УДК 004.89:004.75:004.62

ANDRIY MELNYK, VASYL TYMCHYSHYN, YURII POPYK, VADYM ZABCHUK, VASYL FATIUK

AN INTELLIGENT DATA PROCESSING ARCHITECTURE FOR COMPLEX INFORMATION SYSTEMS: CASE STUDIES IN ENVIRONMENTAL AND ENERGY SYSTEMS

West Ukrainian National University, Ukraine

Abstract. This paper proposes an intelligent data processing architecture for complex environmental and energy systems operating under conditions of high dynamics, heterogeneous data sources, and large-scale information flows. The architecture integrates distributed computing, edge/cloud infrastructure, IoT, stream analytics, and AI/ML models to support real-time data integration, normalization, synchronization, and intelligent analysis. A distinctive feature of the proposed approach is the incorporation of an intelligent anomaly detection method for heterogeneous streaming data. The method is based on multi-component assessment of the system state, taking into account the statistical characteristics of data streams, AI/ML model outputs, contextual rules, data quality, and temporal delays. This enables the detection not only of threshold-based deviations, but also of complex anomalous states associated with atypical parameter combinations, disruptions in temporal dynamics, or inconsistencies with domain-specific constraints. Practical evaluation was conducted using an environmental monitoring system and a smart grid network as case studies. The results confirmed the performance, scalability, adaptability, and effectiveness of the proposed architecture under high-load conditions, as well as its suitability for developing intelligent real-time information systems.

Keywords: software architecture, complex objects, data processing, stream analytics, IoT, anomaly detection, environmental monitoring.

Анотація. У статті запропоновано інтелектуальну архітектуру обробки даних для складних екологічних та енергетичних систем, що функціонують в умовах високої динамічності, гетерогенності джерел і значних обсягів інформаційних потоків. Архітектура поєднує distributed computing, edge/cloud infrastructure, IoT, потокову аналітику та AI/ML-моделі для інтеграції, нормалізації, синхронізації й інтелектуального аналізу даних у режимі реального часу. Особливістю підходу є включення інтелектуального методу виявлення аномалій у гетерогенних потокових даних. Метод ґрунтується на багатокомпонентному оцінюванні стану системи з урахуванням статистичних характеристик потоків, результатів AI/ML-моделей, контекстних правил, якості даних і часових затримок. Це дає змогу виявляти не лише порогові відхилення, а й складні аномальні стани, пов’язані з нетиповими комбінаціями параметрів, порушенням часової динаміки або невідповідністю доменним обмеженням. Практичне оцінювання виконано на прикладі системи екологічного моніторингу та smart grid-мережі. Результати підтвердили продуктивність, масштабованість, адаптивність і ефективність запропонованої архітектури в умовах високого навантаження, а також доцільність її використання для побудови інтелектуальних інформаційних систем реального часу.

Ключові слова: програмна архітектура, складні об’єкти, обробка даних, потокова аналітика, IoT, виявлення аномалій, екологічний моніторинг.

DOI: 10.31649/1681-7893-2026-51-1-41-56

INTRODUCTION

The rapid development of information technologies, intelligent systems, and artificial intelligence methods has led to a substantial increase in the volume, velocity, and heterogeneity of data generated by complex information systems. This issue is particularly relevant for environmental and energy systems, where continuous real-time collection, processing, and analysis of heterogeneous data streams are required.

Environmental monitoring systems, smart grids, distributed energy management systems, and intelligent urban infrastructures depend on effective mechanisms for data integration, state forecasting, anomaly detection, and operational decision-making under conditions of high process dynamics and uncertainty [1-3].

Traditional centralized data processing approaches often fail to provide the scalability, adaptability, and fault tolerance required by modern complex information systems. Their limitations become especially evident when integrating heterogeneous information sources, supporting stream analytics, and applying intelligent data analysis methods in real time. Consequently, the development of intelligent data processing architectures that combine distributed computing, edge/cloud infrastructure, machine learning methods, and adaptive decision-support mechanisms has become an important research direction. Such architectures are especially relevant for environmental monitoring, energy consumption forecasting, optimization of energy system operation, and detection of abnormal states [4].

Although recent studies actively apply artificial intelligence and stream analytics to complex information systems, many existing solutions remain focused on specific application tasks or isolated data processing scenarios. As a result, they do not fully provide an integrated approach to heterogeneous data integration, scalable stream processing, intelligent analysis, and real-time decision support. This creates a need for a universal intelligent architecture capable of ensuring effective interaction between system components, adaptability to changing operating conditions, and reliable decision-making in dynamic environments [5].

The aim of this study is to develop an intelligent data processing architecture for complex information systems and to evaluate its effectiveness using environmental and energy systems as representative case studies. The proposed approach integrates distributed computing, edge/cloud infrastructure, stream analytics, and AI/ML-based anomaly detection to support scalable, adaptive, and real-time processing of heterogeneous data streams.

LITERATURE REVIEW AND PROBLEM STATEMENT

Current research on data processing for complex information systems is primarily focused on improving the efficiency of large-scale data analysis, ensuring system adaptability, and supporting real-time decision-making. Significant contributions to the development of distributed systems, edge computing, fog computing, and intelligent data processing have been made by M. Satyanarayanan, F. Bonomi, H. Gupta, A. Zaslavsky, S. Dustdar, and W. Shi. Their studies demonstrate that distributed, edge, and cloud-based architectures can improve scalability, reduce communication latency, support local data processing, and increase the fault tolerance of information systems [6,7].

At the same time, the literature also indicates several unresolved challenges. Although edge/cloud integration enables more efficient use of network and computational resources, such architectures remain difficult to manage in highly dynamic environments. In particular, the synchronization of distributed components, cybersecurity, data consistency, and the coordination of heterogeneous computing resources remain critical issues. These limitations become especially significant when information systems operate with multiple data sources, high-frequency streams, and strict real-time requirements [8,9].

Despite substantial progress in distributed and intelligent data processing, the development of a universal architecture for complex information systems remains an open problem. Most existing solutions are designed for specific application domains, isolated analytical tasks, or particular data types. As a result, their adaptation to multi-level systems with numerous heterogeneous sources is limited. In practical environmental and energy systems, data may be structured, semi-structured, or unstructured and may originate from sensor networks, IoT devices, geographic information systems, SCADA platforms, smart meters, external APIs, and other information services. This creates a need for effective mechanisms of data integration, temporal synchronization, semantic harmonization, and unified stream processing [10-12].

A further challenge is the need to process data in real time. Environmental monitoring and energy systems require rapid responses to abnormal events, forecasting of system states, and decision support under continuously changing operating conditions. Traditional centralized architectures often fail to provide the required scalability and processing performance when the number of data sources and the intensity of information flows increase. In addition, dependence on central nodes may increase latency and reduce the stability of critical information infrastructures [13-15].

Adaptability and fault tolerance are also essential requirements for intelligent data processing systems. In complex information environments, the structure and intensity of data streams may change dynamically, individual nodes may become unavailable, new sources may appear, and computational loads may become unevenly distributed. Existing architectures often provide only limited mechanisms for automatic adaptation to such changes, which reduces their effectiveness under highly dynamic operating conditions [16,17].

Another important issue concerns the intelligent analysis of heterogeneous data streams. Although machine learning and artificial intelligence methods are widely used for forecasting, anomaly detection, classification, and optimization, their application in environmental and energy systems remains challenging. These challenges are related to the computational complexity of models, the need for high-quality training data, the presence of noise and missing values, and the limited interpretability of AI-based decisions. The latter is particularly important for decision-support systems, where users must understand not only the result of the analysis but also the reasons behind detected anomalies or recommended actions [18-20].

Therefore, the central research problem addressed in this study is the lack of an integrated intelligent data processing architecture capable of combining heterogeneous data integration, scalable stream processing, adaptive anomaly detection, risk assessment, and real-time decision support. This problem justifies the development of a new architectural approach that integrates distributed computing, edge/cloud infrastructure, stream analytics, semantic data management, and AI/ML methods for complex environmental and energy information systems.

PURPOSE AND OBJECTIVES OF THE STUDY

The aim of this study is to develop an intelligent data processing architecture for complex information systems that supports the integration, stream processing, and intelligent analysis of heterogeneous real-time data. The proposed architecture combines distributed computing, edge/cloud infrastructure, IoT components, stream analytics, scalable data storage, and AI/ML models to enable reliable decision support in dynamic information environments.

The study focuses on environmental monitoring and energy systems, where data are generated by multiple heterogeneous sources and differ in format, update frequency, reliability, and semantic context. Under such conditions, the architecture must ensure scalability, adaptability, fault tolerance, low-latency processing, and stable operation under high data-flow intensity.

A key component of the architecture is an intelligent method for anomaly detection in heterogeneous real-time data. The method enables the identification of both simple threshold violations and complex anomalous states caused by atypical parameter combinations, disruptions in temporal dynamics, changes in inter-parameter relationships, or inconsistencies with domain-specific rules.

The proposed method includes preprocessing, integration, normalization, temporal synchronization, feature-space construction, multi-component anomaly scoring, adaptive threshold-based decision-making, risk assessment, and anomaly classification. Its anomaly score combines statistical, AI/ML-based, and contextual components, thereby improving robustness to noise, missing values, and changes in normal operating conditions.

To achieve the aim of the study, it is necessary to analyze existing approaches to distributed computing, edge/cloud architectures, stream analytics, and intelligent data processing; investigate the properties of heterogeneous data streams in environmental and energy systems; develop a multi-level architectural model; and formalize the proposed anomaly detection method.

The architecture and method should be experimentally evaluated using environmental monitoring and smart grid scenarios. The evaluation should consider anomaly detection quality and real-time performance using precision, recall, F_1 -score, false alarm rate, processing latency, throughput, and robustness to partial data loss. Comparison with a baseline threshold-based approach should demonstrate the advantages of the proposed solution in terms of accuracy, stability, and interpretability.

Thus, the expected result is a coherent intelligent data processing architecture enhanced by a formalized anomaly detection method for heterogeneous streaming data. The proposed approach is intended to support scalable stream processing, adaptive anomaly detection, risk assessment, and real-time decision support in environmental, energy, and other critical information systems.

ARCHITECTURE OF AN INTELLIGENT DATA PROCESSING SYSTEM FOR COMPLEX INFORMATION ENVIRONMENTS

The proposed architecture of an intelligent data processing system is designed for operation in complex information environments characterized by multiple data sources, heterogeneous data structures, different update rates, variable data quality, and the need for timely decision-making. Such conditions are typical of environmental and energy systems, where information flows are generated by sensor networks, IoT devices, SCADA systems, geographic information platforms, external APIs, databases, satellite imagery, unmanned aerial vehicles, and other information services.

The central idea of the proposed architecture is to integrate distributed data processing, stream analytics, edge and cloud computing, and artificial intelligence methods within a unified technological framework. This integration enables the system not only to collect and process large volumes of data, but also to generate intelligent conclusions about the current state of the monitored environment, detect anomalous situations, forecast the further dynamics of processes, and support management decision-making. The architecture has a multi-level structure that provides functional separation between components, increases system flexibility, and creates the basis for further scalability (Figure 1).

The first level of the architecture is the data source layer. It includes physical, software-based, and information sources that generate primary data streams describing the state of monitored objects and processes. In environmental systems, such sources may include air pollution sensors, meteorological stations, spatial emission distribution maps, remote sensing data, and external meteorological services. In energy systems, data sources include smart meters, generation controllers, energy consumption monitoring systems, grid load parameters, operational data from energy facilities, and emergency event records.

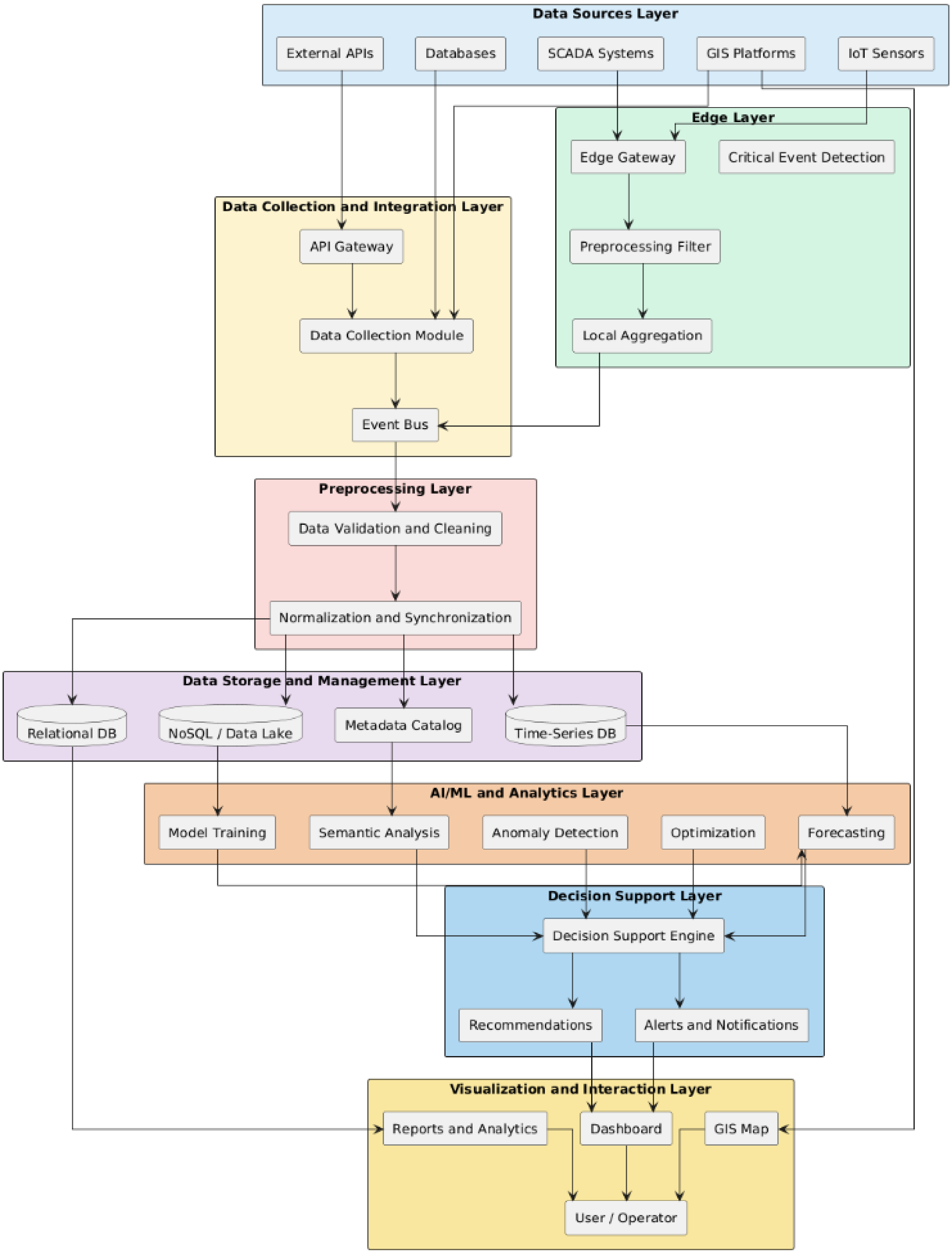

Figure 1 – Architecture of the intelligent data processing system for complex information environments

The second level is the data acquisition and preprocessing layer. Its purpose is to receive streams from heterogeneous sources, verify their correctness, filter noisy values, remove duplicates, process missing values, normalize parameters, and transform data into a unified format. At this stage, timestamps are also harmonized, preliminary aggregation is performed, and data are prepared for subsequent analytical processing. For real-time systems, minimizing the delay between data acquisition and transmission to the next processing levels is of particular importance.

The third level is responsible for data storage and management. It supports the organization of both historical and streaming data in appropriate storage systems. Relational databases can be used for structured data, time-series databases are suitable for high-frequency sensor streams, while NoSQL repositories or data lakes are appropriate for semi-structured and unstructured data. An important component of this level is the metadata subsystem, which describes data provenance, quality, temporal relevance, spatial reference, format, and semantic meaning. The presence of such a subsystem ensures data traceability, improves the reliability of analysis, and simplifies the integration of heterogeneous sources.

The fourth level is the analytical and intelligent processing layer. At this level, methods of machine learning, statistical analysis, anomaly detection, forecasting, classification, and optimization are implemented. AI/ML components of the architecture can be used to forecast pollution levels, assess environmental risks, predict energy consumption, detect abnormal operating modes of energy facilities, and optimize network load. For time-series analysis, recurrent neural networks, LSTM models, gradient boosting, ensemble methods, and other machine learning algorithms may be applied. If necessary, model outputs can be supplemented by expert system rules or semantic relationships between domain objects.

The fifth level is the decision-support and visualization layer. Its function is to present analytical results in the form of dashboards, maps, charts, alerts, forecast scenarios, and recommendations for users. This level should not only display the current state of the system, but also explain the reasons for detected changes or anomalies. For example, in an environmental system, a user may receive a notification about exceeding the permissible pollution level in a specific area, whereas in an energy system, the system may provide a recommendation on load redistribution, generation mode adjustment, or activation of reserve resources.

One of the key advantages of the proposed architecture is its scalability and modularity. These properties allow the system to adapt to an increasing number of data sources, growing volumes of information flows, and higher computational loads without requiring a complete redesign of the infrastructure. The use of distributed computing and the integration of edge/cloud solutions make it possible to distribute workloads among system components, reduce data transmission delays, and increase the speed of real-time information processing.

A particularly important feature of the architecture is the integration of IoT, edge, and cloud layers. IoT devices provide primary data acquisition from the physical environment, the edge layer performs preliminary processing close to the data source, and the cloud layer enables scalable storage, advanced analytics, and training of complex artificial intelligence models. This distribution of functions reduces network load, decreases latency, and improves system reliability. In critical situations, part of the decision-making process can be performed at the edge level without waiting for a response from the cloud infrastructure.

The modules responsible for data acquisition, processing, analysis, and decision support interact through APIs, message brokers, or event-driven infrastructure. The data acquisition module provides connectivity to sensors, external services, and information platforms. The preprocessing module performs data cleaning, transformation, and validation. The analytics module runs forecasting, anomaly detection, and risk assessment algorithms. The decision-support module generates recommendations or automated control actions, while the visualization module provides users with analytical results in an accessible and interpretable form.

Thus, the proposed architecture provides an integrated technological framework for processing heterogeneous data streams in complex information environments. Its main advantage lies in the combination of stream processing, distributed computing, AI/ML models, semantic data description, and decision-support mechanisms. This creates the prerequisites for the effective application of the architecture in environmental monitoring, smart grid systems, and other critical information environments where scalability, adaptability, fault tolerance, and real-time operation are required.

METHODS OF INTELLIGENT ANALYSIS AND ADAPTIVE DATA PROCESSING IN ENVIRONMENTAL AND ENERGY SYSTEMS

The effective operation of contemporary environmental and energy systems increasingly depends on their ability to perform intelligent analysis of large-scale heterogeneous data streams in real time. In this context, the proposed architecture employs an integrated set of intelligent data processing methods that combine machine learning, deep learning, stream analytics, semantic data processing, and adaptive decision-support mechanisms. The primary objective of applying these methods is to identify hidden regularities in data, forecast system states, detect anomalous events, and generate recommendations that support decision-making in complex information environments.

Within the proposed framework, time-series analysis and forecasting constitute one of the fundamental analytical components. Depending on the structure, dimensionality, and dynamics of the input data, regression-based models, Random Forest, XGBoost, Support Vector Machines, and recurrent neural networks, particularly Long Short-Term Memory networks, can be applied. In environmental monitoring tasks, such models enable the prediction of air pollution levels, changes in ecological indicators, and the potential spread of hazardous emissions. In energy systems, machine learning and deep learning models are used to forecast energy consumption, analyse smart grid load profiles, and optimise the operating modes of energy infrastructure. The use of deep neural architectures is particularly relevant in this context, as they are capable of automatically extracting complex nonlinear dependencies from high-dimensional and temporally dynamic data.

A key direction of intelligent data processing within the proposed architecture is the detection of anomalies and critical events. This task is addressed through the application of clustering algorithms, statistical analysis methods, autoencoders, Isolation Forest, and neural models designed to identify behavioural patterns of the system. In environmental systems, these methods make it possible to detect exceedances of permissible pollution thresholds, sensor failures, and atypical changes in environmental parameters in a timely manner. In energy systems, anomaly detection algorithms can be used to identify abnormal equipment behaviour, network overloads, unstable consumption patterns, or potential emergency conditions. Therefore, anomaly detection serves not only as a diagnostic mechanism but also as a preventive component of intelligent system management.

Semantic data processing plays an important role in the proposed architecture, as it enables the intelligent integration of heterogeneous information sources. This component may be implemented using ontological models, knowledge graph-based approaches, and semantic mechanisms for describing data, objects, and relationships within the domain. Semantic processing makes it possible to establish logical links between domain entities, support context-aware analysis, and improve the quality of decision-making. This is particularly important for environmental and energy systems, where data are obtained from numerous heterogeneous platforms and represented in different formats, levels of granularity, and semantic contexts.

To ensure the processing of continuous data flows, the architecture incorporates stream processing and event-driven analytics. Stream processing enables real-time analysis without the need for preliminary accumulation of large datasets, thereby supporting rapid responses to changes in the state of the monitored system. Within the proposed approach, message brokers, event bus mechanisms, and asynchronous data exchange between architectural components may be used to support scalable and fault-tolerant communication. This makes it possible to maintain analytical continuity even under conditions of high data velocity, dynamic system behaviour, and heterogeneous data generation sources.

Adaptive decision-making in the proposed system is based on the integration of forecasting results, stream analytics, and AI/ML-based models. The system is capable of automatically analysing the current state of the environment, assessing risks, and generating recommendations aimed at optimising the operation of environmental or energy systems. In the case of critical events, the architecture supports automated responses based on predefined scenarios or the outputs of intelligent models. As a result, the proposed approach improves operational efficiency, reduces the response time to anomalous situations, and supports stable system performance in highly dynamic information environments.

The AI/ML Decision Flow Diagram shown in Figure 2 represents the logic of intelligent data analysis and decision formation in complex environmental and energy systems. The input layer receives data from IoT sensors, SCADA systems, GIS platforms, and external APIs. These data are then transferred to the AI/ML analysis module, where the main intelligent operations are performed, including system state forecasting, anomaly detection, risk assessment, and the identification of the need for further control actions. Such an approach enables the system not only to analyse the current state of the monitored object but also to predict possible changes in its behaviour.

The first essential analytical direction of the system is anomaly detection. At this stage, AI/ML models analyse incoming data and determine whether the current parameters deviate from expected or permissible values. In environmental systems, such deviations may include a sharp increase in pollution levels, atypical changes in temperature or humidity, or abnormal concentrations of harmful substances. In energy systems, anomalies may include network overloads, unstable energy consumption, atypical equipment behaviour, or deviations from the predicted operating mode. If an anomaly is detected, the system proceeds to the generation of recommendations, alerts, or further risk assessment procedures.

The second important direction is system state forecasting. At this stage, machine learning or deep learning models analyse both historical and current data to predict future values of key parameters. For environmental systems, this may involve forecasting air pollution levels, the spread of emissions, or changes in ecological indicators within a specific area. For energy systems, forecasting may concern load dynamics, electricity consumption, renewable energy generation, or the probability of grid instability. The forecasting results are subsequently transferred to the risk assessment module, where they are interpreted in relation to operational thresholds, system constraints, and domain-specific requirements.

The risk assessment module determines the level of potential threat based on the results of forecasting and anomaly detection. If the predicted risk is high, the system generates recommendations for the operator or for an automated control subsystem. In environmental monitoring, this may involve recommendations to intensify monitoring in a specific area or to issue a warning regarding the exceedance of permissible environmental thresholds. In an energy system, the recommendations may include load redistribution, modification of generation modes, activation of reserve capacities, or other preventive actions aimed at maintaining system stability.

The next stage involves determining whether a control action is required. If the situation requires immediate intervention, the system may initiate an automated action or generate a notification for the operator. Automated actions should be implemented only in scenarios where response rules are clearly defined, verified, and safe. In other cases, the system functions as an intelligent decision-support tool by providing the operator with recommendations, explanations, and predictive scenarios. This hybrid approach ensures a balance between automation, interpretability, and human oversight, which is particularly important in critical environmental and energy infrastructures.

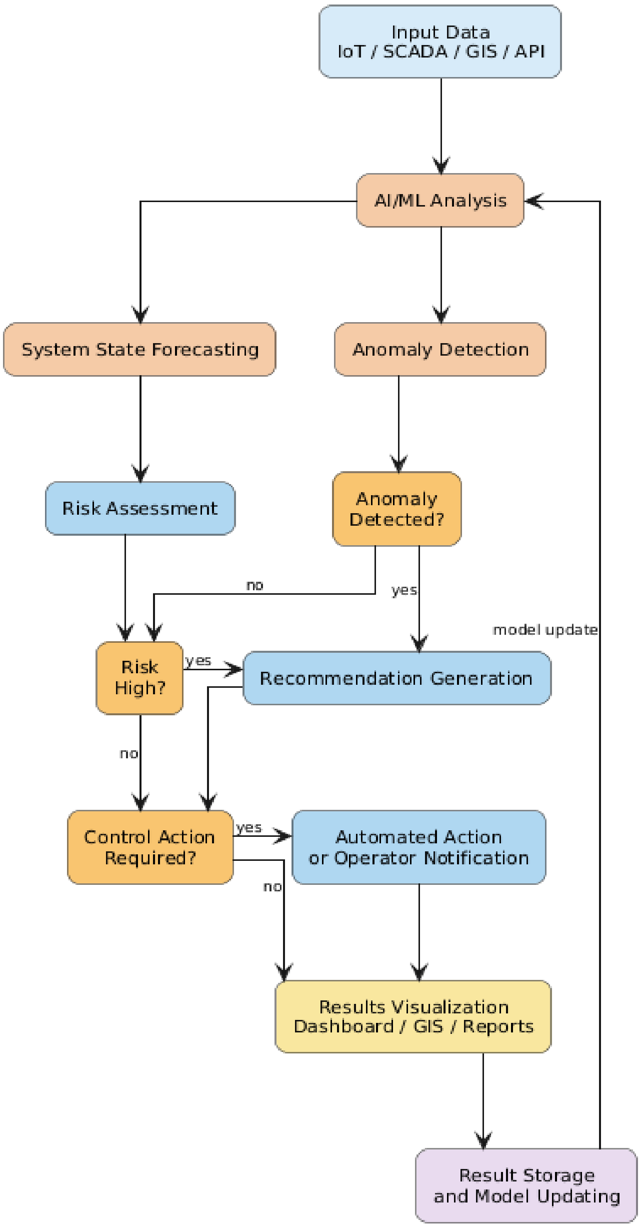

Figure 2 – AI/ML Decision Flow Diagram.

The final stage of the decision-making process is the visualisation of results in the form of a dashboard, GIS map, or analytical report. This allows users to assess the current state of the system, review detected anomalies, analyse predicted risks, and evaluate the proposed recommendations. The results generated by the system are also stored for further analysis and model updating. Through this feedback mechanism, AI/ML components can gradually improve forecasting accuracy and adapt to changes in system behaviour. Thus, the proposed architecture supports continuous learning and long-term analytical improvement.

The application of the proposed approach to intelligent analysis and adaptive data processing makes it possible to improve the efficiency of complex environmental and energy systems by combining stream analytics, AI/ML models, and adaptive decision-making mechanisms. Unlike traditional approaches, the proposed system is capable not only of responding to already observed events but also of forecasting potential risks and forming preventive response scenarios. This is particularly important for real-time systems, where the speed of anomaly detection and the timeliness of decision-making directly affect infrastructure stability, safety, and resource efficiency [21].

Furthermore, the integration of semantic data processing, stream analytics, and self-updating AI/ML mechanisms enables the system to adapt to changing environmental parameters and the emergence of new data sources. As a result, the architecture supports the continuous improvement of analytical processes and enhances forecasting accuracy under conditions of dynamic information flows. The proposed approach provides a methodological and technological basis for the development of modern intelligent monitoring and decision-support platforms in the fields of environmental protection, energy management, smart cities, and other domains that require high scalability, adaptability, and real-time intelligent data processing.

An additional advantage of the proposed architecture is its ability to integrate heterogeneous software and hardware components into a unified information environment without significantly increasing system complexity. The use of a modular approach and service-oriented interaction between components makes it possible to scale the system flexibly in accordance with changes in workload, the number of data sources, and functional requirements. This also enables the gradual extension of the architecture through the integration of new AI/ML models, IoT devices, analytical services, and domain-specific modules without the need for complete system redesign. Consequently, the proposed architecture improves the management of complex information processes, reduces infrastructure maintenance costs, and ensures stable system operation under conditions of continuously increasing data volumes.

AN INTELLIGENT METHOD FOR ANOMALY DETECTION IN HETEROGENEOUS DATA OF REAL-TIME INFORMATION SYSTEMS

The increasing complexity of real-time information systems has led to the continuous generation of large-scale data streams originating from heterogeneous sources. Such systems typically integrate sensor networks, industrial monitoring platforms, edge and cloud infrastructures, external information services, and analytical modules. As a result, the data processed by these systems differ not only in format and sampling frequency, but also in reliability, temporal accuracy, semantic context, and degree of structural organization. This heterogeneity substantially complicates the task of anomaly detection, especially when the system must operate under strict time constraints and maintain stable performance in dynamic environments.

In this context, anomaly detection cannot be reduced to the identification of isolated threshold violations. In complex information systems, an anomalous state may emerge as a local deviation of a single parameter, a disruption of temporal dynamics, an unusual combination of several variables, or a violation of domain-specific dependencies between system components. Consequently, conventional approaches based solely on fixed thresholds or individual machine learning models may be insufficient when data streams are asynchronous, incomplete, noisy, or contextdependent. To address these limitations, this study proposes an intelligent method that combines data integration, temporal synchronization, feature-space construction, multi-component anomaly scoring, adaptive decision-making, risk assessment, and anomaly interpretation within a unified methodological framework.

Formally, the proposed method can be represented as a mapping

$$M_A : D(t) \rightarrow A(t),$$

where M_A denotes the intelligent anomaly detection method, $D(t)$ is the input heterogeneous data stream at time t, and $A(t)$ is the anomaly detection result. The output $A(t)$ is not limited to a binary decision; it also includes the anomaly score, risk level, anomaly type, and explanatory information required for subsequent decision support.

Let the set of data sources in a real-time information system be denoted as

$$S = \{s_1, s_2, \ldots, s_n\},$$

where s_i is the i-th source and n is the total number of sources. Each source generates a data stream

$$D_i(t) = \{d_{i1}(t), d_{i2}(t), \dots, d_{ik}(t)\},$$

and the overall heterogeneous input stream is defined as

$$D(t) = \bigcup_{i=1}^{n} D_i(t)$$

To ensure unified processing, each data element is represented as a tuple

$$d_{ij}(t) = \langle id, s_i, x, \tau, q, m, c \rangle$$

where id is the unique identifier of the data element, s_i indicates its source, x is the value or vector of values, τ is the timestamp, q represents the data quality or reliability indicator, m denotes technical metadata, and c characterizes the contextual or semantic meaning of the data. If the element describes a multidimensional state, the value component is represented as

$$x = (x_1, x_2, \dots, x_p)$$

where p is the number of primary features. Such a representation makes it possible to combine numerical, temporal, technical, and contextual properties within a single data model.

Since the input streams differ in structure and temporal behavior, the method explicitly introduces a heterogeneity profile for each source:

$$H_i = \langle F_i, R_i, T_i, Q_i, C_i, \Delta_i \rangle$$

where F_i denotes the data format, R_i is the arrival frequency or intensity, T_i is the data type, Q_i is the quality or reliability level, C_i is the semantic context, and Δ_i is the admissible temporal deviation or delay. The overall heterogeneity model is therefore defined as

$$H = \{H_1, H_2, \dots, H_n\}$$

This model is used to adapt preprocessing, synchronization, and anomaly scoring procedures to the properties of individual data sources. In particular, sources with higher reliability and lower latency should have a greater influence on the integrated assessment of the system state. For this purpose, a source weight may be introduced as

$$w_i = f_w(Q_i, R_i, \Delta_i)$$

which, in a simplified form, can be expressed as

$$w_i = \frac{Q_i}{1 + \Delta_i}$$

The normalized source weight is then computed as

$$\widetilde{w}_i = \frac{w_i}{\sum_{j=1}^{n} w_j}.$$

After defining the input data model, the proposed method is formalized as a composition of functional transformations:

$$A(t) = F_{\text{det}}\left(F_{\text{risk}}\left(F_{\text{score}}\left(F_{\text{feat}}\left(F_{\text{sync}}\left(F_{\text{norm}}\left(F_{\text{int}}\left(F_{\text{prep}}\left(D(t)\right)\right)\right)\right)\right)\right)\right)\right).$$

This composition reflects the complete processing pipeline, beginning with raw heterogeneous data streams and ending with an interpreted anomaly detection result. The preprocessing function F_{prep} performs initial validation, removal of technically incorrect records, duplicate elimination, and filtering of obvious noise values. The integration function F_{int} maps heterogeneous streams into a unified logical representation:

$$F_{int}: D_{prep}(t) \to D_{int}(t)$$

where

$$D_{\text{int}}(t) = \bigcup_{i=1}^{n} \phi_i(D_i(t)),$$

and ϕ_i is the transformation function for the i-th data source. This stage is essential because anomalies in complex systems often arise not as isolated deviations, but as violations of consistency between several parameters originating from different sources.

The integrated data are then cleaned and normalized. The normalization function is defined as

$$F_{\text{norm}} : D_{\text{int}}(t) \to D_{\text{norm}}(t)$$

For numerical features, min-max normalization is applied:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

where x' is the normalized value, x is the original value, and $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the corresponding feature. If missing values occur, they are restored using an imputation function

$$x_{miss}(t) = f_{imp}(x(t-1), x(t-2), \ldots, x(t-l)),$$

where l is the depth of the retrospective window. In real-time systems, such procedures must be computationally efficient; therefore, local methods such as moving averages, exponential smoothing, or short-term regression models are preferable when low latency is required.

A critical stage of the method is temporal synchronization. Since heterogeneous streams arrive with different frequencies and delays, the system forms synchronized time windows that describe a consistent state of the monitored object or process. For each source, a local window is defined as

$$W_i(t) = \{d_{ij}(\tau) : \tau \in [t - \Delta t, t]\}$$

The general synchronized window is then obtained as

$$W(t) = \bigcup_{i=1}^{n} W_i(t)$$

and the synchronization procedure is represented by

$$F_{\text{sync}} : D_{\text{norm}}(t) \to W(t)$$

The choice of the window size Δt is particularly important. A shorter window improves responsiveness to abrupt anomalies but may increase sensitivity to noise, whereas a longer window provides more stable estimates but can delay detection. Therefore, the window size must be selected according to the dynamics of the monitored process and the admissible response latency.

Based on the synchronized window, the method constructs a feature vector

$$v(t) = F_{\text{feat}}(W(t)).$$

In addition to primary normalized features, the vector may include derived and statistical characteristics:

$$v(t) = \big(x_1(t), x_2(t), \ldots, x_p(t), r_1(t), r_2(t), \ldots, r_b(t)\big).$$

To capture temporal behavior, the method uses feature increments,

$$\Delta x_j(t) = x_j(t) - x_j(t-1)$$

local mean values,

$$\mu_j(t) = \frac{1}{|W(t)|} \sum_{\tau \in W(t)} x_j(\tau)$$

and local standard deviations,

$$\sigma_j(t) = \sqrt{\frac{1}{|W(t)|} \sum_{\tau \in W(t)} \big(x_j(\tau) - \mu_j(t)\big)^2}$$

These characteristics allow the method to detect not only point anomalies, but also contextual, collective, and trend-based deviations.

The central analytical component of the method is the computation of an integral anomaly score. Unlike single-criterion approaches, the proposed method combines statistical, machine-learning-based, and contextual components:

$$AS(t) = \alpha AS_{stat}(t) + \beta AS_{ml}(t) + \gamma AS_{ctx}(t),$$

where $AS(t)$ is the anomaly score, $AS_{\text{stat}}(t)$ is the statistical component, $AS_{ml}(t)$ is the component obtained from an AI/ML model, and $AS_{ctx}(t)$ is the contextual component. The weighting coefficients satisfy the condition

$$\alpha + \beta + \gamma = 1.$$

The statistical component can be defined as a normalized deviation from the local mean:

$$AS_{stat}(t) = \frac{|x(t) - \mu(t)|}{\sigma(t) + \varepsilon}$$

where ε is a small positive constant preventing division by zero. The machine-learning component may be based on reconstruction or prediction error:

$$AS_{ml}(t) = \|v(t) - \hat{v}(t)\|,$$

where $\hat{v}(t)$ is the reconstructed or predicted feature vector. The contextual component evaluates the consistency of the current state with domainspecific rules:

$$AS_{ctx}(t) = f_{ctx}(v(t), R_{domain})$$

where R_{domain} is the set of domain-oriented constraints or rules.

After the anomaly score is obtained, the method applies an adaptive threshold-based decision rule:

$$a(t) = \begin{cases} 1, & AS(t) \geq \theta(t) \\ 0, & AS(t) < \theta(t) \end{cases}$$

The adaptive threshold is computed as

$$\theta(t) = \mu_{AS}(t) + k\sigma_{AS}(t)$$

where $\mu_{AS}(t)$ and $\sigma_{AS}(t)$ are the mean and standard deviation of the anomaly score over the previous time interval, and k is a sensitivity coefficient. This mechanism allows the method to adapt to changes in the normal operating mode, which is especially important for systems affected by seasonality, load variation, external conditions, or structural reconfiguration, Fig. 3.

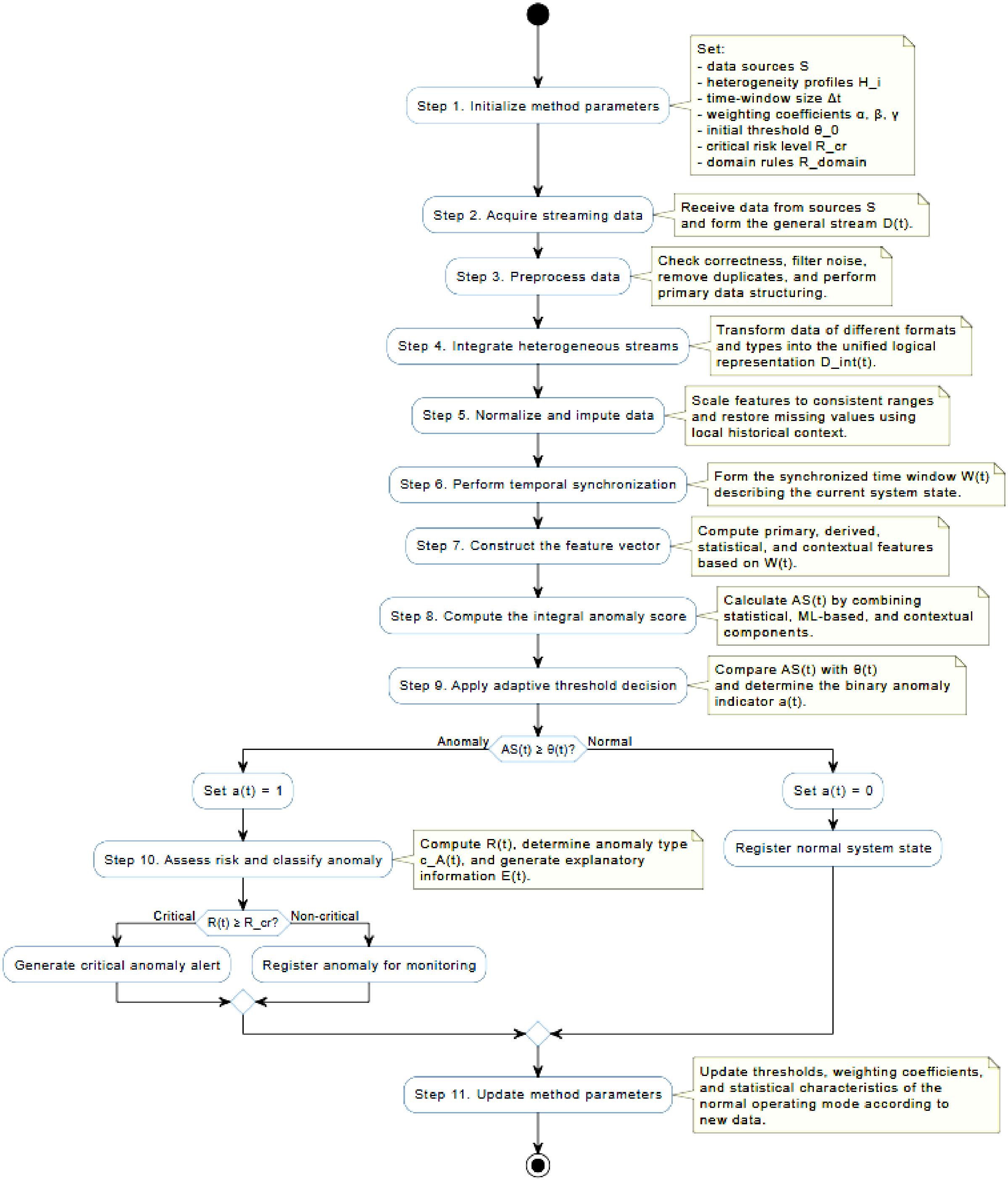

Figure 3 – Algorithmic Implementation of the Intelligent Anomaly Detection Method

Thus, the o Since not all anomalies have the same operational significance, the method additionally evaluates the risk associated with the detected state:

$$R(t) = f_R(AS(t), q(t), c(t), I(t)),$$

where $q(t)$ denotes data reliability, $c(t)$ is the system context, and $I(t)$ represents the potential impact of the anomaly. In simplified form, the risk level can be expressed as

$$R(t) = AS(t) \cdot I(t) \cdot q(t)$$

If

$$R(t) \geq R_{cr},$$

the anomaly is interpreted as critical and requires an operational response. To increase the interpretability of the result, the method also determines the anomaly type. The set of anomaly classes is defined as

$$C_A = \{A_{\mathrm{point}}, A_{\mathrm{context}}, A_{\mathrm{collective}}, A_{\mathrm{trend}}, A_{\mathrm{system}}\},$$

and the classification function is given by

$$c_A(t) = f_{\mathrm{class}}(v(t), AS(t), R(t)).$$

utput of the method is not only a detection signal, but also an interpretable representation of the anomalous state. The final result is defined as

$$A(t) = \langle a(t), AS(t), R(t), c_A(t), E(t) \rangle$$

where $E(t)$ contains explanatory information about the causes and characteristics of the detected deviation. In generalized form, the proposed method can be represented as the tuple

$$M_A = \langle S, D, H, W, V, F, AS, \Theta, R, C_A, A \rangle,$$

where S is the set of data sources, D is the set of heterogeneous data streams, H is the heterogeneity model, W is the set of synchronized time windows, V is the feature space, F is the set of functional transformations, AS is the anomaly scoring function, Θ is the set of adaptive thresholds, R is the risk assessment function, C_A is the set of anomaly classes, and A is the anomaly detection result.

The scientific novelty of the proposed method lies in the formalization of anomaly detection in heterogeneous streaming data as a unified composition of integration, synchronization, feature construction, multi-component scoring, adaptive thresholding, risk assessment, and anomaly interpretation. In contrast to approaches based on fixed thresholds or isolated ML models, the proposed method simultaneously considers statistical deviations, prediction or reconstruction errors, contextual rules, data quality, temporal delays, and the potential impact of anomalous states on system operation.

EXPERIMENTAL EVALUATION AND DISCUSSION

The experimental evaluation was conducted to assess the effectiveness of the proposed intelligent method for anomaly detection in heterogeneous data of real-time information systems. The main objective was to determine whether the method can reliably detect anomalous states under conditions of data heterogeneity, stream asynchrony, missing values, noise, and variable computational load. To provide a representative assessment, two applied scenarios were considered: an environmental monitoring system and an energy smart grid system. These domains were selected because they are typical examples of complex real-time information systems in which data are continuously generated by multiple sources and must be processed with minimal latency.

The experimental environment reproduced a multi-level information system consisting of a data acquisition layer, an edge preprocessing layer, a heterogeneous stream integration module, a temporal synchronization module, a feature construction module, and an intelligent anomaly detection module. The general experimental workflow corresponded to the formal model of the proposed method:

$$D(t) \rightarrow D_{\mathrm{int}}(t) \rightarrow D_{\mathrm{norm}}(t) \rightarrow W(t) \rightarrow v(t) \rightarrow AS(t) \rightarrow A(t).$$

This workflow reflects the transformation of raw heterogeneous data into an interpretable anomaly detection result. In the first scenario, the method was evaluated using an environmental monitoring system. The data streams included measurements of temperature, humidity, atmospheric pressure, PM2.5 and PM10 concentrations, CO_2 level, air quality index, and external meteorological indicators. The set of environmental sources was represented as

$$S_{\mathrm{eco}} = \{s_{\mathrm{temp}}, s_{\mathrm{hum}}, s_{\mathrm{press}}, s_{\mathrm{PM\ 2.5}}, s_{\mathrm{PM\ 10}}, s_{\mathrm{CO2}}, s_{\mathrm{AQI}}, s_{\mathrm{meteo}}\}.$$

The environmental state vector was defined as

$$z_{\mathrm{eco}}(t) = (T\mathrm{emp}(t), \mathrm{Hum}(t), PM2.5(t), PM10(t), CO_2(t), AQI(t)).$$

In this case, anomalous states were associated with abrupt changes in air-quality parameters, atypical increases in pollution indicators, inconsistencies between sensor measurements and meteorological context, and collective deviations involving several environmental variables.

The second scenario concerned an energy smart grid system. The analyzed streams described electricity consumption, generation, voltage, frequency, battery-system state, distributed energy resources, node load, and emergency events. The set of energy-related sources was defined as

$$S_{\mathrm{en}} = \{s_{\mathrm{load}}, s_{\mathrm{gen}}, s_{\mathrm{volt}}, s_{\mathrm{freq}}, s_{\mathrm{bat}}, s_{\mathrm{DER}}, s_{\mathrm{event}}, s_{\mathrm{SCADA}}\}.$$

The smart grid state vector was represented as

$$z_{en}(t) = (\mathrm{Load}(t), \mathrm{Gen}(t), \mathrm{Volt}(t), \mathrm{Freq}(t), \mathrm{Bat}(t), \mathrm{DER}(t)).$$

In this scenario, anomalies included abrupt load changes, voltage or frequency deviations, imbalance between generation and consumption, atypical battery behavior, and potential emergency operating modes.

The experiment was carried out in several stages. First, heterogeneous streams with different arrival frequencies, formats, and reliability levels were generated or collected. To reproduce realistic operating conditions, the streams included noise outliers, missing values, and temporal delays. The data were then integrated into a unified representation, normalized, and synchronized within time windows. For each window, the feature vector $v(t)$ was constructed, after which the integral anomaly score was calculated as

$$AS(t) = \alpha AS_{\mathrm{stat}}(t) + \beta AS_{ml}(t) + \gamma AS_{ctx}(t).$$

The anomaly decision was made using the adaptive rule

$$a(t) = \begin{cases} 1, & AS(t) \geq \theta(t) \\ 0, & AS(t) < \theta(t) \end{cases}$$

where

$$\theta(t) = \mu_{AS}(t) + k\sigma_{AS}(t)$$

The effectiveness of the method was evaluated using both detection-quality metrics and real-time performance indicators. Precision, recall, F_1 score, and false alarm rate were used to assess the quality of anomaly detection:

$$\mathrm{Precision} = \frac{TP}{TP + FP},$$

$$\mathrm{Recall} = \frac{TP}{TP + FN},$$

$$F_1 = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}},$$

$$FAR = \frac{FP}{FP + TN}.$$

In addition, average processing latency and throughput were calculated to evaluate the suitability of the method for real-time operation:

$$L = \frac{1}{N}\sum_{i=1}^{N} \left(t_i^{\mathrm{out}} - t_i^{\mathrm{in}}\right),$$

$$T = \frac{N_{\mathrm{proc}}}{\Delta t}.$$

The experimental results demonstrated that the proposed method provides stable and accurate anomaly detection in both application scenarios (Table 1). In the environmental monitoring case, the method achieved high sensitivity to abrupt changes in air-quality parameters and reduced false alarms by incorporating meteorological context. In the smart grid case, the method enabled timely detection of unstable operating modes, including voltage deviations, frequency disturbances, and generation-consumption imbalance.

Table 1 – Performance Metrics of the Proposed Architecture in Environmental and Energy Case Studies

Metric	**Environmental System**	**Smart Grid System**
Precision	0.92	0.94
Recall	0.89	0.93
F1-score	0.91	0.94
FAR	0.061	0.048
Average latency, ms	185	142
Throughput, events/s	4800	6200
Robustness under 10% data loss	0.96	0.97

The results indicate that the smart grid scenario achieved slightly higher precision, recall, and throughput, as well as lower processing latency. This can be explained by the more structured and regular nature of energy data. By contrast, the environmental monitoring scenario involved greater heterogeneity due to the combination of sensor measurements, external meteorological data, and spatio-temporal context.

To further assess the advantages of the proposed method, its performance was compared with a baseline fixed threshold-based approach. In the baseline approach, anomalies were detected only when predefined boundary values were exceeded. The comparison showed that the proposed method substantially improved detection quality and reduced the false alarm rate (Table 2).

These results confirm that the integration of statistical, AI/ML-based, and contextual components provides a more reliable anomaly detection mechanism than fixed thresholding. The reduction in false alarms is particularly important for real-time information systems, where excessive false alerts may overload operators or trigger unnecessary control actions.

Table 2 – Comparative Evaluation of the Proposed Anomaly Detection Method and Baseline Threshold-Based Approach

Method	Precision	Recall	F1-score	FAR
Fixed threshold-based approach	0.81	0.78	0.79	0.124
Proposed method	0.93	0.91	0.92	0.054

Overall, the experimental findings demonstrate that the proposed method is suitable for heterogeneous real-time environments. Its main advantages include improved detection accuracy, reduced false alarm rate, robustness to missing and noisy data, and acceptable processing latency. At the same time, the effectiveness of the method depends on the correct selection of the time-window size, weighting coefficients, and domain-specific rules. Therefore, future research should focus on automated parameter optimization and the integration of explainable Al mechanisms to improve the interpretability of anomaly detection results.

CONCLUSIONS

This study developed an intelligent data processing architecture for complex information systems, with a particular focus on environmental and energy domains. The proposed approach integrates distributed computing, edge/cloud infrastructure, IoT components, stream analytics, and AI/ML models to support the integration, normalization, synchronization, and intelligent analysis of heterogeneous real-time data. The architecture enables scalable stream processing, adaptive decision support, system-state forecasting, and intelligent interaction among the components of a dynamic information environment.

An important contribution of this study is the integration of an intelligent anomaly detection method for heterogeneous streaming data. The method is based on a multi-component assessment of the system state, taking into account statistical stream characteristics, AI/ML model outputs, contextual rules, data quality, and temporal delays. This makes it possible to detect not only simple threshold-based deviations, but also complex anomalous states associated with atypical parameter combinations, disruptions in temporal dynamics, or inconsistencies with domain-specific constraints.

The proposed architecture was validated using two case studies: an environmental monitoring system and a smart grid system. The experimental results confirmed the effectiveness of the proposed solution in terms of performance, scalability, adaptability, fault tolerance, and anomaly detection quality. The use of the edge layer reduced the load on the cloud infrastructure, decreased processing latency, and improved the system's response to critical events.

The obtained results demonstrate the suitability of the proposed architecture for the development of modern environmental monitoring systems, smart grids, smart city platforms, and other cyber-physical systems operating in highly dynamic information environments. The proposed approach supports the integration of heterogeneous data sources, stream analytics, adaptive anomaly detection, risk assessment, and real-time decision-making.

Future research should focus on the automatic tuning of the anomaly detection method, optimization of the weighting coefficients used in the integral anomaly score, integration of explainable AI mechanisms, and validation of the architecture across a broader range of application scenarios. These directions will improve the interpretability, robustness, and practical applicability of the proposed approach for critical real-time information systems.

FUNDING

This research was supported by the National Research Foundation of Ukraine under project No. 2025.07/0207 «Concept of a Knowledge-Oriented Technology for Modeling and Forecasting the State of the Natural Environment Based on Imprecise Data».

REFERENSES

1. Zhang C., Patras P., Haddadi H. Deep Learning in Mobile and Wireless Networking: A Survey // IEEE Communications Surveys & Tutorials. 2020. Vol. 22(3). P. 2224–2287.
2. Chen Z., Chan I. C. C. Smart Cities and Quality of Life: A Quantitative Analysis of Citizens' Support for Smart City Development // Information Technology & People. 2023. Vol. 36(1). P. 263–285.
3. Khan Z., Anjum A., Soomro K., Tahir M. Towards Cloud Based Big Data Analytics for Smart Future Cities // Journal of Cloud Computing. 2020. Vol. 9(1). P. 1–24.
4. Ahmed E., Yaqoob I., Hashem I. A. T., et al. The Role of Big Data Analytics in Internet of Things // Computer Networks. 2021. Vol. 129. P. 459–471.
5. Aithal P. S. ICT and Digital Technology Based Solutions for Smart City Challenges and Opportunities // International Journal of Applied Engineering and Management Letters (IJAEML). 2022. Vol. 6(1). P. 1–21.
6. Vedantham L. S., Zhou Y., Wu J. Information and Communications Technology (ICT) Infrastructure Supporting Smart Local Energy Systems: A Review // IET Energy Systems Integration. 2022. Vol. 4(4). P. 460–472.
7. Breunig M. M., Kriegel H.-P., Ng R. T., Sander J. LOF: Identifying Density-Based Local Outliers // ACM Transactions on Database Systems. 2020. Vol. 45(2). P. 1–24.
8. Entezari A., Aslani A., Zahedi R., et al. Artificial Intelligence and Machine Learning in Energy Systems // Energy Reports. 2023. Vol. 9. P. 1045–1063.
9. Aleran A., Kumar R., Singh P., et al. Artificial Intelligence and the Internet of Things in Energy Preservation Systems // Computing. 2025. Vol. 107. P. 1–25.
10. Lukashevych V. V., Vakaliuk T. A. Scientific Approaches to Big Data Processing Methods // Actual Problems of Automation and Information Technologies. 2025.
11. Artemchuk V., Iatsyshyn A., Kovach V., Popov O. Intelligent Analysis of Atmospheric Air Monitoring Network Data as a Tool for Environmental Safety Management of Urbanized Territories. 2019.
12. Das D. K., Sharma V., Gupta N., et al. Integrating IoT and AI for Sustainable Energy-Efficient Smart Buildings // Sustainability. 2025. Vol. 17(22). P. 10313–10335.
13. Singh A. R., Patel M., Verma S., et al. A Deep Learning and IoT-Driven Framework for Real-Time Smart Grid Optimization // Scientific Reports. 2025. Vol. 15. P. 1–18.
14. Banad Y. M., Rahman M., Alharbi F., et al. Artificial Intelligence and Machine Learning for Smart Grids // Smart Energy. 2025. Vol. 18. P. 214–233.
15. Nikpour M., Rezaei H., Ahmadi A. Intelligent Energy Management in Smart Cities Using Intelligent Analysis // arXiv preprint. 2023. P. 1–16.
16. Tymchyshyn B., Manzhula V. Mathematical and Algorithmic Support of an Intelligent Software System for Proactive Urban Ecosystem Management // Measurement and Computing Devices in Technological Processes. 2025. No. 4. P. 35–48.
17. Manzhula V., Tymchyshyn B. Semantic Modeling of Heterogeneous Sources of Harmful Atmospheric Emissions for Proactive Urban Ecosystem Management // Scientific Papers of Donetsk National Technical University. Series: Informatics, Cybernetics and Computer Engineering. 2025. No. 2.
18. Zhang J., Tao D. Empowering Things with Intelligence: A Survey of the Progress, Challenges, and Opportunities in AIoT // arXiv preprint. 2020. P. 1–34.
19. Baccour E., Erbad A., Hamila R., Mohamed A. Pervasive AI for IoT Applications: A Survey on Resource-Efficient Distributed Artificial Intelligence // arXiv preprint. 2021. P. 1–29.
20. U.S. Department of Energy. AI for Energy: Opportunities for a Modern Grid and Clean Energy Economy. Washington: U.S. Department of Energy, 2024. 74 p.
21. Bisikalo, O., Kharchenko, V., Kovtun, V., Krak, I., Pavlov, S. Parameterization of the Stochastic Model for Evaluating Variable Small Data in the Shannon Entropy Basis, Entropy, 2023, 25(2), 184.

Дата надходження: 15.02.2026
Дата прийняття до друку після рецензування: 25.03.2026
Дата публікації: 18.06.2026

МЕТОДИ ТА СИСТЕМИ ОПТИКО-ЕЛЕКТРОННОЇ І ЦИФРОВОЇ ОБРОБКИ ЗОБРАЖЕНЬ ТА СИГНАЛІВ

МЕЛЬНИК АНДРІЙ МИКОЛАЙОВИЧ – д.т.н., професор, Західноукраїнський національний університет, Україна, ***e-mail: ame@wunu.edu.ua,*** **http://orcid.org/0000-0001-7799-9877**

ТИМЧИШИН ВАСИЛЬ СТЕПАНОВИЧ – доктор філософії, Західноукраїнський національний університет**, Україна,** ***e-mail: v.tymchyshyn@wunu.edu.ua,*** **https://orcid.org/0000-0001-8574-2561**

ПОПИК ЮРІЙ ІГОРОВИЧ – аспірант, Західноукраїнський національний університет, Україна,
e-mail: yup@wunu.edu.ua, **https://orcid.org/0009-0002-1690-0807**

ЗАБЧУК ВАДИМ ВОЛОДИМИРОВИЧ – доктор філософії, Західноукраїнський національний університет, **Україна,** ***e-mail: vadzab5@gmail.com,*** **https://orcid.org/0009-0008-0648-1446**

ФАТЮК ВАСИЛЬ ІГОРОВИЧ – аспірант, Західноукраїнський національний університет, Україна,
e-mail: vasyl.fatiuk@gmail.com, **https://orcid.org/0009-0001-4328-6796**

А.М. МЕЛЬНИК, В.С. ТИМЧИШИН, Ю.І. ПОПИК, В.В.ЗАБЧУК, В. І. ФАТЮК

ІНТЕЛЕКТУАЛЬНА АРХІТЕКТУРА ОБРОБКИ ДАНИХ ДЛЯ СКЛАДНИХ ІНФОРМАЦІЙНИХ СИСТЕМ: КЕЙС-ДОСЛІДЖЕННЯ ЕКОЛОГІЧНИХ ТА ЕНЕРГЕТИЧНИХ СИСТЕМ

Західноукраїнський національний університет, Україна