DOI 10.34185/1562-9945-5-166-2026-20
UDC: 004.85:004.932:519.246.8

I. Fedorchuk, O. Shkurat

SPATIO-TEMPORAL FORECASTING OF DIGITAL IMAGES WITH IRREGULAR TIME INTERVALS: A REVIEW OF DEEP NEURAL NETWORK APPROACHES

Abstract. This study evaluates four deep learning architectures – a Time-Aware ConvLSTM, Continuous Temporal Transformer, Spatio-Temporal GAN, and Continuous Diffusion Model – for forecasting digital image sequences at irregular time intervals. Tested on Moving MNIST and atmospheric climate records using a stochastic continuous-time masking protocol, the results demonstrate that the ConvLSTM and Transformer severely blur predictions of rigid-body trajectories. In contrast, the GAN achieves robust structural fidelity across both datasets. The CDM excels in modeling rigid-body trajectories but struggles significantly with smooth atmospheric gradients. A Transformer-GAN architecture is proposed to provide high-fidelity spatio-temporal forecasting of digital images with irregular time intervals.
Keywords: digital image processing, continuous-time forecasting, spatio-temporal dynamics, irregular sampling, neural networks, deep learning, Generative Adversarial Networks, Diffusion Models.

Statement of the problem. Spatio-temporal forecasting typically assumes a uniformly sampled sequence of observations. However, in many critical real-world applications, data is acquired at irregular time intervals. Formally, given a sequence of past digital images $X_{t_1}, X_{t_2}, \dots, X_{t_n}$ observed at strictly ascending but non-uniformly spaced timestamps $t_1 < t_2 < \cdots < t_n$ the objective is to model the underlying continuous-time data distribution. The main task is to generate a plausible future frame $X_{t_{target}}$ at an arbitrary, specific target time.

This specific framing of the task introduces significant algorithmic and architectural hurdles. Traditional sequence models are inherently discrete. They process information step-by-step and implicitly assume a constant temporal resolution. When applied to irregularly sampled data, these models often fail to capture the true underlying physical or biological dynamics, as they cannot natively distinguish between an event that occurred one hour ago versus one month ago without complex imputation techniques. Therefore, the primary scientific problem addressed in this research is the analysis of neural network architectures that mathematically treat time as a continuous parameter, enabling the system to learn the spatial transitions between irregularly spaced frames. The model must learn continuous temporal dynamics – often requiring mechanisms like explicit time-embedding, neural

© Fedorchuk I., Shkurat O., 2026

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

differential equations, or continuous diffusion processes – to accurately propagate spatial states across arbitrary time intervals. Furthermore, as the prediction horizon t_{target} increases, the uncertainty inherent in the generation process scales non-linearly, requiring robust probabilistic modeling to avoid blurry or physically impossible predictions.

Solving the spatio-temporal forecasting problem for irregular intervals is of paramount importance because real-world data collection is rarely perfectly synchronized. Such situations occur in medical imaging and disease progression, where patients do not undergo MRI, CT, or X-ray scans at perfectly regular intervals. A model capable of processing irregularly timed historical scans to predict the structural manifestation of a disease at a specific future date is invaluable for personalized treatment planning and early intervention. Another example is the area of earth observation and satellite imagery. Satellite sensors capturing optical imagery are frequently obstructed by cloud cover, resulting in sparse, irregularly sampled time series of the Earth's surface. Predicting environmental changes, agricultural yields, or urban development requires models that can seamlessly bridge these variable temporal gaps.

Beyond these domains, continuous-time forecasting is becoming increasingly critical in the rapidly evolving field of Extended Reality (XR). As users physically navigate their environments, XR headsets must constantly track dynamic positions and orientations via complex camera and sensor arrays to seamlessly render augmented or virtual worlds. However, these systems inherently suffer from data interruptions, asynchronous sensor polling, and network latency, creating highly irregular visual data streams. Robust spatio-temporal forecasting allows the system to correctly "anchor" digital objects in physical space despite these temporal gaps and delays. Furthermore, as the industry transitions toward Cloud XR architectures, mitigating transmission delays through predictive spatial modeling will be essential. Currently, the application of continuous-time neural models in these dynamic XR environments remains limited, presenting a significant opportunity for the architectures explored in this research to provide foundational spatial stability.

Analysis of recent research and publications. Classical statistical forecasting methods and traditional computer vision techniques are fundamentally ill-suited for handling the high dimensionality of digital images combined with unequal time intervals. Consequently, the state-of-the-art has shifted exclusively toward Deep Neural Network architectures.

The Convolutional Long Short-Term Memory network (ConvLSTM) extends traditional recurrent networks by replacing fully connected matrix multiplications with convolution operations, allowing the network to natively capture spatial features while maintaining a temporal memory state. However, because standard ConvLSTMs inherently assume a uniform time step, the architecture must be fundamentally modified into a "Time-Aware" or "Phased" variant to handle irregular intervals. The computational process begins when the model receives an input image at a specific time step alongside the exact temporal gap since the last observation. Firstly, this time delta is broadcasted into a spatial tensor and

Copyright for this paper by its authors. Use permitted under License CC BY 4.0.
ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

concatenated with the input image and the previous hidden state. In the next stage, the network passes this combined tensor through modified convolutional gating mechanisms. The forget controls the decay of past information and is computed mathematically as (1)

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{\Delta f} * \Delta t + b_f, \quad (1)$$

where * denotes the convolution operator, X_t is the input frame, H_{t-1} is the previous hidden state, and Δt is the temporal gap. In the third step, the cell state C_t is updated by multiplying the previous cell state by this time-decayed forget gate and adding the new input information (2):

$$C_t = f_t \circ C_{\{t-1\}} + i_t \circ \tanh\left(W_{\{xc\}} * X_t + W_{\{hc\}} * H_{\{t-1\}} + b_c\right). \quad (2)$$

This mechanism (fig. 1) forces the model's memory to systematically decay over longer periods of missing data, ensuring that stale information does not corrupt distant future predictions.

These time-adapted recurrent architectures have been widely deployed in real-world scenarios requiring continuous spatial monitoring. In the field of meteorology, researchers such as Shi et al. (2015)[1] established the foundation by applying convolutional recurrent models to precipitation nowcasting, and subsequent studies have heavily utilized time-aware variants to predict weather patterns from satellite imagery where intermittent cloud cover causes highly irregular observation gaps. Additionally, in urban planning, these models are frequently employed for traffic flow prediction [3], successfully modeling spatial congestion maps over time while dynamically adapting to irregular data reporting from faulty or asynchronous highway sensor networks.

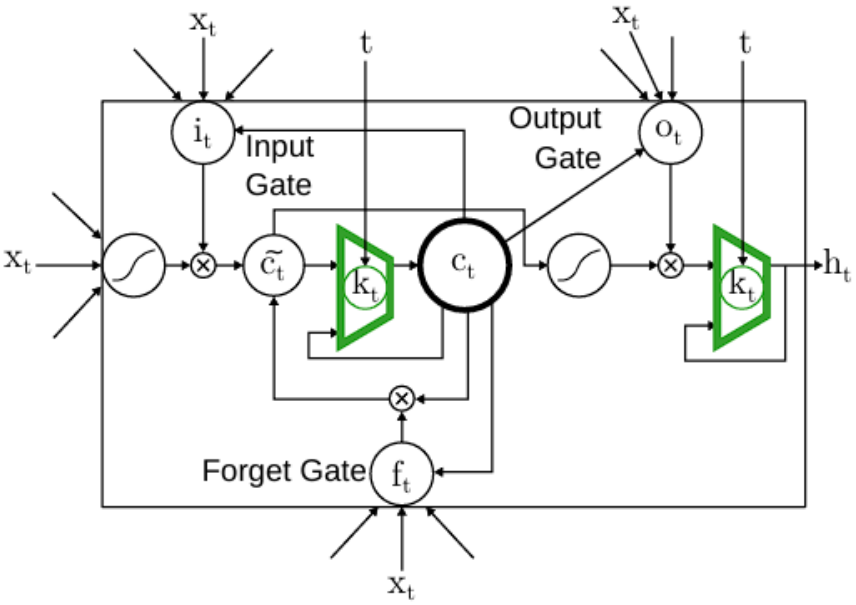


Figure 1 – Phased LSTM model, with time gate k_t controlled by timestamp t [3]

Transformers [4] bypass sequential step-by-step processing entirely by utilizing global self-attention mechanisms, allowing the network to process all available historical frames in parallel. To adapt this architecture for irregular spatio-temporal forecasting, the standard dis-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

crete positional encodings must be replaced with continuous representations [5]. The operational pipeline begins with a spatial encoder, typically a convolutional neural network, which compresses the sequence of high-dimensional images into a sequence of flattened feature vectors. In the second step, the exact continuous timestamps associated with each image are mapped into a high-dimensional vector space using sinusoidal basis functions. This continuous time embedding is calculated as (3):

$$PE(t) = [\sin(\omega_1 t), \cos(\omega_1 t), \sin(\omega_2 t), \cos(\omega_2 t), \ldots], \quad (3)$$

where ω represents different frequency bands. In the third step, these continuous embeddings are added directly to the spatial feature vectors. The transformer then calculates the self-attention weights using the standard formulation (4):

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (4)$$

Because the Queries (Q) and Keys (K) now contain absolute continuous temporal data, the attention mechanism mathematically learns to weight the relevance of past spatial frames based on their exact temporal distance from the target forecasting horizon, rather than their sequential index. Finally, a spatial decoder reconstructs the attention-weighted feature vectors back into the predicted image space.

In Earth observation sciences, continuous transformers are actively utilized for monitoring agricultural crop health and environmental degradation via satellites like Sentinel-2, where orbital revisit times are inherently irregular and data is frequently obscured by atmospheric conditions [6]. Furthermore, in the healthcare domain, spatial-temporal transformers are increasingly applied to electronic health records; while historically used for one-dimensional clinical data, modern adaptations process irregular sequences of patient medical imaging to accurately forecast long-term physiological trajectories [7].

Neural Ordinary Differential Equations (Neural ODEs) represent a radical mathematical departure from traditional networks by entirely abandoning discrete layers in the temporal dimension [8]. Instead of learning a discrete step-by-step state transition, a Neural ODE parameterizes the continuous derivative of the hidden state with respect to time using a neural network. The process executes in three distinct steps. First, a convolutional encoder maps the initially observed image into a highly compressed continuous latent vector $z(t_0)$. Next, the core dynamics of the system are defined by a neural network f_θ which models the derivative (5):

$$\frac{dz(t)}{dt} = f_\theta(z(t), t). \quad (5)$$

To generate a prediction at any arbitrary future timestamp t_{target}, the model employs a black-box differential equation solver (such as the Runge-Kutta method). The solver mathematically integrates the derivative starting from the initial state across the irregular time gap (6):

$$z(t_{target}) = z(t_0) + \int_{t_0}^{t_{target}} f_\theta(z(t), t)\, dt. \quad (6)$$

In the final step, a convolutional decoder projects the newly integrated latent state $z(t_{target})$ back into the high-dimensional pixel space. This integration guarantees a mathematically continuous and physically smooth trajectory of the spatial features over time, solving the temporal averaging issues seen in recurrent models.

Because Neural ODEs natively model continuous physical and biological processes, they have become a dominant architecture in longitudinal medical imaging. Researchers frequently utilize Latent ODEs to predict the future structural manifestation of diseases, such as the spatial progression of Alzheimer's disease or the volumetric growth of oncology tumors from sporadic MRI and CT scans, handling the highly irregular gaps between patient hospital visits [8]. Beyond healthcare, these dynamical systems are utilized in biomechanical engineering for motion tracking, successfully reconstructing continuous human or animal movement trajectories from camera feeds suffering from dropped frames or varying sampling rates [9].

While standard regression models are optimized to minimize absolute pixel error, they frequently produce blurred predictions when confronted with the inherent uncertainty of long, irregular time gaps. Generative Adversarial Networks (GANs) resolve this by learning the exact data distribution to synthesize realistic predictions through a two-player architecture [10]. The process involves two simultaneous steps driven by opposing networks. First, the Generator network receives the irregular historical frames and a specific target time gap. It attempts to synthesize a realistic future frame. Concurrently, the Discriminator network evaluates this generated frame against actual ground-truth sequences. The Discriminator outputs a probability representing whether the sequence is real or artificially generated. The entire architecture is trained using a minimax mathematical formulation, frequently augmented with a pixel-level penalty to enforce sequence accuracy (7):

$$V(D,G) = \mathbb{E}[\log D(x)] + \mathbb{E}[\log(1 - D(G(z,\Delta t)))] + \lambda \left|\left|G(z,\Delta t) - y\right|\right|_2. \quad (7)$$

This adversarial game forces the Generator to focus on preserving both spatial realism and physical motion dynamics over the requested continuous time interval. These conditionally timed GANs have found significant success in domains requiring high visual fidelity. In media and broadcasting, they are utilized for high-fidelity video frame interpolation, smoothly generating missing frames in archived footage or slow-motion conversions where the temporal gaps are arbitrary. In the physical sciences, conditional GANs are employed for fluid dynamics simulations [11], generating highly detailed future states of turbulent flows based on sporadically captured experimental sensor data.

Continuous-Time Diffusion Models (CDM) represent the current state-of-the-art in generative forecasting, relying on the mathematics of Stochastic Differential Equations (SDEs) to manage extreme uncertainty [12]. The training process operates by slowly corrupting a true future frame with Gaussian noise over a continuous time variable $t \in [0,1]$. This forward corruption step is defined by a stochastic differential equation (8):

$$dx = f(x,t)dt + g(t)dw, \quad (8)$$

where w represents standard Brownian motion. The model is trained to predict the score function, which is the gradient of the log probability density of the noisy data. During the infer-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

ence step, the model reverses this process to generate the missing frame at the targeted irregular timestamp. Starting from a state of pure noise, the model utilizes a numerical SDE solver to integrate the reverse-time SDE, conditioned on the historical context frames (9):

$$dx = [f(x,t) - g(t)^2 \nabla_x \log p_t(x)]dt + g(t)d\bar{w}. \quad (9)$$

The general process is portrayed in figure 2.

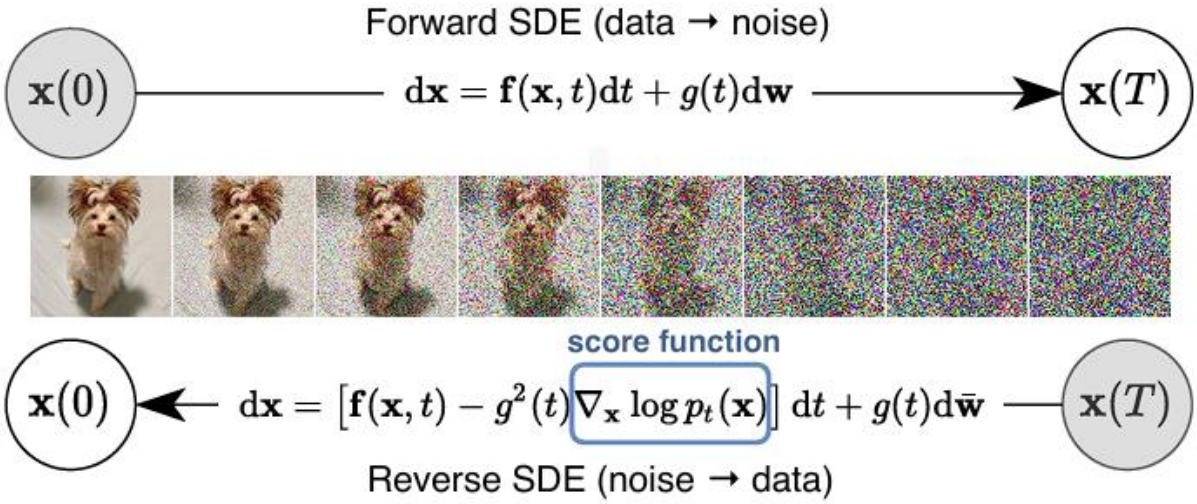


Figure 2 – The bidirectional SDE process [12]

In advanced meteorology, score-matching diffusion models are utilized to generate realistic, high-resolution future radar sequences of severe storm systems. Furthermore, in materials science, these continuous diffusion frameworks are used to predict the structural deformations and failure points of complex materials under stress over irregular observational periods, capturing the stochastic nature of physical degradation [13].

Purpose of the research. The object of the research is the process of spatio-temporal forecasting and reconstruction of digital image sequences captured at irregular time intervals. The subject of the research encompasses deep learning architectures – recurrent, transformer-based, generative adversarial, and diffusion models – and their capacity to interpolate spatial states across time.

The aim of the research is to conduct an analysis of the effectiveness of these architectures in forecasting spatio-temporal dynamics of varying natures, and to identify the most robust methods for systems with high structural accuracy requirements.

To achieve this aim, the following tasks were defined:

- investigate mechanisms for integrating time embeddings into neural network structures to overcome the limitations of discrete models;
- conduct a series of experiments on the Moving MNIST and meteorological air mass datasets to compare model accuracy on rigid and fluid structures;
- analyze the trade-offs between pixel precision and structural stability of the generated frames.

Presentation of the main research material. To empirically evaluate the theoretical capabilities of the selected architectures, a standardized experimental pipeline was developed. The primary objective of these experiments was to test the models' capacity to handle continuous-time dynamics and accurately interpolate missing spatial data across irregular observation intervals. Traditional architectures operate under the strict assumption of uniform temporal sampling, making this setup a rigorous test of how effectively these networks can deduce underlying physical trajectories and continuous atmospheric gradients when observations are scattered or corrupted. To ensure a robust evaluation across different types of motion, the experiments utilized two distinct datasets representing both the linear momentum of rigid bodies and the chaotic thermodynamics of atmospheric climate records.

The first dataset, Moving MNIST [14], features sequences of handwritten digits moving and colliding within a 64×64 pixel grid, providing a baseline for predicting the linear trajectories and elastic collisions of rigid objects. A substantial subset of 3,000 sequences was utilized, with each sequence consisting of 20 frames normalized to a floating-point range of [0.0, 1.0].To elevate the mathematical rigor of the evaluation, the second dataset introduced real-world meteorological data to evaluate continuous, non-rigid behavior. Specifically, tracking data of near-surface air temperature gradients, derived from the NCEP/NCAR Reanalysis project [15], was extracted. The raw temperature values were min-max normalized to a [0.0, 1.0] range to simulate visual pixel intensities and spatially interpolated to a fixed 64×64 resolution. Sequences for this meteorological dataset were extracted using a sliding window protocol to a fixed length of 10 frames. To rigorously evaluate the models' capacity to generalize, both datasets were strictly partitioned into a 70% training, 15% validation, and 15% testing split.

To simulate the irregular time intervals a stochastic continuous-time masking strategy was applied equally to both datasets. For every extracted sequence, a normalized continuous time vector representing the temporal progression from start to finish was generated. A binary mask was subsequently applied to the sequence where each frame had a 40% probability of being completely dropped. The initial frame of every sequence was explicitly excluded from this dropout probability and permanently preserved to ensure the networks always had a starting spatial context. During training and inference, the models were provided with only the sparse observed frames, the continuous timestamp vector, and the binary mask.

Four distinct architectures were implemented to process the continuous-time masked sequences, all optimized using the Adam optimizer with a batch size of 16. The first model, a Time-Aware ConvLSTM, served as a recurrent baseline modified to ingest explicit temporal deltas (Δt), optimized with a learning rate of 10^{-3} to minimize Mean Squared Error (MSE). The second model, a Continuous Temporal Transformer, provided an attention-based baseline utilizing sinusoidal continuous-time embeddings, similarly optimized at a learning rate of 10^{-3}. The third approach was a Spatio-Temporal GAN, which utilized the Transformer as a Generator paired with a deep 3D-CNN Discriminator. This adversarial model utilized an alternating training scheme with a learning rate of 2×10^{-4} balancing a Binary Cross-Entropy

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

adversarial loss with a heavily weighted MSE pixel loss ($\lambda = 100$) to anchor predictions to reality. Finally, a Continuous Diffusion Model (CDM) relying on a customized Mini U-Net was trained using a Variance Preserving Stochastic Differential Equation framework. Optimized with a learning rate of 10^{-4}, the CDM utilized a stable x_0-prediction DDIM reverse loop during inference to inherently bypass the mathematical singularities associated with continuous-time score matching.

The evaluation phase of the experiments was strictly isolated to the task of continuous-time interpolation. Performance metrics were calculated exclusively on the frames that were dropped by the stochastic mask. Two primary metrics were utilized: Mean Squared Error (MSE) to quantify absolute pixel-level deviation, and the Structural Similarity Index Measure (SSIM) to evaluate perceptual fidelity, structural cohesion, and contrast. To ensure statistical reliability and account for the stochastic nature of the masking process, the metrics were calculated across all test batches and aggregated. The final reported results represent the statistical mean accompanied by a 95% confidence interval (CI), calculated as (10):

$$CI = \bar{x} \pm 1.96\left(\frac{s}{\sqrt{n}}\right), \tag{10}$$

where $\bar{x}$ is the sample mean, s is the standard deviation, and n is the number of test batches.

The plots of the training dynamics of the models are depicted on Fig.3-6 (MNIST dataset) and Fig.7-10 (Meteorological Fluid Dynamics dataset). For MNIST dataset the total training duration for each architecture was determined based on the validation runs.

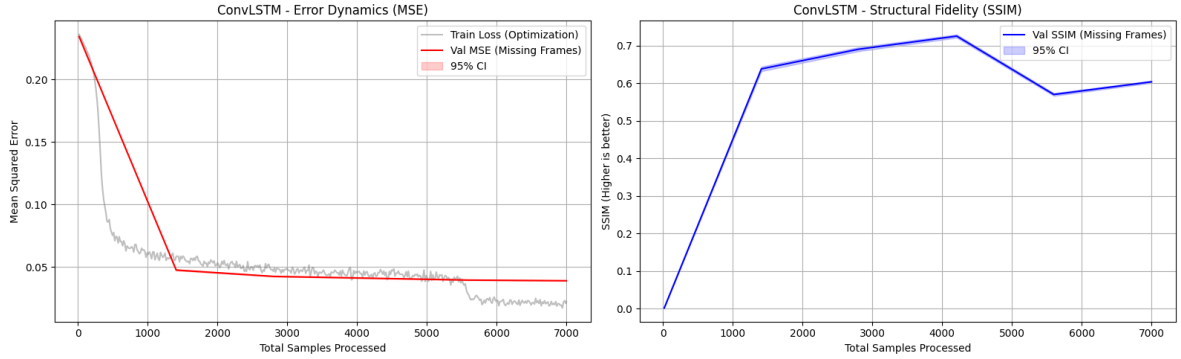


Figure 3 – Time-Aware ConvLSTM training dynamics on MNIST dataset

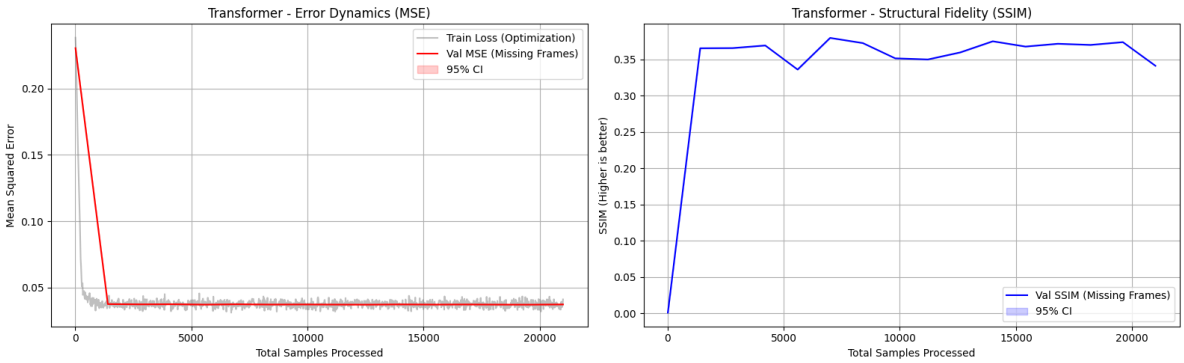


Figure 4 – Continuous Temporal Transformer training dynamics on MNIST dataset

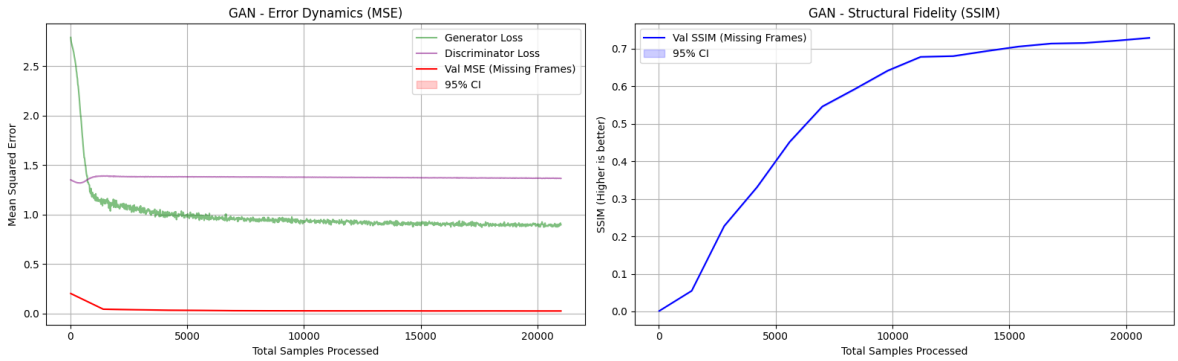


Figure 5 – GAN training dynamics on MNIST dataset

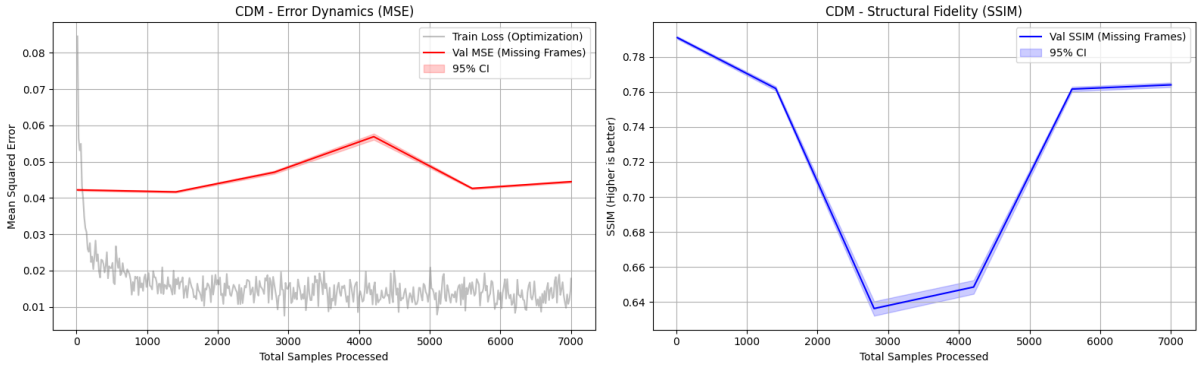


Figure 6 – CDM training dynamics on MNIST dataset

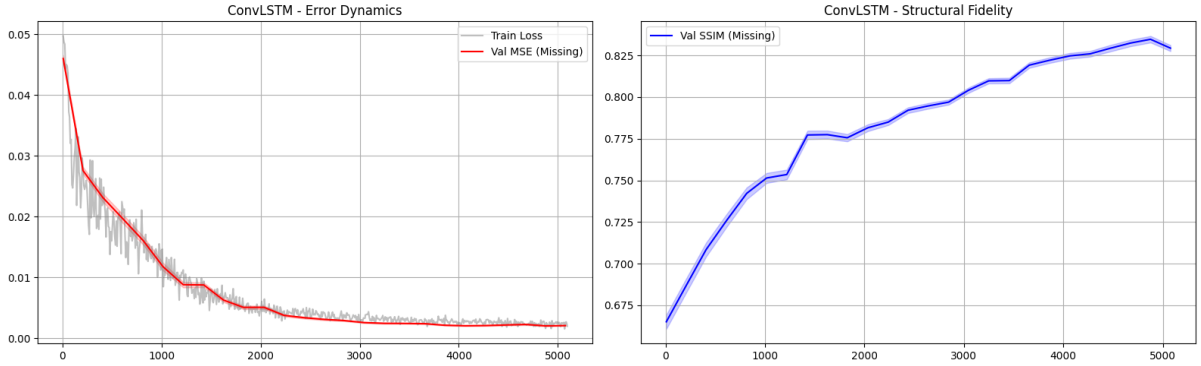


Figure 7 – Time-Aware ConvLSTM training dynamics on air temperature dataset

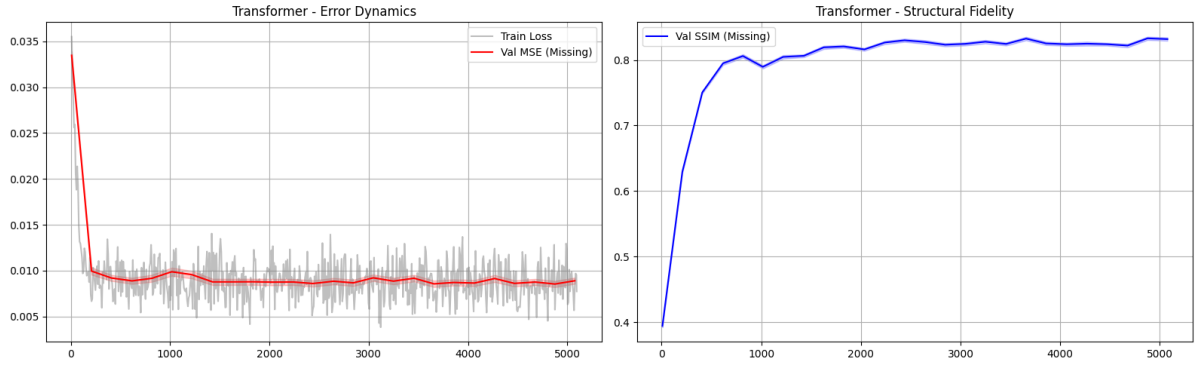


Figure 8 – Continuous Temporal Transformer training dynamics on air temperature dataset

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

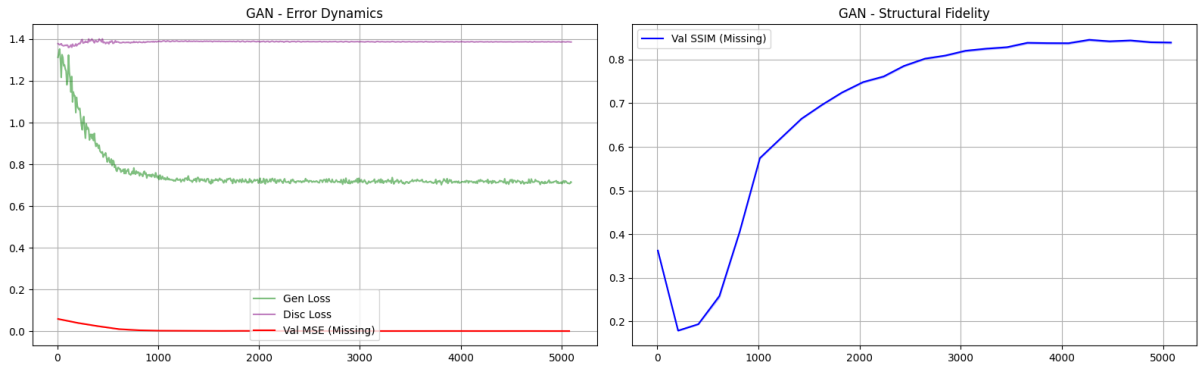


Figure 9 – GAN training dynamics on air temperature dataset

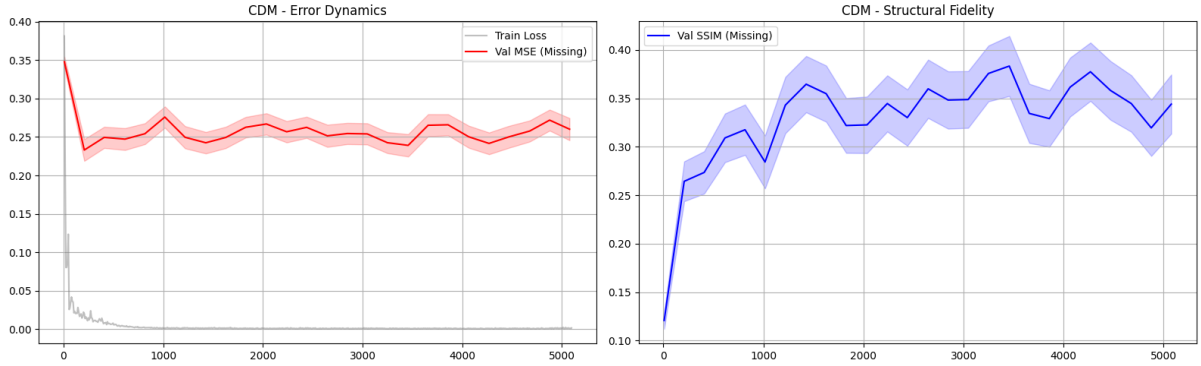


Figure 10 – CDM training dynamics on air temperature dataset

Table 1 and Table 2 summarize the final quantitative evaluation of all architectures across the two datasets.

Table 1

Forecast performance on irregular Moving MNIST

	Mean Squared Error (MSE) ± 95% CI	Structural Similarity (SSIM) ± 95% CI
Time-Aware ConvLSTM	0.0390 ± 0.0002	0.6037 ± 0.0030
Continuous Temporal Transformer	0.0374 ± 0.0002	0.3413 ± 0.0002
Spatio-Temporal GAN	0.0257 ± 0.0002	0.7289 ± 0.0011
Continuous Diffusion Model (CDM)	0.0445 ± 0.0003	0.7640 ± 0.0013

Table 2

Forecast performance on irregular Meteorological Fluid Dynamics dataset

	Mean Squared Error (MSE) ± 95% CI	Structural Similarity (SSIM) ± 95% CI
Time-Aware ConvLSTM	0.0021 ± 0.0001	0.8294 ± 0.0018
Continuous Temporal Transformer	0.0089 ± 0.0003	0.8320 ± 0.0025
Spatio-Temporal GAN	0.0018 ± 0.0001	0.8391 ± 0.0021
Continuous Diffusion Model (CDM)	0.2601 ± 0.0144	0.3441 ± 0.0305

On the meteorological dataset comprising atmospheric climate records, the ConvLSTM demonstrated highly stable, monotonic convergence. Across the 5,076 processed samples, the Validation MSE underwent a rapid exponential decay, ultimately settling at an 0.0021 ± 0.0001. Correspondingly, the structural fidelity climbed in a smooth, near-linear trajectory, achieving a final SSIM of 0.8294 ± 0.0018 on the missing frames. Because ConvLSTMs are optimized via MSE, they inherently produce spatial averages when uncertain about exact future states. When forecasting the shifting gradients of near-surface air temperatures, a slightly blurred spatial average still maintains good structural cohesion and gradient continuity, thereby yielding low MSE and high SSIM scores. Conversely, the limitations of deterministic MSE-optimized sequence modeling became apparent on the Moving MNIST dataset. ConvLSTM plateaued quite early, achieving a final MSE of 0.0390 ± 0.0002 and SSIM of 0.6037 ± 0.0030. When the ConvLSTM is uncertain about the exact linear trajectory or the precise angle of an elastic collision during a temporal gap, it hedges its prediction by outputting an overlapping average of all possible future positions. While this mathematically minimizes the worst-case MSE penalty, it results visually in severe blurring or "ghosting" of the final prediction.

The Continuous Temporal Transformer served as the non-recurrent, attention-based baseline. The training dynamics of the Transformer on the Moving MNIST dataset exhibited a rapid initial convergence followed by a plateau. As observed in the error dynamics, the Validation MSE plummeted almost instantly within the first 1,500 samples, dropping to its minimum boundary and remaining entirely flat for the remainder of the 21,000-sample training cycle. The final recorded MSE on the missing frames was 0.0374 ± 0.0002, which is mathematically competitive with the recurrent ConvLSTM baseline. However, an analysis of the SSIM reveals a critical vulnerability in the attention-based approach for predicting the trajectories of rigid bodies. While the SSIM initially spiked alongside the dropping MSE, it quickly plateaued and stagnated, ultimately finishing at a score of 0.3413 ± 0.0002. In contrast to its performance on rigid body dynamics, the Continuous Temporal Transformer demonstrated high capability when applied to the meteorological dataset of atmospheric climate records.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

The training dynamics reveal a highly efficient convergence profile. Within the first 1,000 samples, the Validation MSE plummeted to its near-minimum, while the SSIM simultaneously spiked from 0.4 to over 0.8. By the conclusion of the 5,076-sample training cycle, the Transformer achieved a final MSE of 0.0089 ± 0.0003 and a SSIM of 0.8320 ± 0.0025 on the stochastically dropped frames.

The training dynamics of the GAN on the Moving MNIST dataset illustrate the mechanics of adversarial equilibrium. Looking at the error dynamics, the Generator loss initially drops sharply as it quickly learns the general spatial distribution of the moving objects in the image sequences, while the Discriminator loss remains relatively stable, indicating a healthy adversarial tension without mode collapse. Correspondingly, the Validation MSE plummets in the first few thousand samples, flatlining at 0.0257 ± 0.0002 across the 21,000-sample training cycle. The structural fidelity climbs monotonically and smoothly throughout the training process, ultimately achieving a final SSIM of 0.7289 ± 0.0011 on the stochastically dropped frames. When applied to the meteorological dataset, the Spatio-Temporal GAN produced the most mathematically precise and structurally cohesive reconstructions of any evaluated architecture. The training dynamics reveal a sequence of adversarial adaptation. In the initial phases of training (the first 1,000 samples), the Validation SSIM experienced a sharp, transient dip, falling below 0.20. The Generator loss steadily decayed to a stable plateau, and the Validation MSE plunged to near absolute zero. Simultaneously, the SSIM reversed its initial decline and executed a smooth, monotonic climb, ultimately reaching a peak Validation SSIM of 0.8391 ± 0.0021 alongside the MSE of 0.0018 ± 0.0001 by the end of the 5,076-sample training cycle.

Unlike the smooth, monotonic curves of the ConvLSTM or GAN, the CDM's structural fidelity exhibits a "V-shape" trajectory on the Moving MNIST dataset. In the earliest stages of training, the Validation SSIM starts high, before plummeting to roughly 0.63 near the 3,000-sample mark. Initially, the untrained U-Net fails to properly estimate the noise gradients, actively destroying the image during the reverse diffusion loop. However, once the network's gradients align with the true data manifold, the model experiences a rapid recovery. The SSIM climbs, ultimately finishing at 0.7640 ± 0.0013 by the end of the 7,000-sample cycle. The Validation MSE for the CDM does not achieve the near-zero values seen in the ConvLSTM or GAN, instead fluctuating and ultimately settling at a higher 0.0445 ± 0.0003. While a deterministic model outputs a blurry average that strictly minimizes MSE, the CDM tries to generate a highly realistic prediction. However, if that generated trajectory is shifted by even one or two pixels from the exact ground-truth trajectory, the rigid MSE metric penalizes it heavily. While the CDM demonstrated a remarkable capacity to reconstruct the rigid boundaries and linear momentum of Moving MNIST, its application to the atmospheric climate records revealed a severe architectural vulnerability. When tasked with interpolating chaotic, low-frequency temperature gradients, the CDM exhibited the poorest mathematical and structural performance of all evaluated models. By the end of the 5,076-sample training cycle, the model achieved a Validation MSE of 0.2601 ± 0.0144 and a Validation SSIM of

0.3441 ± 0.0305 on the missing frames. The training loss decayed rapidly to near-zero, proving that the U-Net successfully learned to minimize the denoising objective on the training set. However, the Validation MSE hit a hard plateau almost immediately, hovering at an exceptionally high error rate. Simultaneously, the structural fidelity (SSIM) trajectory exhibited extreme instability, oscillating with large 95% CI indicating a complete lack of predictive consensus across the validation batches. When the CDM attempts to denoise these atmospheric states across an irregular temporal gap, it struggles to differentiate between the artificial Gaussian noise it must remove and the natural, dispersion of the weather system it must preserve. Furthermore, because unguided diffusion is inherently stochastic, the model generates plausible but mathematically incorrect weather states that diverge largely from the true physical trajectory, leading to the high absolute pixel error. These results indicate that while standard continuous-time diffusion is highly effective for predicting trajectories of rigid objects, it is fundamentally ill-suited for interpolating chaotic, low-frequency atmospheric climate data without explicit physics-informed regularization.

Conclusions. The empirical results confirm that the efficacy of continuous-time forecasting architectures is fundamentally linked to the frequency domain and physical nature of the target data. Adapted discrete-time models, such as Time-Aware ConvLSTMs, struggle significantly with high-frequency rigid body dynamics, frequently suffering from severe structural degradation due to the averaging of uncertain futures. Similarly, pure Continuous Temporal Transformers fail to preserve sharp boundaries when predicting linear trajectories, smearing predictions to minimize absolute error. However, for chaotic, low-frequency atmospheric climate records, the Transformer's global attention mechanism excels. Ultimately, the Spatio-Temporal GAN proved to be the robust architecture, utilizing adversarial regularization to prevent blurring and achieve state-of-the-art structural fidelity and pixel precision in across both rigid-body trajectories and complex atmospheric environments. Conversely, the CDM demonstrated high structural reconstruction for rigid body trajectoris but failed on smooth atmospheric temperature gradients due to the unintended injection of structural noise.

While the experiments conducted on the stochastic Moving MNIST and atmospheric climate record datasets successfully highlight the foundational differences between these algorithmic paradigms across both linear and non-linear physics, they operate within 2D constraints. A critical next step for future research is to validate these findings on high-dimensional, real-world datasets that inherently feature unequal time intervals and volumetric geometry. Specifically, extending these continuous-time architectures into 3D spatial environments is essential for emerging XR systems, where robust spatio-temporal forecasting is strictly required to anchor digital objects seamlessly despite sensor interruptions, asynchronous polling, and network latency. Additionally, testing these models on longitudinal 3D MRI medical scans to track disease progression, or on high-resolution satellite imagery, is necessary to fully assess their computational scalability and practical clinical or environmental utility.

Furthermore, the contrasting strengths and vulnerabilities of the evaluated models suggest a highly promising avenue for future architectural development: a forecasting-generation

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

framework. Because Continuous Temporal Transformers excel at modeling long-range global dependencies and tracking smooth, interconnected gradients, they are well suited to act as the temporal encoder. This Transformer could process the irregularly sampled historical context to output a highly accurate conditional latent vector for a specific target time. To overcome the Transformer's inability to render sharp boundaries, this condition could then be fed into a GAN acting as the spatial decoder. By utilizing the Transformer to resolve the continuous-time temporal dynamics and the GAN to solve the reverse stochastic differential equation, such model could theoretically achieve both the rigorous temporal accuracy of attention mechanisms and the high structural fidelity of score-based generative modeling.

REFERENCES

1. Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in Neural Information Processing Systems*, *28*.

2. Neil, D., Pfeiffer, M., & Liu, S. C. (2016). Phased LSTM: Accelerating Recurrent Network Training for Long or Event-based Sequences. *Advances in Neural Information Processing Systems*, *29*.

3. Cui, Z., Henrickson, K., Ke, R., & Wang, Y. (2019). Traffic Graph Convolutional Long Short-Term Memory Neural Network for Urban Traffic Flow Prediction. *Transportation Research Part C: Emerging Technologies*, 105, 361-374.

4. Vaswani, A., Shazeer, N., Parmar, N.,Uszkoreit, J., Jones, L.,Gomez, A.N.,& Polosukhin,I.(2017).Attention is all you need. *Advances in neural information processing systems, 30*.

5. Fonseca, A. H. D. O., Zappala, E., Caro, J. O., & Van Dijk, D. (2023). Continuous spatio-temporal transformers. *arXiv preprint* arXiv:2301.13338.

6. Rußwurm, M., & Körner, M. (2020). Self-Attention for Raw Optical Satellite Time Series Classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169, 421-435.

7. Tipirneni, S., & Reddy, C. K. (2022). Self-Supervised Transformer for Sparse and Irregularly Sampled Multivariate Clinical Time Series. *ACM Transactions on Knowledge Discovery from Data*, 16(6), 1-17.

8. Rubanova, Y., Chen, R. T. Q., Bettencourt, J., & Duvenaud, D. (2018). Neural Ordinary Differential Equations. *Advances in Neural Information Processing Systems*, 31.

9. De Brouwer, E., Simm, J., Arany, A., & Moreau, Y. (2019). GRU-ODE-Bayes: Continuous Modeling of Sporadically-Observed Time Series. *Advances in Neural Information Processing Systems*, 32.

10. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative Adversarial Networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/https://doi.org/10.1145/3422622

11. Xie, Y., Franz, E., Chu, M., & Thuerey, N. (2018). tempoGAN:A Temporally Coherent, Volumetric GAN for Super-resolution Fluid Flow. *ACM Transactions on Graphics*, 37(4), 1-15.

12. Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint* arXiv:2011.13456.

13. Price, I., Sanchez-Gonzalez, A., Alet, F., et al. (2023). GenCast: Diffusion-based Ensemble Forecasting for Medium-Range Weather. *arXiv preprint* arXiv:2312.15796.

14. Srivastava, N., Mansimov, E., & Salakhutdinov, R. (2015). Unsupervised learning of video representations using LSTMs. Proceedings of the 32nd International Conference on Machine Learning (ICML). http://www.cs.toronto.edu/~nitish/unsupervised_video/
15. Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., & Joseph, D. (2018). The NCEP/NCAR 40-year reanalysis project. *Renewable energy*, 1, 146-194.

Received 24.06.2026.
Accepted 25.08.2026.
Published 30.10.2026

Часово-просторове прогнозування цифрових зображень з неоднорідними часовими інтервалами: огляд глибоких нейромережевих підходів

У дослідженні оцінено чотири архітектури глибокого навчання (ConvLSTM, Transformer, Spatio-Temporal GAN та дифузійна модель) для прогнозування послідовностей зображень через нерегулярні часові інтервали. Тестування на наборах Moving MNIST таатмосферних кліматичних даних показало, що детерміновані моделі розмивають траєкторії жорстких об'єктів, тоді як GAN забезпечує високу структурну точність на обох наборах даних. Дифузійна модель чудово генерує траєкторії жорстких об'єктів, але не справляється з плавними атмосферними градієнтами. Запропоновано концепцію Transformer-GAN архітектури для забезпечення високоточного просторово-часового прогнозування цифрових зображень з неоднорідними часовими інтервалами.

Емпіричні результати підтверджують, що ефективність архітектур прогнозування в безперервному часі фундаментально пов'язана з частотною областю та фізичною природою цільових даних. Адаптовані дискретні моделі, такі як Time-Aware ConvLSTM, значно стикаються з труднощами при прогнозуванні високочастотної динаміки твердого тіла, часто страждаючи від серйозної структурної деградації через усереднення невизначеного майбутнього.

Федорчук Іван Васильович – аспірант кафедри програмного забезпечення комп’ютерних систем, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Україна.
ORCID: http://orcid.org/0009-0009-4429-1683
Шкурат Оксана Сергіївна – доцент кафедри програмного забезпечення комп’ютерних систем, Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ.
ORCID: http://orcid.org/0000-0001-7633-9121

Fedorchuk Ivan – PhD student of the Department of Computer Systems Software, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ukraine.
ORCID: http://orcid.org/0009-0009-4429-1683
Shkurat Oksana – Associate Professor of Computer Systems Software Department, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv.
ORCID: http://orcid.org/0000-0001-7633-9121

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)