ISSN 2411-3336; e-ISSN 2541-9404

JOURNAL OF MINING INSTITUTE

Zapiski Gornogo instituta

Journal homepage: pmi.spmi.ru

Research article *Geotechnical Engineering and Engineering Geology*

Centrifugal pump and electrical motor fault detection with motor current signature analysis and automated machine learning

Roman R. Khalikov[1], **Mikhail Yu. Chernetskiy**[1], **Ilia E. Revin**[2], **Vadim A. Potemkin**[2]✉

[1] *ROTEC Digital Solutions JSC, Moscow, Russia*

[2] *ITMO University, Saint Petersburg, Russia*

***How to cite this article*:** Khalikov R.R., Chernetskiy M.Yu., Revin I.E., Potemkin V.A. Centrifugal pump and electrical motor fault detection with motor current signature analysis and automated machine learning. Journal of Mining Institute. 2025. Vol. 275, p. 42-55.

Abstract

Centrifugal pumps, as key components of hydraulic systems, play a fundamental role in ensuring the reliable operation of numerous industrial processes in sectors such as energy, chemical production, and oil refining, where uninterrupted equipment performance is of critical importance. Failures of centrifugal pumps can result in substantial financial losses due to costly repairs and unplanned production downtime. This paper presents an innovative approach to diagnosing and detecting faults in centrifugal pumps. The method is based on the application of Motor Current Signature Analysis (MCSA) in combination with automated machine learning (AutoML) technologies. Such an approach enables efficient and highly accurate identification of early signs of equipment malfunction. The experimental study was conducted using an open dataset collected under conditions close to real-world operation. The proposed method achieved a fault detection accuracy of 89 %, which significantly exceeds the performance of the traditional gradient boosting method. This confirms the advantage of a comprehensive model developed through AutoML. Further improvement in diagnostic accuracy was made possible by applying an enhanced Park's vector transformation to the raw current and voltage data. This approach makes it possible to detect even subtle anomalies in pump operation, thereby strengthening the capability for early fault prediction. The study not only highlights the potential of MCSA as a non-invasive and scalable tool for equipment condition monitoring but also demonstrates the promise of AutoML for technical diagnostics of industrial pumps.

Keywords

machine learning; electric motor; gradient boosting; composite model; AutoML; fault detection; time series

Received: 09.04.2025 ***Accepted: 25.08.2025*** ***Online: 13.10.2025*** ***Published: 31.10.2025***

Introduction

The global pump market reached a volume of USD 59.2 billion in 2023*, with centrifugal pumps representing its largest segment. These pumps are critically important for combined-cycle power plants, coal-fired thermal power stations, nuclear reactors, chemical facilities, and other industries [1, 2]. According to data from the European Association of Pump Manufacturers, an oil refinery with a capacity of 300,000 barrels per day may operate up to 650 pumps, each of which requires thorough monitoring to prevent failures. Without intelligent systems capable of automatically analyzing sensor data and detecting anomalies [3], timely monitoring of such a fleet of equipment becomes impossible.

* Global Pump Market Outlook. Spring 2024. Sample Material. URL: https://www.oxfordeconomics.com/resource/global-pump-market-outlook-2024/ (accessed 10.04.2025).

One of the promising methods for pump condition monitoring is Motor Current Signature Analysis (MCSA). This approach enables the assessment of equipment condition based on the analysis of consumed current using low-cost sensors and well-established signal processing techniques. MCSA is particularly effective for centrifugal pumps, as it allows the detection of problems at an early stage without direct access to the pump, in contrast to vibration, acoustic or pressure analysis methods.

Despite the potential of MCSA, its manual or semi-automated application requires significant effort from experts and developers, which hinders scalability across a large number of units. A possible solution lies in diagnostic systems based on machine learning (ML), which have been investigated for more than 15 years. Most MCSA-related studies focus on motor diagnostics – such as rotor bar breakage [4, 5], stator winding faults [6], and bearing defects [7-9]. Typical approaches include feature extraction in the time, frequency, and time-frequency domains [10-12], demodulation transformations [13, 14] and their combinations with various ML models. However, research dedicated specifically to pumps is very limited due to the lack of high-quality data. The majority of studies rely on synthetic datasets [8] or laboratory test rigs [15, 16], which constrains the applicability of their findings under real operating conditions. An exception is the study by C.E.Sunal et al. [9], which employed industrial data from centrifugal pumps. Nevertheless, its technical implementation details are insufficiently disclosed.

Classical machine learning methods and deep learning (DL) approaches are promising [17], yet they are constrained by the limited availability of labeled data. In practice, traditional ML is capable of solving tasks under conditions of data scarcity. However, in such cases the creation of an informative feature space becomes critically important. An effective method for generating additional informative signals for equipment analysis and feature construction remains the use of the Extended Park's Vector Approach, which mitigates the influence of supply frequency and highlights fault-related features [18]. The development of automated machine learning (AutoML) further opens new opportunities. AutoML simplifies data preprocessing, model construction, and hyperparameter tuning, enabling efficient utilization of data even in the presence of class imbalance [19-21].

The objective of this study is to develop a methodology for fault detection in pump units with electric motors based on current signature analysis under conditions of limited data availability. To this end, the research addressed the task of identifying an optimal combined approach that integrates MCSA techniques, the generation of an informative feature space, and AutoML for early-stage fault classification of centrifugal pumps and electric motors using an open dataset simulating real operating conditions. The study considers a scenario in which a fault with similar signal manifestations may occur either in the motor or in the pump. A series of experiments evaluated the impact of the Extended Park's Vector Approach on diagnostic accuracy. The results demonstrated that AutoML is capable of automatically generating models with accuracy exceeding 89 %, outperforming optimized gradient boosting. This confirms the potential of AutoML for industrial diagnostic applications.

Related works

Motor current signature analysis of pumps and electric motors. Several approaches have been proposed for anomaly detection and fault classification of pumps and motors using MCSA [22]. One of the more recent methods, presented by Y.Han et al. [15], focuses on unsupervised anomaly detection through a comprehensive framework combining MCSA, the Extended Park's Vector Approach, a CNN-LSTM attention model, and spectral analysis. The model attempts to reconstruct the instantaneous current amplitude by leveraging the original phase signals along with additional voltage features. The authors conducted an in-depth analysis and manually identified three levels of cavitation as well as the number of damaged impeller blades. Decision-making was based on a statistically computed threshold in the frequency domain and on the residual difference between measured and predicted values. All tasks were performed under variable flow conditions.

C.E.Sunal et al. [9] applied a classical approach by visualizing the components of the Park's vector. They achieved an accuracy ranging from 85.5 to 100 % (depending on the signal sampling frequency) in the image classification task using a fine-tuned ResNet-34 model. The visual representation addressed the data imbalance problem by effectively augmenting the existing dataset. The model was trained and validated at frequencies of 1500 and 3000 Hz, while the test data included 1500, 3000, and 4500 Hz. The data were collected from pumps operating under different working conditions. This approach enabled the supervised detection of anomalies by identifying defective pumps. A detailed review by the same authors [8] demonstrates that deep learning models are capable of solving motor and pump fault detection tasks with high accuracy. Only a small fraction of classical machine learning models achieve accuracy above 90 %.

However, these studies lack information regarding the severity of faults. Signals of advanced defects can be identified visually through frequency analysis, yet experts may encounter difficulties in detecting weak anomalies. Faults may be masked within the sidebands of the carrier frequency or bearing frequencies, remaining unnoticed after demodulation due to the low spectral power of defect harmonics and high noise levels. Machine learning methods allow for the prediction of emerging faults, but the limited volume of available data can critically affect deep learning models, leading to overfitting. An additional challenge arises when the operating conditions of the pump or motor change, necessitating the collection of new data.

Application of AutoML in fault detection. Initially, AutoML was focused on automating typical tasks of ML engineers, thereby making machine learning more accessible to domain experts. However, the evolution of the field has demonstrated that AutoML can outperform humans in designing model architectures for both classical ML [23, 24], and deep learning [25]. A variety of AutoML frameworks exist, but to the best of the authors' knowledge, only two of them – FEDOT and TPOT [26]) – are capable not only of automating model search and hyperparameter optimization, but also of generating composite models using genetic algorithms to improve performance. A composite model is a special type of ensemble, resembling stacking. It can be represented as a graph, where each node corresponds to a model or a data processing method. By combining nodes and connections, it is possible to obtain an optimal structure tailored to the specific dataset.

Previous studies on fault diagnosis using AutoML demonstrate that such approaches enable the development of complex models with higher accuracy than individual algorithms. J.Zhang et al. [12] showed that TPOT outperformed SVM and XGBoost, achieving accuracy between 95.8 and 99.3 % under different signal-to-noise ratios in vibration signals. Using wavelet decomposition, A.S.Maliuk et al. [27] successfully classified three types of bearing faults (inner race, outer race, and ball) as well as the normal condition. Similarly, R.H.Hadi et al. [28] reported that the AutoML framework PyCaret achieved 95.6 % accuracy on new data without employing composite models. M.Cerrada et al. [29] compared the results of TPOT and H2O [30] in the task of classifying the severity of three types of gear faults. Both frameworks demonstrated comparable accuracy (exceeding 96 %) across all scenarios, on par with other ML and DL methods. The authors noted that the key features used by the models were consistent, despite differences in their architectures.

The main advantage of FEDOT over TPOT and other frameworks is its ability to generate models with high variability. High variability of composite models refers to the capability of constructing a directed acyclic graph, where the nodes represent data preprocessing operations or trained models, and the edges define the flow of results between nodes. FEDOT allows the connection of preceding nodes to any subsequent nodes. In this way, FEDOT generalizes the approach to composite model construction implemented in TPOT. Moreover, FEDOT supports diverse data types and tasks, offering a flexible interface for interaction. Studies employing FEDOT for motor diagnostics are very

limited. In the work of I.Revin et al. [31] FEDOT was tested on time series classification tasks using public datasets. In 90 % of cases, the performance of FEDOT's composite models exceeded or approached that of state-of-the-art algorithms, confirming their applicability to tasks analogous to fault detection.

Methods

Experimental data. Publicly available datasets for analyzing motor current and voltage, particularly for motors coupled with pumps, are limited. Most existing datasets cover only a narrow range of fault types, motor power ratings, and operating modes. However, S.Bruinsma et al. [7] introduced a dataset that partially addresses these issues. The dataset includes three-phase currents and voltages recorded after a frequency converter for induction motors rated at 11 and 22 kW, as well as vibration signals measured at five points on the motor and pump. The data were collected at a sampling frequency of 20 kHz. Electrical signals were recorded for 15 s with a 2 min interval between measurements. The signals were not processed in any way; the raw 24-bit data were stored directly in CSV files. In total, 20 fault types were recorded, each with three severity levels. Data for healthy states were measured 94 times, while the number of records for other faults varied from 3 to 10. Based on this, the dataset is unbalanced, with the proportion of healthy states exceeding faulty ones by approximately 5:1. Data labeling was performed automatically at the time of acquisition, with each fault recorded separately. The dataset was not validated by independent experts, but it was collected at different motor speeds, which increases its applicability to real industrial conditions. Despite the authors' efforts, this dataset has rarely been used in studies on current-based diagnostics of pumps.

For the experiments, the minimum severity level of faults was selected, as the primary objective was early detection. Measurements were carried out at 100 % motor speed. The task was formulated as a multiclass classification problem with seven classes: healthy state, motor bearing faults (inner race, outer race, ball), pump bearing fault (simultaneous inner and outer race), and looseness of motor and pump mountings. The data were pre-analyzed to examine signal properties before and after pre-processing, as well as to identify outliers. The seven selected classes represent the most challenging cases for defect analysis. The faults in the dataset were artificially induced. Bearing race damages were introduced from the rolling element side using a milling cutter, resulting in grooves spanning the entire width of the race with a length of 1 mm and a depth of 350 μm. The degree of damage varied depending on the number of such scratches. The rolling elements themselves were damaged with a rotary engraver across the entire ball surface, producing scratch-like defects over the ball surfaces [7]. Looseness of the mountings was simulated by reducing the tightness of already fastened bolts.

Data preprocessing and feature extraction. The original dataset contains files for each phase and signal type, where the columns correspond to 15-second measurements. The files were transformed into an array of size $N_{\text{samp}} \times N_{\text{chan}}$, where $N_{\text{chan}} = 6$ (three current phases and three voltage phases). The array includes all available data for both healthy and faulty states, with corresponding class labels. Thirty percent of the measurements for each state were reserved for the test set.

In part of the experiments, additional data channels were introduced using the Extended Park's Vector Approach (up to six channels). This transformation enhances the diagnostics of three-phase electrical machines by minimizing the influence of the fundamental supply frequency (50 Hz) and highlighting fault-related features. For example, the extended Park's transformation for current is defined as a linear combination of the three phases, which simplifies the analysis of asymmetry and harmonics:

$$\begin{cases} i_a = \left(2i_u - i_v - i_w\right)/3; \\ i_b = \left(i_u - i_w\right)/\sqrt{3}, \end{cases}$$

where i_u, i_v, i_w correspond to the three current phases, and i_a, i_b represent the real and imaginary components of the Park's vector.

The linear transformation makes it possible to obtain up to four data channels (two for current and two for voltage). These vectors can also be used to generate up to four additional channels by computing the squared modulus of the Park's vector:

$$i_{\text{inst}} = \left| i_a + j i_b \right|,$$

where i_{inst} denotes the instantaneous amplitude.

This transformation removes the fundamental 50 Hz component from the signal. After generating the additional channels, the sample array was divided into time windows of 1 s (20,000 samples per window) and converted into a three-dimensional tensor. The tensor dimensions corresponded to the number of windows, window size, and number of data channels. For each window, feature extraction was performed according to standard MCSA practices [32]. Features in the time and frequency domains were calculated, including mean value, variance, root mean square (RMS), crest factor, kurtosis, and skewness.

The energy of the approximation coefficients from the first, second, and third levels of the wavelet packet decomposition was added in the time-frequency domain. In total, 28 features were calculated for each data channel, resulting in up to 224 features. The complete list of features is presented in Table 1. At the final stage, min-max scaling was applied to both the training and test datasets.

Table 1

Generated functions

Time domain		Frequency domain	
Mean value	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	Mean value	$\overline{F(x)} = \frac{1}{n}\sum_{i=1}^{n} F_i(x)$
Variance	$\sigma^2_{\text{time}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$	Variance	$\sigma^2_{\text{freq}} = \frac{1}{n}\sum_{i=1}^{n}(F_i(x) - \overline{F(x)})^2$
RMS	$RMS_{\text{time}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$	RMS	$RMS_{\text{freq}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} F_i(x)^2}$
Peak value	$PV_{\text{time}} = \max(\lvert x \rvert)$	Peak value	$PV_{\text{freq}} = \max\left(\left\lvert F(x) \right\rvert\right)$
Crest-factor	$CF_{\text{time}} = \frac{PV_{\text{time}}}{RMS_{\text{time}}}$	Crest-factor	$CF_{\text{freq}} = \frac{PvT_{\text{freq}}}{RMS_{\text{freq}}}$
Kurtosis	$Kurt_{\text{time}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n(\sigma^2_{\text{time}})^2} - 3$	Kurtosis	$Kurt_{\text{freq}} = \frac{\sum_{i=1}^{n}(F_i - \overline{F(x)})^4}{n(\sigma^2_{\text{freq}})^2} - 3$
Skewness	$Sk_{\text{time}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{3}{2}}}$	Skewness	$Sk_{\text{freq}} = \frac{\sum_{i=1}^{n}(F_i(x) - \overline{F(x)})^3}{n\left(\frac{1}{n}\sum_{i=1}^{n}(F_i(x) - \overline{F(x)})^2\right)^{\frac{3}{2}}}$
Clearance factor	$Clf = \frac{PV_{\text{time}}}{\left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\lvert x_i \rvert}\right)^2}$	FFT peak value frequency	$PF = \arg\max\left(\left\lvert F(x) \right\rvert\right)$
Line integral	$LI = \sum_{i=1}^{n}\lvert x_{i+1} - x_i \rvert$	Spectrum energy	$E = \sum_{i=1}^{n} F_i(x)^2$
Impulse factor	$IF = \frac{PV_{\text{time}}}{\frac{1}{n}\sum_{i=1}^{n}\lvert x_i \rvert}$	Time and frequency domain	
Shape factor	$SF = \frac{RMS_{\text{time}}}{\frac{1}{n}\sum_{i=1}^{n}\lvert x_i \rvert}$	Wavelet packet decomposition detail coefficient energy	$cD_i = \sum_{j=1}^{n_i} D_{i,j},\ i = 1, 2, 3$
Peak-to-peak value	$PtP = \max x_i - \min x_i$	Wavelet packet decomposition approximate coefficient energy	$cA_i = \sum_{j=1}^{n_i} A_{i,j},\ i = 1, 2, 3$
Shannon's entropy	$H = -\sum_{i1}^{n} x_i^2 \log\left(x_i^2\right)$		

As part of the study, four datasets were constructed to investigate the impact of additional channels on the effectiveness of fault diagnostics using MCSA. The first dataset (NoPark) contained only the preprocessed raw current and voltage data and served as the baseline for evaluating classification accuracy improvements obtained by adding Park's vector components. The second dataset consisted solely of instantaneous current and voltage vectors. This dataset was used to assess the demodulation effect of the Extended Park's Vector Approach and to compare the expressiveness of the data against unprocessed current and voltage signals. The third dataset combined the previous two and served as the primary dataset for evaluating the effectiveness of the proposed algorithms, as it contained the largest amount of signal information. The fourth dataset included Park's vector components as well as the instantaneous amplitude. This dataset was used to evaluate the amount of fault-related information contained in the Park's vector components (each component representing a linear combination of three phases) and to reduce the dimensionality of the feature space. Table 2 summarizes the characteristics of each dataset used in the experiments.

Table 2

Datasets

Experiment	Data	Training dataset size (number of samples × number of features)
NoPark	Three current and voltage phases	1680 × 168
OnlyInstAmp	Instantaneous current and voltage amplitude	1680 × 56
WithInstAmp	Three phases and instantaneous current and voltage amplitude	1680 × 224
FullPark	Park's vectors and instantaneous amplitude	1680 × 168

Classification algorithms. Two approaches were used for fault classification. During preliminary training, the gradient boosting algorithm demonstrated superior performance compared to other classical machine learning methods. This result is consistent with the general consensus that gradient boosting is among the most effective techniques for such tasks. Thus, the first approach involved the use of gradient boosting on decision trees implemented through the CatBoost library [33]. The CatBoost classifier was trained using hyperparameters optimized with the Optuna tuning framework [23].

For hyperparameter sampling in Optuna, the Tree-structured Parzen Estimator algorithm [24, 34] was applied with 10 initial trials. As the adaptive pruning algorithm (pruner), Hyperband [34] was used with a minimum resource allocation of 50, a maximum of 400, a reduction factor of 2, and 10 bootstrap samples. Thanks to Optuna's multithreading capabilities, hyperparameter optimization was significantly accelerated, with approximately 3000 configurations tested in each experimental scenario. The maximum number of training iterations for CatBoost was set to 400.

During the optimization process, Optuna tuned the following hyperparameters: loss function – MultiClass (multiclass) or MultiClassOneVsAll (multiclass OneVsAll); learning rate – from 0.01 to 1 with a step of 0.001; L2 regularization on leaves – from 1 to 200 with a step of 1; boosting type – Ordered or Plain; class weight calculation formula – Balanced or SqrtBalanced; tree depth – from 1 to 10 with a step of 1; and feature fraction (percentage of random subspace) – from 0.01 to 1 with a step of 0.01.

In the second approach, the open-source AutoML platform FEDOT [20] was applied to solve the same classification task. The FEDOT platform is capable of automatically generating, optimizing, and tuning composite models using a directed acyclic graph representation and

genetic algorithms. In this study, the best quality preset was used, which allows the inclusion of any available machine learning models in the construction of the composite model.

The early stopping criterion was set to 10 generations without improvement, and the maximum number of generations was limited to 100. During hyperparameter tuning with Optuna, the objective was to maximize the macro-averaged F1-score on the test dataset. In the case of composite model construction, the FEDOT framework autonomously attempted to maximize the macro-averaged F1-score using five-fold cross-validation. To minimize the effect of class imbalance, macro-averaging was applied. Since the FEDOT framework does not support macro-averaging natively, a custom metric was implemented.

Results

Exploratory data analysis. The original recorded signal is shown in Fig.1. The waveform deviates substantially from an ideal sinusoid due to the use of a frequency converter, with the largest distortions observed in the voltage. Analysis of the signal distributions revealed no significant differences between the raw data for healthy and faulty states. However, in the healthy state data (healthy 1 at 100 % speed, 15th measurement), a motor shutdown process was detected, which was excluded as an outlier.

The Extended Park's Vector Approach enabled an alternative representation of the data. The effect of the transformation on the signal spectrum is shown in Fig.2. As severity level 3, a motor signal with a damaged bearing outer race was used. Demodulation redistributes the spectral amplitude to other frequencies, thereby facilitating fault detection. Park's transformation suppresses the supply carrier frequency while amplifying other significant harmonics. This alters the distribution of the signal when comparing healthy and faulty states, which in turn affects the distribution of the extracted features.

A visual analysis of the Park's vector components was conducted. Figure 3 illustrates the difference between a bearing outer race fault (BPFO) at severity levels 1 and 3 at 100 % motor speed. The main distinction between the healthy state and both faulty states is driven by the larger volume of healthy state data. However, it is evident that the pattern at severity level 3 is more distorted and asymmetric compared to severity level 1, which remains closer to the healthy state. This indirectly indicates the difficulty of detecting faults at early stages, even when applying machine learning methods.

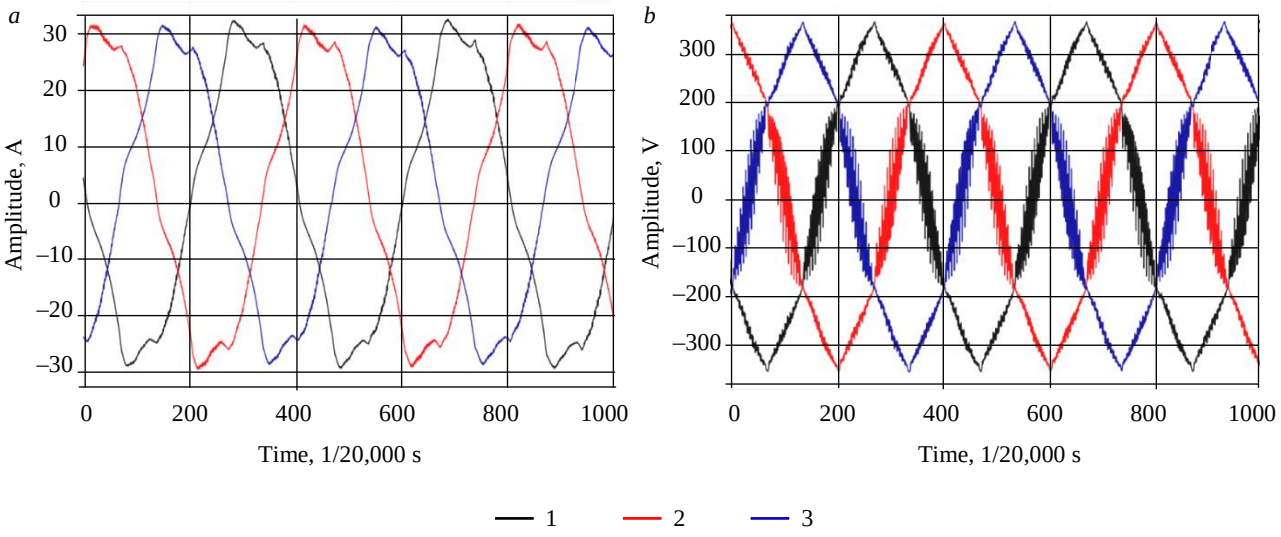

Fig.1. Original current signal (*a*) and voltage signal (*b*)
1, 2, 3 – phases

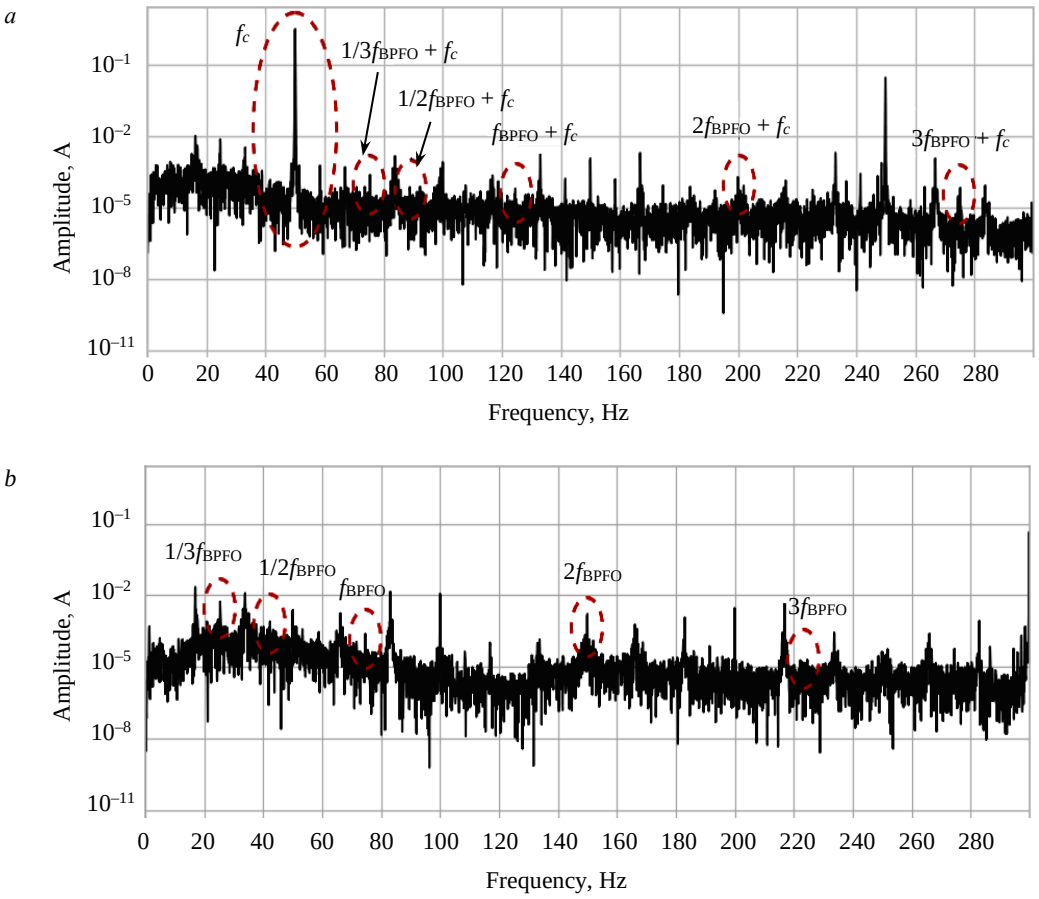

Fig.2. Comparison of the spectra of the original current signal and the instantaneous amplitude:
a – spectrum of the instantaneous current phase; *b* – spectrum of the generalized current after Park's transformation

f_c – carrier frequency 50 Hz; $f_{BPFO} \approx 74{,}7$ Hz depending on the bearing and motor rotational frequency

Analysis of feature distributions using boxplots confirmed that the proposed features capture class differences and can improve classification performance. Figure 4 presents three informative features: one generated from the raw voltage data; one obtained from the spectrum of the α-component of Park's transformation; and one computed through wavelet packet decomposition of the instantaneous current. Each of these features can be effectively used for training ML models.

Analysis of AutoML results. A time budget of 30 min was allocated for generating composite machine learning models, and 5 min for hyperparameter tuning. The experiments were carried out in three stages:

• Baseline: a combination of a quantile feature generator and a random forest model.

• Comparison with SOTA: the best result of the proposed approach was compared with the performance of other time series classification models using the F1-score metric.

• Ensembling: features from the top-performing models were combined into a single matrix, after which the model selection process was repeated.

Figure 5 visualizes the process of model composition. An improvement in model quality is observed with each generation, followed by metric stabilization, which confirms FEDOT's ability to converge to an optimal solution within a reasonable time. Analysis of the diversity within the model population showed that, at the beginning, the distribution of metrics was wide, reflecting variability in performance. At later stages, the distribution narrows as the target metric increases and the standard deviation decreases. Evolutionary optimization thus ensures both high model quality and the preservation of diversity, even in the final generations.

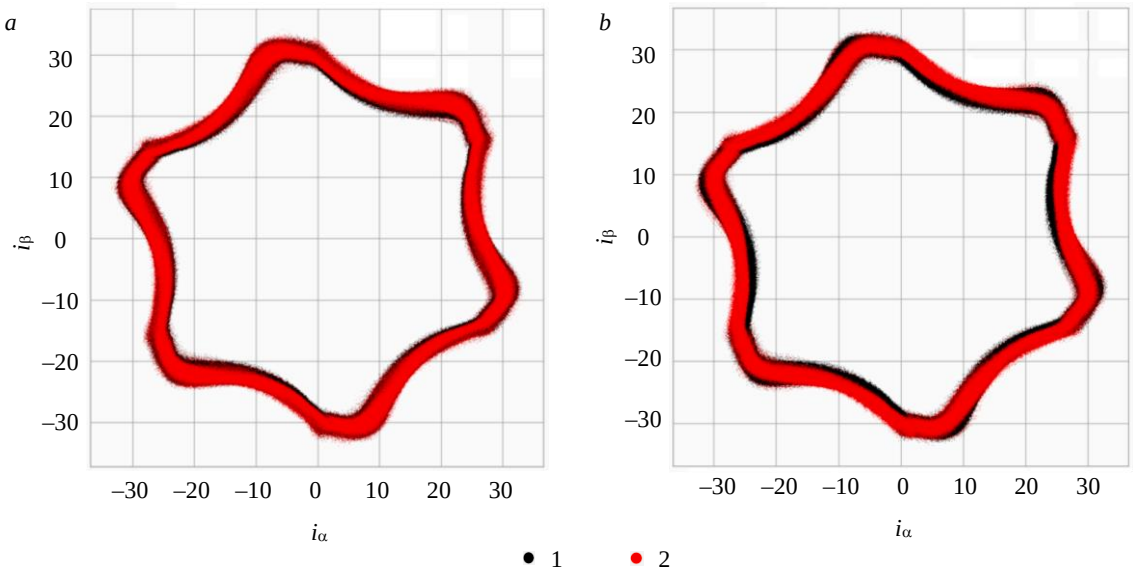

Fig.3. Comparison of the structure of Park's vector components for severity level 1 (*a*), severity level 3 (*b*), and the fault-free state

1 – components of Park's transformation unrelated to faults, identical in both plots;
2 – fault-related components that vary with the severity level in each case;
i_α and i_β are real and imaginary components of Park's vector

Advantages of the approach over traditional gradient boosting:

• Flexible model composition – examples of the generated pipelines (Fig.6) demonstrate the framework's ability to combine heterogeneous preprocessing methods and algorithms.

• Automation of labor-intensive stages – automatic feature selection, model choice, and hyperparameter tuning significantly reduce development time.

• Integration of AutoML and MCSA – the results confirm that combining AutoML with MCSA enables the development of effective diagnostic systems for industrial equipment.

Model evaluation. The macro F1-scores for both algorithms, measured on the test dataset, are presented in Table 3. The CatBoost algorithm achieved a maximum macro F1-score of 0.68 when trained on data without Park's transformation. Models trained on other datasets showed slightly lower performance with nearly comparable scores. The FEDOT framework demonstrated superior results, achieving a macro F1-score above 0.89. The best-performing model was trained on data combining raw and instantaneous amplitude signals. The composite model begins with a resampling node, which does not apply any transformations in the case of a multiclass task. The second node applies the FastICA algorithm with unit variance whitening, without dimensionality reduction, and passes the result to the next node. The third node is a logistic regression model with an inverse regularization strength of C = 5.88. The normalization node applies min-max scaling to the probabilities produced by the logistic regression. The final node performs quadratic discriminant analysis (QDA) on the normalized probabilities. The resulting pipeline is simple to understand and interpret; it confirms that the data are well-prepared and provides all the necessary information on the faults. Even in cases where pump and motor faults produce similar spectral characteristics, the algorithm clearly distinguishes them. A comparison of the confusion matrices for the best CatBoost and composite models is shown in Fig.7.

Table 3

F1-score for the CatBoost algorithm and the composite model

Experiment	CatBoost	Composite model	Experiment	CatBoost	Composite model
NoPark	0.68	0.76	WithInstAmp	0.65	0.89
OnlyInstAmp	0.66	0.65	FullPark	0.64	0.86

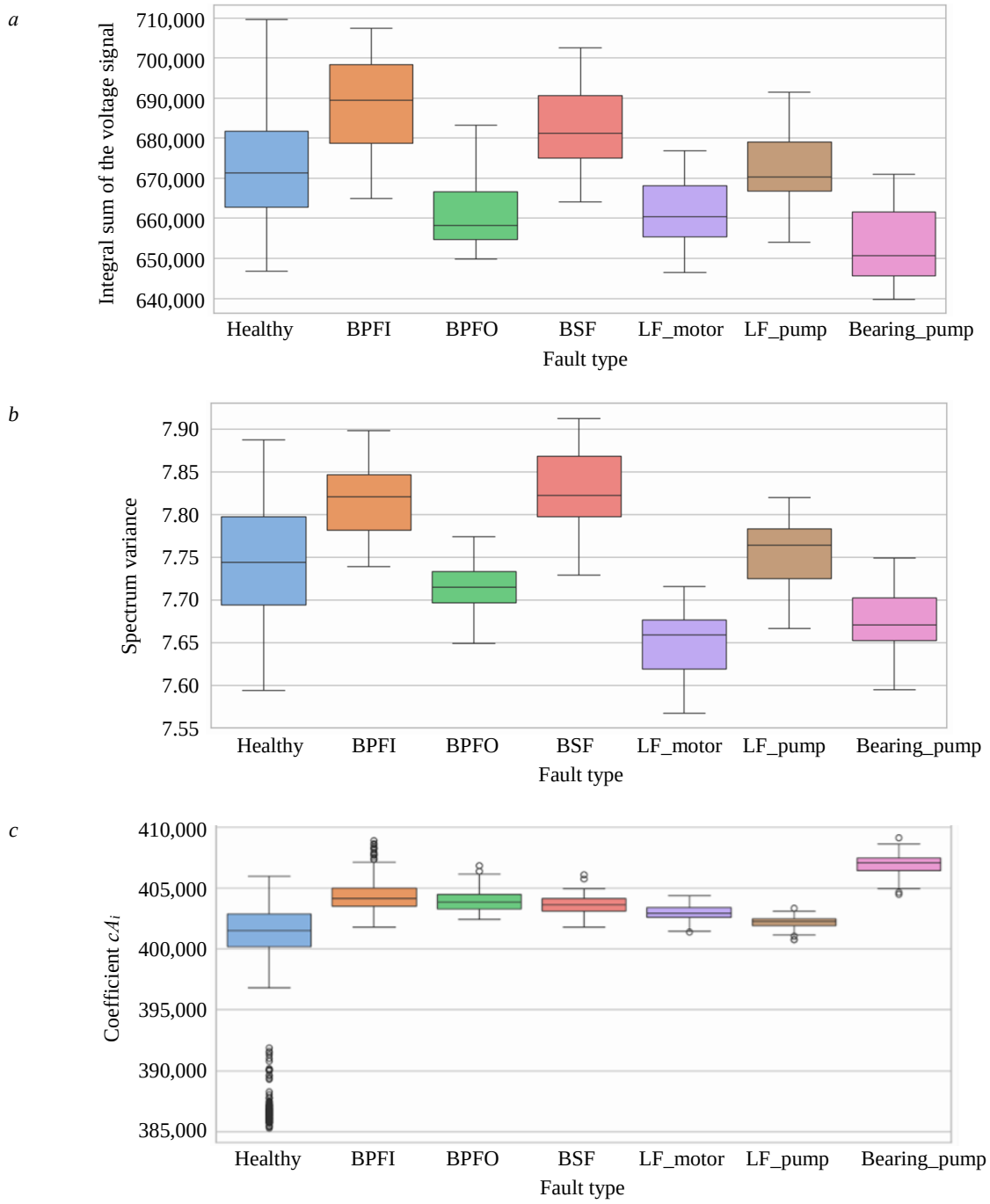

Fig.4. Example distributions of extracted features: *a* – first voltage phase, linear integral; *b* – computed spectrum variance; *c* – instantaneous current amplitude, coefficient *cAi* (energy of approximation coefficients from the wavelet packet decomposition)

Healthy – absence of any faults; BPFI, BPFO, and BSF – faults associated with inner race, outer race, and ball defects of the motor bearing; LF_motor and LF_pump – looseness in motor and pump mountings; Bearing_pump – pump bearing fault with damage to both races

The models experience the greatest difficulty in accurately classifying outer and inner race faults. Both models exhibit confusion between inner race defects and rolling element defects. This can be explained by the fact that the induced faults represent an early stage of defect development and are weakly expressed in the signals, while the extracted features showed similar distributions for these confused fault types. For the composite model, inner race faults proved to be the most challenging to recognize. Similar behavior was observed during the preliminary selection of the best classical ML

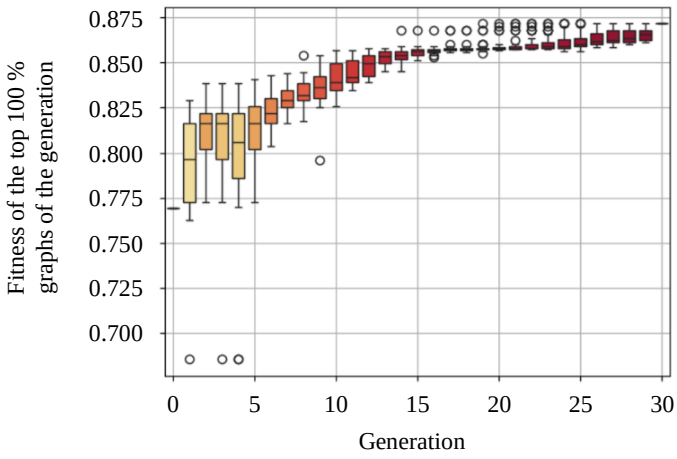

Fig.5. Evolution of model composition performance

model, indicating the general difficulty of identifying this type of defect. The gradient boosting algorithm was able to clearly identify the inner race fault, but encountered problems with other fault types, which may suggest the existence of a trade-off between the decision boundaries separating the classes. It is worth noting that both classifiers did not confuse most of the identical faults of the motor and pump, which indicates a clear separation of these defects in terms of the generated features. The composite model also demonstrated the lowest false positive rate in detecting anomalies associated with healthy state labeling.

Discussions and limitations

The application of the CatBoost algorithm with various combinations of additional features did not achieve better results in detecting such pump unit faults as motor bearing defects (inner race, outer race, ball), pump bearing defect (simultaneous inner and outer race), and looseness of motor and pump mountings, compared to the approach proposed in this paper. The composite model demonstrated superior performance when using features generated from instantaneous current and voltage amplitudes. The results of the composite model confirm the preliminary analysis and clearly show that the Extended Park's Vector increases the amount of useful information. The study confirmed that this approach should be applied in combination with raw data to achieve the best results.

An important outcome of the study is the demonstration that the classifiers were able to distinguish between similar faults in the electric motor and the pump. The obtained results suggest that the proposed model can be trained to detect anomalies with a low false positive rate, even when using data corresponding to low fault severity levels.

It should be noted that several practical issues remain unresolved within the scope of this study and require further investigation. In this work, we used the maximum available motor speed, whereas in real applications the motor may operate under varying conditions. Variable frequency drives are often employed to regulate motor speed, and such drives were present during the recording of the dataset used. Variable frequency drives alter the amplitude and frequency of the motor supply, which fundamentally changes the recorded signals and the extracted features – resulting in data drift. The same effect would occur when replacing the motor or pump model. Due to data drift, the model's performance may deteriorate, and therefore it must be retrained, and in some cases modified, to maintain its accuracy. The modification of pretrained models is possible within the FEDOT framework, and the fine-tuning of such models is simpler and faster compared to deep learning models.

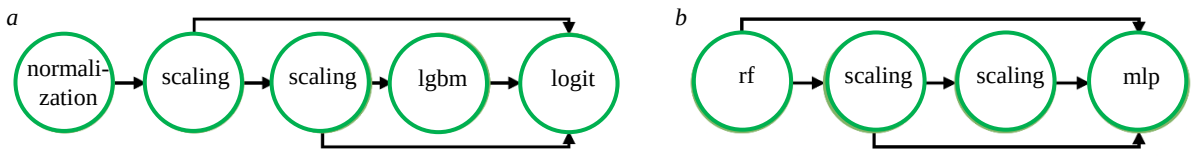

Fig.6. Examples of composite pipelines generated by AutoML: *a* – linear pipeline with boosting; *b* – linear pipeline with perceptron

Computational nodes: normalization; scaling; lgbm – light gradient boosting machine; logit – logistic regression; rf – random forest; mlp – multilayer perceptron

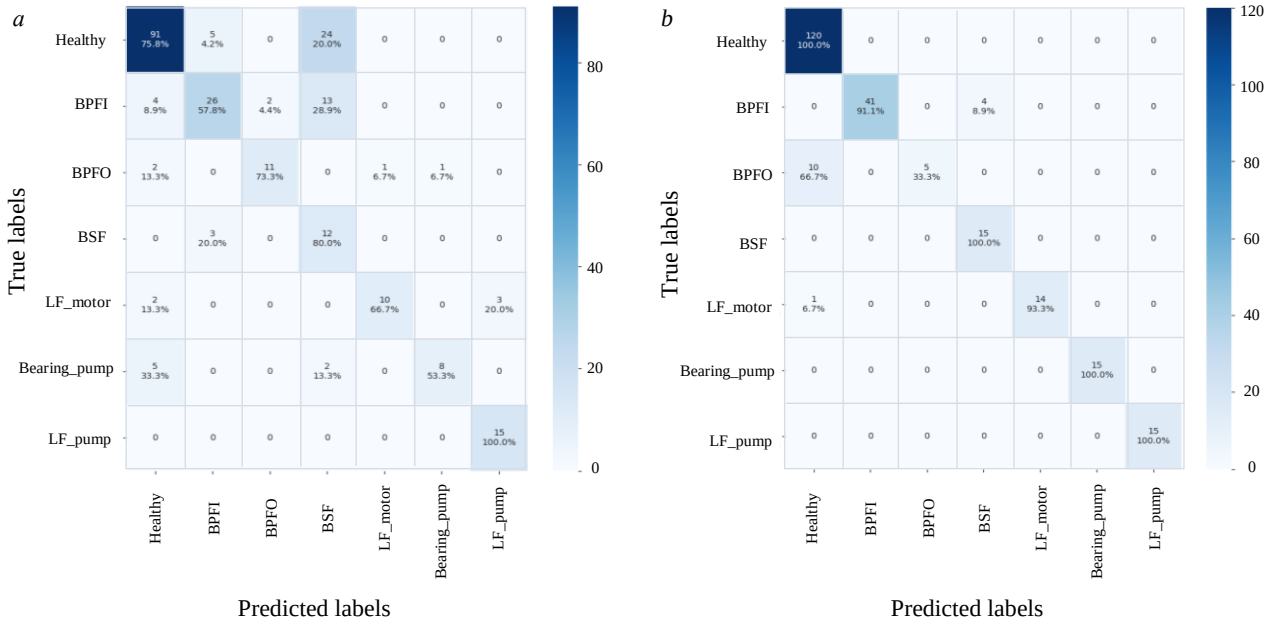

Fig.7. Comparison of confusion matrices between CatBoost (*a*) and the composite model (*b*).
The upper value indicates the total number of classified samples; the lower value shows the percentage of samples normalized with respect to the true set (100 % for each row)

Healthy – absence of any faults; BPFI, BPFO, and BSF – faults associated with inner race, outer race, and ball defects of the motor bearing; LF_motor, LF_pump – looseness in motor and pump mountings; Bearing_pump – pump bearing defects

The authors believe that these limitations can be overcome either by normalizing the signal with respect to the speed and load of the diagnosed units, or by deriving features that are invariant to signal amplitude and frequency.

It is important to note the limitations of data collection, which represent a common challenge in industry. Recording large amounts of fault data is difficult, since damaged electric motors and pumps must be serviced immediately. The classical machine learning approach helps to mitigate this limitation compared to deep learning methods, even under conditions where only a small amount of fault data is available. Generative neural networks may improve this situation; however, they still face the same shortage of training data. Nevertheless, such methods are beginning to emerge.

Conclusions

An approach to classifying a range of industrial pump and electric motor faults using a model composition framework is presented, achieving a macro F1-score of 0.89. It has been shown that the use of the Extended Park's Vector in combination with raw data yields superior results under conditions of limited training data availability.

The study demonstrated that the proposed algorithm outperforms the gradient boosting algorithm when using the same additional features for the considered faults (motor bearing defects: inner race, outer race, ball; pump bearing defect: simultaneous inner and outer races; and looseness of motor and pump mountings).

It was established that, through current and voltage analysis using the proposed approach, it is possible to distinguish between bearing faults associated with both the electric motor and the pump.

This research highlights the potential of MCSA as a non-invasive and scalable tool for equipment condition monitoring, the importance of generating additional features, and the promise of AutoML for technical diagnostics of pump units.

With a view to practical implementation of the developed approach, further studies are planned, focusing on changes in pump rotational speed and the resulting data drift.

REFERENCES

1. Zhukovskiy Y., Buldysko A., Revin I. Induction Motor Bearing Fault Diagnosis Based on Singular Value Decomposition of the Stator Current. *Energies*. 2023. Vol. 16. Iss. 8. N 3303. DOI: 10.3390/en16083303

2. Korolev N.A., Zhukovskiy Y.L., Buldysko A.D. et al. Energy resource evaluation from technical diagnostics of electromechanical devices in minerals sector. *Mining Informational and Analytical Bulletin.* 2024. N 5, p. 158-181 (in Russian). DOI: 10.25018/0236_1493_2024_5_0_158

3. Bolón-Canedo V., Morán-Fernández L., Cancela B., Alonso-Betanzos A. A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing*. 2024. Vol. 599. N 128096. DOI: 10.1016/j.neucom.2024.128096

4. Garcia-Calva T., Morinigo-Sotelo D., Fernandez-Cavero V., Romero-Troncoso R. Early Detection of Faults in Induction Motors – A Review. *Energies*. 2022. Vol. 15. Iss. 21. N 7855. DOI: 10.3390/en15217855

5. Atta M.E.E.-D., Ibrahim D.K., Gilany M.I. Broken Bar Fault Detection and Diagnosis Techniques for Induction Motors and Drives: State of the Art. *IEEE Access*. 2022. Vol. 10, p. 88504-88526. DOI: 10.1109/ACCESS.2022.3200058

6. Ghanbari T., Mehraban A., Farjah E. Inter-turn fault detection of induction motors using a method based on spectrogram of motor currents. *Measurement*. 2022. Vol. 205. N 112180. DOI: 10.1016/j.measurement.2022.112180

7. Bruinsma S., Geertsma R.D., Loendersloot R., Tinga T. Motor current and vibration monitoring dataset for various faults in an E-motor-driven centrifugal pump. *Data in Brief*. 2024. Vol. 52. N 109987. DOI: 10.1016/j.dib.2023.109987

8. Sunal C.E., Dyo V., Velisavljevic V. Review of Machine Learning Based Fault Detection for Centrifugal Pump Induction Motors. *IEEE Access*. 2022. Vol. 10, p. 71344-71355. DOI: 10.1109/ACCESS.2022.3187718

9. Sunal C.E., Velisavljevic V., Dyo V. et al. Centrifugal Pump Fault Detection with Convolutional Neural Network Transfer Learning. *Sensors*. 2024. Vol. 24. Iss. 8. N 2442. DOI: 10.3390/s24082442

10. Chao Zhao, Zio E., Weiming Shen. Domain generalization for cross-domain fault diagnosis: An application-oriented perspective and a benchmark study. *Reliability Engineering & System Safety*. 2024. Vol. 245. N 109964. DOI: 10.1016/j.ress.2024.109964

11. Kim M.-C., Lee J.-H., Wang D.-H., Lee I.-S. Induction Motor Fault Diagnosis Using Support Vector Machine, Neural Networks, and Boosting Methods. *Sensors*. 2023. Vol. 23. Iss. 5. N 2585. DOI: 10.3390/s23052585

12. Jiusi Zhang, Ke Zhang, Yiyao An et al. An Integrated Multitasking Intelligent Bearing Fault Diagnosis Scheme Based on Representation Learning Under Imbalanced Sample Condition. *IEEE Transactions on Neural Networks and Learning Systems*. 2024. Vol. 35. Iss. 5, p. 6231-6242. DOI: 10.1109/TNNLS.2022.3232147

13. Kumar P., Hati A.S. Review on Machine Learning Algorithm Based Fault Detection in Induction Motors. *Archives of Computational Methods in Engineering*. 2021. Vol. 28. Iss. 3, p. 1929-1940. DOI: 10.1007/s11831-020-09446-w

14. Yakhni M.F., Cauet S., Sakout A. et al. Variable speed induction motors' fault detection based on transient motor current signatures analysis: A review. *Mechanical Systems and Signal Processing*. 2023. Vol. 184. N 109737. DOI: 10.1016/j.ymssp.2022.109737

15. Yuejiang Han, Jiamin Zou, Bo Gong et al. The use of model-based voltage and current analysis for torque oscillation detection and improved condition monitoring of centrifugal pumps. *Mechanical Systems and Signal Processing*. 2025. Vol. 222. N 111781. DOI: 10.1016/j.ymssp.2024.111781

16. Chen Liang, Yan Hao, Xie Tengzhou, Li Zhiguo. Identification of cavitation state of centrifugal pump based on current signal. *Frontiers in Energy Research*. 2023. Vol. 11. N 1204300. DOI: 10.3389/fenrg.2023.1204300

17. Gospodarikov A.P., Revin I.E., Morozov K.V. Composite model of seismic monitoring data analysis during mining operations on the example of the Kukisvumchorrskoye deposit of AO Apatit. *Journal of Mining Institute.* 2023. Vol. 262, p. 571-580. DOI: 10.31897/PMI.2023.9

18. Nath A.G., Udmale S.S., Singh S.K. Role of artificial intelligence in rotor fault diagnosis: a comprehensive review. *Artificial Intelligence Review*. 2021. Vol. 54. Iss. 4, p. 2609-2668. DOI: 10.1007/s10462-020-09910-w

19. Salehin I., Islam S., Saha P. et al. AutoML: A systematic review on automated machine learning with neural architecture search. *Journal of Information and Intelligence*. 2024. Vol. 2. Iss. 1, p. 52-81. DOI: 10.1016/j.jiixd.2023.10.002

20. Baratchi M., Can Wang, Limmer S. et al. Automated machine learning: past, present and future. *Artificial Intelligence Review*. 2024. Vol. 57. Iss. 5. N 122. DOI: 10.1007/s10462-024-10726-1

21. Alsharef A., Aggarwal K., Sonia et al. Review of ML and AutoML Solutions to Forecast Time-Series Data. *Archives of Computational Methods in Engineering*. 2022. Vol. 29. Iss. 7, p. 5297-5311. DOI: 10.1007/s11831-022-09765-0

22. Barandier P., Mendes M., Cardoso A.J.M. Comparative analysis of four classification algorithms for fault detection of heat pumps. *Energy and Buildings*. 2024. Vol. 316. N 114342. DOI: 10.1016/j.enbuild.2024.114342

23. Barbudo R., Ventura S., Romero J.R. Eight years of AutoML: categorisation, review and trends. *Knowledge and Information Systems*. 2023. Vol. 65. Iss. 12, p. 5097-5149. DOI: 10.1007/s10115-023-01935-1

24. Bischl B., Binder M., Lang M. et al. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*. 2023. Vol. 13. Iss. 2. N e1484. DOI: 10.1002/widm.1484

25. Morales-Hernández A., Van Nieuwenhuyse I., Rojas Gonzalez S. A survey on multi-objective hyperparameter optimization algorithms for machine learning. *Artificial Intelligence Review*. 2023. Vol. 56. Iss. 8, p. 8043-8093. DOI: 10.1007/s10462-022-10359-2

26. Dahiya S., Nanda H., Artwani J., Varshney J. Using Clustering techniques and Classification Mechanisms for Fault Diagnosis. *International Journal of Advanced Trends in Computer Science and Engineering*. 2020. Vol. 9. N 2, p. 2138-2146. DOI: 10.30534/ijatcse/2020/188922020

27. Maliuk A.S., Ahmad Z., Kim J.-M. A Technique for Bearing Fault Diagnosis Using Novel Wavelet Packet Transform-Based Signal Representation and Informative Factor LDA. *Machines*. 2023. Vol. 11. Iss. 12. N 1080. DOI: 10.3390/machines11121080

28. Hadi R.H., Hady H.N., Hasan A.M. et al. Improved Fault Classification for Predictive Maintenance in Industrial IoT Based on AutoML: A Case Study of Ball-Bearing Faults. *Processes*. 2023. Vol. 11. Iss. 5. N 1507. DOI: 10.3390/pr11051507

29. Cerrada M., Trujillo L., Hernández D.E. et al. AutoML for Feature Selection and Model Tuning Applied to Fault Severity Diagnosis in Spur Gearboxes. *Mathematical and Computational Applications*. 2022. Vol. 27. Iss. 1. N 6. DOI: 10.3390/mca27010006

30. Hutter F., Kotthoff L., Vanschoren J. Automated Machine Learning. Methods, Systems, Challenges. Springer, 2019, p. 219. DOI: 10.1007/978-3-030-05318-5

31. Revin I., Potemkin V.A., Balabanov N.R., Nikitin N.O. Automated machine learning approach for time series classification pipelines using evolutionary optimization. *Knowledge-Based Systems*. 2023. Vol. 268. N 110483. DOI: 10.1016/j.knosys.2023.110483

32. Javed K., Gouriveau R., Zerhouni N. State of the art and taxonomy of prognostics approaches, trends of prognostics applications and open issues towards maturity at different technology readiness levels. *Mechanical Systems and Signal Processing*. 2017. Vol. 94, p. 214-236. DOI: 10.1016/j.ymssp.2017.01.050

33. Khan A.A., Chaudhari O., Chandra R. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications*. 2024. Vol. 244. N 122778. DOI: 10.1016/j.eswa.2023.122778

34. Ozaki Y., Tanigaki Y., Watanabe S. et al. Multiobjective Tree-Structured Parzen Estimator. *Journal of Artificial Intelligence Research*. 2022. Vol. 73, p. 1209-1250. DOI: 10.1613/jair.1.13188

Authors: **Roman R. Khalikov**, *Specialist* (*ROTEC Digital Solutions JSC, Moscow, Russia*), *https://orcid.org/0009-0009-2369-4926*, **Mikhail Yu. Chernetskiy**, *Candidate of Engineering Sciences, Head of Department* (*ROTEC Digital Solutions JSC, Moscow, Russia*), *https://orcid.org/0000-0001-7444-6660*, **Ilia E. Revin**, *Candidate of Engineering Sciences, Researcher* (*ITMO University, Saint Petersburg, Russia*), *https://orcid.org/0000-0002-4459-8724*, **Vadim A. Potemkin**, *Candidate of Engineering Sciences, Researcher* (*ITMO University, Saint Petersburg, Russia*), *vadim_potemkin@itmo.ru, https://orcid.org/0000-0002-8019-7282*.

The authors declare no conflict of interests.