2. Cloud Security Alliance. Security Guidance for Critical Areas of Focus in Cloud Computing. URL: https://cloudsecurityalliance.org/ (accessed: 14.10.2026).

3. National Institute of Standards and Technology. Attribute Based Access Control (ABAC) Definition and Considerations. NIST SP 800-162. Gaithersburg: NIST, 2014. DOI: https://doi.org/10.6028/NIST.SP.800-162.

УДК 004.8:681.5

DATA MINING AND MACHINE LEARNING IN AUTOMATED PROCESS CONTROL SYSTEMS

Olena Polonska

Institute of Special Communications and Information Protection,
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Kyiv, Ukraine

Language advisor: Zhytska S. A., Senior Lecturer
National Technical University of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»

Modern critical infrastructure encompasses energy, water supply, chemical industry, and other strategically important sectors, the stable operation of which depends on the efficiency of automated process control systems (APCS). As the number of industrial devices, sensors, and monitoring systems grows, so does the volume of process data, creating a need for intelligent data processing methods. One promising avenue is the integration of Big Data, Data Mining, and Machine Learning technologies, which enable a shift from traditional monitoring to predicting deviations and enhancing process control efficiency[1]. This approach is underpinned by the Big Data concept, characterized by large data volumes (Volume), high rates of data ingestion and processing (Velocity), and a diversity of formats (Variety). In APCS environments, this data may include historical records of temperature, pressure, resource consumption, equipment operating modes, and industrial network events. Data reliability and the ability to derive practical value from analysis are also crucial, as errors, omissions, and information noise can compromise the accuracy of analytical models[2].

Data Mining or intelligent data analysis serves as a key tool for uncovering hidden patterns within vast datasets of processed information. Its application makes it possible to identify interrelationships between process parameters, detect anomalous equipment operating modes, pinpoint the causes of malfunctions, and identify opportunities for optimizing resource utilization[3]. The primary methods include classification, clustering, association analysis, and forecasting. Specifically, classification helps distinguish between normal and potentially critical states; clustering identifies typical equipment operating modes; and association analysis detects combinations of events linked to the occurrence of process deviations. Within APCS architecture, data mining can be used to analyze data accumulated in historical repositories (Industrial Historian) and transmit the resulting insights to analytical systems and decision-support tools. Unlike standard SCADA functions such as filtering, data aggregation, and trend visualization, intelligent analysis enables detection of complex relationships that are not always apparent during routine monitoring. This makes it an effective complement to existing automation infrastructure, without requiring the replacement of core components[4].

Machine learning is a key area of development in intelligent analysis; it enables the construction of models for forecasting and pattern recognition based on data. While data mining focuses primarily on discovering and interpreting relationships, machine learning allows training of models that apply these relationships to new data. For instance, supervised learning methods such as logistic regression and random forests can be used to classify equipment states[5], whereas unsupervised learning methods such as K-means can be employed to group operational modes and detect anomalous behavior. Combining data mining and machine learning creates opportunities to analyze both historical data and real-time information streams. Such historical data may include thermal, hydrodynamic, mechanical, and electrical parameters, as well as

event logs, discrete status signals, raw material characteristics, etc. For active technological processes, this approach enables detection of deviations from normal operating modes, forecasting of peak loads, and support of efficient resource allocation. Regarding the accumulated data analysis, these technologies help optimize computing resource usage, identify trends in operational metrics, and improve maintenance planning. Notably, data mining can be applied to streaming data and machine learning to historical datasets; thus, the distinction between them is defined not by the data type itself, but by the specific objectives and methods employed. It is possible to compare a system utilizing these two tools against one that is not (i.e., in their absence) based on performance evaluation criteria. The evaluation criteria are straightforward: anomaly detection accuracy, the number of false positives, detection time, and forecast accuracy. Operators rely on them to determine whether a particular tool or technology is useful based on specific priorities and whether the innovation will integrate with existing methodologies.

Each specific criterion is calculated in a different way. For instance, detection accuracy is calculated using a confusion matrix within a MATLAB environment via supervised learning[6]. This matrix comprises four metrics: True Positive (TP) - the correct tool; it detects something that actually exists. False Positive (FP) - a false alarm; the tool indicates detection, but nothing is there. True Negative (TN) - the correct tool; it ignores something that does not exist. False Negative (FN) - a miss; the tool detects nothing, even though an object is present. This metric is expressed as a percentage. The number of false triggers is calculated by counting instances where the tool signals by mistake the presence of a problem or event that does not actually exist; this is measured in absolute units. Detection time is derived by subtracting the time the deviation occurred from the time it was detected, and is therefore measured in units of time. Prediction accuracy is determined by comparing predictions against actual data, expressed in the units of the numerical parameter itself. From an industrial automation architecture perspective, these technologies can be integrated at the levels of data historians, analytics servers, SCADA, and MES, in accordance with ISA-95 principles. Their specific placement depends on requirements regarding processing speed, data availability, and production process safety. At the same time, solutions generated by intelligent models must be applied with due regard for the constraints of the industrial environment, while critical control functions must remain under the supervision of appropriate automation systems and established safety mechanisms.

This combined approach is currently being actively employed in modern Industrial Control Systems (ICS). It enhances operational safety within critical infrastructure facilities and represents a promising avenue for improving the efficiency and reliability of industrial systems. Combining accumulation and processing of large datasets, identification of hidden patterns, and forecasting of process deviations creates conditions for timely fault detection, efficient resource utilization, and improved decision-making. Practical implementation of this approach will facilitate transition to intelligent process control, providing safety and reliability requirements for critical infrastructure[7].

References

1. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd ed. Waltham: Morgan Kaufmann, 2012. URL: https://www.sciencedirect.com/book/monograph/9780123814791/data-mining-concepts-and-techniques (дата звернення: 14.10.2026).

2. Chang W. L., Grady N. NIST Big Data Interoperability Framework: Volume 1, Big Data Definitions. Version 2. Gaithersburg : National Institute of Standards and Technology, 2018. (NIST Special Publication 1500-1). URL: https://doi.org/10.6028/NIST.SP.1500-1r2 (дата звернення: 14.10.2026).

3. Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer C., Wirth R. CRISP-DM 1.0: Step-by-Step Data Mining Guide. 2000. URL: https://www.dataprix.com/files/CRISP-DM.pdf (дата звернення: 14.10.2026).

4. International Society of Automation. ISA-95 Series of Standards: Enterprise-Control System Integration. URL: https://www.isa.org/standards-and-publications/isa-95-standard (дата звернення: 14.10.2026).

5. Scikit-learn developers. Ensemble methods: Random forests and other randomized tree ensembles. URL: https://scikit-learn.org/stable/modules/ensemble.html (дата звернення: 14.10.2026).

6. Scikit-learn developers. *scikit-learn: Machine Learning in Python. User Guide*. URL: https://scikit-learn.org/stable/user_guide.html (дата звернення: 14.10.2026).

7. National Institute of Standards and Technology. Guide to Industrial Control Systems (ICS) Security. NIST Special Publication 800-82 Rev. 3. Gaithersburg: NIST, 2023. URL: https://doi.org/10.6028/NIST.SP.800-82r3 (дата звернення: 14.10.2026).

УДК 004.85:004.89

РОЗРОБКА ІНФОРМАЦІЙНОЇ СИСТЕМИ АНАЛІЗУ ДОСТОВІРНОСТІ ТА ПОЯСНЮВАНОСТІ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

Даніїл Чулов

Миколаївський національний аграрний університет, м. Миколаїв, Україна

Науковий керівник: В. В. Співак,
асистент Миколаївський національний аграрний університет

Інтенсивне впровадження методів штучного інтелекту та машинного навчання (МН) у системи підтримки прийняття рішень в інженерії, медицині, енергетиці та агропромисловому виробництві зумовлює необхідність суворого контролю надійності автоматизованих прогнозів [1; 2]. Сучасні нелінійні моделі, зокрема глибокі нейронні мережі та ансамблі дерев прийняття рішень, демонструють високу точність класифікації, проте функціонують за принципом «чорної скриньки» та характеризуються систематичною гіпервпевненістю (overconfidence), коли прогнозовані оцінки ймовірностей суттєво перевищують емпіричну частоту правильних відповідей [3]. Існуючі програмні засоби зазвичай розглядають задачу пояснюваності (Explainable AI – XAI) ізольовано від задачі квантифікації невизначеності та калібрування імовірностей. Зокрема, методи SHAP [4] і LIME [5] призначені для атрибуції ознак і не оцінюють надійності самих імовірностей. Як наслідок, інтерпретація невідкаліброваного прогнозу за допомогою методів важливості ознак формує у кінцевого користувача хибну довіру до потенційно помилкового рішення. Отже, актуальним науково-прикладним завданням є розробка комплексної інформаційної системи (ІС), яка інтегрує методи оцінювання статистичної достовірності, пост-обробного калібрування та адитивної пояснюваності моделей МН у єдиному програмному контурі.

Мета роботи – розробити таку ІС та оцінити її ефективність на тестових наборах даних класифікації за показниками калібрування, покриття конформних множин і діагностики ознак, що спричиняють невизначеність.

Архітектура запропонованої інформаційної системи (рис. 1) базується на чотирьох функціонально зв’язаних підсистемах: модулі інтеграції моделей та наборів даних, модулі діагностики калібрування та рекалібрування ймовірностей, модулі конформного прогнозування та модулі адитивної інтерпретації рішень XAI. Робочий цикл ІС: розбиття даних на калібрувальну, верифікаційну та тестову частини; обчислення ECE, MCE, EntCE; за потреби – рекалібрування; побудова конформних множин і аудит покриття за підгрупами; обчислення SHAP-атрибуцій для об’єктів із розширеною множиною.