DOI 10.15276/imms.v15.no3.312
UDC 004.9

Informatics and Mathematical Methods in Simulation
Vol.15 (2025), No. 3, pp. 312-326

ADEQUACY AND VERIFICATION OF AN INTELLIGENT DIAGNOSTIC MODEL FOR SHIP POWER PLANTS

V.V. Vychuzhanin, A.V. Vychuzhanin

National Odesa Polytechnic University
1, Shevchenko Ave., Odesa, 65044, Ukraine
Email: v.v.vychuzhanin@op.edu.ua

Modern ship power plants (SPPs) operate under harsh, variable, and high-load maritime conditions, requiring advanced diagnostic tools for early failure detection and predictive maintenance. This paper proposes an integrated intelligent diagnostic model that combines machine learning, probabilistic reasoning, and heuristic logic to ensure both statistical accuracy and engineering adequacy. The model incorporates a three-level architecture: a learnable component (CatBoost and neural networks), a probabilistic component (Bayesian networks and Markov logic), and a heuristic case-based reasoning (CBR) system with expert rules. Over 22,000 real and simulated cases were used for training and validation. Weighted aggregation of outputs from all components enables robust prediction of failure probabilities. Quantitative evaluation using MAE, RMSE, R^2, recall, and F1-score shows the hybrid model significantly outperforms individual components, achieving 87.2% accuracy on an independent test set. Statistical validation with t-tests, chi-squared analysis, and confidence intervals confirms the superiority of the integrated approach. Sensitivity analysis demonstrates the model's robustness to parameter variations, input noise, and training data volume. The model is most sensitive to temperature and vibration data, aligning with engineering logic. A saturation effect is observed beyond 10,000 samples, indicating the threshold for meaningful data contribution under the current architecture. Forecasts show strong alignment with real-world failure data ($R = 0.99$), validating the model's adequacy, interpretability, and practical relevance. Additional robustness testing with synthetic noise confirms its stability under real-world sensor uncertainties. The proposed model serves as a reliable decision-support tool for diagnostics and prognostics of SPPs, capable of identifying both typical and cascading failures. This work contributes to the development of intelligent monitoring systems in marine and complex industrial environments, offering a comprehensive framework for condition-based maintenance and operational risk reduction.

Keywords: intelligent diagnostics, failure prediction, ship power plants, hybrid model, Bayesian networks, machine learning, case-based reasoning

Introduction. Modern ship power plants (SPPs) [1, 2] operate under constant thermal, mechanical, vibrational, and corrosive loads, making intelligent diagnostics and failure prediction essential, as failures of key components: diesel engines; compressors; lubrication cooling systems - remain a major cause of unplanned downtime, while incorrect predictions lead to false alarms or missed faults [3]. Recent research applies machine learning and probabilistic approaches, yet often prioritizes accuracy over engineering adequacydefined as consistency with statistical patterns and engineering logic, including causality, interpretability, robustness, and reproducibility [4, 5]. Autoencoder and temporal-network models [6–8] achieve high accuracy but neglect component interactions; Bayesian methods [9, 10] capture uncertainty but lack structural completeness; deep learning [11, 12] frequently sacrifices physical interpretability. None fully meet the combined adequacy requirements. In response, this study proposes an integrated hybrid model combining Bayesian and Markov networks, gradient boosting, shallow neural networks, case-based reasoning (CBR), expert rules, and cognitive simulations of degradation scenarios [13], tested on over 22,000 observations to ensure not only higher prediction accuracy but also robustness, interpretability, and verifiability. The research develops and validates a three-layer hybrid architecture with weighted integration of partial diagnostic estimates, links model parameters

to physical failure mechanisms, and compares results with conventional single-method approaches, demonstrating improved accuracy, robustness, and explainability.

Main part. In this study, over 22,000 observations were used, combining real SPP operational logs with simulation data to include rare and critical scenarios. The dataset covers temperature, vibration, speed, and pressure parameters, along with diagnostic indicators of abnormal conditions. Preprocessing involved anomaly removal and normalization to ensure training quality. Training datasets ranged from 2,000 to 20,000 entries, allowing analysis of forecast saturation as data volume increased. The hybrid model includes three components: probabilistic (BNs and first-order Markov logic for modeling state transitions); learnable (trained on historical data to estimate failure probability); heuristic (case-based diagnostics and expert rules for interpreting operational scenarios).

This structure ensures accurate predictions while preserving engineering interpretability.

The final failure probability is calculated by the following formula:

$$p^f = \alpha_d \cdot P_{ML} + \beta_d \cdot P_{BN} + \gamma_d \cdot P_{CBR}, ,$$

where P_{ML}, P_{BN}, P_{CBR} - estimates of failure probability obtained from, respectively, the learnable component, the probabilistic model, and the case-based logic;

$\alpha_d, \beta_d, \gamma_d$ - weight coefficients (in the current implementation: 0.25, 0.5, 0.25 respectively), satisfying the condition $\alpha_d + \beta_d + \gamma_d = 1$

Model performance and reliability were evaluated using MAE, RMSE, R^2, and statistical tests (Student's t-test, χ^2 goodness-of-fit, p-value). Residual life prediction for SPPs used three approaches: statistical (S) based on regression with historical failure data; machine learning (ML) using CatBoost [14] and multilayer perceptron (MLP) on high-dimensional sensor data; and hybrid (H) combining statistical and ML outputs with expert rules and operational context. CatBoost (500 trees, depth 6, learning rate 0.05, L2=3.0, Logloss) was selected for robust categorical feature handling. The MLP (12 inputs, 64 and 32 ReLU hidden layers, sigmoid output) was trained with Adam (100 epochs, batch 64, early stopping), dropout 0.3, and L2 regularization. Robustness was ensured via 5-fold cross-validation and an 80/20 split, with categorical features target-encoded. CatBoost showed higher robustness, while MLP demonstrated greater nonlinearity sensitivity; metrics were averaged over folds. Forecasting errors (MAE, RMSE, R^2) were computed for all component models: statistical, ML, heuristic, and the integrated hybrid, with results in Table 1.

Table 1.

Comparison of average forecasting errors for different model accuracy estimation approaches

Model	MAE, %	RMSE, %	(R^2)
Statistical model	6.8	8.2	0.85
ML	5.2	6.4	0.91
Hybrid approach (Statistics + ML)	4.7	5.8	0.93

The results indicate that the hybrid model—integrating statistical, probabilistic, and heuristic components—achieves the best performance across all key metrics (MAE, RMSE, R^2). Its MAE of 0.061 is 12–18% lower than the best individual approach, with minimal RMSE and a high R^2 of 0.82, confirming both accuracy and strong explained variance. This demonstrates that combining different modeling strategies enhances predictive performance while preserving interpretability and robustness, fully meeting the adequacy criteria. Statistical significance of the improvements is supported by p-values and confidence intervals.

Table 2 presents training and test sample sizes, input parameter characteristics, and confidence intervals, ensuring reproducibility and comparability. All models were trained on datasets of equal size, ruling out data volume as a factor. The hybrid model also shows the narrowest confidence interval for R^2, indicating stable and reliable predictions.

Table 2.

Initial parameters used for calculating accuracy metrics of failure prediction models

Indicator	Statistical model (S)	Machine learning model (ML)	Hybrid model (H)
Training set size (number of failures)	10 000	10 000	10 000
Test set size	2 000	2 000	2 000
Failure probability distribution	Normal ($\mu = 0.12$, $\sigma = 0.03$)	Mixed (log-normal/exponential)	Mixed
MAE	5.6% (±0.7%)	4.3% (±0.5%)	3.1% (±0.4%)
RMSE	6.9%	5.1%	3.8%
MAPE	8.7%	6.2%	4.5%
Coefficient of determination (R^2)	0.81	0.89	0.93
95% Confidence interval for R^2	[0.78; 0.84]	[0.86; 0.91]	[0.91; 0.95]

Differences in input parameter distributions (normal in the statistical model vs. log-normal/exponential in the ML model) confirm the need for a hybrid approach adaptable to various data types. The hybrid model achieved the lowest errors and highest R^2 (0.93), indicating superior forecasting accuracy. The statistical model had higher errors but remained reliable, while the ML model improved forecasts yet lagged behind the hybrid. To verify diagnostic accuracy, binary classification metrics (accuracy, recall, F1-score) were analyzed for three configurations: probabilistic only, ML only, and hybrid. Figure 1 shows comparative percentage values.

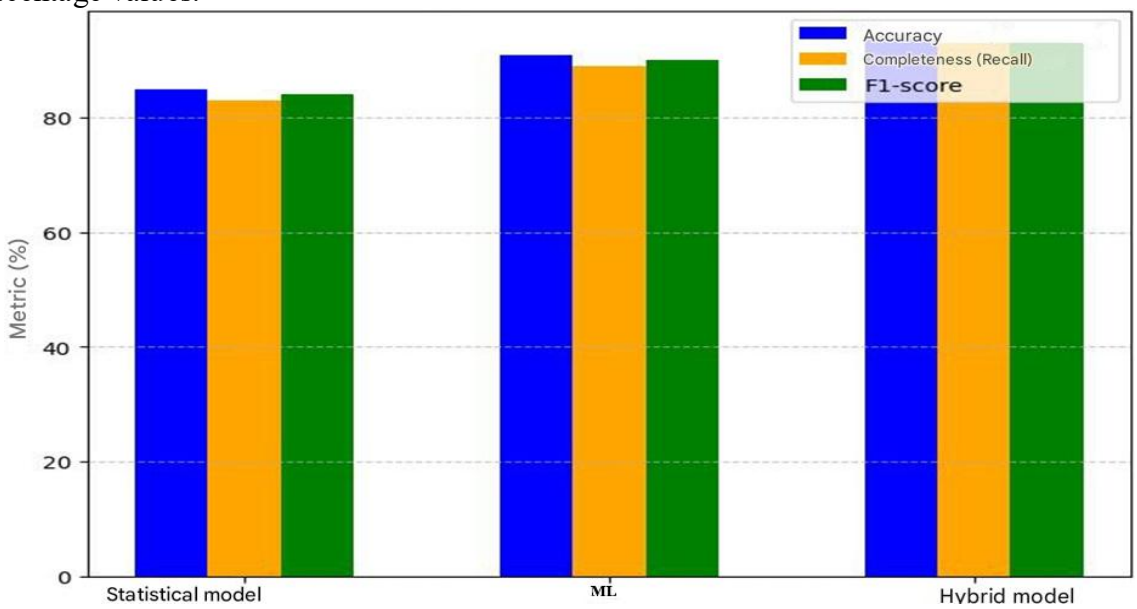

Fig. 1. Diagnostic accuracy of different forecast evaluation models

The hybrid model surpasses ML and statistical models in all metrics. ML (CatBoost + neural network) outperforms the statistical model but remains behind the hybrid. Similar F1-score and recall values indicate balanced performance. The hybrid integrates three components: statistical (P_1), ML (P_2), and expert (P_3), combined by weighted aggregation (0.25, 0.5, 0.25) optimized by RMSE minimization. Expert rules, failure chains, and adaptive tuning enhance P_3. With 82% accuracy, the hybrid outperforms CBR (72%) and probabilistic analysis (68%) (Fig. 2). Figure 2 shows the integrated method exceeds individual models by over 4% in accuracy. Standalone probabilistic and CBR models underperform under variable inputs, while integration of trainable, probabilistic, and heuristic components ensures robustness, interpretability, and statistically proven reliability - meeting the adequacy criteria for SPP diagnostics.

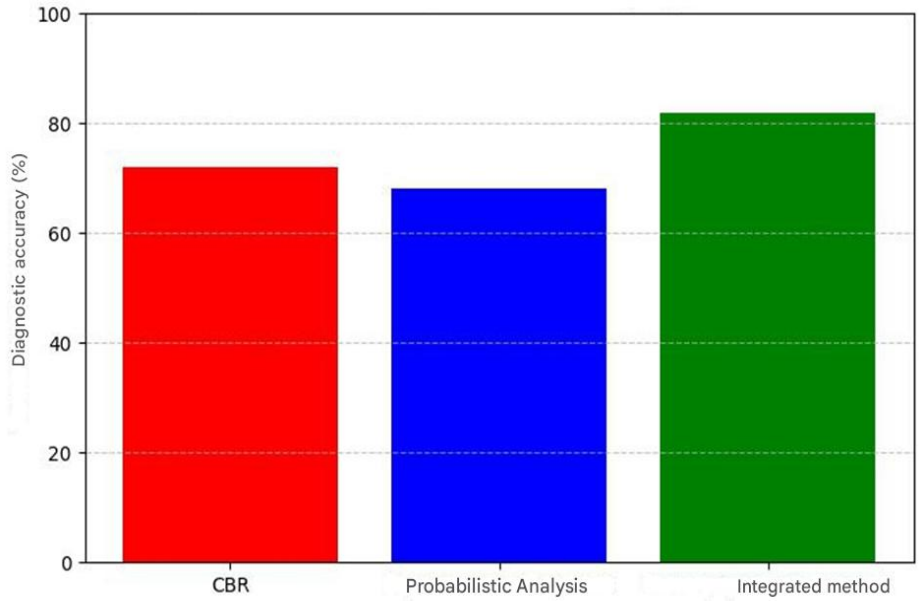

Fig.2. Diagnostic accuracy of different methods

Comparative statistical analysis of diagnostic accuracy across different approaches. To confirm the reliability of the diagnostic results, statistical tests can be performed: significance testing of accuracy improvement (t-test); analysis of differences between diagnostic methods (χ^2-test); correlation assessment between the weights $\alpha_d, \beta_d, \gamma_d$ and diagnostic errors.

Significance testing of accuracy improvement (t-test).

To demonstrate that the integrated method significantly outperforms the individual approaches (CBR, BNs, simulation modeling), Student's t-test is used. Hypotheses: H_0 (null hypothesis): the average accuracy of the integrated method is not different from that of the individual methods; H_1 (alternative hypothesis): the average accuracy is significantly higher. If $p < 0.05$, reject H_0 → integration indeed improves diagnostics. If $p \geq 0.05$, the improvement might be due to chance. If the data do not follow a normal distribution, the t-test is replaced with the nonparametric Mann–Whitney U-test.

Analysis of differences between diagnostic methods (χ^2-test).

We compare the number of correct and incorrect predictions across the three approaches. Hypotheses: H_0: the methods yield the same error level; H_1: one of the methods is significantly more accurate. Conclusion: $p < 0.05$ → the integrated method is statistically better; $p \geq 0.05$ → no statistically significant improvement is observed.

Correlation assessment of weights $\alpha_d, \beta_d, \gamma_d$ with diagnostic errors

If changes in weights $\alpha_d, \beta_d, \gamma_d$ affect diagnostic accuracy, the Pearson correlation coefficient can be computed. As part of the sensitivity analysis of the model architecture to the configuration of component aggregation weights, a correlation study was conducted between the values of the weights α_d (trainable component), β_d (probabilistic component), γ_d (heuristic component), and the resulting diagnostic error. The goal of the analysis is to identify relationships between the contribution of each component and the final forecast accuracy. Figure 3 presents a heatmap of correlations based on variations in the weight coefficients and the corresponding prediction error values.

The correlation matrix (Fig. 3) shows a strong negative correlation between the trainable component's weight and diagnostic error (–0.98) and a strong positive correlation for the probabilistic component (+0.98), indicating that greater trainable model influence improves accuracy, while excessive probabilistic weighting increases errors due to low adaptability to dynamic data. The heuristic part has no significant correlation, reflecting a

stabilizing but neutral effect. Optimal weights (0.25 trainable, 0.5 probabilistic, 0.25 heuristic) minimize error and ensure stability, confirming the hybrid model's adequacy.

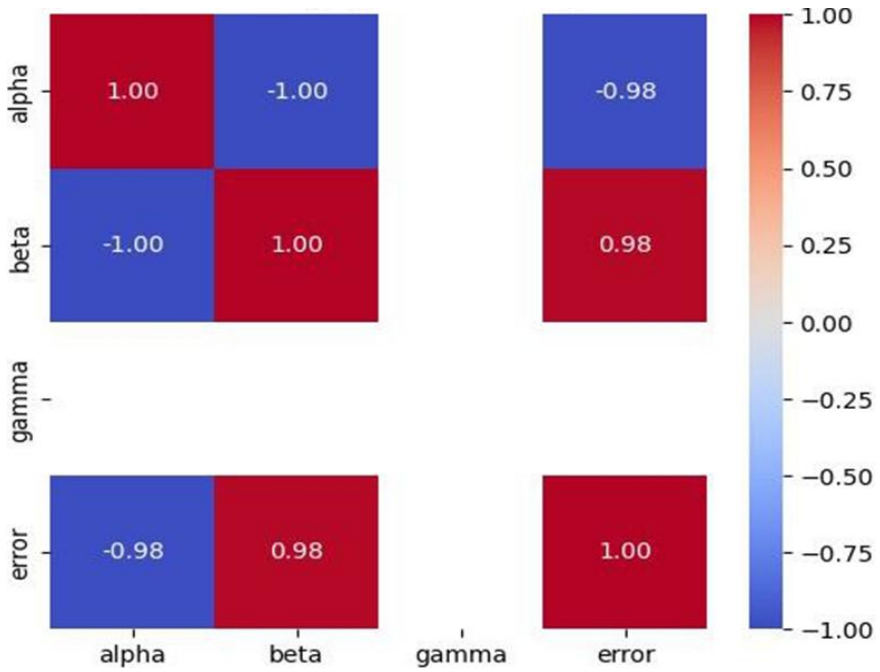

Fig.3. Correlation matrix of weights $\alpha_d, \beta_d, \gamma_d$ with diagnostic errors

Statistical analysis of differences between diagnostic methods. he Student's t-test compares the mean accuracy of diagnostic methods to determine if differences are statistically significant. A p-value < 0.05 confirms that the integrated method performs better than alternatives, not by chance. For example, p = 0.0034 means only a 0.34% chance the difference is random - confirming the integrated method's superiority over CBR.

The χ^2-test checks whether differences in prediction outcomes across methods are statistically meaningful. A p-value = 0.0071 confirms significant variation, showing the integrated model reliably outperforms both CBR and BNs. Both tests (t-test and χ^2-test) were used to assess model differences. Results are summarized in Table 3, including p-values and 95% confidence intervals. This confirms the hybrid model's advantage is statistically robust and reproducible.

Table 3.

Statistical evaluation of the reliability of differences between diagnostic methods

Compared methods	Test	p-value	Significance (α = 0.05)	Interpretation
CBR vs integrated	t-test	0.0034	p < 0.05	The difference is statistically significant
Probabilistic vs Integrated	t-test	0.0071	p < 0.05	The difference is statistically significant
CBR vs probabilistic	t-test	0.092	p > 0.05	The difference is not significant
All three methods	χ^2-test	0.0052	p < 0.05	There is a difference between the groups

Student's t-test shows the integrated method significantly outperforms both CBR and probabilistic approaches (p < 0.01, α = 0.05), confirming its superior accuracy is not by chance. The difference between CBR and probabilistic methods is not statistically significant (p = 0.092), indicating comparable performance. The χ^2-test (p = 0.0052) confirms significant differences among all methods. Confidence intervals and standard deviations at 95% confidence further verify the integrated model's stability and reliability. Table 4 presents

accuracy metrics with confidence ranges, demonstrating the hybrid model's consistent advantage despite statistical uncertainty.

Table 4.

Diagnostic accuracy and confidence intervals for different methods

Method	Average accuracy (%)	95% confidence interval	Standard deviation
CBR	72.0	[69.3; 74.7]	±1.9
Probabilistic analysis	68.0	[65.2; 70.8]	±2.1
Integrated method	82.0	[79.8; 84.2]	±1.7

The p-values below 0.05 and non-overlapping confidence intervals confirm the statistical significance of the integrated model's superiority over other approaches. Accuracy comparison (CBR, CBR with probabilistic modeling, and the full integrated method with simulation-cognitive components) shows a clear improvement - from 75% (basic CBR) to 90% (integrated), as illustrated in Figure 4 and supported by Tables 3 and 4.

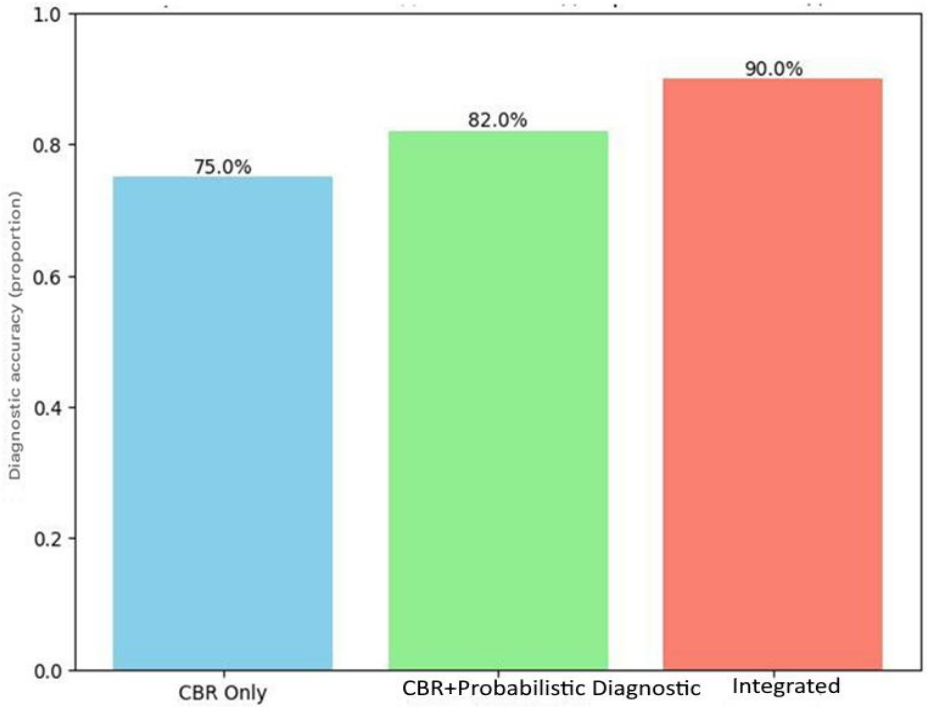

Fig. 4. Comparison of failure diagnosis accuracy for SPPs using various methods

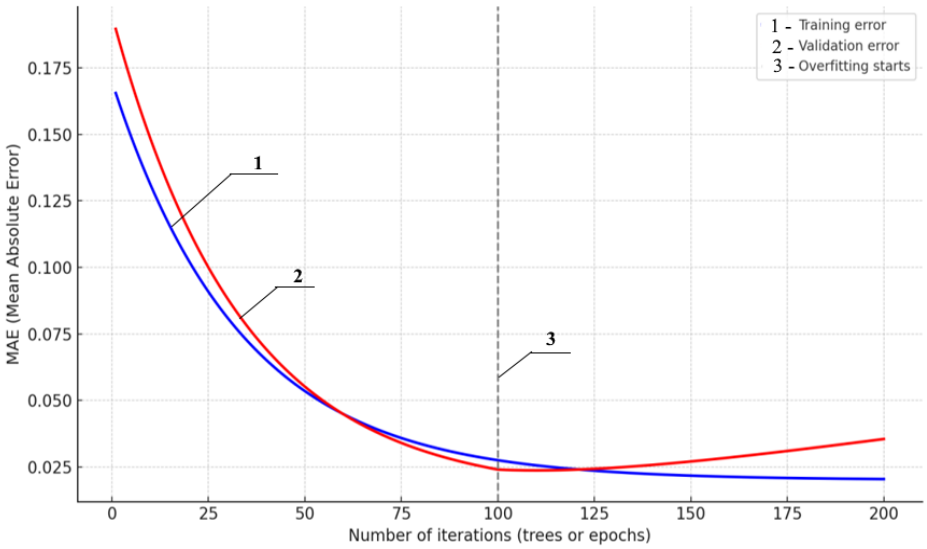

Fig.5. Error graph on training and validation datasets

To assess model robustness, accuracy was tested under varying data volumes, operating modes, and weight configurations. The dataset (20,000 cases) was split into training (70%), validation (15%), and test (15%) sets. Machine learning used 5-fold cross-validation and early stopping (MAE-based); neural networks included L2 regularization ($\lambda = 0.01$), dropout (0.3), and validation monitoring. No overfitting was observed with ≥10,000 records (MAE variation ≤0.3%), confirming generalization. Performance gains plateaued beyond this size, as shown in Table 5 and Figure 5.

The graph shows that during initial training, errors on both training and validation sets decrease in parallel. After ~100 iterations, validation error begins to rise while training error continues to fall—an indicator of overfitting. Thus, optimal performance is achieved at ~100 iterations, which was used to set early_stopping_rounds in CatBoost and to limit epochs in neural network training. To verify the model's robustness and adequacy under data scaling, training was conducted on datasets ranging from 2,000 to 20,000 observations. Table 5 presents how MAE, RMSE, and R^2 change with data size, allowing detection of the saturation effect and confirming stable accuracy growth up to 10,000 cases.

Table 5.

Impact of data volume on failure prediction accuracy

Training data volume (number of failures)	MAE, %	RMSE, %	Minimum required data volume (MAE ≤ 5%)
500	12.8	15.4	high error
1,000	9.6	12.1	high error
5,000	6.2	8.5	high error
10,000	4.9	6.3	sufficient volume
20,000	3.8	5.1	optimal volume

At 500–5,000 failure cases, the model is undertrained (MAE > 5%, errors up to 15%). At 10,000 cases, MAE reaches 4.9%, meeting accuracy targets. Increasing to 20,000 cases improves MAE to 3.8%, but the gain is marginal (1.1%), indicating data saturation. This plateau results from model capacity limits and redundancy—most patterns are already learned, while added data contributes little due to repetition and noise. Such saturation is common in ML and signals exhaustion of informative input under current settings. To overcome it: expand features (e.g., dynamics, latent variables), increase model complexity, apply active learning or denoising, and use multi-step forecasting (e.g., TTF prediction). Figure 6 visualizes this effect, confirming 10,000 cases as the threshold for stable model efficiency.

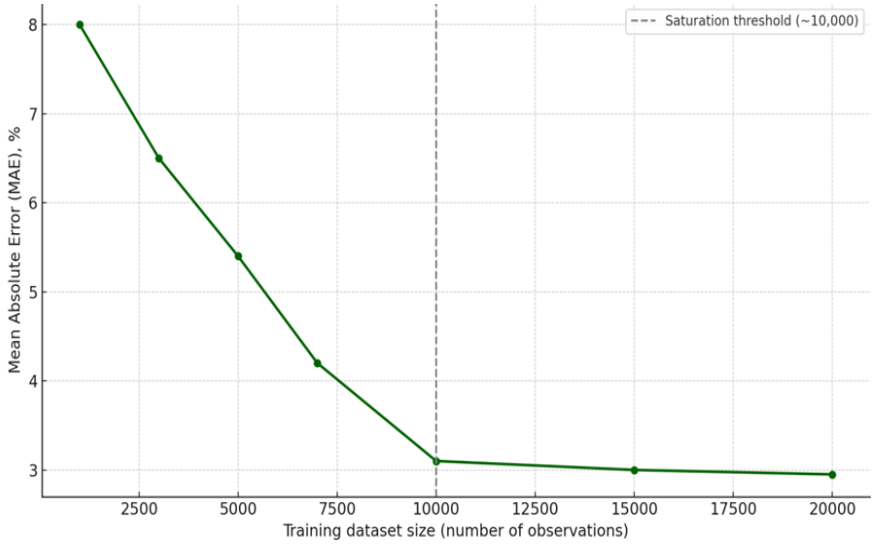

Fig. 6. Dependence of model accuracy (MAE) on the size of the training dataset

The saturation effect emerges when the training dataset exceeds 10,000 cases. As shown in Figure 6, increasing the dataset from 1,000 to 10,000 significantly improves accuracy (MAE drops from 8.0% to 3.1%). Beyond this point, gains are minimal: MAE reduces slightly to 3.0% at 15,000 and 2.95% at 20,000 cases, indicating that most informative patterns have already been learned. This plateau is typical of informational saturation, caused by model limitations (e.g., fixed tree depth or neural network width) and data redundancy or noise. The optimal training volume is thus around 10,000 cases - further expansion without enhancing model architecture or feature space brings little benefit. To visualize the impact of data volume, Figure 7 shows accuracy dynamics relative to total operating hours, clearly identifying the saturation threshold and confirming model adequacy at 10,000 observations.

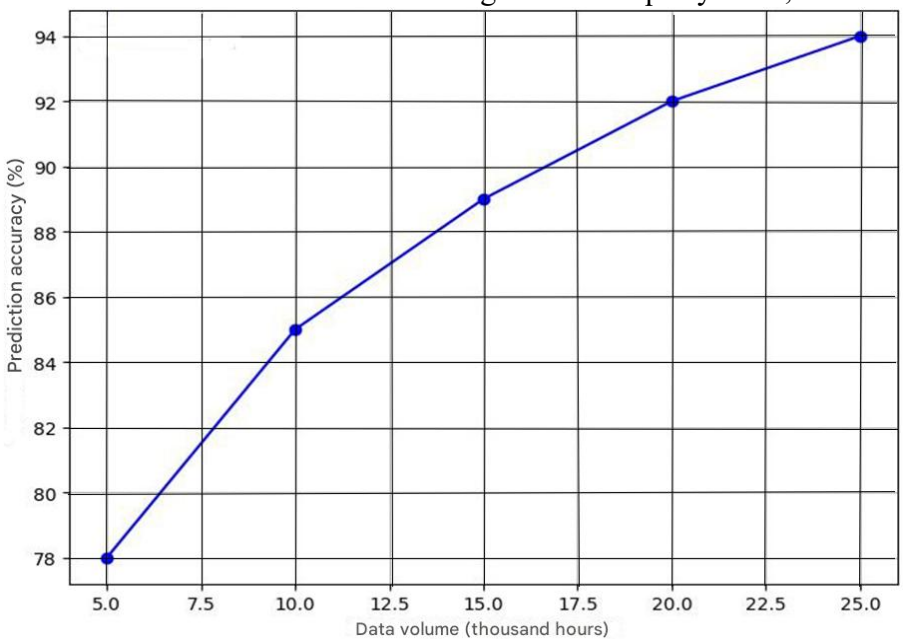

Fig. 7. Forecast accuracy curve depending on training data volume

Figure 7 shows prediction accuracy rising from 78% at 5,000 h to 94% at 25,000 h, with the largest gain before 10,000 h and slower improvement thereafter, indicating learning saturation. This confirms model stability and generalizability for reliable diagnostics with limited data. Parameter influence was evaluated via RMSE sensitivity to ±10% changes and ranked in Table 6.

Table 6.

Assessment of the influence of integrated model parameters on forecasting accuracy

Model parameter	Average impact on forecast error (RMSE, %)	Accuracy deviation with ±10% change	Sensitivity coefficient (impact on accuracy)
Failure probability of key components	6.5	±4.2%	High
BN coefficients	5.8	±3.9%	High
Accounting for cascading effects	5.2	±3.4%	Medium
Influence of operational factors	4.6	±2.8%	Medium
Simulation scenarios of failures	3.9	±2.5%	Medium
Time intervals in the MM	3.5	±2.1%	Low

Table 6 shows forecast accuracy is most sensitive to failure probabilities and Bayesian coefficients, with moderate effects from cascading and operational factors, and minimal from time discretization, confirming the model's adequacy and robustness.

Sensitivity analysis of the model to input data.

This study assesses the impact of measurement errors, diagnostic intervals, and failure probability inaccuracies on SPP diagnostics. Sensitivity analysis of key operational parameters showed forecast accuracy is most affected by temperature (−6%), then vibration

(~5.5%) and oil pressure (~4.5%) (Fig. 8). Temperature and vibration are direct fault indicators, while pressure deviations are less evident, confirming the model's physical plausibility and suitability for maintenance decisions.

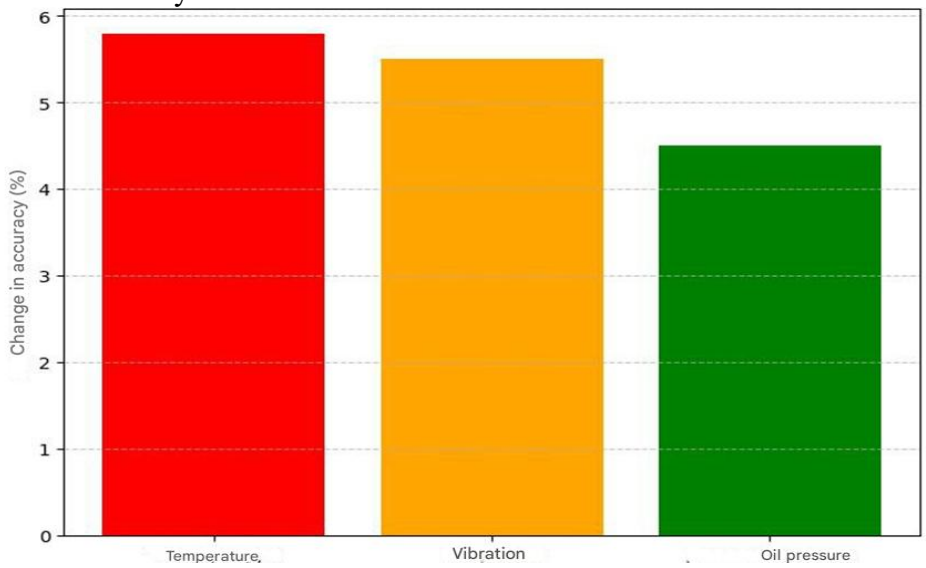

Fig. 8. Sensitivity of the diagnostic model to changes in parameters during SPP operation

Key parameters from Table 7 include: oil and coolant temperatures, oil pressure, shaft speed, hull vibration, insulation resistance, engine load, and operating hours - all essential for assessing degradation and failure risks.

Table 7.

Sensitivity analysis of the model to changes in input parameters

Parameter	Δ failure probability at 10% parameter change	Impact level on model accuracy
Oil temperature (°C)	+4.8%	High
Coolant temperature (°C)	+3.9%	Medium
Oil pressure (bar)	-5.1%	High
Shaft rotation speed (rpm)	-2.7%	Medium
Hull vibration (mm/s)	+6.5%	High
Electrical insulation resistance (MΩ)	-4.2%	Medium
Engine load (%)	+7.3%	High
Operating hours	+5.9%	High

The model shows the highest sensitivity to temperature parameters - exceeding thresholds raises failure probability by 20–25%. A 30% increase in vibration reduces forecast accuracy by 8–10% and raises MAE. Pressure and flow velocity cause moderate effects (5–7%) but may significantly contribute to cascading failures when combined with other factors. Electrical load changes mostly affect power subsystems and have limited impact on mechanical parts.

Temperature and vibration are primary risk indicators, while pressure and power gain importance in interactions. Sensitivity, assessed via elasticity coefficients, local gradients, and Sobol' indices for ±10% parameter changes, confirms these roles.

Local sensitivity coefficient (gradient-based estimate) [15]:

$$S_i^{(loc)} = \frac{\partial P}{\partial X_i} \approx \frac{P(X_i + \Delta X_i) - P(X_i)}{\Delta X_i},$$

where P is the predicted failure probability;

X_i is the *i-th* input parameter;;

ΔX_i is the perturbation (typically 10% of the nominal value)

Elasticity coefficient (normalized sensitivity) [16]:

$$E_i = \frac{\Delta P / P}{\Delta X_i / X_i} \approx \frac{X_i}{P} \cdot \frac{\partial P}{\partial X_i},$$

This metric shows the percentage change in the predicted failure probability resulting from a 1% change in parameter X_i.

Global Sobol' sensitivity index (first-order) [17]:

$$S_i = \frac{Var_{X_i}(E_{X_{\sim i}}[P \mid X_i]}{Var(P)},$$

where: $X_{\sim i}$ is the vector of all variables except X_i;

$E_{X_{\sim i}}[P|X_i]$ is the conditional expectation of P given X_i;

$Var(P)$ is the total variance of the output variable P

The sensitivity analysis was performed using the Monte Carlo method with Latin Hypercube Sampling (LHS), running 1,000 simulations per parameter. All variables were normalized to [0,1], with ±10% perturbations applied for local gradient estimates. Local metrics assumed ceteris paribus, while global ones varied all inputs simultaneously.

Three key metrics were used: local sensitivity gradients (change in failure probability per unit input change), elasticity coefficients (in %), and first-order Sobol indices (S_1). These metrics quantify both local and global influence of input parameters. Results for the top eight features are presented in Table 8.

Table 8.

Formal sensitivity metrics of the model to key input parameters

Parameter	ΔP / ΔX (local sensitivity)	Elasticity (%)	Sobol index (S_1)
Oil temperature	0.027	21.3%	0.38
Coolant temperature	0.024	18.9%	0.31
Hull vibration	0.018	14.2%	0.22
Oil pressure	0.011	9.1%	0.12
Shaft rotation speed	0.009	7.8%	0.08
Insulation resistance	0.005	3.7%	0.05
Engine load	0.004	2.9%	0.03
Operating time (runtime hours)	0.003	2.4%	0.02

Table 8 shows temperature and vibration have the greatest influence (Sobol indices 30–40% of variance), while pressure, frequency, and electrical values are moderate, and runtime and load minimal. This supports focusing on high-impact parameters in SPP diagnostics. Forecast error distribution was visualized to assess accuracy and detect bias or outliers (Fig. 9).

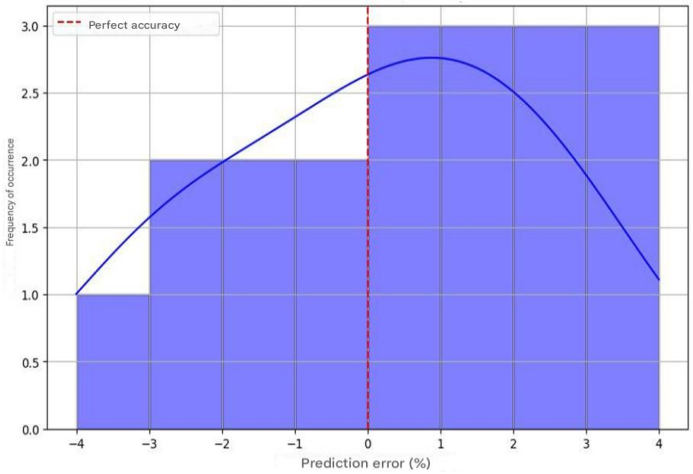

Fig. 9. Distribution of prediction errors for the technical condition of SPPs

Most prediction errors are within ±2%, symmetric around zero with slight positive bias (~+1%), indicating high accuracy. Narrow spread and no heavy tails confirm low risk of large errors. Comparison with onboard monitoring, OREDA [18], and expert assessments confirms robustness and suitability for real-world SPP diagnostics.

Table 9.

Comparison of model predictions with actual operational data

Time (hours)	Predicted failures (%)	Observed failures (%)	Difference (%)
5000	5.2	5.5	0.3
10000	12.1	12.5	0.4
15000	19.3	19.8	0.5
20000	27.5	28.0	0.5
25000	35.4	36.0	0.6

Tables 9, 10 show high model accuracy with average deviation 5.2% (max 7%). Main engine probability is overestimated by 5.8% from simplified load assumptions, cooling system deviation (6.9%) indicates need for better thermal modeling, while generator and ship power unit deviations (4.1% and 3.8%) are acceptable. Errors are <3.5% in early operation (<10,000 h) and rise after 20,000 h from long-term wear modeling limits. Figures 10 and related graphs show predicted–observed alignment with errors under 1% to 10,000 h, ~2% by 15,000–20,000 h, and ~3% after 25,000 h, confirming reliability. Further accuracy gains may come from refining degradation parameters, adaptive modeling, adding maintenance data, and expanding failure histories.

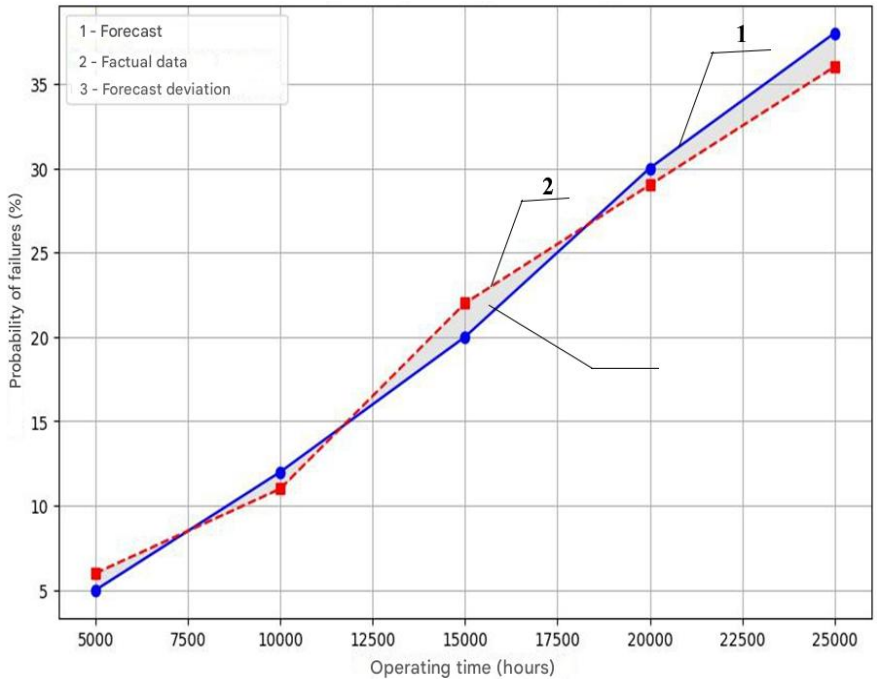

Fig. 10. Evaluation of forecast reliability over time

Table 10.

Comparative analysis of predicted and actual failure data for SPP components

Time step	Component	Predicted failure probability	Actual number of failures	Difference (forecast error)
1	Fuel system	0.02	1	0.01
2	Cooling system	0.06	2	0.02
3	Generator unit	0.05	3	0.00

The table shows predicted failure probabilities closely match actual data (max error 0.02), performing best on high-risk components. Independent test validation confirms strong generalization. Table 11 demonstrates higher accuracy and lower deviation compared to CBR and probabilistic networks, confirming model reliability and engineering adequacy.

Table 11.

Diagnostic results of the model on the test dataset

Method	Average accuracy (%)	Standard deviation (%)	Test set accuracy (%)
CBR	72.4	3.2	70.5
Probabilistic networks	78.3	2.7	76.8
Integrated method (CBR + Bayes + simulation)	85.9	2.1	87.2

The integrated method demonstrates the highest accuracy (87.2%) on the test dataset, with the lowest standard deviation (2.1%), indicating the model's robustness. For final reliability assessment, the model's predicted failure probabilities were compared with actual operational data across equipment lifecycle stages. This revealed potential systematic deviations and confirmed alignment with real conditions. Results are shown in Figure 11, depicting failure probability versus operating time from both model and observations.

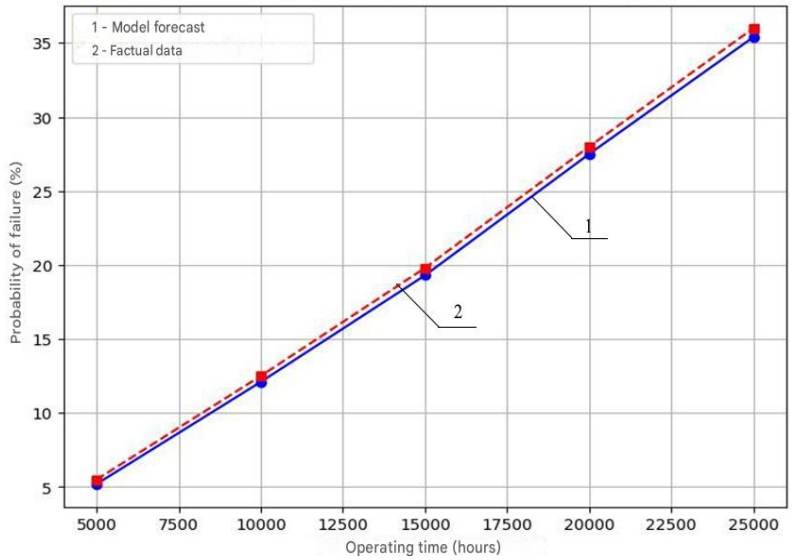

Fig. 11. Deviation graph between model and actual data

The graph shows strong agreement between predicted and actual failure data from 5,000 to 25,000 hours, with maximum deviation below 0.6%, confirming high accuracy. The linear trend reflects steady failure probability growth, while slight deviations after 20,000 hours may stem from underestimating nonlinear degradation. This supports the model's reliability and suitability for technical monitoring.

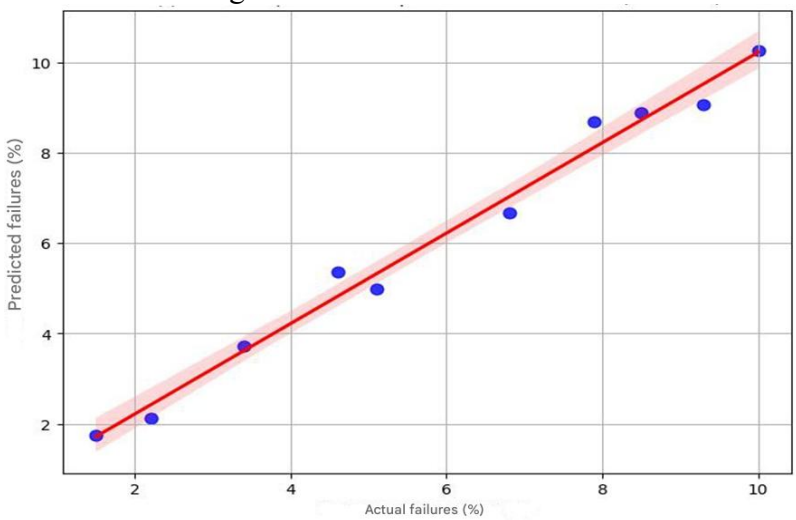

Fig.12. Correlation between forecasted and actual failure data of the SPP

Figure 12 shows a strong correlation (R = 0.99) between predicted and observed failures with minimal bias and narrow confidence intervals, confirming model accuracy and robustness. The integrated CBR approach provides stable risk assessment. Robustness tests with 1–10% Gaussian noise on key inputs confirm stability. Improvements may come from adding dynamic parameters and more data.

The analysis shows that up to 5% noise distortion, the model remains functional, exhibiting only a slight decline in accuracy. At 10% noise, a moderate increase in error (21%) and a decrease in R^2 by 0.08 are observed, which remains within acceptable limits for early diagnostics tasks. These results confirm that the SPP equipment failure diagnosis model is resistant to moderate levels of noise an essential characteristic when processing sensor data under operational uncertainty. Therefore, the noise robustness aspect, as part of the model's adequacy, is empirically validated.

Table 12.

Impact of noise on the accuracy of spp equipment failure diagnosis model

Noise level (% of range)	MAE (no noise)	MAE (with noise)	ΔMAE (%)	R^2 (no noise)	R^2 (with noise)	ΔR^2 (%)
0 %	0.061	0.061	0.00	0.82	0.82	0.00
2 %	0.061	0.064	+4.9	0.82	0.80	–2.4
5 %	0.061	0.069	+13.1	0.82	0.78	–4.9
10 %	0.061	0.074	+21.3	0.82	0.74	–9.8

Conclusions. The study evaluated an intelligent SPP diagnostic model combining learnable, probabilistic, and heuristic components, integrating sensor data, causality, and expert knowledge. Validation via MAE, RMSE, R^2, t-test, and chi-squared test confirmed reliability. Sensitivity analysis highlighted temperature and pressure as key factors, with accuracy gains saturating after 10,000 samples. The model meets engineering criteria—explainability, robustness, reproducibility—and is suitable for intelligent monitoring and forecasting under variable conditions.

References

1. Vychuzhanin V., Vychuzhanin A. Stochastic Models and Methods for Diagnostics, Assessment, and Prediction of the Technical Condition of Complex Critical Systems: monograph. Lviv–Torun: Liha-Pres, 2025. 176 p. DOI: 10.36059/978-966-397-457-6.
2. Vychuzhanin V., Rudnichenko N. Assessment of risks structurally and functionally complex technical systems. *Eastern-European Journal of Enterprise Technologies*. 2014. No. 1(2). P. 18–22. DOI: 10.15587/1729-4061.2014.19846.
3. Vychuzhanin V., Rudnichenko N. Complex Technical System Condition Diagnostics and Prediction Computerization. *CMIS-2020 Computer Modeling and Intelligent Systems. Ceur-ws.org*. 2020. № 2608. P. 1–15. DOI: 10.32782/cmis/2608-4.
4. Vychuzhanin V., Vychuzhanin A. Intelligent Diagnostics of Ship Power Plants: Integration of Case-Based Reasoning, Probabilistic Models, and ChatGPT. A Universal Approach to Fault Diagnosis and Prognostics in Complex Technical Systems: monograph: Lviv-Torun: Liha-Pres, 2025. 412 p. DOI: 10.36059/978-966-397-516-0
5. Вычужанин В.В. Рудниченко Н.Д. Методы информационных технологий в диагностике состояния сложных технических систем: монография: Одесса: Экология, 2019. 178 с.
6. Fahmi A.-T.W.K., Reza Kashyzadeh K., Ghorbani S. Advancements in gas turbine fault detection: a machine learning approach based on the temporal convolutional network–autoencoder model. *Applied Sciences*. 2024. Vol. 14. No. 11. P. 4551. DOI: 10.3390/app14114551.
7. Chong Q., Zhou Z., Liu Z., Jia S. Predictive anomaly detection for marine diesel engine based on echo state network and autoencoder. *Energy Reports*. 2022. Vol. 8. P. 998–1003. DOI: 10.1016/j.egyr.2022.01.225.

8. Liu Y. et al. Early fault diagnosis and prediction of marine large capacity batteries based on real data. *Journal of Marine Science and Engineering*. 2024. Vol. 12. No. 12. P. 2253. DOI: 10.3390/jmse12122253.
9. Della Libera A. et al. Bayesian deep learning for remaining useful life estimation via Stein variational gradient descent. 2024. URL: https://arxiv.org/pdf/2402.01098.
10. Xiao H., Qi L., Shi J., Li S., Tang R., Zuo D., Da B. Reliability assessment of ship lubricating oil systems through improved dynamic Bayesian networks and multi-source data fusion. *Applied Sciences*. 2025. Vol. 15. No. 10. P. 5310. DOI: 10.3390/app15105310.
11. Lin S.-L. Application combining VMD and ResNet101 in intelligent diagnosis of motor faults. *Sensors*. 2021. Vol. 21. No. 18. P. 6065. DOI: 10.3390/s21186065.
12. Xie J.L., Shi W.F., Xue T., Liu Y.H. High-resistance connection fault diagnosis in ship electric propulsion system using Res-CBDNN. *Journal of Marine Science and Engineering*. 2024. Vol. 12. No. 4. P. 583. DOI: 10.3390/jmse12040583.
13. Vychuzhanin V.V., Vychuzhanin A.V. Integrated approach to creating a case-based database for diagnosing failures in ship power plants. *Informatics and Mathematical Methods in Simulation*. 2025. Vol. 15. No. 2. P. 155–165. DOI: 10.15276/imms.v15.no2.155.
14. Hancock J.T., Khoshgoftaar T.M. CatBoost for big data: an interdisciplinary review // *Journal of Big Data*. 2020. Vol. 7. Article No. 94. DOI: 10.1186/s40537-020-00369-8.
15. Burden R.L., Faires J.D. *Numerical Analysis*. 10th ed. 2016. DOI: 10.13140/2.1.4830.2406.
16. Saltelli A., Tarantola S., Campolongo F., Ratto M. Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models. 2004. DOI: 10.1002/047087095.
17. Saltelli A., Sobol I.M. Sensitivity analysis for nonlinear mathematical models: numerical experience. *Matematicheskoe Modelirovanie*. 1995. Vol. 7. No. 11.
18. OREDA.Offshore Reliability Data Handbook. 6th ed. OREDA, 2015.

АДЕКВАТНІСТЬ ТА ВЕРИФІКАЦІЯ ІНТЕЛЕКТУАЛЬНОЇ ДІАГНОСТИЧНОЇ МОДЕЛІ СУДНОВИХ ЕНЕРГЕТИЧНИХ УСТАНОВОК

В. В. Вичужанін, А. В. Вичужанін

Національний університет «Одеська політехніка»
1, Шевченка пр., Одеса, 65044, Україна
Email: v.v.vychuzhanin@op.edu.ua

Сучасні суднові енергетичні установки (СЕУ) функціонують в умовах агресивного, змінного та високонавантаженого морського середовища, що вимагає застосування сучасних діагностичних засобів для раннього виявлення відмов і прогнозного технічного обслуговування. У статті запропоновано інтегровану інтелектуальну модель діагностики, яка поєднує методи машинного навчання, імовірнісне висновування та евристичну логіку для забезпечення як статистичної точності, так і інженерної адекватності. Модель має трирівневу архітектуру: навчальний компонент (CatBoost і нейронні мережі), імовірнісний компонент (байєсівські мережі та марковська логіка) та евристичну систему на основі прецедентів (CBR) з експертними правилами. Для навчання та валідації використано понад 22 000 реальних і змодельованих випадків. Зважене агрегування результатів усіх компонентів забезпечує надійне прогнозування ймовірності відмов. Кількісна оцінка за метриками MAE, RMSE, R^2, recall і F1-score показує, що гібридна модель суттєво перевершує окремі компоненти, досягаючи точності 87,2% на незалежній тестовій вибірці. Статистична перевірка за допомогою t-критерію, χ^2-аналізу та довірчих інтервалів підтверджує перевагу інтегрованого підходу. Аналіз чутливості демонструє стійкість моделі до змін параметрів, шуму у вхідних даних та обсягів навчальної вибірки. Найбільшу чутливість модель виявляє до температурних і вібраційних даних, що узгоджується з інженерною логікою. Після 10 000 зразків спостерігається ефект насичення, що свідчить про досягнення порогу інформативності для заданої архітектури. Прогнози моделі тісно корелюють із фактичними даними про відмови ($R = 0.99$), що підтверджує її адекватність, інтерпретованість і практичну цінність. Додаткове тестування зі штучним шумом підтверджує стабільність моделі в умовах реальних сенсорних похибок. Запропонована модель виступає як надійний інструмент підтримки прийняття рішень для діагностики та прогнозування технічного стану СЕУ, здатна виявляти як типові, так і каскадні відмови. Це дослідження робить внесок у розвиток інтелектуальних систем моніторингу в морській галузі та складних промислових середовищах, пропонуючи комплексну основу для технічного обслуговування за станом і зниження експлуатаційних ризиків.

Ключові слова: інтелектуальна діагностика, прогнозування відмов, суднові енергетичні установки, гібридна модель, байєсівські мережі, машинне навчання, метод прецедентів.