I.J. Intelligent Systems and Applications, **2025, 4, 69-77**
Published Online on August 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2025.04.07

Determining the Number of Effective Distributions Based on Neural Network Ensemble

Nazarov Fayzullo
Department of Artificial Intelligence and Information Systems, Samarkand State University named after Sharof Rashidov, Samarkand, 140107, Uzbekistan
E-mail: fayzulla-samsu@mail.ru
ORCID iD: https://orcid.org/0000-0003-3925-406X

Rashidov Akbar*
Department of Artificial Intelligence and Information Systems, Samarkand State University named after Sharof Rashidov, Samarkand, 140107, Uzbekistan
E-mail: researcher.are@gmail.com
ORCID iD: https://orcid.org/0000-0001-7510-5841
*Corresponding author

Yarmatov Sherzodjon
Department of Artificial Intelligence and Information Systems, Samarkand State University named after Sharof Rashidov, Samarkand, 140107, Uzbekistan
E-mail: sherzod2601@gmail.com
ORCID iD: https://orcid.org/0000-0001-8257-235X

Received: 05 March 2025; Revised: 28 April 2025; Accepted: 07 June 2025; Published: 08 August 2025

Abstract: Since big data streams contain hidden meanings, there is a permanent motivation to store and process them. However, storing and processing this data requires special methods and tools. Today, the most effective approach for this situation is distributed computing mechanisms. However, this approach is economically expensive, since it requires a lot of computing resources. Therefore, users who do not have economic capabilities strive to solve problems with large data streams on a single server. In this situation, a sharp drop in efficiency in terms of time is observed. However, even for a single computing machine, the use of an internal distribution mechanism can lead to efficiency in terms of time. In this case, efficiency depends on several indicators, the most important of which is determining the number of effective distributions. However, determining the number of effective distributions is a complex process. To solve this problem, this research paper considers the use of artificial intelligence algorithms. First of all, the research methodology is developed and processes that are in it are explained. In the next step, Random Forest, XGBoost, Support Vector Regression, and Multiple Linear Regression algorithms are tested to determine the number of effective distributions. In order to improve the accuracy of the study, a multilayer neural network is improved, that is, a neural network ensemble method is developed that combines the above machine learning algorithms. At the end of the study, the research results are presented and explained in detail.

Index Terms: Big Data, Internal Distribution Mechanism, Number Of Effective Distributions, Neural Network Ensemble.

1. Introduction

Today, storing, processing, and analyzing big data streams is one of the most urgent issues in the field of computer technology [1, 2]. Although extracting results from big data streams requires complex processes, the fact that these data streams contain valuable hidden meanings has led to continued research on them [3, 4].

It is known that processing large data streams requires the use of specific special algorithms. The use of special algorithmic tools in data stream management allows for fast and high-quality data processing. Traditional methods for fast processing of large data do not show effective results [5-7]. Today, the use of distributed methods for fast and high-quality processing of large data is considered to be purposeful [8, 9]. In general, several servers are needed to organize distributed computing systems. However, it is possible to apply this mechanism to a single-server database [10, 11]. For this, the tables of the database contained in a single server can be used as distributed systems [12, 13]. As a result, storing

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

and processing data based on distributed tables leads to effective results. In this case, that is, when applying distribution mechanisms to database tables, the main issue is to develop a data management algorithm based on internal distribution. A special module for managing data streams in tables created based on the internal distribution mechanism will need to be developed. This module can be organized as a set of modules that perform two interrelated functions. The first module performs the task of distributing tables according to the characteristics of data flows and computing machines, organizing the overall structure [14, 15]. The second module processes the data stored in the selected tables, and correctly organizes the queries submitted to it [16]. The first module performs complex processes in managing data flows. One of these processes is determining the number of effective distributions. As mentioned above, distributed processing of data leads to time efficiency. However, excessive distribution leads to an increase in the time of processes related to data distribution and has a negative impact on the overall time efficiency. Or, if a sufficient number of distributions is not implemented, the maximum efficiency expected cannot be achieved with a small number of distributions. This situation can be interpreted as an analogy of having more people working on a small problem than necessary or having too few workers on a very large problem. In short, determining the number of distributions appropriate to the situation is one of the important problems in research on distributed systems [17]. Since this process depends on a number of factors, it is advisable to automate it using intelligent systems. The solution to this problem will be considered during this resaerch work.

2. Method and Materials

2.1. Research Methodology

Data flow management based on the application of distribution mechanisms to database tables depends on the number of distributed tables. Therefore, the object of this study is to determine the number of distributions required for the system to operate in the most efficient way. This goal is achieved based on the following methodology (Fig. 1).

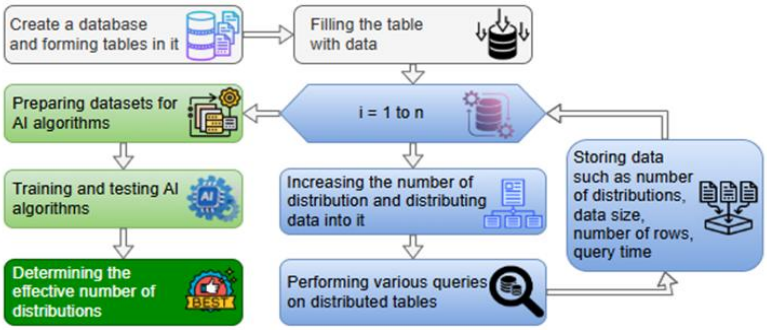


Fig.1. Methodology for intelligently determining the number of effective distributions

The number of distributions is determined by the system's ability to respond quickly to requests. In other words, the number of distributions and the time of the request are closely related to each other. Therefore, in the research methodology, after creating a database and filling the necessary tables with data, the number of distributions and the time spent on the corresponding requests are determined on a cycle-by-cycle basis. Then, based on the data collected during the cycle, artificial intelligence algorithms determine the optimal number of distributions. The processes reflected in this research methodology are covered in detail during the study.

2.2. Formation of Preliminary Data for the Development of Distribution Rules

Based on the developed research methodology, some experimental queries were performed on specially distributed tables of the database. A table with 10,853,312 rows of records stored in the database of a single computer server was used for the experimental tests. The experiments on the selected tables are mainly carried out in the following order.

- In the first experimental tests, separate search queries are performed on the table with 10,853,312 rows of records (Table 1).
- In the next experimental test, the 10,853,312 tables are divided into 1,2,3,5,7,10 distributed tables based on a certain established rule. In the next step, search queries are performed separately in each group of distributed tables, as in the first experiment. The results of the experimental tests can be seen in Table 1.
- In this step, the first and second experimental tests are performed in the query streams. That is, queries are made to the selected table or set of tables simultaneously by several computers (Table 2).

Table 1. Query time metrics for various distributed tables

Number of distributions	Average number of records in a single table	Average time taken for a query
1	10 853 312	7.3912
2	10 853 312/2	3.5083
3	10 853 312/3	3.0222
5	10 853 312/5	1.8281
7	10 853 312/7	1.6150
10	10 853 312/10	0.0073

Table 2. Time metrics for querying distributed tables in a stream by multiple users at the same time.

Number of distributions	**2**	**3**	**5**	**7**	**10**
Query sending time (seconds)	955	453	344	145	80
Number of queries in the stream	468	586	678	251	2430
Number of queries per second	0.49	1.29	1.97	1.73	30.37
Average time taken for a single query per stream (seconds)	12,7761	4.1783	3.0746	2.8973	0.0077

According to Table 1 above, when a table with an average of 10 million records is divided into 5, 7, and 10 tables, the query time decreases to 1.8, 1.6, and 0.007 seconds, which is quite efficient. Table 2 shows that even though the number of queries in the form of streams increases, the average query time decreases due to the increase in the number of partitions.

It can be seen from the above tables that the use of distributed methods on a single server leads to improved time efficiency even when the size of data and streams increases. However, this is not enough, because the highest efficiency is achieved when the optimal number of tables (x) is determined to process a query on a table with n records in m seconds. The most efficient approach to determining x is to use artificial intelligence models. Therefore, the next steps of the research will focus on finding a solution to the problem based on artificial intelligent algorithms.

2.3. Development of Machine Learning-based algorithms to Determine the Number of Effective Distributions

Determining the number of database table distributions based on experience is not always an effective approach. In addition, the number of distributions also depends on the characteristics of the servers. The number of distributions determined for a particular computing machine may not be suitable for a computing machine with other characteristics. In such cases, one of the effective methods for determining the number of distributed tables based on the specified characteristics is the use of artificial intelligence and machine learning approaches. To determine the number of distributed tables, a data set was initially formed for training and testing machine learning algorithms. Based on correlation analysis, scaling, normalization, and encoding methods, the data set was prepared for machine learning models [18-21]. In this case, the method of determining Pearson correlation coefficients is used to extract the desired features from the data set (1).

$$r_{uv} = \frac{k \cdot \sum uv - \sum u \cdot \sum v}{\sqrt{[k \cdot \sum u^2 - (\sum u)^2] \cdot [k \cdot \sum v^2 - (\sum v)^2]}} \quad (1)$$

here r_{uv}- correlation coefficient of u value and v value, k – number of observations, $\sum u$- sum of u sets, $\sum v$- sum of v sets, $\sum uv$ - sum of the product of u and v sets, $\sum u^2$ - sum of squares of u set, $\sum v^2$ - sum of squares of v set [22]. As a result, the following features were found to be the most important in the data set:

- the volume of data to be processed;
- the number of records in the database under consideration;
- the number of table partitions;
- the time spent on processing data in distributed tables.

In addition, in data preprocessing step all the data had been scaled by minmax scaling method (2):

$$u_{ijnew} = \frac{u_{ij} - u_{jmin}}{ujmin_{jmax}} \quad (2)$$

here u_{ij}- the value at the intersection of i - row and j - column, u_{jmin} is the minimum value in column j, u_{jmax} is the maximum value in column j [23].

Also, one hot encoding method was used to convert textual data in the dataset into digital form [23].

In the research work, the following artificial intelligence algorithms were used to intellectualize the need to divide a query to a table with n records into x tables in order to process it in m seconds:

Random Forest, XGBoost, Support Vector Regression, Multiple Linear Regression.
The general steps of the algorithms based on these artificial intelligence models are performed in the following steps.

Step 1. Start;
Step 2. Receive the data set;
Step 3. Prepare the data set;
Step 4. Calculate the correlation coefficients of the fields;
Step 5. Split the data set into test and training sets;
Step 6. Train the machine learning model;
Step 7. Test the machine learning model;
Step 8. Model evaluation (RMSE, MAE);
Step 9. If (RMSE < 2 and MAE < 0.80) then go to step 10;
otherwise go to step 6;
Step 10. Accept new data for distribution;
Step 11. Train the new data to the Machine Learning model;
Step 12. Determine the distribution boundaries based on the prediction;
Step 13. Print the result;
Step 14. Done.

It is known that Random Forest, XGBoost, Support Vector Regression, Multiple Linear Regression algorithms are traditional algorithms of artificial intelligence, and there are several shortcomings in their use. These problems include [24-26]:

- adaptation of models to the training sets;
- drift of prediction results from extreme values;
- emergence of instability in models.

To prevent drift of prediction results from extreme values, instability of models or adaptation to sets, it is necessary to smooth the results obtained from the most effective machine learning ensemble models. Therefore, in order to increase the accuracy during the study, a Neural Network Ensemble is used.

2.4. Development of a Neural Network Ensemble Method Based on Machine Learning Algorithms to Determine the Number of Effective Distributions

To intellectualize the process of distributing big data streams, it is desirable to generalize machine learning ensembles based on neural networks. In this case, the prediction results are smoothed to obtain results based on the best algorithms. The study developed a method for smoothing the prediction results and generalizing the metadata generated from them based on neural networks by training several machine learning algorithms together and combining their strengths. The developed method is based on modern ensemble approaches, which combine the results of each model using the strengths of Random Forest, XGBoost, Support Vector Regression and Multiple Linear Regression algorithms [27]. This data is transmitted to a multilayer perceptron neural network, which is a meta-model, and the results are obtained. The processed data set is presented in the form of the following set:

$$D = \{(\bar{x}_i, y_i)\}_{i=1}^{N} \tag{3}$$

Where $\bar{x}_i$ is the feature vector, y_i is the target values, N is the number of samples. In this case, the training data set is divided into two parts. $D_t = \{(\bar{x}_i, y_i)\}_{i=1}^{N_t}$ is for training the base models, $D_m = \{(\bar{x}_j, y_j)\}_{j=N_t+1}^{N_j}$ is for training the Meta-model.

The set of base models that combines the strengths of machine learning is defined as below (4).

$$M = \{f_{RF}, f_{XGB}, f_{SVR}, f_{LR}\} \tag{4}$$

For each model, $f_k \in M$, $k \in \{RF, XGB, SVR, LR\}$ are satisfied and these models are trained on the D_t data. This set of models is used to generate input data for the neural network meta-model. Each model is trained separately and given prediction values on the test set. These predictions are then combined as a feature vector in formula (6) and fed to the meta-model. The generation of predictions of the main models is done on the D_m data and is expressed by (5).

$$\hat{y}_k = f_k(\bar{x}_j) \;\; \forall \bar{x}_j \in D_m, \forall f_k \in M \tag{5}$$

To combine the strengths of the models used to smooth the prediction results, the input data for the Meta-model are created. To do this, the predictions of the main models are combined, and for each *j*-sample, a new feature vector of the form (6) is created:

$$Z_j = [\hat{y}_{RF}^{(j)}, \hat{y}_{XGB}^{(j)}, \hat{y}_{SVR}^{(j)}, \hat{y}_{LR}^{(j)}] \tag{6}$$

The vector (4) above constitutes the input data for the meta-model. Each sample Z_j and the corresponding true target value y_j are combined and expressed by (7).

$$Z_j = \{(Z_j, y_j)\}_{j=1}^{N_j} \tag{7}$$

As a result, the new metadata set is represented as follows (8):

$$Z_j = \begin{bmatrix} \hat{y}_{RF}^{(1)} & \hat{y}_{XGB}^{(1)} & \hat{y}_{SVR}^{(1)} & \hat{y}_{LR}^{(1)} & y_1 \\ \hat{y}_{RF}^{(2)} & \hat{y}_{XGB}^{(2)} & \hat{y}_{SVR}^{(2)} & \hat{y}_{LR}^{(2)} & y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{y}_{RF}^{(N_j)} & \hat{y}_{XGB}^{(N_j)} & \hat{y}_{SVR}^{(N_j)} & \hat{y}_{LR}^{(N_j)} & y_{N_j} \end{bmatrix} \tag{8}$$

Formula (8) represents the dataset created for training the meta-model (i.e., the neural network). In each sample (i.e., each j) — the predictions given by the main models (e.g., RF, XGBoost, SVR) in the study are combined and they are paired with the actual results. A multilayer perceptron (MLP) neural network is used as the meta-model. The meta-model aims to combine the strengths of the underlying models and smooth their predictions to approximatey_j. To optimize this process, the following objective function is formulated (7):

$$\min_{\theta} L(g(Z;\theta), y) \tag{9}$$

Here, θ is the training parameters (coefficients and biases) of the MLP model, $g(Z;\theta)$ is a function representing the output of the neural network. The main task of the neural network used as a meta-model is to maximize the accuracy of the final result by combining the predictions of the main machine learning models. To do this, the neural network is based on a loss function that measures the difference between the predictions given by the models and the actual target values. This means that the parameters of the meta-model (i.e. the weights and biases in the neural network) should be chosen in such a way that the amount of loss (error) between the predictions and the actual values for all training samples is minimal. The reason why minimization is necessary is that this way the model makes fewer errors in its predictions, i.e. the predictions are as close as possible to the actual values. In this way, the neural network not only combines the main models, but also balances their individual weaknesses and, as a result, improves the generalization ability. The minimization process is carried out using gradient-based optimization methods (e.g. SGD, Adam). *L* is the loss function, and its general form is given in the following form (10).

$$L(g(Z;\theta), y) = \frac{1}{N_m}\sum_{i=1}^{N_m}(g(\hat{y}_{RF}^{(j)}, \hat{y}_{XGB}^{(j)}, \hat{y}_{SVR}^{(j)}, \hat{y}_{LR}^{(j)};\theta)) - y_j)^2 \tag{10}$$

A neural network consists of several layers, each layer processing input data and generating output. This is represented as a *C*-layer neural network as follows (11).

$$g(Z^{(j)};\theta) = h^{(C)}, c = 1, \ldots, C,\ h^{(c)} = \sigma(W^{(c)}h^{(c-1)} + b^{(c)}) \tag{11}$$

Here $\theta = \{W^{(c)}, b^{(c)}\}_{c=1}^{C}$ – is the training weights ($W^{(c)}$) and biases ($b^{(c)}$). The final prediction function of the neural network is expressed by (12).

$$\hat{y}^{(j)} = g(Z^{(j)}, \theta) \tag{12}$$

When applying the neural network composite regression method on the test data, the test data is defined by $D_s = \{(\bar{x}_k)\}_{k=1}^{N_s}$, and the predictions $\hat{y}_k = f_k(\bar{x}_k)\ \forall f_k \in M_t, \forall \bar{x}_k \in D_s$ are obtained from the base models for the test data. Based on the combination of the strengths of the base models, test metadata is created to smooth the predictions and is expressed by (13).

$$Z_s = \{(\hat{y}_{RF}^{(i)}, \hat{y}_{XGB}^{(i)}, \hat{y}_{SVR}^{(i)}, \hat{y}_{LR}^{(i)})\} \tag{13}$$

The final predictions are calculated using the Meta-model to perform prediction smoothing based on combining the strengths of the underlying models and are expressed by (14).

$$\hat{y} = g(Z_s) \tag{14}$$

Formula (14) represents the function of providing the final prediction by the meta-model. That is, the predictions given by the main models in the previous stage are passed to the neural network. The neural network, based on its trained parameters, provides the final value for each sample in the form of $\hat{y}$. In this process, the MLP learns how to combine the weights between the main models, and outputs the result in the form of a smoothed, optimal prediction.

The regularization method is also used to optimize the applied neural network model. This process describes the neural network regression method. The selection of the main and meta-models and the optimization of their parameters have a significant impact on the result.

The main reason for choosing algorithms such as Random Forest, XGBoost, and Support Vector Regression (SVR) in this study is their high accuracy and flexibility in solving complex regression problems:

Random Forest is based on an ensemble of multiple decision trees and is resistant to the problem of overfitting, has high accuracy and stability [28].

XGBoost is a powerful algorithm based on gradient boosting, which is used at the industrial level in terms of accuracy and speed, and is especially effective in studying complex relationships.

Support Vector Regression allows for high-precision modeling of linear and nonlinear regressions and works especially well on smaller training sets [29].

When these models are used together, their individual strengths — stability (Random Forest), accuracy and speed (XGBoost), and smoothness and generalization ability (SVR) — are combined. This significantly increases the overall accuracy and robustness of the final model.

3. Results

To intellectualize the need to divide a query to a developed table with n records into x tables in order to process it in m seconds, experimental results were obtained on machine learning algorithms based on Random Forest, XGBoost, Support Vector Regression, and Multiple Linear Regression methods. The experimental results are presented in Table 3.

Table 3. Machine learning algorithms used to partition database tables based on data training and their effectiveness

Algorithms based on Machine learning	RMSE	MAE
Algorithm based on the Random Forest model	1.64	0.96
Algorithm based on the XGBost model	1.73	0.88
Algorithm based on the Support Vector Regression	1.65	1.05
Algorithm based on Multiple Linear Regression	1.71	1.008

As a result of data training, the algorithm based on the Random Forest model of machine learning, which was used to divide the tables in the database, recorded the most effective results in terms of RMSE of 1.64. Using the Random Forest model of machine learning, it was determined how many tables should be divided into to process the table data in a database with 10,853,312 records in 2.31 seconds, and as a result, the Random Forest model-based algorithm responded that it was necessary to divide it into 4.48, that is, 5 tables.

The developed neural network regression method is designed to work with complex and large data. The method mainly allows you to train a combination of Random Forest, XGBoost, Support Vector Regressor, and Multiple Linear Regression models together through linear and nonlinear relationships and creates the possibility of smoothing the prediction results, which significantly increases the accuracy of the model. Experimental results were obtained based on the neural network method to intellectualize the need to divide data in a table with n records into x tables in order to process them in m seconds. The accuracy of the method based on the experiment is presented in Table 4.

Table 4. Results of the neural network ensemble method for determining the number of effective distributions

	RMSE	MAE
The neural network ensemble	1.642	0.862

Table 5. Advantages of the proposed method

№	Main theoretical advantage	The proposed method	Traditional methods (RF, XGBoost, SVR, LR)
1	Balancing bias and variance	Reduces model deviation and increases accuracy through smoothing	Relies on a single model, variance or bias may be high
2	Studying nonlinear relationships	It studies any nonlinear relationship (universal approximator)	Most of them depend on linear or finite kernel functions.
3	Ensemble advantage – generalization through stacking	Each model combines its strengths, smoothed by a metamodel	Each model works separately, the combination is limited
4	High parameter flexibility and scalability	Flexible with dropout, normalization, and regularization	Model parameters are limited, difficult to scale

It can be seen from Table 4 that the performance of the proposed approach is slightly improved over the results of traditional machine learning algorithms. Although the performance has a small change, this method can be used as a reliable model because it overcomes the shortcomings of traditional approaches.

4. Conclusions

In conclusion, the proposed artificial intelligence model has several advantages, as can be seen from the comparison results presented in Table 5.

In general, the research work is devoted to one of the problems encountered in managing big data streams based on the internal distribution mechanism. In other words, the study investigated the issue of finding the optimal number of distributions for efficient storage and processing of big data streams. First of all, the research methodology was covered. In the next steps, the number of effective distributions was determined based on the data collected on the basis of experiments using the Random Forest, XGBoost, Support Vector Regression, and Multivariate Linear Regression machine learning algorithms. As a result of these experiments, the algorithm based on the XGBoost model recorded the best result (0.88) in terms of the MAE evaluation indicator. In general, this result could also satisfy the study, but since traditional machine learning models have problems such as adaptation to the training set, deviation of prediction results from extreme values, and the appearance of instability in the models, a new method was developed for the study. The neural network ensemble method developed in the study solved the above problems, and the error rates were reduced compared to previous methods. That is, the neural network ensemble method achieved values of 1.642 in terms of RMSE evaluation indicator and 0.862 in terms of MAE evaluation indicator.

References

[1] Rawat R., Yadav R. Big Data: Big Data Analysis, Issues and Challenges and Technologies. IOP Conference Series: Materials Science and Engineering, 2021, DOI 10.1088/1757-899X/1022/1/012014

[2] Badshah, A., Daud, A., Alharbey, R. et al. Big data applications: overview, challenges and future. Artif Intell Rev 57, 290 (2024). https://doi.org/10.1007/s10462-024-10938-5

[3] A. Rashidov, D. Mardonov and A. Soliev, "Diagnosis of Diabetes Mellitus Based on Artificial Intelligence Algorithms," 2025 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2025, pp. 349-353, doi: 10.1109/SmartIndustryCon65166.2025.10986060

[4] Singh, Nitin. Emerging Trends in Technologies for Big Data. The International Technology Management Review. 5. 202. 10.2991/itmr.2015.5.4.4.

[5] D. Junaydullaev, S. Tursunov and A. Rashidov, "An Approach Based on Data Profiling at the Preparing a Dataset for Cleaning," 2025 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2025, pp. 578-583, doi: 10.1109/SmartIndustryCon65166.2025.10986179

[6] Gonzalez A. et. al. Profiling Hyperscale Big Data Processing. In Proceedings of the 50th Annual International Symposium on Computer Architecture (ISCA '23). Association for Computing Machinery, New York, NY, USA, Article 47, 1–16. https://doi.org/10.1145/3579371.3589082

[7] Janev V. et. al. Knowledge graphs and big data processing. Lecture Notes in Computer Science; Information Systems and Applications, incl. Internet/Web, and HCI, 12072. DOI: 10.1007/978-3-030-53199-7

[8] Rashidov A., Akhatov A., Mardonov D.. The Distribution Algorithm of Data Flows Based on the BIRCH Clustering in the Internal Distribution Mechanism. 2024 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2024, pp. 923-927, doi: 10.1109/SmartIndustryCon61328.2024.10516193.

[9] X. Sun, Y. He, D. Wu and J. Z. Huang. Survey of Distributed Computing Frameworks for Supporting Big Data Analysis. in Big Data Mining and Analytics, vol. 6, no. 2, pp. 154-169, June 2023, doi: 10.26599/BDMA.2022.9020014.

[10] Rashidov, A., Akhatov, A. R., & Nazarov, F. M. Real-Time Big Data Processing Based on a Distributed Computing Mechanism in a Single Server. In C. Ananth, N. Anbazhagan, & M. Goh (Eds.), Stochastic Processes and Their Applications in Artificial Intelligence (pp. 121-138). IGI Global. https://doi.org/10.4018/978-1-6684-7679-6.ch009

[11] Akhatov, A. R., & Nazarov, F. M., Rashidov A. Mechanisms of information reliability in big data and blockchain technologies . ICISCT 2021: Applications, Trends and Opportunities, 3-5.11.2021, doi: 10.1109/ICISCT52966.2021.9670052

[12] Mazumder, Sourav & Bhadoria, Robin & Deka, Ganesh. Distributed Computing in Big Data Analytics: Concepts, Technologies and Applications [J]. 10.1007/978-3-319-59834-5.

[13] Steen Maarten, Tanenbaum Andrew. A brief introduction to distributed systems. Computing. 98. 10.1007/s00607-016-0508-7.

[14] Akhatov A., Renavikar A., Rashidov A. Optimization of the database structure based on Machine Learning algorithms in case of increased data flow. Proceedings of the International Conference on Artificial Intelligence, Blockchain, Computing And Security (ICABCS 2023), Gr. N01 Da, Up, India, 24-25 February 2023

[15] Akhatov A., Renavikar A., Rashidov A., Nazarov F. Optimization of the number of databases in the Big Data processing [J]. Проблемы информатики, № 1(58) 2023, DOI: 10.24412/2073-0667-2023-1-33-47

[16] Rashidov, A., Akhatov, A., Nazarov, F. The Same Size Distribution of Data Based on Unsupervised Clustering Algorithms [J]. In: Hu, Z., Zhang, Q., He, M. (eds) Advances in Artificial Systems for Logistics Engineering III. ICAILE 2023. Lecture Notes on Data Engineering and Communications Technologies, vol 180. Springer, Cham. https://doi.org/10.1007/978-3-031-36115-9_40

[17] J.L. Gustafson “Reevaluating Amdahl’s law” // Communications of ACM 31(5), 1988, 532-33.

[18] Makhmadiyarovich N.F. Sherzodjon Y. Development of algorithms for predictive evaluation of investment projects based on machine learning. Artificial Intelligence, Blockchain, Computing and Security Volume 2. eBook ISBN: 9781032684994, pages 681 – 685, January 2023.

[19] L. Y. Xuramov, S. Xafizova and M. Mustaffaqulov, "Calculating Singular Integrals with Cauchy Kernels in Digital Processing of Gastroenterological Medical Signals," 2025 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2025, pp. 97-103, doi: 10.1109/SmartIndustryCon65166.2025.10985974.
[20] S. Yarmatov, X. Bustanov and L. Safarova, "Optimization and Improvement of Reliability of Machine Learning Algorithms Based on Regularization Methods," 2025 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2025, pp. 394-398, doi: 10.1109/SmartIndustryCon65166.2025.10986242.
[21] L. Y. Xuramov, A. B. Baxromo and M. E. Sanayev, "Advanced Noise-Resistant Electrogastroenterological Classification Employing Convolutional Neural Networks and Hybrid Wavelet Transform Denoising," 2025 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2025, pp. 89-96, doi: 10.1109/SmartIndustryCon65166.2025.10986224.
[22] N. Fayzullo, S. Sariyev and Y. Sherzodjon, "Analyzing the Effectiveness of Ensemble Methods in Solving Multi-Class Classification Problems," 2025 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2025, pp. 788-793, doi: 10.1109/SmartIndustryCon65166.2025.10986248.
[23] Bolikulov F. et. al. Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms. Mathematics 2024, 12, 2553. https://doi.org/10.3390/math12162553
[24] Nazarov F. M., Yarmatov S. Optimization of Prediction Results Based on Ensemble Methods of Machine Learning. 2023 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2023, pp. 181-185, doi: 10.1109/SmartIndustryCon57312.2023.10110726.
[25] Ali, Zeravan & Abduljabbar, Ziyad & Tahir, Hanan & Sallow, Amira & Almufti, Saman. Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: a Review [J]. Academic Journal of Nawroz University. 12. 320-334. 10.25007/ajnu.v12n2a1612.
[26] S. Rana, H. M. Rai, L.Khuramov and D. Mardonov,Shaping the Future with Quantum Computing: An Exploration of its Emerging Field and Revolutionary Potential, Procedia Computer Science, Volume 259, 2025, Pages 844-853, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2025.04.036.
[27] Nazarov F. M., Yarmatov S., Xamidov M. Machine Learning Price Prediction on Green Building Prices. 2024 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2024, pp. 906-911, doi: 10.1109/SmartIndustryCon61328.2024.10515790.
[28] Hari Mohan Rai, Aditya Pal, Rashidov Akbar Ergash o'g'li, Bobokhonov Akhmadkhon Kholmirzokhon Ugli, Yarmatov Sherzojon Shokirovich. "Advanced AI-Powered Intrusion Detection Systems in Cybersecurity Protocols for Network Protection" Procedia Computer Science, Volume 259, 2025, Pages 140-149, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2025.03.315
[29] Mekhriddin Nurmamatov, Shokhrukh Sariyev and Izhar Uddin, "Methods of Using Artificial Intelligence Algorithms in Human Resource Management", International Russian Smart Industry Conference (SmartIndustryCon), 2025, doi.org/10.1109/SmartIndustryCon65166.2025.10986087

Authors' Profiles



Nazarov Fayzullo Makhmadiyarovich holds the academic degree of Doctor of Philosophy in Technical Sciences and the academic title of Associate Professor. Currently, Nazarov Fayzullo Makhmadiyarovich is the Dean of the Faculty of Intellectual Systems and Computer Technologies of Samarkand State University. His research interests include problems related to parallel computing, blockchain technologies, distributed systems, and deep learning.



Rashidov Akbar Ergash ugli - Head of the Department of Artificial Intelligence and Information Systems of Samarkand State University named after Sh. Rashidov, PhD. Field of interests: Artificial intelligence, Big Data, distributed computing technologies, Data analyses, Data mining. Author of more than 50 scientific articles.
Scopus: https://www.scopus.com/authid/detail.uri?authorId=57463410100.
Google Scholar: https://scholar.google.com/citations?user=_zl-wjYAAAAJ&hl=ru.
e-mail: researcher.are@gmail.com



Yarmatov Sherzodjon Shokir o'g'li holds a Doctor of Philosophy degree in Technical Sciences. Currently, Yarmatov Sherzodjon works as an associate professor at the Department of Artificial Intelligence and Information Systems at Samarkand State University. His research interests include problems related to artificial intelligence and deep learning.

How to cite this paper: Nazarov Fayzullo, Rashidov Akbar, Yarmatov Sherzodjon, "Determining the Number of Effective Distributions Based on Neural Network Ensemble", International Journal of Intelligent Systems and Applications(IJISA), Vol.17, No.4, pp.69-77, 2025. DOI:10.5815/ijisa.2025.04.07