UDC 004.896
IRSTI 28.23.27

https://doi.org/10.55452/1998-6688-2024-21-3-90-115

[1]Samigulina Z.I.,
PhD, Professor, ORCID ID: 0000-0002-5862-6415,
e-mail: z.samigulina@kbtu.kz
[1]*Baikadamova S.S.
Master student, ORCID ID: 0009-0003-1734-8548,
*e-mail: tassbulatova@gmail.com

[1]Kazakh-British Technical University, 050000, Almaty, Kazakhstan

THE INFLUENCE OF DATA SAMPLING ON SOLVING THE PROBLEM OF PATTERN RECOGNITION FOR DIAGNOSTICS OF INDUSTRIAL EQUIPMENT

Abstract

With the sophisticated technology that modern industrial organizations are equipped with, state prediction and diagnostics are essential duties. The current research aims to develop a more accurate modified artificial intelligence system for industrial equipment diagnostics in the oil and gas industry. Researching faulty signals and processing methods utilized by equipment in the oil and gas industry, as well as assessing the advantages and disadvantages of different signal extraction strategies, are the first steps in the process. The second is the application of artificial intelligence to decision-making and equipment defect detection. This method widely used by the oil and gas sectors to lower equipment failure rates. The recommended diagnostic system helps organizations reduce the financial risks associated with equipment defects by increasing production dependability, enabling for maintenance planning, predicting probable failures, and expediting equipment repairs. The article is devoted to the study of the data sampling influence on the classifier's predictive ability in diagnosing of the industrial equipment. Various types of data samples were considered, such as: simple random sample, cluster sample, systematic sample. According to the results of listed data samples were built classifiers based on particle swarm optimization and ensemble models (bagging and voting type). The best results were achieved using the systematic sampled dataset and an ensemble modeling strategy with voting, which combines forecasting based on a neural net, gradient boosted trees and naive Bayes models: accuracy 93.6%; classification error 8%; recall 94.32%; precision 93.87%. The resulting best strategy for diagnosing equipment based on data sampling and an ensemble model was used for implementation in FMEA (Failure Mode and Effects Analysis) technology in order to obtain an improved version, which is adapted for working with big data.

Key words: industrial equipment diagnostics, data sampling, simple random sample, cluster sample, systematic sample, particle swarm optimization, ensemble methods, FMEA improved model.

Introduction

Currently, the digitalization of production is actively developing, which in turn makes it possible to introduce new innovative techniques for data processing into real industrial systems. Large oil refineries built on the basis of modern microprocessor technology with a remote data input/output architecture generate a huge amount of production information per day, the analysis of which in real time is not possible. Most of the information is archived and is not used in any way to predict system behavior. A simple programmable logic controller can handle up to 200 points, while scanning times vary from 10 seconds to 1 minute. Thus, the automation reads 36,000 data per hour. For large distributed control systems, the number of points is about 2,000, depending on the type of production, which leads to exponential growth in the generated data. In order to make a correct forecast, experts recommend analyzing data for the year, taking into account the seasonality of work, failures and various modes, which leads to a huge amount of production data. Classic statistical algorithms for

data processing do not give the desired results, since behind the data there are specific units and their states, an error in the analysis of which can lead to an emergency stop at the enterprise.

To solve such problems, large oil companies are actively implementing artificial intelligence that can make such appropriate data sampling in order to maintain the dynamics of the industrial system's behavior and maintain high accuracy analysis.

A common task is timely intelligent diagnostics of equipment to prevent costly downtime in enterprises. However, not all artificial intelligence methods can successfully cope with the processing of production data, which has its own specifics. Issues of equipment diagnostics, risk assessment and failure severity are resolved using methods such as HASOP (Hazard and Operability Study), HASID, FTA, FMEA (Failure Mode and Effects Analysis), etc. These methods allow to eliminate errors at an early stage, also they are simple and intuitive, and do not contain complex calculations, but can be labor-intensive for large systems. Errors accumulate in a redundancy system when several levels of a hierarchical system need to be monitored, and one big problem is the lack of ability to assess the reliability of the entire complex system.

The problem can be solved using an extended FMEA equipment failure model using modern predictive analytics methods. However, to integrate artificial intelligence into the FMEA model, it is necessary to develop the most efficient intelligent algorithm that can deal with a large amount of production data [1, 2].

Thus, it is relevant to develop new, modern methods for processing industrial data for diagnosing expensive equipment in order to minimize economic costs in production and timely maintenance of a complex technical equipment.

Currently, artificial intelligence methods are actively used to analyze production data. For example, in work [3] a system of preventive repair and diagnostics of various types of equipment is considered. Neural network technology is used to recognize bearing defects. A convolutional neural network and wavelet transform are used to obtain 2D images of vibration signals. Research [4] focuses on the detection and classification of electric motor faults to ensure equipment operational integrity. A data preprocessing method based on wavelet transform and short-time Fourier transform is used to identify hidden patterns in the vibration data of electric motors. Data analysis was carried out based on a convolutional neural network. The study highlights the effectiveness of combining artificial intelligence and data preprocessing techniques for diagnosing equipment failures. The work [4] covers research on swarm intelligence algorithm based on differential evolution (DE) and gray wolf optimization (GWO) using a convolutional neural network classifier to diagnose rolling bearing faults. An improved 1D-CNN algorithm with hyperparameter optimization is proposed. The effectiveness of the algorithm was confirmed by research on the databases of Case Western Reserve University (CWRU) and Jiangnan University (JNU). Study [5] examines data collected from protective relays to diagnose faults in power system components. The authors presented a hybrid support vector machine and artificial neural network (SVM-ANN) model for fault diagnosis.

Currently, combined algorithms based on several artificial intelligence algorithms have shown high efficiency. For example, [6] discusses a deep learning method for diagnosing faults in rolling bearings using artificial immune systems. An adaptive enhanced deep convolution algorithm is used to extract the feature set. Research [7] is devoted to the analysis of the hybrid CSA-DEA method (Clonal Selection Algorithm with a Differential Evolution Algorithm) for flaw detection of structures with cracks. The input data of the hybrid system is the relative frequency values of the damaged structure, and the output data is the relative crack locations and depth. The work [8] considers a bioinspired anomaly detection algorithm based on the surface dendritic cell algorithm (Cursory Dendritic Cell Algorithm, CDCA). Immune models such as the Dendritic Cell Algorithm (DCA) are a promising area of research, since the problem of detecting abnormalities in CPPS is reminiscent of the work of dendritic cells in protecting the human body from dangerous pathogens. A new variant of DCA, the CDCA algorithm, is proposed, designed for continuous monitoring of industrial processes and online anomaly detection. The experimental results showed their effectiveness and were carried out on the basis of two industrial datasets for detecting physical anomalies and network intrusions (Skoltech Anomaly Benchmark (SKAB) and M2M using OPC UA).

A promising area of research is the use of ensembles of models for equipment diagnostics [9]. For example, [10] solves the problem of bearing fault detection using a graphical autoencoder and ensemble learning. A new approach to bearing fault detection based on graph neural networks and ensemble learning is proposed. The study [11] focuses on the development of a weighted ensemble model based on LightGBM for fault diagnosis with a small number of pulses. With an ensemble, the potential diagnostic error of each classifier can be reduced, thereby increasing the generalization ability of the entire model. Work [12] considers a hybrid method for diagnosing faults in power transformers based on the classification of ensemble trees and training subsets using Rogers and Gouda coefficients. The study [13] proposed an ensemble prediction model for analyzing equipment failures in the oil and gas industry. The work implements an ensemble approach to combine different classifiers in order to improve the performance of the SVM training classifier.

Modifications based on artificial intelligence of industrial equipment diagnostic models are interesting. For example, [14] discusses the role of artificial intelligence in improving the efficiency of failure modes and effects analysis (FMEA). The study [15] presents a modification of the FMEA model by replacing the expert assessment of the "probability of occurrence" criterion with a machine learning model based on the support vector machine. Work [16] is devoted to an intelligent model for FMEA risk assessment based on fuzzy logic, nearest neighbor method and support vector machine.

The stable and safe operation of mechanical equipment is becoming increasingly significant in current industry, which aims to reduce unnecessary routine shutdown, maintenance costs and even sudden person casualties [17]. Thus more and more attention has been paid to fault diagnosis of machinery. Meanwhile, with the rapid development of Internet of Things, sensing technology, and big data, a new revolution is quietly sprouting up in this field, in which a major feature is ever-increasing mass of data [18].

In general, majority of industrial equipment lacks the capability of built-in self-diagnostics and prognostics. In addition, the nominal characteristics, along with the failure modes, change over time due to wear off, maintenance, and the repair/replacement of parts and components. This reality calls for alternative approaches that can minimize the need for analysis of the specific machine failure modes [19]. Fault diagnosis approaches can be classified into three categories: model-based [20, 21], signal-based [22–23], and data-driven approaches [24–25]. In model-based approach, focus is on establishing mathematical models of complex industrial systems. These models can be constructed by various identification methods, physical principles, etc. Signal-based approach uses detected signals to diagnose possible abnormalities and faults by comparing detected signals with prior information of normal industrial systems [26]. Usually, difficulty occurs in building accurate mathematical models and obtaining accurate signal patterns for complex industrial and process systems. Data-driven fault diagnosis approach requires large amount of historical data, rather than models or signal patterns [27]. Therefore, data-driven methods are suitable for complex industrial systems.

Over the past few years, there has been substantial advancement in the field of applying AI algorithms for diagnostics. The oil and gas business may now use less time and money thanks to the adoption of artificial intelligence techniques. By adopting the appropriate neural algorithms, machine learning has advanced this field and plays a crucial role in diagnosing machinery and correctly forecasting results said in their work Andrey Ostroukh, Leonid Berner, Maria Karelina [28]. Condition monitoring and fault diagnostic systems are crucial for lowering the likelihood that this equipment may malfunction. In this work [29] Stefania Santini, Francesco Flammini - students of Malardalen university in Sweden, University of Naples Federico 2nd, Italy, conducted defect diagnostics in rotating machinery using artificial intelligence, signal processing, and permutation entropy.

Only the initial phase in condition monitoring and maintenance involves the collecting of signals. In order to identify fault formation, it is also important to analyze the gathered monitoring data and extract features.

Finally, artificial intelligence models and methodologies are utilized to detect and forecast defects based on the retrieved information. Traditional, contemporary, and intelligent diagnosis techniques all fall under the category of signal processing technologies [30–31].

The obtained results require further study at the level of management decisions on the state of the equipment for the current production [32]. These days, using a project-oriented approach to manage an industrial organization is frequently linked to such solutions. In that case, various techniques were explored for the development of the classificatory according to their pros and cons in the theoretical analysis provided. For example, the Ensemble voting method provides numerous benefits within industrial equipment diagnostic systems. Benefits of using the ensemble voting technique [33–34]: Increased precision: when multiple models are combined, collective voting typically yields more stable and precise predictions compared to using individual models alone. Also, there is a greater complexity, which is combining various models and developing synthesis strategies may lead to an increase in the computational complexity and resource demands of the diagnostic system [35–36].

Thus, an analysis of the literature proves the relevance of developing new equipment diagnostic methods both for analyzing databases of equipment failures and for improving existing diagnostic models.

The problem statement of the research is formulated as follows: it is necessary to develop an intelligent algorithm for diagnosing industrial equipment, taking into account the large amount of production information and the possibility of integration into the FMEA equipment diagnostic model.

Materials and Methods

The initial dataset of the research is full with unsystematic and unclear data and values, which is also consisted of enormous amount of information. In this section, all the data preparation methods and further classification development techniques are provided.

Sampling methods

A set of data plays a huge role in solving problems associated with real industrial production. During the operation of control systems, microprocessor technology generates a huge amount of production data, so the average time for sensors' information scanning by the controller is 20 m/s, while if a programmable logic controller on a production line serves up to 200 points, then the size of the technological process observation database per day will be enormous in size. Thus, the issue of reducing the dimensionality of the source data and correct data sampling is an urgent task.

In order to observe what kind of sampling is better and more efficient for particle swarm optimization and machine learning methods, it is necessary to choose correct data sampling types. All of them are probability sampling methods, because they are the best solutions for quantitative research. Let's get through each of them briefly.

Simple random sample

There are lots of sampling methods, for instance Simple random sample. Simple random sample is a type of sampling where all the dataset information used and every member of that data has a chance to be selected for further manipulations. The sampling frame must include the entire dataset information. To perform this type of sampling, usually used such tools as random number generators or other techniques that rely entirely on chance.

In Python there is a specific library to perform such sampling methods. Algorithm 1 for the simple random sampling is as follows:

Algorithm 1. Data sampling method: simple random sample.

Step 1. Read the csv file with the number of equipment characteristics.

Step 2. Import Python library for sampling techniques by "import random".

Step 3. Declare the range of the initial dataset.

Step 4. Specify the random sample size as 100.

Step 5. Perform Simple Random Sampling by the following "random.sample(x,y)" command.

Step 6. Get the result of the random sampling.

Here demonstrated the simple explanation of the random sampling method (Figure 1).

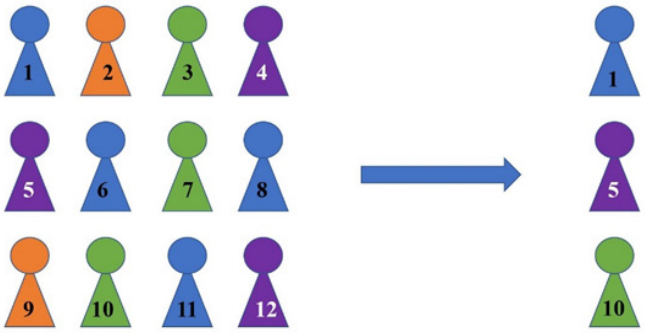


Figure1 – Simple random sampling illustration

It is clear from the Figure1 that by using random sampling technique it could be guaranteed that every member of the population or dataset can be chosen for the sampling group without any systematic or statistical rules.

Cluster sample

Another one frequently used data sampling method is cluster sampling. Cluster sampling is similar to stratified sample by dividing the whole data into subgroups. However, in cluster sampling each subgroup must have similar characteristics to the entire sample. And also, by using that method it randomly selects entire subgroups not individuals. The Algorithm 2 for Cluster sampling is described below:

Algorithm 2. Data sampling method: cluster sampling.
Step 1. Read the csv file with the number of equipment characteristics.
Step 2. Import Python library for sampling techniques by “import random”.
Step 3. Declare the range of the initial dataset.
Step 4. Define the number of clusters and cluster size.
Step 5. Perform randomly selecting of some clusters.
Step 6. Create a list to store the cluster samples.
Step 7. Sample all elements within the selected clusters.
Step 8. Get the result of the cluster sampling.
In Figure 2 presented the structure of the cluster data sampling method.

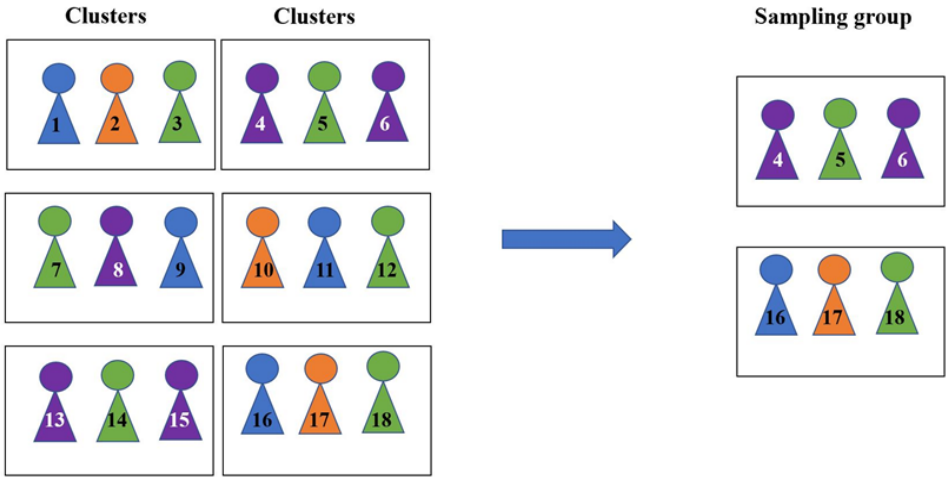


Figure 2 – Cluster sampling illustration

This data sampling method could be used in industrial equipment diagnostic systems.

Systematic sample

Systematic sample is quite similar to simple random sampling, but is usually a little easier to perform. Each member of the population: in our terms is each failure of the equipment is numbered and then individuals are selected at regular intervals not by a random generator. Systematic sampling method's Algorithm 3.

Algorithm 3. Data sampling method: systematic sample

Step 1. Read the csv file with the number of equipment characteristics and machine values.

Step 2. Import Python library for sampling techniques by the following "import random".

Step 3. Declare the range of the initial dataset.

Step 4. Specify the random sample size as 100.

Step 5. Calculate the sampling interval.

Step 6. Perform Systematic Sampling.

Step 7. Get the result of the random sampling.

The systematic sampling method is more effective in this study, because of its interval dividing technique, which gives an opportunity to analyze each equipment's value and not drop any important data. In Figure 3 the systematic sampling method is illustrated.

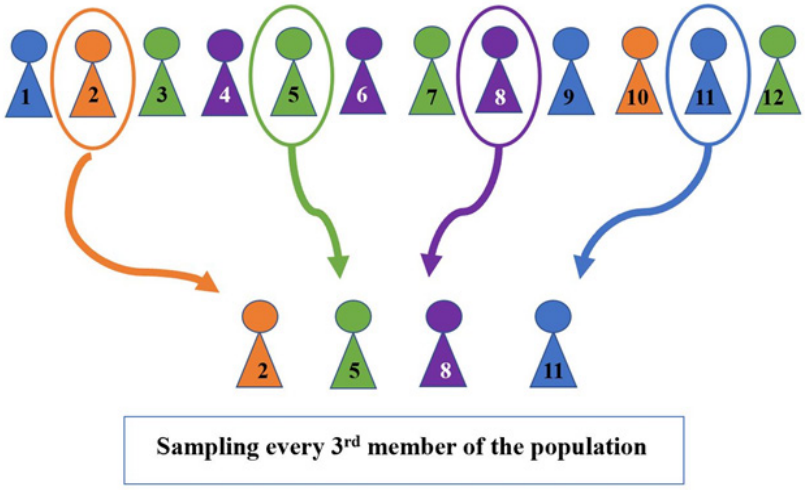


Figure 3 – Systematic sampling illustration

The advantages of this method are its simplicity and quick implementation.

Pattern recognition methods

As pattern recognition methods were applied Particle swarm optimization (PSO) and Ensemble method. Those methods are better explained in following paragraphs.

Particle swarm optimization

Particle swarm optimization (PSO) is one of the bioinspired algorithms, and it searches the solution space for the best possible solution in a straightforward manner. It differs from other optimization techniques in that it does not depend on the gradient or any differential form of the objective and simply requires the objective function. There are also not many hyperparameters.

A particle swarm optimization operates in this manner: it begins with a number of random locations on the plane (referred to as particles) and let them search for the minimum point in random directions. Every particle should look around the lowest position it has ever found as well as the lowest point the entire swarm of particles has ever found at each step. Regard the minimal point of the function to be the least point that this swarm of particles has ever investigated after a specific number of iterations. For better understanding in a practical vision here the pseudocode of the PSO technique is described below [37].

Pseudocode of the PSO algorithm:

Input: a_{nm} – position of the particle; v_{nm} – velocity of the particle; i – number of iteration;
Output: the optimal solution for the particle position

For each particle n
For each dimension m
Initialize position a_{nm} randomly with possible interval
Initialize velocity v_{nm} randomly with possible interval

End_for
End_for

Iteration i = 1 Do
For each particle n calculate suitable value
If the suitable value is better that $TheBest_{nm}$ in history
Set current suitable value as the $TheBest_{nm}$

End_if
End_for

Chose the particle with the best suitable value as the $TheBest_m$

For each particle n

For each dimension m calculate velocity with the equation

$$\vartheta_{nm}^{t+1} = \omega^t \vartheta_{nm}^t + \varphi_{nm}^t (TheBest_{nm}^t - x_{nm}^t) + \varphi_{2m}^t (TheBest_n^t - x_{nm}^t)$$

Update particle position with the equation:

$$\alpha_{nm}^{t+1} = x_{nm}^t + \vartheta_{nm}^{t+1}$$

End_for

End_for

i = i+1

While max iteration or min error criteria are not attained.

Particle Swarm Optimization is a metaheuristic optimization algorithm inspired by the social behavior of flocks of birds and schools of fish. It is often used to solve optimization problems, such as those encountered in research on diagnostics of industrial equipment. In this context let's present some advantages and disadvantages of using PSO.

Advantages of the method [38]:

1. Global Optimization: PSO is good at finding global optimization in complex search spaces. In the context of diagnostic methods PSO could effectively explore the space of possible diagnostic models or parameters to find the optimal solution.

2. Simple Implementation: PSO is relatively easy to implement and requires minimal parameter tuning compared to other optimization algorithms. This is beneficial in research environments where time and resources are limited.

3. Convergence Speed: often converges to a solution relatively quickly, especially for simple or moderately complex optimization problems. This is beneficial when working with large datasets or performing repeated experiments in device diagnostic studies.

4. Robustness: tends to be robust to noise and is less likely to fall into local optima compared to other optimization algorithms. This is important in research on diagnostics of industrial systems where noisy sensor data and complex interactions between components can make optimization problems difficult.

5. Scalability: it could be easily parallelized, allowing efficient optimization on high-performance computing platforms and distributed systems. This is advantageous when dealing with large industrial facilities or performing experiments that require large amounts of calculations.

Disadvantages of the method:

1. Premature Convergence: PSO prematurely converges to suboptimal solutions, especially in highly multimodal optimization environments or misleading optimization landscapes. This can be problematic in device diagnostic studies, where the lack of a global optimum can result in inaccurate or unreliable diagnostic models.

2. Limited Search: you can have difficulty effectively exploring the search space, especially when dealing with high-dimensional or non-convex optimization problems. This limitation could impact the ability to find optimal diagnostic models or parameters in complex industrial systems.

3. Parameter Sensitivity: Although PSO requires fewer parameters to be tuned than other optimization algorithms, PSO's performance is influenced by the selection of parameters such as inertia weights and acceleration factors. It may be easy to receive. Finding optimal parameter settings may require experimentation and can be computationally intensive in research environments.

4. Noisy optimization: PSO may not perform well for optimization problems with noisy or uncertain objective functions. In research on diagnostics of industrial equipment, sensor data can be noisy or incomplete, and this limitation can affect the reliability of diagnostic models obtained with his PSO.

5. Lack of Guarantees: method does not guarantee convergence to the global optimum, especially in non-convex or discontinuous optimization landscapes. This lack of warranty can be problematic in critical industrial applications where accurate diagnostics are essential for safety and efficiency.

Overall, PSO could be useful tool when considering diagnostic methods for industrial equipment, offering advantages such as global optimization capabilities, ease of implementation, and speed of convergence.

However, researchers should be aware of its limitations, especially regarding premature convergence, limited exploration, and sensitivity to parameters, and carefully consider whether PSO is suitable for a particular optimization problem.

Ensemble (bagging and vote types)

Ensemble techniques are a significant technique in computer science and machine learning that combines numerous base models to produce a better, more robust predictive model. These strategies frequently outperform individual models by utilizing the diversity of the basis models and combining their predictions.

Ensemble approaches often rely on a set of basis models, known as weak learners or base classifiers/repressors. These basis models can be the same type (homogeneous ensembles) or different types (heterogeneous ensembles), including decision trees, neural networks, support vector machines, and any other machine learning algorithm [39].

1. Bagging (Bootstrap Aggregating) is the process of training multiple base models independently on distinct subsets of training data (sampled with replacement) and then aggregating their predictions. Random Forest is a common bagging-based ensemble approach that uses decision trees as its basic model. However, in this research Neural Network methods are used for bagging, which is more efficient.

Advantages of ensemble bagging method [40]:

◆ Variance reduction: By training numerous base models on distinct subsets of the training data (sampled with replacement), bagging helps minimize the final model's variance. Each base model learns slightly different parts of the data, resulting in a more robust and stable ensemble.

◆ Bagging improves the generalization performance of the ensemble model. Bagging decreases the risk of overfitting by aggregating predictions from many models trained on different subsets of data and capturing more generalizable patterns in the data.

◆ Robustness to Noise: Because bagging integrates predictions from numerous models, it is more resistant to noisy or outlier data. Outliers may have less of an impact on the final prediction due to the averaging or voting procedure.

◆ Bagging base model training is easily parallelizable because each model is trained independently. This makes bagging appropriate for distributed computing systems and can result in considerable speedups in model training.

Disadvantages of the Ensemble bagging [41–42]:

◆ Increased Computational Cost: Training multiple base models in bagging can be computationally expensive, especially if the base model is complicated or a high number of models are included in the ensemble. This can limit bagging's scalability in large datasets or resource-constrained contexts.

◆ Loss of Interpretability: Because the ensemble incorporates predictions from numerous models, the final model's interpretability may be lower than that of the individual base models.

Understanding the underlying decision-making process of the ensemble may become increasingly difficult, especially with sophisticated bagging schemes or with a high number of base models.

2. Voting: In classification tasks, each base model predicts a class label, and the final prediction is selected by a majority vote. In this research for voting method were used following techniques: Naïve Bayes, Neural Net and Gradient Boosted Trees.

In that article, our aim is to collect the dataset of equipment failures and apply various sampling techniques to get the desired results. Artificial intelligence and statistical methods are currently used to solve a variety of practical challenges.

3 FMEA model

FMEA, or Failure Mode and Effects Analysis, is table, which contains the structured way to classifying all errors that can occur during the design, manufacture or assembly of a product or service. It is especially useful for identifying and mitigating risks early in the development process [43].

In Table 1 is a breakdown of common FMEA table headings and their meanings represented:

Table 1 – FMEA table representation

Input/Name/or the Step in process	The Failure Mode (Potential)	Failure Effects	Severity	Potential Causes of error	Occurence	Detec-tion	RPN
The process step, name, changes in process, which is under the investigation?	Where and how the process or feature could go wrong?	Explanation of the impact, which can be caused by this failure on the customer or process itself?	Range from 1 to 10	The main possible causes of the failure or error	Range from 1 to 10		Risk Priority Number

There are the explanation of the headers in the Table 1:

◆ Process Step/Input – What device or process step exactly is under the investigation. List of components or functions being analyzed. It could be helpful to identify the exact part of the process which being evaluated and focuses more on the analysis of specific elements.

◆ Potential Failure Mode – Exactly in what step or changes the process go wrong? This feature describes the possibilities in which a component or function device might fail to get the desired outcomes or goals.

Basic for distinguishing the different ways of process failure, which is the primary step in preventing such downfalls.

◆ Failure Effects (potential) – Describes in what ways the failure could impact the operation of the system, including its impact on customer satisfaction, safety, and adherence to regulations. Recognizing the consequences of failures is key in determining which issues to address first, allowing for resources to be allocated to tackle the most important issues.

◆ Severity (1–10) – Typically ranging from 1 to 10, assess the gravity of the impacts caused by each failure mode. It offers a measurement for estimating potential harm, which is vital for evaluating and controlling risks.

◆ Potential Causes – Identifies factors or shortcomings that could result in the failure mode. Identifying reasons for failures is essential in order to create successful prevention plans.

◆ Occurrence (1–10) - Approximates the chances of a failure happening, usually measured on a numerical scale. Assists in determining the probability of failures, which is essential for prioritizing and managing risks.

◆ Detection (1–10) – Rated on a scale of 1 to 10, evaluates the effectiveness of existing controls in identifying or stopping a potential failure mode. Ensures that potential issues are identified early on, reducing risks and enhancing dependability.

◆ RPN – Risk Priority Number – A numerical rating derived from multiplying the ratings for Severity, Occurrence, and Detection. Offers a numerical assessment of the risks linked to every potential mode of failure, helping in determining which risk reduction strategies should be prioritized.

After the data sampling there is need to apply optimization methods on 4 different datasets: initial data without any changes, simple random sampled data, cluster sampled data and systematic sampled data each one respectfully. Then, it could be more convenient to chose the best suitable sampling method for further experiment of developing the classificatory.

Initial dataset explanation

The database is taken from the kaggle equipment diagnostics data repository. It contains machine failures and process characteristics [https://opendatacommons.org/licenses/dbcl/1-0/].

As a dataset there is a csv file with the equipment characteristics and its failure happened or not with the range R = 10x8091, 80910 data attributes. It has columns (headers) as:

◆ Number – unique data identification;

◆ Product ID – consisting of a letter L, M, or H for low (50% of all products), medium (30%) and high (20%) as product quality variants and a variant-specific serial number;

◆ Type – type of the equipment L, M or H (described above);

◆ Air temperature – generated using a random walk process later normalized to a standard deviation of 2 K around 300 K;

◆ Process temperature – generated using a random walk process normalized to a standard deviation of 1 K, added to the air temperature plus 10 K;

◆ Rotational speed – calculated from a power of 2860 W, overlaid with a normally distributed noise;

◆ Torque – torque values are normally distributed around 40 Nm with a SD = 10 Nm and no negative values;

◆ Machine failure – indicates, whether the machine has failed in this particular datapoint for any of the following failure modes;

◆ HDF – heat dissipation failure: heat dissipation causes a process failure, if the difference between air and process temperature is below 8.6 K and the tools rotational speed is below 1380 rpm;

◆ PWF – power failure: the product of torque and rotational speed (in rad/s) equals the power required for the process. If this power is below 3500 W or above 9000 W, the process fails. Here is the initial dataset of the research Table 2.

Table 2 – Fragment of the equipment diagnostic database before applying data sampling

N	Product ID	Type	Air tempe-rature	Process tempe-rature	Rotational speed	Torque	Machine failure	HDF	PWF
1	M14860	M	298.1	308.6	1551	42.8	0	0	0
2	L47181	L	298.2	308.7	1408	46.3	0	0	0
3	L47182	L	298.1	308.5	1498	49.4	0	0	0
…	…	…	…	…	…	…	…	…	…

In our dataset there are 1,000 rows of various failure modes. It would be more appropriate to divide this amount of failure and make a research on every classified part of them.

The 3D scatter plot of the initial database is presented (Figure 4).

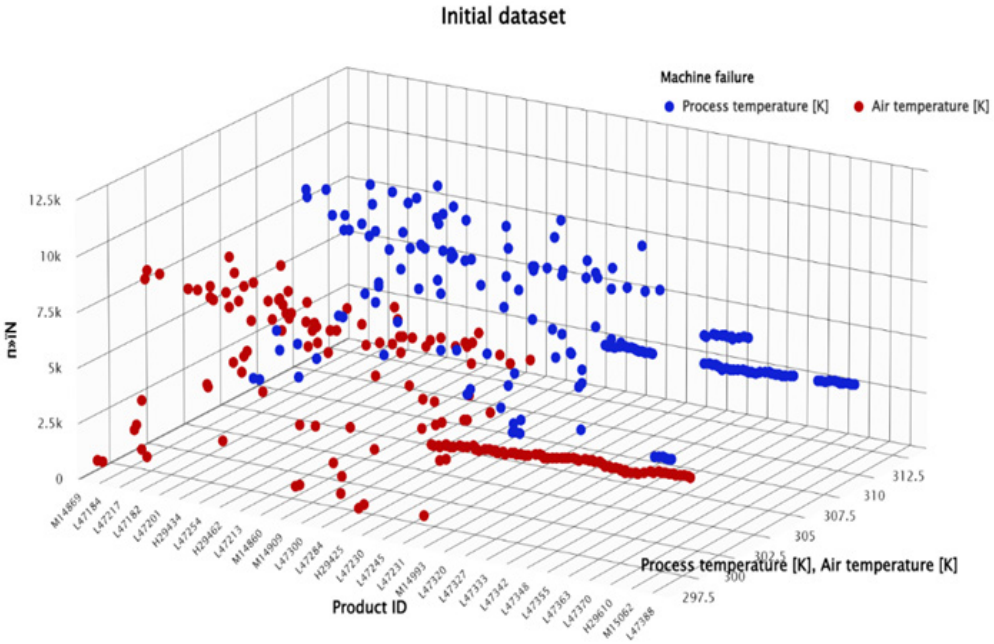


Figure 4 – The 3D scatter illustration of the initial dataset

In the Figure 4 it is observed that the values of the sensors are similar and too many of them are located closer to each other, because the initial dataset is huge and consists of 10,000 values of each equipment on various condition. In order to make an appropriate experiment, it is need to sample this dataset and split it into short but effective ones.

Simple random sampled dataset

Here is the Table 3, where the random sampled data is performed. In this table "N" – is the randomized serial number of the equipment, they are written in random sampled way.

Simple random sampling was implemented by Python algorithm, which was explained on previous sections. Using the random sampling was generated a new dataset with range R = 10x200.

Also, in Figure 5 the 3D scatter plot of the random sampled data is shown below.

Table 3 – Simple random sampling dataset

N	Product ID	Type	Air tempe-rature	Process temperature	Rotatio-nal speed	Torque	Machine failure	HDF	PWF
894	M15753	M	295.7	306.2	1423	42.5	0	0	0
1673	L48852	L	298.1	307.8	1432	49.8	0	0	0
2344	L49523	L	299.1	308.3	1305	61.4	0	0	0
…	…	…	…	…	…	…	…	…	…

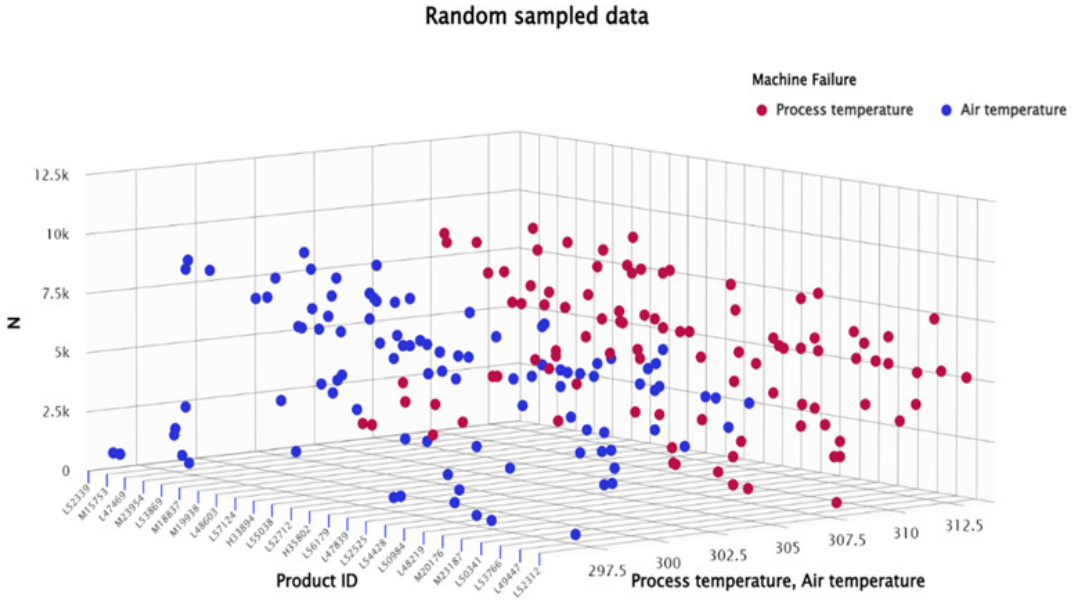


Figure 5 – The 3D scatter illustration of the simple random sampled dataset

In the figure above, it is seen that the data values are located separately from each other, which means that the random sampling is implemented correctly.

Cluster sampling dataset

Cluster sampling was implemented by the algorithm explained on the previous sections in Python platform. In Table 4 the result of the cluster sampling method is shown:

Table 4 – Cluster sampling dataset

N	Product ID	Type	Air tempe-rature	Process tempe-rature	Rotational speed	Torque	Machine failure	HDF	PWF
1	M14860	M	298.1	308.6	1551	42.8	0	0	0
7	L47186	L	298.1	308.6	1558	42.4	0	0	0
8	L47187	L	298.1	308.6	1527	40.2	0	0	0
…	…	…	…	…	…	…	…	…	…

On the Table above it is seen that values are selected differently from random sampling, because in cluster sampling it uses dividing by the cluster group technique and then only the randomize the selected clusters. Cluster sampling method also have generated a new dataset with range R = 10x200. That range is more specific and convenient while using it in further optimization techniques. Here is the 3D scatter representation of the cluster sampled dataset (Figure 6).

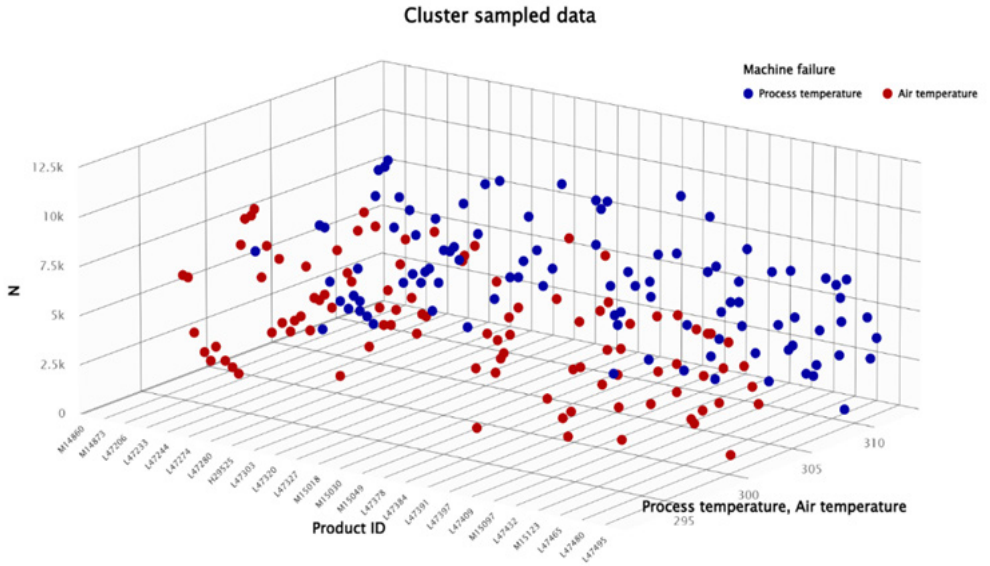


Figure 6 – The 3D scatter illustration of the cluster sampled dataset

In this Figure 6 it is shown that the values of the sensors are located separate but also huddled together in some places as they sampled by clusters. It is necessary to admit that the cluster selected group are randomized before getting the final result, that's why the values on the Figure 6 are separated as random sampling.

Systematic sampling dataset

Systematic sampling was also implemented in Python platform with explained algorithm. Here is the Table 5, where systematic sampled data is presented.

Table 5 – Systematic sampling dataset

N	Product ID	Type	Air temperature	Process temperature	Rotational speed	Torque	Machine failure	HDF	PWF
8	L47187	L	298.1	308.6	1527	40.2	0	0	0
18	M14877	M	298.7	309.2	1410	45.6	0	0	0
29	L47208	L	299.1	309.4	1439	44.2	0	0	0
…	…	…	…	…	…	…	…	…	…

By applying systematic sampling on the initial dataset, it was given a new different dataset with the range R = 10x200 also. In systematic sampling there is need to divide the initial dataset by the specific interval. The 3D scatter plot of the systematic sampled dataset is shown in the Figure 7.

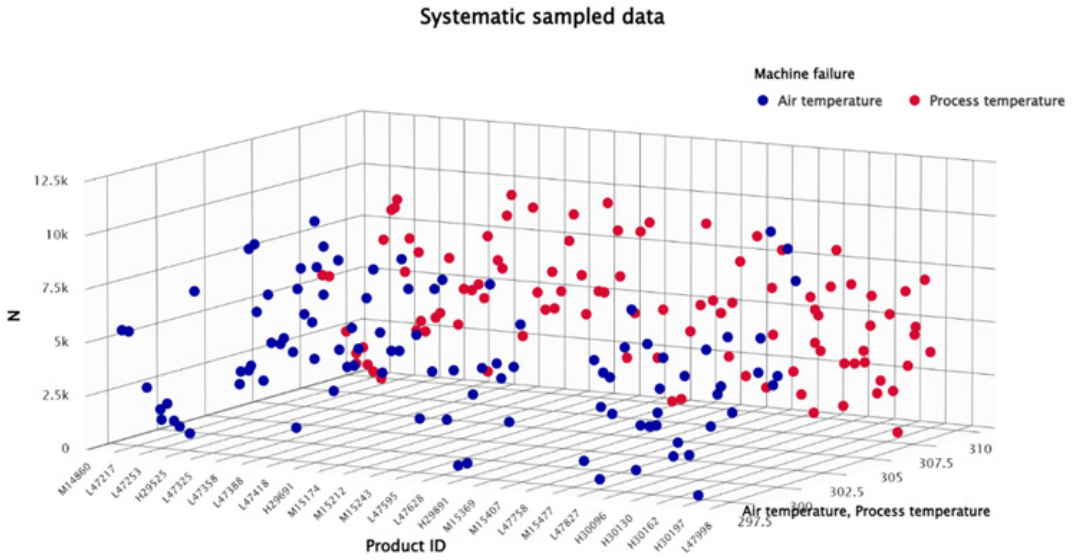


Figure 7 – The 3D scatter illustration of the initial dataset

The systematic sampling is close to the random sampling by the logic but the sampling is implemented by the specific number or character, so each there is more probability that all the values from every condition can be selected.

Results and discussion

After implementing the sampling techniques on the initial dataset, were developed the classifiers based on PSO and ensemble methods, which results are provided in this section.

Development of a classifier based on the PSO algorithm

After applying those three methods of sampling, it is necessary to apply particle swarm optimization on each subgroup of dataset. By comparing them with each other it could be possible to investigate which method of sampling gets better results with PSO techniques. The Figure 8 represents the implementation of the PSO process.

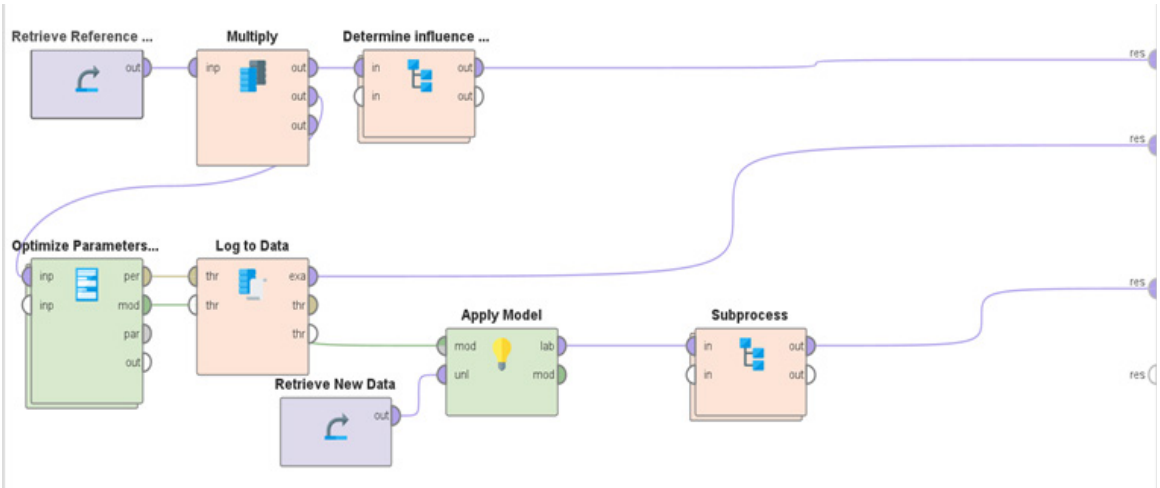


Figure 8 – PSO modeling process

In equipment diagnostic systems it is crucial to determine influence factors of the dataset, because not all the equipment values cause a failure. In PSO modeling the weighting algorithm is used to determine the important factors, which are also presented in Figure 8.

Development of a classifier based on an ensemble bagging method

In Figure 9 represented the bagging modeling process, where except the bagging were used cross validation function to connect with the dataset and its labels.

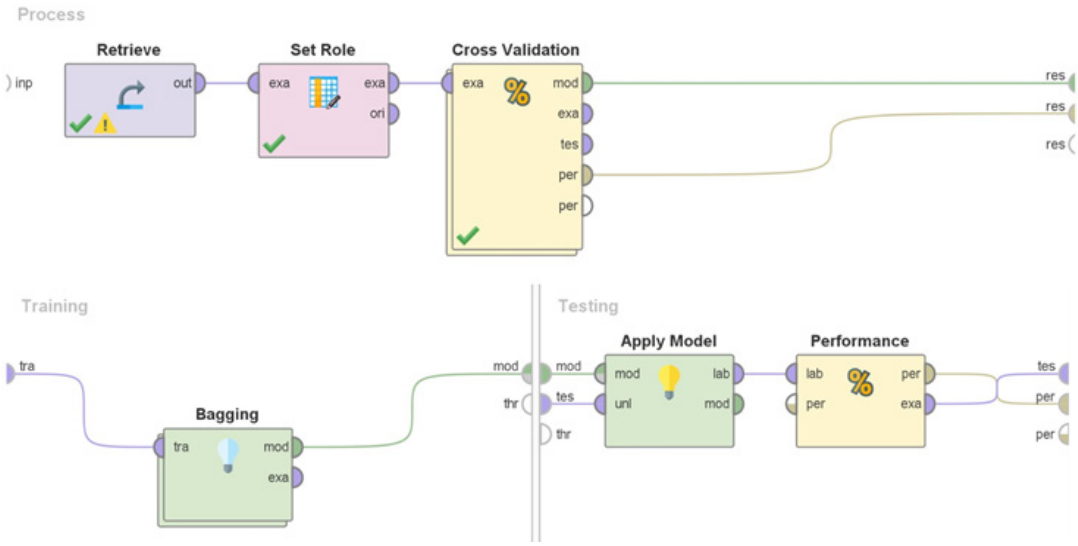


Figure 9 – Ensemble bagging modeling process

In the training of the method are used 100 till 200 training cycles to observe how fast could the method learn by time. All the applied methods' results shown in the next section.

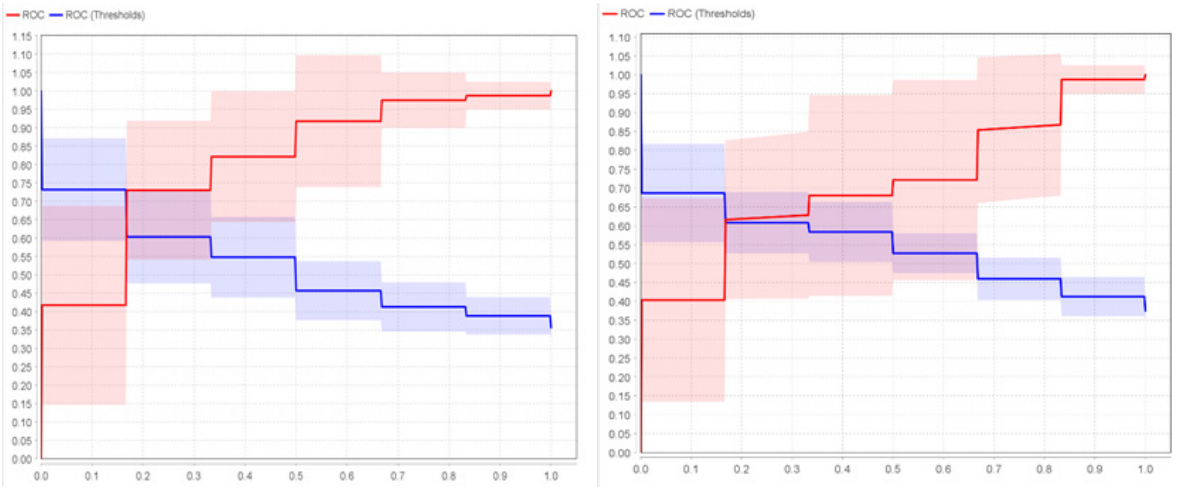
Modelling results and metrics

At first, in the Table 6 there are the results of the PSO and ensemble bagging methods divided by each sampled dataset. The results of modeling and experiments were evaluated based on the following metrics: Accuracy, Classification error, Recall, Precision, Recall. The problem of binary classification was considered, where the first class with the value "1" is the normal operation of the equipment, the second class "0" is the equipment failure.

Table 6 – The results of the PSO and Ensemble bagging methods implemented with initial and sampled datasets

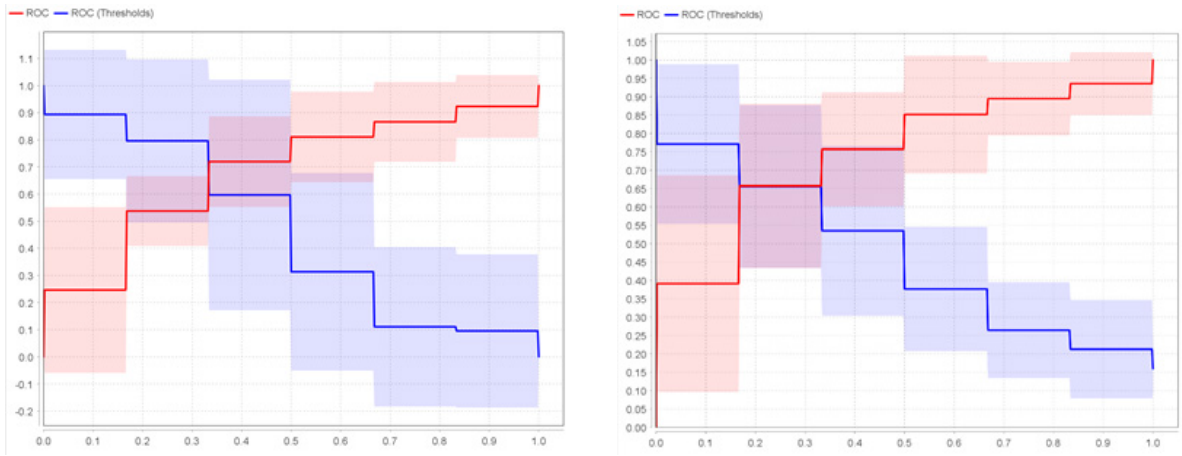
Dataset	Classification algorithm	Accuracy	Classification error	Recall	Precision
Initial raw dataset (before data reduction)	PSO	77,63 %	22,37 %	76%	81%
	Ensemble (bagging)	89,72 %	10,28 %	80%	83%
Dataset after random sampling	PSO	81,88 %	18,12 %	83,06%	83,76%
	Ensemble (bagging)	90,54 %	9,46 %	89%	91%
Dataset after cluster sampling	PSO	85,73 %	14,27 %	89,42%	82,57%
	Ensemble (bagging)	91,62%	8,38%	92,07%	89,5%
Dataset after systematic sampling	PSO	89,17%	10,83%	85%	89%
	Ensemble (bagging)	93,808%	6,19%	88%	91%

The ROC (Receiver Operating Characteristic) comparison is also represented, because it is efficient way to compare the classification methods. The area under the ROC curve (AUC-ROC) is used to quantify a classifier's overall performance. A higher AUC value (closer to 1) suggests that the model distinguishes between positive and negative cases more accurately.



a) b)

Figure 10 – ROC thresholds of a) initial dataset and
b) random sampled dataset by ensemble bagging model



c) d)

Figure 11 – ROC thresholds of c) cluster sampled dataset and
d) systematic sampled dataset by bagging model

In Figures 10 and 11, there is seen that the b) random sampled and d) systematic sampled data's bagging model are more efficient and faster tends to one in its learning model.

Comparing different models: In machine failure prediction research, numerous models are frequently constructed and tested to determine the best successful strategy. The ROC comparison method allows researchers to objectively examine and compare the performance of various models. Researchers can decide which model has the best prediction ability by studying its ROC curves and AUC values.

Ensemble VOTE

Ensemble voting combines the strengths of various distinct models, which may mitigate the faults of any single model. Ensemble approaches, which aggregate forecasts from multiple models, frequently produce more accurate predictions than any particular model alone. In our research were used Neural net, Gradient Boosted trees and Naïve Bayes as a component of the ensemble vote method. In Figure 12 is shown the modeling process of ensemble vote technique.

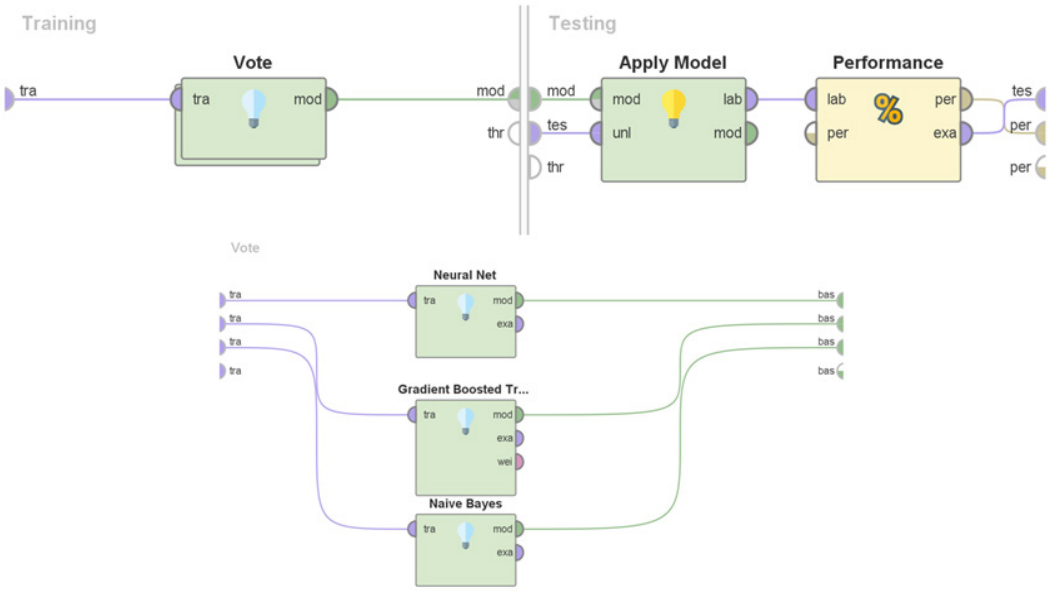


Figure 12 – Ensemble Vote modeling process

On top of the predictions made by the basic learners in its subprocess, this operator applies a majority vote (for classification) or an average (for regression). The Ensemble voting results with initial and sampled dataset are shown in the Table 7.

Table 7 – The results of the Ensemble vote method implemented with initial and sampled datasets

Dataset	Method type	Accuracy, %	Classification error, %	Recall, %	Precision, %
Initial	Ensemble (vote)	86.94	13.06	89.07	85.78
Random sampled	Ensemble (vote)	92.2	7.8	90.06	91.66
Cluster sampled	Ensemble (vote)	90.44	9.56	91.2	91.66
Systematic sampled	Ensemble (vote)	93.6	6.4	94.32	93.87

From the above table it is clear that the ensemble vote methods are more effective than the other, also the combination of the vote method with the systematic sampling techniques gives the best result for the machine failure prediction experiment.

Furthermore, for better experiment results comparison in this research the ROC comparison techniques were extracted from the implemented vote model. The ROC (Receiver Operating Characteristic) comparison is an important technique for evaluating the performance of machine learning models, especially in tasks such as machine failure prediction. It does a thorough examination of the trade-offs between true positive rate (sensitivity) and false positive rate (1-specificity) at various thresholds. Here is the ROC comparison of the best given result - vote model with systematic dataset is shown in the Figure 13.

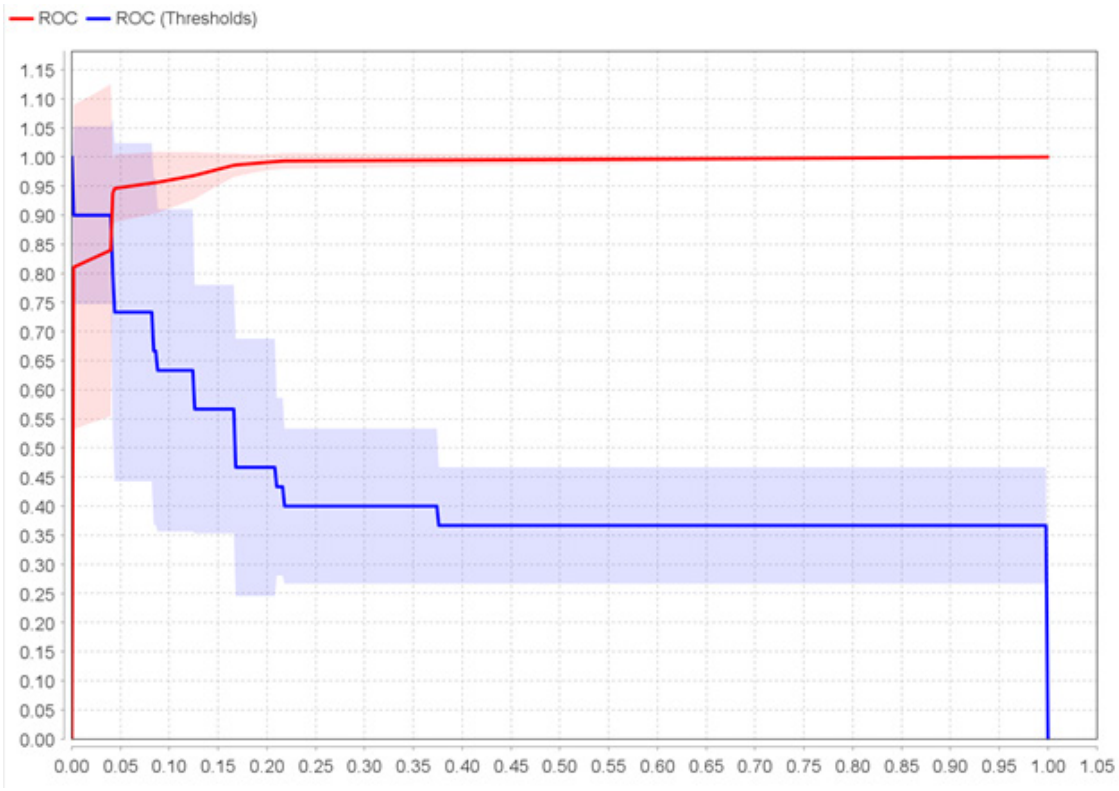


Figure 13 – ROC comparison of vote model with systematic sampled dataset

In this research was implemented Naïve Bayes and Gradient Boosted Trees in the Ensemble vote modeling. Here is the results of the ROC comparison below, where also added the Last Large Margin, Deep learning and Random forest methods just for the comparison (Figure 14).

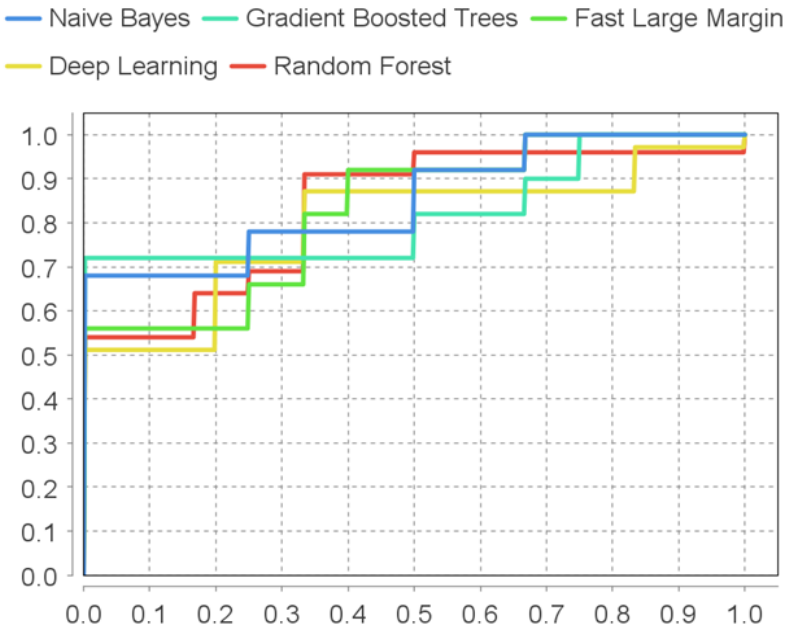


Figure 14 – ROC comparison of various types of classification

As it's illustrated, the ROC curve is a graphical representation of a classification model's performance under different threshold settings. The ROC curve for models below rises sharply from the lower-left corner, showing that it achieves high true positive rates (sensitivity) while maintaining relatively low false positive rates (1-specificity) across various threshold settings. This shows that Naïve Bayes and Gradient boosted trees are highly predictive and successfully distinguishes between positive and negative examples.

Finally, in this research, various machine-learning algorithms for predicting equipment failure in industrial production were studied. Throughout our research, a variety of predictive modeling strategies were investigated, including particle swarm optimization (PSO), ensemble bagging, and ensemble voting. Each technique was assessed based on its capacity to distinguish between normal functioning and failure occurrences, with an emphasis on maximizing predicted accuracy and resilience. Our experiment shows that, while PSO and ensemble bagging showed promise in capturing underlying patterns in the data and generating individual predictive models, the ensemble vote modeling approach emerged as the most effective and reliable method for machine failure prediction in our setting.

The ensemble vote modeling strategy, which combines the predictions of Neural net, Gradient boosted trees and Naïve Bayes models via a voting process, outperformed other methods tested.

Thus, the developed strategy can be applied to production data and integrated into classical equipment diagnostic models, for example, the FMEA model for solving problems associated with large volumes of production data, with the ability to determine the impact of only individual failures, but not their combinations, and also replace the routine work of an expert for automated scanning using artificial intelligence.

FMEA model integration

In this work the integration of FMEA is important, because the effectiveness of using FMEA table in diagnostic processes is crucial. At first, there is need to explain all the table headers and their necessity. After that, by using previous artificial intelligent methods such as Ensemble vote type techniques the FMEA table's "Severity" character would be predicted. For this experiment only systematic sampled dataset is used, because after all previous training sets exactly the systematic sampling method have shown the most accuracy and effectiveness.

Description of the TCO Dataset

The object of the research is unit 300, which's obligation is to purify crude high- and medium-pressure associated petroleum gas coming from unit 200 from hydrogen sulfide H2S, carbonyl sulfide COS and carbon dioxide CO2. The overall structure of the process is shown in Figure 15.

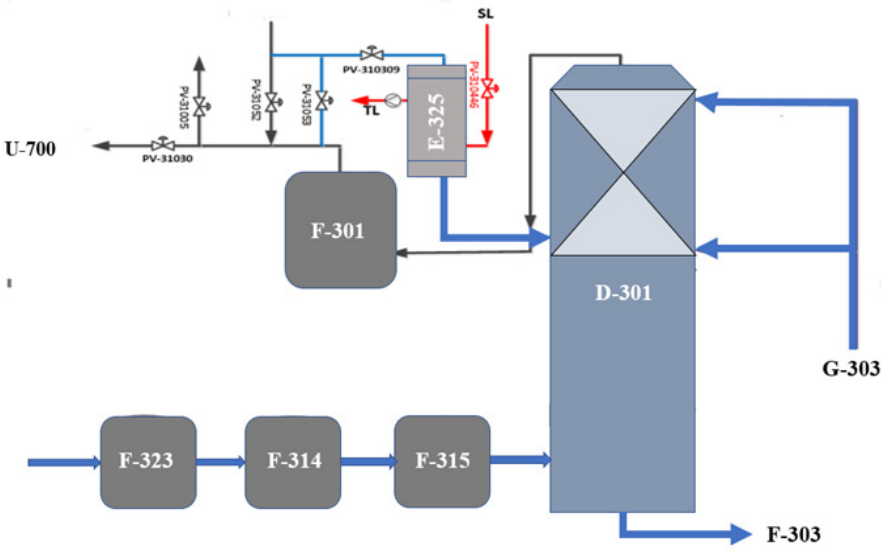


Figure 15 – Gas purification process in high pressure

When cleaning, a selective process of absorption (absorption) of the above components into a diethanolamine solution is used.

The products of installation 300 are:

- purified VD gas – raw material for the gas fractionation unit (U-700);
- purified SD gas returned to U-200 for compression;
- purified LP gas supplied to the U-500 tail gas afterburner as fuel gas (through the exhaust gas collector).
- acid gas with a high content of H2S – raw material for sulfur extraction at U-400.
- Incoming high- and medium-pressure gas streams contain large amounts of moisture and sulfur-containing components.

The dataset for FMEA prediction system is represented in Table 8.

Table 8 – FMEA model for technological process installation 300, TengizChevroil plant

ID	Type	Air temp.	Process temp.	Rota-tional speed	Torque	Process In/ Out	Potential Failure mode	Potential failure effects	SEVE-RITY	Potential causes
M14860	M	298.1	308.6	1551	42.8	F-301- High Pres. purified gas sep.	FAL - 03004 Low alarm output flow	Material without reworked	6	Control loop failure
L47187	L	298.1	308.6	1527	40.2	Trans-port material	Scratched material or Damaged material	Material without reworked	6	Moving Materials
M14877	M	298.7	309.2	1410	45.6	Anodization	FAL - 03004 Low alarm output flow	Material without reworked	7	Handling materials
L47217	L	298.8	308.1	1439	39.1	F-301- High Pres. purified gas sep.	FAL - 03004 Low alarm output flow	Material without reworked6	6	Control loop failure
M14904	M	298.8	308.1	1472	47.5	Trans-port material	Scratched material or Damaged material	Material without reworked	8	Moving Materials

The following Figure 16 represents the main Ensemble vote modeling process for FMEA table prediction:

FMEA intellectual diagnostic results

Let’s consider the results of modifying the classical FMEA model using artificial intelligence based on the strategy discussed in sections 5.1–5.4.

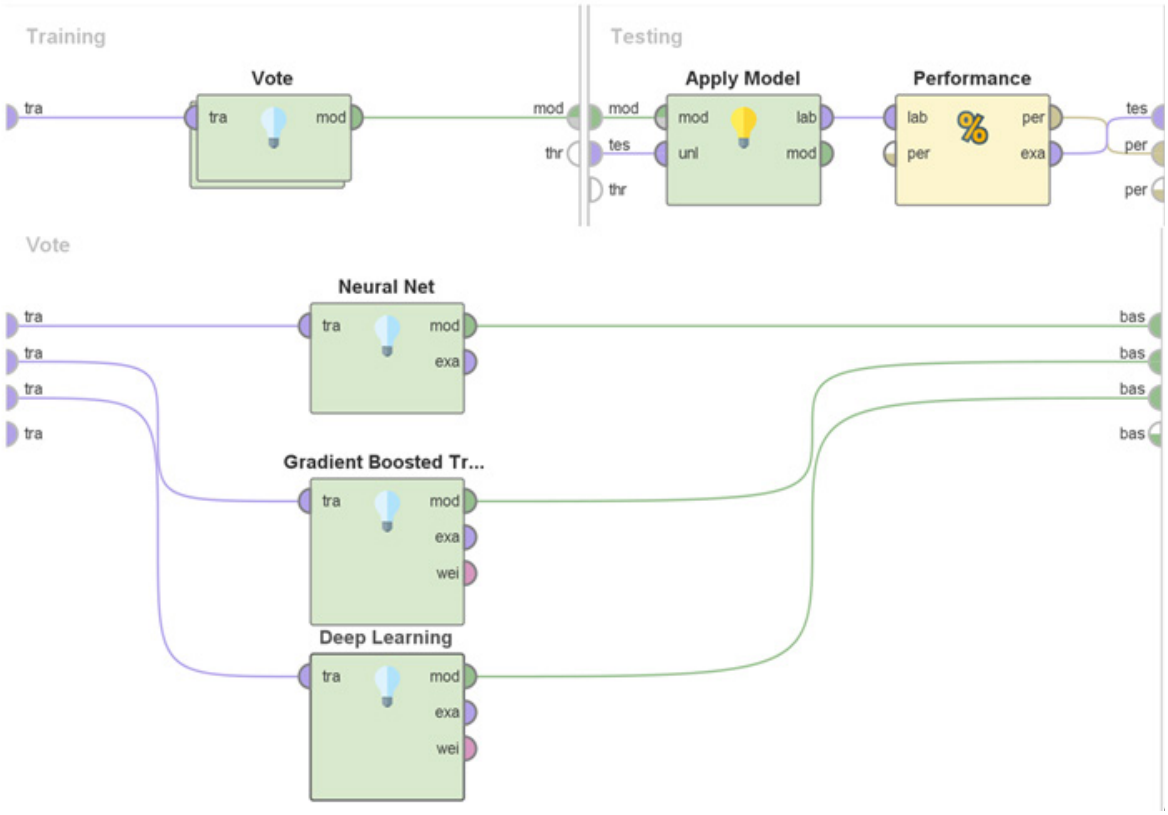


Figure 16 – Ensemble Vote type modeling for FMEA table Severity prediction

Table 9 – the classification results of FMEA Severity prediction system:

Dataset	Method type	Accuracy, %	Classification error, %	Recall, %	Precision, %
Systematic sampled FMEA table dataset	Ensemble (vote) – Neural Net, Gradient boosted tree, Deep learning	88.75	11.25	92.11	91.24

The table above displays the classification outcomes of a combined model that combines the predictions of a neural network, a gradient-boosted tree, and a deep learning algorithm. This method, also known as a voting ensemble or ensemble classifier, combines different predictive models to potentially enhance the overall accuracy and reliability of the predictions. Here is an in-depth analysis of the metrics that have been given:

In general, the ensemble voting model shows strong performance in all analyzed metrics, indicating its efficiency in managing the predictive tasks related to FMEA. The model performs well in both recall and precision, indicating it is effective at detecting positive cases and accurately predicting them. This makes it a dependable tool for predictive analytics in safety-critical situations.

In the following Figure 16 is shown the AUC-ROC comparison of the “Severity” coefficient’s prediction model:

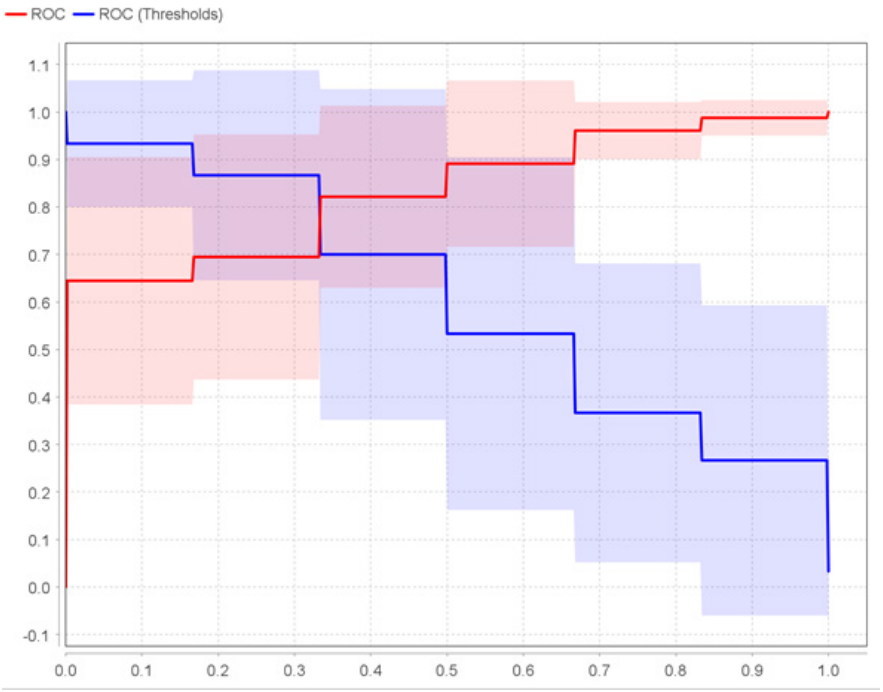


Figure 17 – AUC (optimistic) of FMEA Severity prediction process

In studies on predicting machine failures, researchers often create and evaluate various models in order to identify the most effective strategy. Using the ROC comparison method enables researchers to conduct a fair and unbiased evaluation of different models' performances. By analyzing the ROC curves and AUC values, researchers can determine which model has the highest predictive accuracy.

The process of integration of the FMEA table's coefficient prediction achieved dignified results, which means that the ensemble vote model with the combination of gradient boosted tree, neural net and deep learning gets the efficient and correct predicting values for the "severity" coefficient.

Conclusion

Industrial automation systems are characterized by a large amount of production data generated in real time. For example, the scanning cycle of a programmable logic controller averages 20 ms; if the control loop contains 200 points, then the automation system reads 12,000 data per minute. Most of the generated data is archived and is not used to predict the condition of equipment due to the large dimension of the data. Thus, the development of new and improved classification models using different data samples is relevant. Industrial equipment has its own operating specifics and the task of data sampling is to preserve the properties and dynamics of the control object as much as possible. The scientific novelty of this research is in the development of an improved classifier based on systematic sampling data and the construction of an ensemble of models, including Neural Net, Gradient boosted tree, Naïve Bayes.

Throughout the research were completed next operations:

- The initial database was processed using three different data samples: Simple random, cluster and systematic sampling;
- he best data sampling method was selected in combination with which the classifier achieves the best results in modeling;
- The particle swarm method and ensemble models were chosen as the classifier. In the process of studying the properties of algorithms, their advantages and disadvantages, the particle swarm method was chosen as the most suitable for working with specific production data.

◆ Several types of ensemble construction based on Bagging and Voting models were considered. It was proven that the database after using Systematic sampling and the ensemble with the voting type showed the best results.

◆ The FMEA diagnostic table's coefficient's prediction system was integrated by using Ensemble vote type in a combination of Gradient boosted tree, neural net and deep learning techniques.

Building an ensemble allows to compensate for the shortcomings of the previous algorithm with the advantages of the next one in the ensemble. A comparative analysis of the application of these methods based on metrics was carried out. It has been proven that a classifier based on the systematic data sampling and an ensemble with a voting type is the most effective for diagnosing industrial equipment.

Ensemble voting substantially reduced the limits of individual models, such as overfitting and model variability, while improving prediction accuracy and generalization capabilities. Finally, the final classifier provides a robust and dependable framework for preventative maintenance methods by leveraging the collective wisdom of multiple models, resulting in improved operational efficiency, reduced downtime, and increased productivity in industrial settings.

Information on funding

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic Kazakhstan (Grant No. AP23486386).

REFERENCES

1 Samigulina G., Samigulina Z. Diagnostics of industrial equipment and faults prediction based on modified algorithms of artificial immune systems. Journal of Intelligent Manufacturing, Springer, 2022, vol. 33, pp.1433–1450.

2 Samigulina G., Samigulina Z. Biologically Inspired Unified Artificial Immune System for Industrial Equipment Diagnostic. In: Nicosia, G., et al. Machine Learning, Optimization, and Data Science. LOD 2022. Lecture Notes in Computer Science, 2023, vol 13811. Springer, Cham. https://doi.org/10.1007/978-3-031-25891-6_7.

3 Elton P De Souza, Lis Moura, Thiago Barroso Costa, João Lucas Lobato Soares. Convolutional neural networks for pattern-based fault diagnosis in low-rotation equipment. International congress of mechanical engineering, 2023.

4 Leandro Ventricci, Ronny Francis Ribeiro Junior, Guilherme Ferreira Gomes. Motor fault classification using hybrid short-time Fourier transform and wavelet transform with vibration signal and convolutional neural network. Journal of the Brazilian Society of Mechanical Sciences and Engineering, 2024, vol. 6, p. 46.

5 Qiushi Wang, Zhicheng Sun, Yueming Zhu, Chunhe Song. Intelligent fault diagnosis algorithm of rolling bearing based on optimization algorithm fusion convolutional neural network. Mathematical Biosciences & Engineering, 2023, vol. 11, pp. 63–82.

6 Renjie Wang, Ningyuan Yu, Bin An. Research on Power Equipment Fault Diagnosis Based on Improved SVM Algorithm. Journal of Electrical Systems, 2024, vol. 5, pp. 112–125.

7 Tian Y., Liu X. A Deep Adaptive Learning Method for Rolling Bearing Fault Diagnosis Using Immunity. Tsinghua Science and Technology, 2019, vol. 24, no. 6, pp. 1–14.

8 Sahu S., Kumar P.B., Parhi D.R. Analysis of hybrid CSA-DEA method for fault detection of cracked structures. Journal of Theoretical and Applied Mechanics, 2019, vol. 57, no. 2, pp. 369–382.

9 Pinto C., Pinto R., Gonçalves G. Towards. Bio-Inspired Anomaly Detection Using the Cursory Dendritic Cell Algorithm. Algorithms, 2022, vol. 15, no. 1, pp. 1–28.

10 Xiaochen Zhang, Chen Wang, Wei Zhou, Jiajia Xu. Trustworthy Diagnostics With Out-of-Distribution Detection: A Novel Max-Consistency and Min-Similarity Guided Deep Ensembles for Uncertainty Estimation. IEEE Internet of Things Journal, 2024, vol. 1.1, pp. 99–120.

11 Meng Wang, Jiong Yu, Hongyong Leng, Xusheng Du. Bearing fault detection by using graph autoencoder and ensemble learningю Scientific Reports, 2024, vol.14, no. 1.

12 Weihua Li, Jingke He, Huibin Lin, Ruyi Huang. A LightGBM-based Multi-scale Weighted Ensemble Model for Few-shot Fault Diagnosis. IEEE Transactions on Instrumentation and Measurementm, 2023, vol. 1, p. 99.

13 Arnaud Nanfak, Charles Hubert Kom, Samuel Eke. Hybrid Method for Power Transformers Faults Diagnosis Based on Ensemble Bagged Tree Classification and Training Subsets Using Rogers and Gouda Ratios. International Journal of Intelligent Engineering and Systems, 2022, vol. 5, pp. 12–24.

14 Zhiyuan Chen, Olugbenro. O. Selere, Nicholas Lu Chee Seng. Equipment Failure Analysis for Oil and Gas Industry with an Ensemble Predictive Model. Proceedings of the 9th International Conference on Computational Science and Technology, 2023, pp. 569–581.

15 Hezla L., Gurina R., Hezla M., Rezaeian N. The Role of Artificial Intelligence in Improving Failure Mode and Effects Analysis (FMEA) Efficiency in Construction Safety Management. AI Technologies and Virtual Reality, 2024, pp. 397–411.

16 Podoplelova E.S., Knyazev I.I. Modification of the fmea method using machine learning algorithms. Izvestiya SFedU engineering sciences, 2023.

17 Jiao J., Zhao M., Lin J. and Ding C. Deep Coupled Dense Convolutional Network With Complementary Data for Intelligent Fault Diagnosis. IEEE Trans. Ind. Electron, 2019, vol. 6, pp. 92–98.

18 Lei Y., Jia F., Lin J., Xing S. and Ding S.X. An Intelligent Fault Diagnosis Method Using Unsupervised Feature Learning Towards Mechanical Big Data. IEEE Trans. Ind. Electron., 2019, p. 78.

19 Dimitar P. Filev, Ratna Babu Chinnam, Finn Tseng and Pundarikaksha Baruah. An Industrial Strength Novelty Detection Framework for Autonomous Equipment Monitoring and Diagnostics. IEEE Transactions on Industrial Informatics, 2010, vol. 4, pp. 61–78.

20 Venkatasubramanian V., Rengaswamy R., Yin K. and Kavuri S.N. A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. Computers & Chemical Engineering, 2003, vol. 27, no. 9, pp. 293–311.

21 Hwang I., Kim S., Kim Y. and Seah C.E. A Survey of Fault Detection, Isolation, and Reconfiguration Methods. IEEE Transactions on Control Systems Technology, 2010, vol.18, no. 3, pp. 636–653.

22 Lei Y., Lin J., He Z. and Zuo M.J. Condition monitoring and fault diagnosis of planetary gearboxes: A review. Measurement journal, 2014, vol.35, pp. 108–126.

23 Henriquez P., Alonso J.B., Ferrer M.A. and Travieso C.M. Automatic Fault Diagnosis Systems Using Audio and Vibration Signals. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2012, vol. 5, pp. 642–652.

24 Yan R., Gao R.X. and Chen X. Non-stationary signal processing for bearing health monitoring. Signal Processing, 2014, vol. 1.

25 Venkatasubramanian V., Rengaswamy R., Kavuri S.N. and Yin K. A review of process fault detection and diagnosis. Compuers & chemical engineering, Part III: Process history based methods, 2003, pp. 327–346.

26 Yin S., Ding S.X., Xie X. and Luo H. A Review on Basic Data-Driven Approaches for Industrial Process Monitoring. IEEE Transactions on Industrial Electronic, 2014, vol. 11, pp. 6418–6428.

27 Ding S.X. A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. Journal of Process Control., 2014, vol. 24, no. 2, pp. 431–449.

28 Gao Z., Cecati C. and Ding S.X. Fault Diagnosis and Fault-Tolerant Techniques, Part I: Fault Diagnosis With Model-Based and Signal-Based Approaches. IEEE Transactions on Industrial Electronics, Part I, 2015, vol. 62, pp. 3757–3767.

29 Gao Z., Cecati C. and Ding S.X. Fault Diagnosis and Fault-Tolerant Techniques, Part II, IEEE Transactions on Industrial Electronics, 2015.

30 Evstifeev A.A. and Zaeva M.A. A hybrid teaching factory model towards personalized education. Method of Applying Fuzzy Situational Network to Assess the Risk of the Industrial Equipment Failure, 2021.

31 Mourtzis D., Angelopoulos J. and Panopoulos N., 2020, vol. 5, pp. 166–171.
32 Saeed Rajabi, Mehdi Saman Azari, Stefania Santini, and Francesco Flammini. Expert Systems with Applications, 2022.
33 Teerawat Thepmanee, Sawai Pongswatd, Farzin Asadi, and Prapart Ukakimaparn. Implementation of control and scada system: Energy Reports, 2022, vol. 8, pp. 934–941.
34 Dietterich T.G. Ensemble methods in machine learning. Multiple Classifier Systems, 2015, pp. 1–15.
35 Zhou Z.H. Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC, 2012.
36 Kuncheva L.I. Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons, 2014.
37 Rokach L. Ensemble-based classifiers. Artificial Intelligence Review, 2010, vol. 33, no. 1–2, pp. 1–39.
38 Kadi Mohamed, Amine Naim, Akkouche Naim, AkkoucheSary, Awad Sary and Awad Show., 2010.
39 Imran Rahman Pandian, Vasant Balbir, Singh Mahinder and Abdullah-Al-Wadud. Alexandria Engineering Journal, 2016.
40 Ponni Ponnusamy and Prabha Dhandayudam. Journal of Electrical Engineering and Technology journal, 2023.
41 Ali Aldrees, Hamad Hassan Awan, Arbab Faisal and Abdeliazim Mustafa Mohamed, Process Safety and Environmental Protection journal, 2022.
42 Sinem Bozkurt and Kemal Keskin, 2022.
43 Sriparna Saha and Asif Ekbal. Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. Data & Knowledge Engineering journal, 2018.

[1]Самигулина З.И.,
PhD, профессор, ORCID ID: 0000-0002-5862-6415,
e-mail: z.samigulina@kbtu.kz
[1]*Байкадамова С.С.
магистр, ORCID ID: 0009-0003-1734-8548,
e-mail: tassbulatova@gmail.com

[1]Қазақстан-Британ техникалық университеті, 050000, Алматы қ., Қазақстан

ДЕРЕКТЕРДІ ІРІКТЕУДІҢ ӨНДІРІСТІК ЖАБДЫҚТАРДЫ ДИАГНОСТИКАЛАУДАҒЫ ҮЛГІЛЕРДІ ТАНУ МІНДЕТІН ШЕШУГЕ ӘСЕРІ

Аңдатпа

Мақалада өнеркәсіптік жабдықты диагностикалау кезінде классификатордың болжамдық қабілетіне деректерді іріктеу әдістерінің әсері зерттеледі. Қарастырылған деректерді іріктеу әдістеріне қарапайым кездейсоқ таңдау, кластерлік таңдау және жүйелі таңдау жатады. Іріктеу әдістерінің нәтижелеріне сәйкес, бөлшектер тобын оңтайландыру және ансамбльдік үлгілерге (қаптау және дауыс беру түрлері) негізделген классификаторлар әзірленді. Нейрондық желі, градиентті күшейтілген ағаштар және Бейс үлгілері негізіндегі болжауды біріктіретін дауыс беру ансамбльдік модельдеу стратегиясы ең үздік нәтижені көрсетті. Ең жоғары дәлдікке нейрондық желі, градиентті күшейтілген ағаштар және аңғал Бейс үлгілері негізіндегі болжауды біріктіретін дауыс беру ансамбльдік модельдеу стратегиясы мен деректерді жүйелі таңдау әдісін қолдану арқылы қол жеткізілді: дәлдік (accuracy) 93,6%; классификация қателігі (classification error) 8%; еске түсіру (recall) 94,32%; дәлдік (precision) 93,87%. Соңғы кезеңде, деректерді іріктеу негізінде жабдықты диагностикалаудың ең тиімді стратегиясы мен ансамбльдік модель үлкен деректермен жұмыс істеуге бейімделген жақсартылған нұсқаны әзірлеу мақсатында FMEA (Failure Mode and Effects Analysis) технологиясына енгізілді.

Тірек сөздер: диагностикалық жүйе, мәліметтерді іріктеу, қарапайым кездейсоқ іріктеу, кластерлік таңдау, жүйелі таңдау, бөлшектер тобын оңтайландыру, ансамбль әдістері, FMEA жақсартылған технологиясы.

[1]Самигулина З.И.,
PhD, профессор, ORCID ID: 0000-0002-5862-6415,
e-mail: z.samigulina@kbtu.kz
[1]*Байкадамова С.С.,
магистр, ORCID ID: 0009-0003-1734-8548,
e-mail: tassbulatova@gmail.com

[1]Казахстанско-Британский технический университет, 050000, г. Алматы, Казахстан

ВЛИЯНИЕ ВЫБОРКИ ДАННЫХ НА РЕШЕНИЕ ЗАДАЧИ РАСПОЗНАВАНИЯ ОБРАЗОВ ДЛЯ ДИАГНОСТИКИ ПРОМЫШЛЕННОГО ОБОРУДОВАНИЯ

Аннотация

Статья посвящена исследованию влияния выборки данных на прогностическую способность классификатора при диагностике промышленного оборудования. Рассматривались различные типы выборок данных, такие как простая случайная выборка, кластерная и систематизированная выборка. По результатам различных выборок данных были построены классификаторы на основе методов роя частиц и ансамблевых моделей (бэггинг и тип с голосованием). Наилучшие результаты показала стратегия ансамблевого моделирования с голосованием, которая сочетает в себе прогнозирование на основе нейронной сети, деревьев с градиентным усилением и наивных Байесовских моделей. Наилучшие результаты были достигнуты с использованием систематического метода выборки данных и стратегии ансамблевого моделирования с голосованием, которая сочетает в себе прогнозирование на основе нейронной сети, деревьев с градиентным усилением и наивных моделей Байеса: accuracy 93,6%; classification error 8%; recall 94,32%; precision 93,87%. Полученная лучшая стратегия диагностики оборудования на основе выборки данных и ансамблевой модели была использована для реализации в технологии FMEA (Failure Mode and Effects Analysis) с целью получения улучшенной и адаптированной версии для работы с большими данными.

Ключевые слова: диагностика промышленного оборудования, выборка данных, простая случайная выборка, кластерная выборка, систематическая выборка, оптимизация роя частиц, ансамблевые методы, улучшенная модель FMEA.

Article submission date: 02.04.2024.