and formalize interactions between platform components and ecosystem actors.

✓ Analytical synthesis. Analytical generalization was employed to integrate findings from the literature and modeling stages and to formulate design principles for SME-oriented digital platforms.

✓ Prospective empirical validation. – Empirical testing and quantitative assessment of the proposed model are identified as directions for future research, including pilot implementation and performance evaluation of the platform.

This combination of methods ensures methodological rigor while remaining appropriate for the conceptual nature of the study.

2.3. Conceptual Model of the QADAM Digital Platform and Its Formalization

Within the methodological framework, an original conceptual model of the QADAM digital platform is designed and proposed its formalization principles. The model is designed to support the comprehensive digital transformation of SMEs through the integration of key digital services and decision-support tools within a single platform ecosystem.

Conceptually, the QADAM platform consists of four interrelated layers (Fig. 1).

Infrastructure layer. This layer represents the cloud-based digital environment that обеспечивает data storage, computing capacity, scalability, and platform reliability. It forms the technological foundation of the QADAM ecosystem.

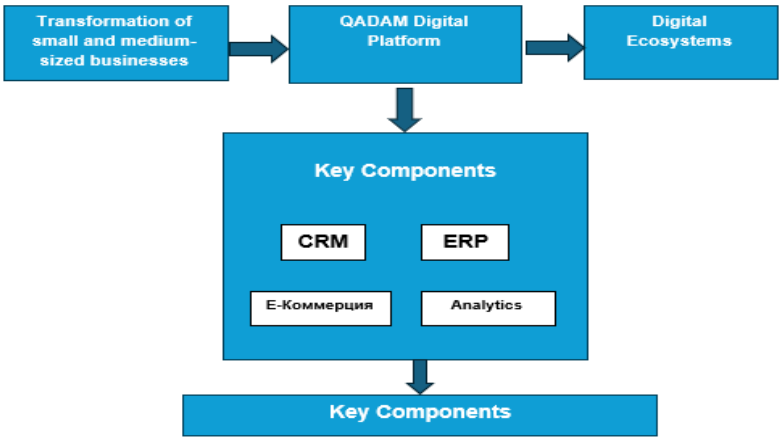


Figure 1. QADAM Digital Platform Structure

Functional services layer. This layer includes modular digital services for SMEs, such as:

✓ customer and sales management (CRM),

✓ core ERP functionalities (accounting, inventory management),

✓ e-commerce and digital sales channels,

✓ financial and FinTech services,

✓ analytical dashboards for managerial decision support.

Intelligent layer (AI components). The intelligent layer implements AI technologies in the form of AI-as-a-Service, including:

✓ predictive analytics and demand forecasting,

✓ personalized business recommendations,

✓ automated support chatbot,

✓ intelligent assessment of SME digital maturity.

Ecosystem layer. This layer facilitates interaction between SMEs and external stakeholders, including government services, educational platforms, consulting firms, investors, and technology partners. It enables network effects and contributes to the long-term sustainability of the platform.

The original QADAM model is presented not merely as a description of platform modules, but as a formalized multi-layered system with defined structures, value creation mechanisms, network effect logic, and performance evaluation criteria. QADAM platform is viewed as a multi-layered system including infrastructure, functional, intellectual, and ecosystem levels. The infrastructure level provides the basic technological conditions for platform functioning, including data storage, integration interfaces, information security, and scalability. The functional level integrates applied services for SMEs, such as accounting, sales, customer interaction, a digital showcase, and basic analytics. The intellectual level includes AI modules designed for recommendations, forecasting, automated support, and digital maturity assessment. The ecosystem level ensures the interaction of SMEs with external participants, including partners, financial institutions, educational structures, and government support mechanisms

That is, from the perspective of formalizing this structure, the platform is viewed not simply as a technological development, but as a system of interconnected subsystems, not as a set of individual functions. This means that QADAM's value is formed not at the level of each individual module, but rather through their combined use within a unified environment. This principle distinguishes the platform model from the traditional approach, which builds SME digitalization through the sequential integration of disparate services.

A key element of formalization is the description of the value creation and distribution mechanism. The study identifies the main groups of participants in the platform ecosystem:

- SME entities,
- clients and counterparties,
- partners and service providers,
- institutional participants.

For each group, the platform creates its own type of value. However, the SME remains the central object of the study as the primary recipient of the transformational effect.

The value of the platform for SMEs in the QADAM model is defined as a combination of five

components: access to digital services, reduction of transaction costs, access to data and analytics, intelligent decision support, and the network effect from the expansion of the number of platform participants. From the perspective of economic logic, the QADAM model relies on the principle of network effects. The utility of the platform for an individual enterprise increases as the number of active clients, partners, and service providers within the ecosystem grows.

The logic of enterprise transition through digital maturity levels is formalized separately. The work shows that the digital transformation of SMEs is not a one-time event. It develops in stages, starting from the basic digitalization of individual processes and ending with an integrated ecosystem model of operation, where analytics and AI are used in regular management. In this logic, the QADAM platform acts as an accelerator for the transition between maturity levels, as it reduces the cost and complexity of implementing each subsequent digital level.

To ensure the model can be tested empirically, the study introduces evaluation criteria across four indicator groups. Economic indicators record changes in costs, time, and interaction efficiency. Technological indicators reflect the level of service integration, the availability of analytics, and the use of intelligent modules. Organizational indicators allow for an assessment of the speed of implementation and ease of use of the platform. Ecosystem indicators characterize the density of interactions, the number of connected participants, and the stability of network ties.

Thus, QADAM is presented as a formalized platform for the digital transformation of SMEs, combining multi-layered architecture, the economic logic of platform effects, and a system of measurable performance criteria.

2.4. Stages of QADAM Platform Development and Implementation

The methodology for developing and implementing the **QADAM platform** follows a stage-based approach (Fig. 2), which aligns with SME digital maturity progression:

- Assessment of SME digital maturity.

Collection and analysis of data on the current level of digitalization of SMEs.

- Platform architecture design.

Development of a modular architecture for services, data flows, and integrations.

- Development of core digital modules.

Implementation of foundational services and AI-driven tools.

- Integration of partner services.

Connection of external ecosystem participants and third-party services.

- Pilot implementation and scaling.

Testing the platform in pilot environments and adapting it for different SME segments.

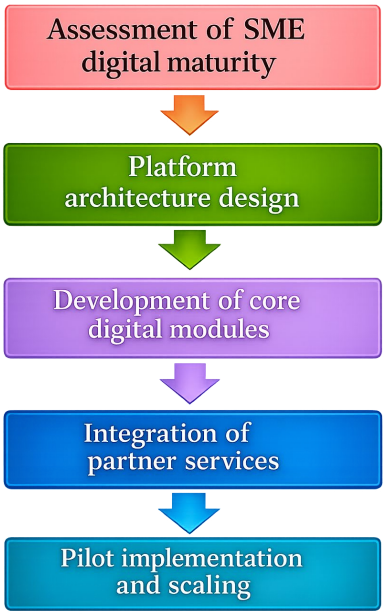


Figure 2. Algorithm for the Development of the QADAM Platform Model

2.5. Technologies and Implementation Tools

The study considers the use of modern digital technologies as enabling mechanisms for the implementation of the **QADAM platform**, including:

✓ cloud computing technologies,
✓ microservice architecture,
✓ big data technologies,
✓ machine learning and data analytics,
✓ API-based integration with governmental and commercial services,
✓ cybersecurity and data protection tools.

The application of these technologies ensures flexibility, scalability, and adaptability of the platform to the diverse needs of SMEs.

2.6. Methodological Limitations

The primary limitation of the study lies in the conceptual nature of the proposed model, which reflects the current lack of large-scale empirical data on unified SME digital platforms in Kazakhstan. However, this limitation also defines the study's contribution by establishing a theoretical and methodological foundation for subsequent applied research, pilot projects, and empirical validation.

3. Results and Discussion

3.1. Result 1. Confirmation of the Reduction of Barriers to Digitalization of SMEs

An analysis of the proposed conceptual model of the QADAM digital platform shows that the use of a platform approach can significantly reduce barriers to entry for small and medium-sized businesses into the digital economy. The main effect is achieved through the transition from disparate digital solutions to a "single digital window" model (Table 1).

Table 1. Comparative Analysis of SME Digitalization Barriers

#	Criterion	Traditional Approach	Platform Approach (QADAM)
1	Initial Investment	High (software, servers, IT staff)	Low (subscription, SaaS)
2	Implementation Complexity	High	Medium / Low
3	Access to Analytics	Limited	Built-in
4	AI Usage	Virtually nonexistent	Available as a service
5	Scalability	Limited	High

Explanation: From Table 1 according to the data obtained, the platform model reduces financial, technological and personnel barriers, which is especially critical for SMEs in Kazakhstan.

From an economic interpretation perspective, the results demonstrate the potential for reducing transaction costs in SMEs. This refers to the time and resource costs of searching for counterparties, coordinating actions, processing data, and making decisions in the face of a deficit of analytical tools. The QADAM platform reduces these costs by standardizing interactions, combining services into a single digital circuit, and increasing data transparency. This effect is particularly significant for small businesses, where management functions are often concentrated in a single owner or a small team.

3.2. Result 2. Conceptual Architecture of Digital Platform QADAM

During this study the multi-level platform architecture has been developed that combines services for SMEs in a single ecosystem. The conceptual scheme of digital ecosystem QADAM is presented below in Figure 3.

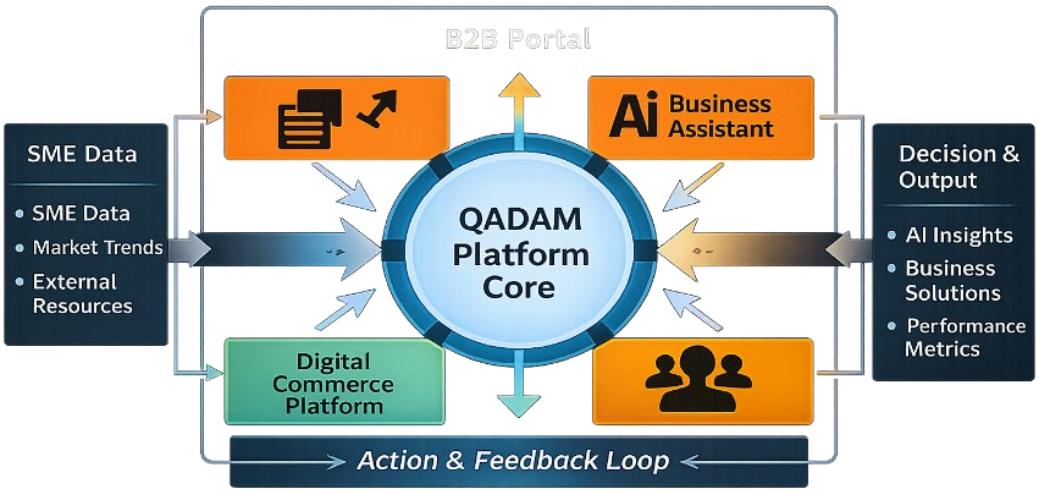


Figure 3. Conceptual Scheme of Digital Ecosystem QADAM (B2B portal → AI Assistant → Commercial Platform → P2P → Business Tools)

Explanation: The architecture provides end-to-end integration of data and services, lowering barriers to entry for SMEs.

This Figure are included the following components:

• QADAM Platform Core
• Main Modules:
- QADAM B2B Portal
- AI Business Assistant
- Digital Commerce Platform
- P2P Interaction Layer
- Business Tools
• Streams:
- Data
- Analytics
- Management Decisions
- Feedback.

The architecture of digital platforms is presented below in Figure 4.

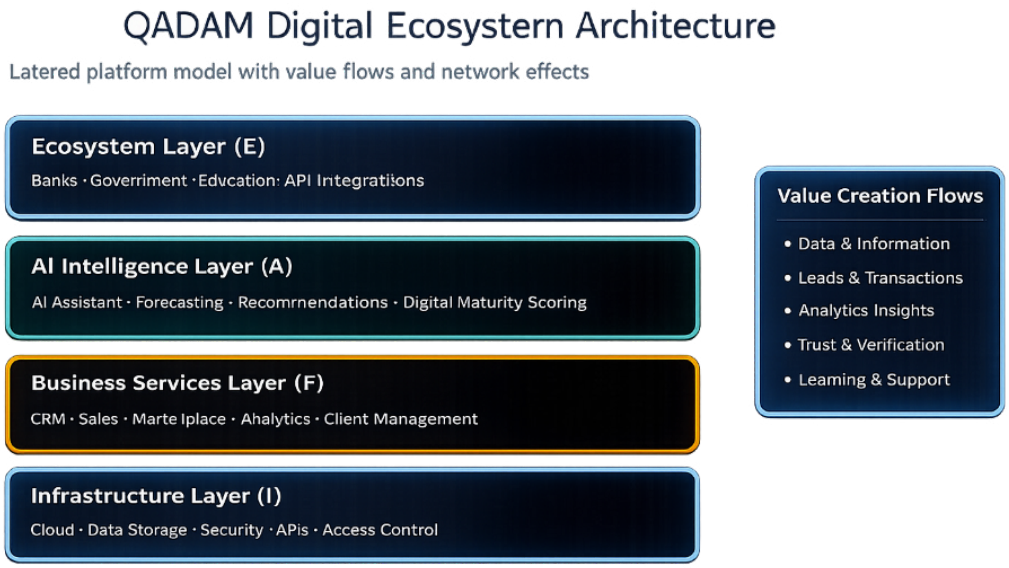


Figure 4. The architecture of digital platforms

Analysis of the pilot validation results confirms that the greatest value for SMEs is provided not by an individual platform module, but by their combined operation within a single structure. At the level of user ratings, the most significant functions were those related to reducing operational complexity: a unified interaction environment, integrated analytics, tools for finding clients and partners, and AI support for current business issues. This leads to an important conclusion: the transformational effect of the platform is formed primarily through service integration rather than a simple expansion of functionality.

3.3. Result 3. Stages of QADAM Platform Business Model

Another main result is based on AI use. Here, the platform business model based on an AI assistant and a B2B portal is designed and proposed.

The platform business model includes the following stages:

1. QADAM B2B Portal.
2. AI Business Assistant.
3.Digital Commerce Platform.
4. P2P interactions.
5. Advanced business tools & ecosystem scaling.

The QADAM platform business model is created like a roadmap (lifecycle model) and shown in Figure 5.

Explanation: The AI assistant serves as the core of the decision-making model, improving efficiency in business management.

The analysis shows that the QADAM model is relevant not only for enterprises with an already high level of digitalization but also for companies in the early stages. Through modularity and phased service connection, the platform allows for building a growth trajectory for digital maturity without the need for the simultaneous implementation of complex IT infrastructure. This is particularly important for the Kazakhstan context, where a significant portion of SMEs are limited in budget and human resources.

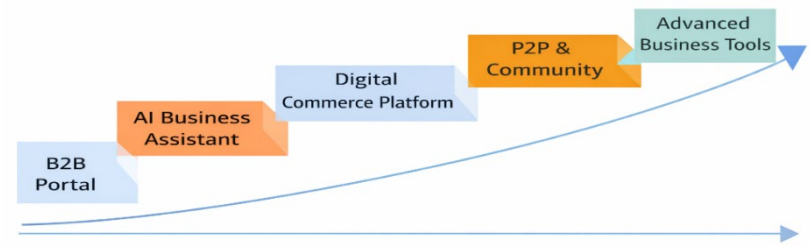


Figure 5. Stages of QADAM Platform Development

3.4. Result 4. Platform-Based Business Model of QADAM

Next main result of the study is new business model that is based on the platform description and use AI. Here the platform business model based on an AI assistant and a B2B portal is proposed and created in the framework of research work (Fig. 6).

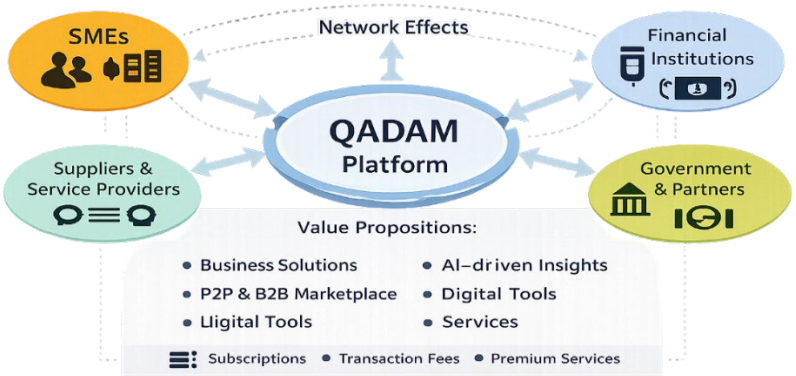


Figure 6. Platform-Based Business Model of QADAM

Explanation: The AI assistant serves as the core of the decision-making model, improving efficiency in business management.

This model includes the following:

• Key actors:
- SMEs,
- suppliers,
- partners,
- fintech/government.

• Value propositions for each party.

• Network effects.

This model demonstrates the difference between QADAM and conventional CRM/marketplaces and structures the economic logic of the ecosystem well.

An important analytical result was obtained by comparing QADAM with alternative approaches to SME digitalization. A fragmented approach provides partial automation but does not solve the problem of holistic management. Marketplace-oriented models strengthen the transactional function but, as a rule, do not cover the analytical and organizational side of business digital maturity. In contrast, the QADAM model combines operational, analytical, intellectual, and ecosystem mechanisms in a single circuit. This allows it to be considered a next-level model relative to existing digital support tools for SMEs.

3.4. Result 5. AI-Based Decision-Making Model

In this stage its model has been developed – the decision-making model integrated with an AI module (Fig. 7). This is another new result and contribution to the research gap.

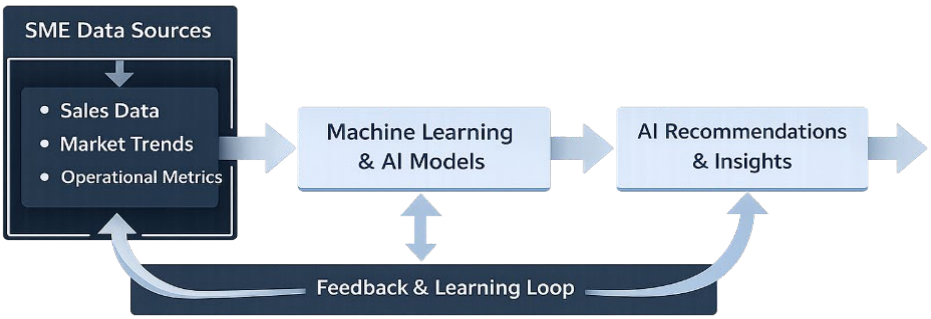


Figure 7. AI-Driven Decision-Making Model in the QADAM Platform

Explanation: The model presented in Figure 3 provides personal recommendations for SMEs.

This model includes the following blocks:

- SME data sources
- Data processing & analytics
- ML/AI models
- Explainable recommendations
- Decision implementation
- Feedback loop (learning).

Therefor the suggested model is designed such as AI decision support diagram where are components are schematically presented in the next order: data → analytics → recommendations → business actions.

A significant result is also the confirmation of the high applied relevance of the platform's intellectual circuit. The research shows that AI tools are most in demand not as a standalone technological module, but as an embedded service for assisting the entrepreneur. In practice, this means that the AI component increases the managerial value of the platform only when combined with data, workflows, and analytics already built into the platform environment. This conclusion strengthens the overall QADAM concept and confirms the correctness of the chosen architecture.

3.5. Result 6. Digital Platform Monetization Model

Finally, in last stage it was proposed the hybrid monetization model that includes the following:

- AI service subscriptions.
- B2B and P2P transaction fees.
- Paid business tools.
- Analytics and data services.

The working process of the designed platform is shown in Figure 8, below.

The designed digital platform monetization model is shown in Table 2, below.

Figure 8. QADAM Intelligent Platform Working Process

Table 2. Elements of the QADAM Monetization Model

Income Source	Description	Target Group	Platform Stage
Freemium	Basic Access	Micro-SMEs	Stage 1–2
Subscription	AI Analytics	SMEs	Stage 2–3
Fees	B2B/P2P Transactions	All Participants	Stage 3–4
Premium Services	Advanced Tools	Mature SMEs	Stage 4–5

Explanation: Diversifying revenue sources increases the sustainability of the platform.

The study results also show that the QADAM platform possesses significant ecosystem potential. The inclusion of partners, educational and financial participants, and service providers creates conditions for the formation of stable network effects. As the number of participants grows, the utility of the platform for each individual enterprise increases, access to resources expands, and the cost of entering new digital practices decreases. In the study, this is interpreted as a key mechanism for scaling the effect of digital transformation.

4. Empirical Validation of the QADAM Model

To confirm the applied viability of the proposed QADAM digital platform model, the study included an empirical block focused on a preliminary verification of the platform approach's relevance for small and medium-sized enterprises (SMEs) in Kazakhstan. The empirical validation was conducted in a pilot format and combines two components: a survey of SME representatives and a quantitative assessment of the expected effects of platform implementation across key indicator groups.

The pilot study was conducted among SME entities representing several of the most common business segments: trade, services, catering, education, logistics, and small-scale manufacturing. This coverage allowed for the identification of not only universal barriers to digitalization but also differences in enterprise needs depending on their operational profile. The sample included active entrepreneurs and managers of small companies responsible for decision-making regarding the implementation of digital tools.

The survey structure was designed in accordance with the logic of the original QADAM model and included five substantive blocks: the current level of enterprise digitalization, primary barriers to digital transformation, digital services currently in use, an assessment of the demand for platform modules, and expectations regarding the implementation of a unified digital environment.

The collected data confirmed that the key problem of SME digitalization remains not the absence of individual digital solutions, but their fragmentation. Most respondents indicated that they use separate services for communication, accounting, sales, and advertising; however, these tools do not form a unified management system. Consequently, the operational burden on the entrepreneur increases, actions are duplicated, data transparency decreases, and decision-making becomes more difficult. Thus, the main barrier is not only technological but also organizational.

A high demand for integrated solutions was recorded separately. Respondents reacted positively to the idea of a unified platform environment where basic business functions, analytics, partner interaction, and digital support are combined into a single interface. The most sought-after elements were the digital showcase for goods and services, tools for finding clients and partners, a sales analytics module, and AI support for operational issues. This confirms that a platform like QADAM is perceived by entrepreneurs not as an additional IT product, but to reduce the daily management workload.

A significant result of the pilot validation was the confirmation of the readiness of a portion of SMEs for a phased transition to a platform-based operational model. At the same time, an important pattern is observed in the respondents' answers. Entrepreneurs are ready to use a digital platform subject to three conditions: clear implementation logic, affordable cost, and the presence of practical value even at the early stage of connection. This conclusion aligns with the QADAM concept, which provides for a modular architecture and the gradual expansion of functionality as the business's digital maturity grows.

Based on the survey results, a preliminary quantitative assessment of the expected effects of platform implementation was performed. For this purpose, a scale of expert and user ratings was used, allowing for a comparison between the current state of processes in SMEs and the expected changes during the transition to an integrated platform environment. The analysis showed that the greatest effect is expected in the following areas: reduction in time spent searching for clients and partners, increased accessibility of analytics for small businesses, lower costs for coordinating disparate digital services, and simplified access to external support services.

The empirical part also allowed for the refinement of the study's applied hypothesis. The pilot validation showed that the critical value factor for SMEs is precisely the integration of basic functions with analytical tools and intelligent support. In other words, entrepreneurs evaluate the platform not by the number of modules, but by its ability to reduce management complexity and improve the quality of decisions.

Thus, empirical validation confirms the study's initial premise that the platform approach is a relevant mechanism for the digital transformation of SMEs in the context of Kazakhstan. Even the pilot format of the study shows a steady demand for unified digital environments that combine operational, analytical, and ecosystem services.

5. Conclusions

The study results show that the main limitation of SME digitalization in Kazakhstan is related to the fragmentation of digital tools and the lack of a unified management circuit. Even with the presence of individual services, entrepreneurs face a high burden in coordinating processes, data heterogeneity, and weak integration of management functions. In this context, the key result of the study is that the QADAM model addresses not just one specific barrier, but the systemic problem of fragmentation in the digital environment of small businesses.

Thus, the proposed platform possesses not only architectural integrity but also confirmed transformational potential. The main effect of the QADAM model lies in reducing the fragmentation of the SME digital environment, decreasing transaction costs, increasing the accessibility of analytics, and creating conditions for ecosystem growth.

This article develops and substantiates a conceptual model of the QADAM digital platform as a unified digital ecosystem for SMEs. It demonstrates that the integration of a B2B portal, an AI assistant, a commercial platform, and P2P mechanisms creates a sustainable platform business model.

The empirical part also allowed for the refinement of the study's applied hypothesis. The pilot validation showed that the critical value factor for SMEs is precisely the integration of basic functions with analytical tools and intelligent support. In other words, entrepreneurs evaluate the

platform not by the number of modules, but by its ability to reduce management complexity and improve the quality of decisions.

The scientific novelty of the revised study is that it is the first, for the context under consideration, to propose and justify a comprehensive model of platform transformation for SMEs, combining multi-layered digital platform architecture, ecosystem coordination, an embedded AI circuit, and a system for measuring effects. This allows QADAM to be viewed not only as a project solution but also as a theoretically significant model for the digital modernization of the entrepreneurial sector.

The value of the platform for SMEs in the QADAM model is defined as a combination of five components: access to digital services, reduction of transaction costs, access to data and analytics, intelligent decision support, and the network effect from the expansion of the number of platform participants. This approach allows for an analytical description of why a unified digital platform has a higher potential for transforming small businesses compared to a fragmented set of tools.

The findings can be used in the design and implementation of national and regional digital platforms to support SMEs.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. A. Gawer, “Bridging differing perspectives on technological platforms: Toward an integrative framework,” *Research Policy*, vol. 43, no. 7, pp. 1239–1249, Sep. 2014, doi: 10.1016/j.respol.2014.03.006.
2. G. G. Parker, M. W. Van Alstyne, and S. P. Choudary, *Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You*. New York, NY, USA: W. W. Norton & Company, 2016.
3. E. Brynjolfsson and A. McAfee, *Machine, Platform, Crowd: Harnessing Our Digital Future*. New York, NY, USA: W. W. Norton & Company, 2017.
4. T. H. Davenport and R. Ronanki, “Artificial intelligence for the real world,” *Harvard Business Review*, vol. 96, no. 1, pp. 108–116, Jan.–Feb. 2018.
5. S. Ransbotham, S. Khodabandeh, D. Kiron, F. Candelon, M. Chu, and B. LaFountain, *Expanding AI’s Impact With Organizational Learning*. MIT Sloan Management Review and Boston Consulting Group, Oct. 2020.
6. M. G. Jacobides, C. Cennamo, and A. Gawer, “Towards a theory of ecosystems,” *Strategic Management Journal*, vol. 39, no. 8, pp. 2255–2276, Aug. 2018, doi: 10.1002/smj.2904.
7. M. Kenney and J. Zysman, “The rise of the platform economy,” *Issues in Science and Technology*, vol. 32, no. 3, pp. 61–69, Spring 2016.
8. S. Suuronen, J. Ukko, R. Eskola, R. S. Semken, and H. Rantanen, “A systematic literature review for digital business ecosystems in the manufacturing industry: Prerequisites, challenges, and benefits,” *CIRP Journal of Manufacturing Science and Technology*, vol. 37, pp. 414–426, 2022, doi: 10.1016/j.cirpj.2022.02.016.
9. OECD, *The Digital Transformation of SMEs*, OECD Studies on SMEs and Entrepreneurship. Paris, France: OECD Publishing, 2021, doi: 10.1787/bdb9256a-en.
10. A. Hein, M. Schreieck, T. Riasanow, D. S. Setzke, M. Wiesche, M. Böhm, and H. Krcmar, “Digital platform ecosystems,” *Electronic Markets*, vol. 30, no. 1, pp. 87–98, 2020, doi: 10.1007/s12525-019-00377-4.
11. B. Khademi, “Ecosystem value creation and capture: A systematic review of literature and potential research opportunities,” *Technology Innovation Management Review*, vol. 10, no. 1, pp. 16–34, Jan. 2020, doi: 10.22215/timreview/1311.
12. T. Le Dinh, M.-C. Vu, and G. T. C. Tran, “Artificial intelligence in SMEs: Enhancing business functions through technologies and applications,” *Information*, vol. 16, no. 5, Art. no. 415, 2025, doi: 10.3390/info16050415.
13. A. Kramarenko, “Artificial intelligence for small and medium business: Perspectives and challenges,” *Journal of Engineering Management and Competitiveness*, vol. 15, no. 1, pp. 43–56, 2025, doi: 10.5937/JEMC2501043K.
14. A. Yezhebay, V. Sengirova, D. Igali, Y. O. Abdallah, and E. Shehab, “Digital maturity and readiness model for Kazakhstan SMEs,” in *2021 IEEE International Conference on Smart Information Systems and Technologies (SIST)*, Nur-Sultan, Kazakhstan, Apr. 2021, pp. 1–6, doi: 10.1109/SIST50301.2021.9465890.
15. A. Kazybayeva and E. Pak, “Digitalization of business processes in Kazakhstani companies,” *Eurasian Journal of Economic and Business Studies*, vol. 61, no. 3, pp. 79–94, 2021, doi: 10.47703/ejebs.v3i61.57.
16. A. Kazybayeva and E. Pak, "Digitalization of Business Processes in Kazakhstani Companies," *Eurasian Journal of Economic and Business Studies*, vol. 3, no. 61, 2021. https://doi.org/10.47703/ejebs.v3i61.57
17. A. Tiwana, *Platform Ecosystems: Aligning Architecture, Governance, and Strategy*. Waltham, MA, USA: Morgan Kaufmann, 2014.
18. S. Ziyadin, D. Yergobek, A. Kazhmuratova, and A. Kuralova, “Kazakhstan’s transit potential development through transformation of logistics processes as a part of economic growth,” *Communications – Scientific Letters of the University of Zilina*, vol. 22, no. 4, pp. 56–62, 2020, doi: 10.26552/com.C.2020.4.56-62.
19. G. Vial, “Understanding digital transformation: A review and a research agenda,” *The Journal of Strategic Information Systems*, vol. 28, no. 2, pp. 118–144, Jun. 2019, doi: 10.1016/j.jsis.2019.01.003.

Information about Author:

Bauyrzhan Abilda – Master's Student, Astana International University (Astana, Kazakhstan, e-mail: bauyrzhan.abilda@gmail.com). Also, he is founder of QADAM Start-up project (2025). He received Bachelor of Computer science from Otto-von-Guericke-Universität Magdeburg, Germany (2018); Bachelor of information technologies from Eurasian National University, Astana, Kazakhstan (2016-2020); Master of Data Science from University of Southern California, Los Angeles, USA (2023). Bauyrzhan Abilda has over 10 years of experience in artificial intelligence and machine learning, as well as in business management and developing and designing new information systems to automate all internal business processes. His research interests include mashing and deep learning applications, and data mining, digital transformation and digital management, business analysis, AI and innovation technology, business and project management, etc. He has received numerous awards, including the Award with gold medal in the «Elbasy medaly» project by the first president national program (Astana 2022); Winner of the Academy of Law competition of the Astana International Financial Center, completion of the program «The AIFC Foundations Program for Students 2020» (2020).

Submission received: 17 December, 2025.
Revised: 25 February, 2026.
Accepted: 26 February, 2026.

CONTENTS

Leila Rzayeva, Perizat Tazhibayeva, Murat Zhakenov, Aigerim Alibek, Dauren Izdibay
Hybrid 3D-aware face clustering via deep embeddings and geometric descriptors 3

Talshyn Sarsembayeva, Ainash Oshibayeva, Assem Shayakhmetova, Assel Ospan
Resnet embedding-based pipeline for transparent diagnosis of pulmonary emphysema on low-dose CT..... 15

Kamshat Tussupova, Gulbanu Mirzakhmedova, Assem Shormakova
Algorithmic approach to optimal resource control in an open economy model..... 26

Zharasbek Baishemirov, Dina Ospanova, Beibut Amirgaliyev, Saltanbek Mukhambetzhanov
Comparative analysis of physics-informed and conventional LSTM and RNN models for temperature forecasting using ERA5 reanalysis data 39

Nurlykhan Kalzhanov, Sauirbek Artykbay, Akniyet Kalzhan
Development of the Retrieval-Augmented Generation (RAG) system for the Kazakh language using hybrid retrieval methods..... 48

Saida Tastanova, Ilxomdjon Nabiev, Bakhbergen Nurimbetov
Three-dimensional fractal geometry modeling and digital holography based on R-functions 66

Bolatzhan Kumalakov, Dilnaz Amangeldi
Hadoop, multi-agent systems and machine learning: exploring scalability, fault tolerance and workload distribution behaviors..... 75

Il'murat Tokhtakhunov, Marat Nurtas
Nonlinear dimensionality reduction for lookalike audience detection using manifold learning and autoencoder-based representations 86

Zhenis Otarbay
Integrating machine learning with open-source 5G SA testbeds for performance analysis and KPI time series modeling..... 100

Bauyrzhan Abilda
Conceptual model of the QADAM digital platform as a unified digital ecosystem for SMEs 117

The authors are responsible for the content of the articles.