DOI: 10.23977/jnca.2025.100105
ISSN 2371-9214 Vol. 10 Num. 1

# *Research on Efficiency Driven Classification in Petroleum Engineering Based on Big Data Algorithm*

**Xiaoyu Jia**

*Sinopec Huabei Oil & Gas Company Gas Production Plant No.1, Zhengzhou, 450006, Henan, China*
*545906643@qq.com*

***Keywords:*** Petroleum Engineering Big Data; Application Field Classification; Deep Learning; Neural Network; Optimization Algorithm

***Abstract:*** Existing petroleum engineering big data algorithms have issues like poor efficiency, insufficient classification accuracy, and poor algorithm adaptability in classification tasks by application field. In order to resolve these issues, this paper proposes an RNN (Recurrent Neural Network) algorithm, which improves the performance of the model in multi-category classification by introducing a combination of ReLU activation function and Softmax output layer. By extracting features and optimizing data from different application scenarios in the field of petroleum engineering, the algorithm effectively improves the classification accuracy and application efficiency of the model. Specifically, this paper uses different application fields in petroleum engineering big data as classification labels, uses the architecture of a neural network with multiple layers, and combines it with the Adam optimizer to improve the training speed and stability of the model by adjusting and fine-tuning the model parameters layer by layer. In the training process at each stage, special emphasis is placed on the adjustment of hyperparameters and the alleviation of the gradient vanishing problem, guaranteeing the effectiveness and precision of the classification results in multi-domain data. The findings from the experiments demonstrate that the enhanced algorithm has strong future potential and practical value, and that it can effectively boost the computation efficiency of huge amounts of data in oil and gas engineering as well as the accuracy of classification assignments in real-world applications. In the comparison of different activation functions, the ReLU activation function (improved model) performed best, with a classification accuracy of 0.852, a training time of 125 seconds, an F1-Score of 0.81, and an AUC-ROC of 0.9.

## 1. Introduction

In the field of petroleum engineering, given how quickly big data technologies is developing, the scale and complexity of data are increasing. Petroleum engineering big data covers a large amount of information from various links such as geological exploration, production, equipment monitoring, etc. How to efficiently analyze and process this data has become a key factor in improving production efficiency, optimizing decision-making, reducing risks and reducing costs. However, traditional classification algorithms often face problems such as low classification efficiency, insufficient accuracy and poor algorithm adaptability when processing petroleum engineering big data. In order to improve the accuracy and efficiency of big data classification, deep learning and neural network technology have emerged as a powerful tool to solve these problems.

This study suggests a deep learning-based neural network classification algorithm to address the issues of precision and productivity in the big data identification task for petroleum engineers by industrial field. Through comparative experiments on different activation functions (such as ReLU, Sigmoid, Tanh), their effects in the classification of petroleum engineering big data are explored, thereby optimizing the performance of the classification model. It is believed that this research will offer new ideas and methods for the efficient classification and application of petroleum engineering data, and promote the application and development of big data technology in the petroleum industry.

This paper is structured as follows: First, the research background and objectives are introduced, and the application requirements and limitations of existing methods in petroleum engineering big data classification are explained. Then, the design and optimization process of the proposed deep learning neural network model is described in detail, focusing on the selection of activation function, loss function and optimization algorithm. Subsequently, a series of experiments are conducted to evaluate the impact of different activation functions on classification accuracy and training efficiency, and the advantages of the model are demonstrated by comparison with traditional machine learning methods. Finally, the main findings of the study are summarized, and future research directions and potential improvement plans are proposed.

## 2. Related Work

The field of chemical engineering has seen both new possibilities and obstacles as a result of the big data and intelligence-based innovations' lightning-fast growth. More and more studies have explored how to apply these advanced technologies to the petroleum industry to improve performance in terms of prediction accuracy, optimize production processes, and improve equipment maintenance efficiency. Sadat et al. outlined the content, characteristics, and related topics of big data, systematically reviewed big data analysis techniques and tools, and explored its application in the chemical industry, aiming to promote the implementation of big data in engineering processes and provide a practical perspective on big data applications [1]. The development of big data and artificial intelligence had provided new ways of exploration for petroleum engineering research. Based on data and physical guidance, combined with machine learning models, it provided higher prediction accuracy and computational efficiency. Xie et al. summarized four physical information embedding mechanisms [2]. Bahaloo et al. reviewed the application of AI in upstream process optimization, evaluated the challenges of conventional methods, and explored the impact of AI on the future development of the oil industry [3]. Ozowe et al. explored the impact of advanced gas injection technology on reservoir management strategies and its role in oil and gas recovery, evaluated the effectiveness of injection technologies such as $CO_2$, nitrogen and hydrocarbon gas, and discussed technical advances such as real-time reservoir simulation, 4D seismic monitoring, intelligent well technology and machine learning [4]. Mohammed et al. introduced the basic concepts and applications of PCA (Principal components analysis), explored PCA variants and its role in classification algorithms, and demonstrated how PCA improves performance through experiments [5]. Ali et al. optimized KNN (k-nearest neighbor) connections by utilizing multithreading and multiprocessors to save computing resources, and evaluated the sequential and parallel performance of KNN using six commonly used data sets. The study used Spark Radoop to verify the effectiveness and scalability of the method [6].

These studies have provided valuable experience and inspiration for the use of artificial intelligence with large data in petroleum engineering. Wang et al. proposed a modified framework for a nearby data lake based on creation of a model includes geographical characteristics of coalbed methane reservoirs, and established a coupling theoretical model of "field data, laboratory data, management data and data management system". The study showed that the model improved the computing efficiency by 12% and has broad application prospects[7]. James et al. integrated big data with existing business intelligence systems and used deep learning algorithms for analysis. The system can automatically analyze reports and generate prediction results to provide managers with better decision support. The system helps identify and avoid potential problems through predictive analysis and improves management efficiency[8]. Ohalete et al. reviewed the application and development of preventative upkeep in the oil and gas sector using data science plus neural networks (AI), focusing on evaluating how AI and data science can switch from conventional techniques to more sophisticated predictive maintenance plans. The study showed that AI algorithms and data analysis significantly improved the accuracy of equipment failure prediction, optimized maintenance scheduling, and reduced downtime and operating costs [9]. Melberg and Gressgard analyzed how the implementation and application of digital technology in the oil industry changes work content, organizational structure and management methods. The study showed that in the context of mature technology, efficiency and cost pressure, digitalization bring about fundamental changes in the way employees and managers work [10]. Arinze et al. used machine learning, deep learning and predictive analysis to improve decision-making efficiency, maximize the techniques of exploration, manufacturing, logistics, and refinement, reduce downtime and enhance safety. However, challenges remain in terms of technical complexity, regulatory framework, network security and data privacy [11]. Weijermars et al. aimed to explore how geologists and petroleum engineers can use artificial intelligence tools to improve work efficiency, especially with the help of personalized tools provided by commercial platforms [12]. Existing research faces bottlenecks in data quality, algorithm interpretability, technical complexity, data privacy and security, which limit the extensive use of intelligent machines and big data in the petroleum engineering field.

## 3. Method

### 3.1 Model Architecture Design

In order to better meet the needs of petroleum engineering big data classification tasks by application field, this paper proposes a classification algorithm based on RNN. RNN is particularly suitable for processing engineering data with time series and correlation, such as data from various stages of oil exploration, development, production, maintenance, etc., and can better capture the sequence characteristics and time series changes in the data.

The input layer, hidden layer, stimulation operate, production level, and various additional aspects work together to create a neural network model for the task of automatic classification of petroleum engineer big data by utilization field. The model can then be optimized by utilizing the proper formulas alongside adjusting the parameters associated with them.

Assuming that the feature vector of the input data is X, which contains various information in petroleum engineering data, such as geological data, production data, or equipment monitoring data. The input layer's node count is the same as its feature count N, that is, the input vector $X \in R^N$ .

In the hidden layer, assuming that there are L layers, each containing $H_l$ neurons (where l represents the number of hidden layers). Each concealed layer's output is the output of the previous layer multiplied by the weight matrix $W_l$ , plus the bias $b_l$ , and then transformed by the activation function. It is expressed by the formula:

$$H_l = f(W_l.H_{1-l} + b_l) \tag{1}$$

Among them: $H_l$ is the output vector of the l-th layer; $W_l$ is the weight matrix of the l-th layer; $b_l$ is the bias term of the l-th layer; $f(.)$ is the activation function;

For the hidden layer, the formula of the ReLU activation function is:

$$\mathrm{Re}\,LU(x) = \max(0, x) \tag{2}$$

The advantage of this function is that it is easy to calculate and can effectively alleviate the gradient vanishing problem.

The softmax activate function, which transforms the network's final result into the distribution of probability and works well for multi-category classification duties, is typically employed in the outcome plane when the model itself performs tasks like classification. The outcome of this layer's function is:

$$y_i = \frac{e^{zi}}{\sum_{j=1}^{C} e^{z_j}} \tag{3}$$

Among them, $y_i$ is the likelihood of the i-th group; $z_i$ is the input of the :th neuron in the output layer (that is, the original output of the network); C is the total number of classification categories.

$$L = -\sum_{i=1}^{C} y_i \log(p_i) \tag{4}$$

Among them: L is the loss value; $y_i$ is the actual label (one-hot encoding); $p_i$ is the probability predicted by the model

The optimization algorithm uses Adam (adaptive moment estimation algorithm), which updates the weights by calculating the gradient's refer to or initial-order instance, and deviation, or third-order period. The formula is:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla L \tag{5}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)\nabla L^2 \tag{6}$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{7}$$

$$\theta_t = \theta_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \tag{8}$$

Among them: $\theta_t$ is the weight at the current moment; $\nabla L$ is the losses function's gradients in relation to the amount of fat; $m_t$ and $v_t$ are decay rate hyperparameters, usually set close to 1 (such as 0.9, 0.999); $\eta$ is the learning rate, and $\varepsilon$ is a constant to prevent division by zero.

Through these designs, the model can be trained and optimized according to the characteristics of the input data, and finally perform multi-category classification according to the requirements of the classification task. The combination of the decline in the correlates with cross- Entropy and the Adam optimizer ensures the stability and efficiency of the training process.

### 3.2 Classification Task Design

In petroleum engineering big data, data usually comes from different application fields, which have different characteristics and requirements. Therefore, the labels of the classification task should be defined according to these application fields.

In petroleum engineering, different application areas involve their own unique data types and requirements. In the exploration stage, data related to geological exploration, such as geological data, seismic wave data, etc., are mainly processed. These data are used to evaluate the location and characteristics of potential oil and gas reservoirs. In the development stage, data usually involves the exploitation process of the oil field, including well location design, water injection process, drilling data, etc. The goal is to increase oil and gas production and recovery through reasonable development plans. The production stage focuses on the production data of oil wells, such as flow monitoring, pressure monitoring, equipment status, etc. These data help optimize production efficiency, reduce costs, and ensure the stability and safety of the production process. Finally, in the maintenance stage, data involves equipment maintenance records, detection data, and fault warnings, etc., which are used to ensure the reliability of equipment operation, reduce sudden failures, and improve overall operation and maintenance efficiency. Data in these different fields have their own characteristics and application requirements, so the specific characteristics of these fields need to be considered when classifying.

According to the characteristics of these application fields, a label can be defined for each application field. These labels represent different data application scenarios, and each sample is labeled with a corresponding label according to the application field to which it belongs.

When performing multi-category classification, neural networks usually use a softmax output layer. The Softmax function can transform the result of the neural network into its likelihood shipping, so that the predicted value of each category is between 0 and 1, and the sum of the predicted values of all categories is 1. The goal of the network is to adjust the network parameters through training so that the probability of the final output is as close as possible to the category corresponding to the true label.

## 4. Results and Discussion

### 4.1 Experimental Environment

Hardware: NVIDIA GPU (such as RTX 3080);

Software: Python 3.x, TensorFlow/Keras, Scikit-learn, Pandas, NumPy;

Dataset: Large data set of petroleum engineering (including geological data, production data, equipment monitoring data, etc., annotated data from different application fields).

### 4.2 Comparative Experiments

Experiment 1: Using traditional SVMs and decision trees and other classic machine learning algorithms to classify the same data set, and recording indicators such as classification accuracy, training time, and model stability.

In Figure 1, the experiment compares the performance of traditional machine learning algorithms (such as SVM, decision tree, KNN and random forest in the petroleum engineering big data classification task. A comprehensive evaluation is performed based on multiple indicators such as classification accuracy, training time, F1-Score and AUC-ROC. In terms of classification accuracy, SVM performs best, reaching 0.852, slightly higher than random forest (0.837). Decision tree (0.823) and K-nearest neighbor (0.805) are relatively low, indicating that these traditional algorithms perform relatively poorly in complex petroleum engineering data classification tasks. In terms of training time, decision tree is clearly ahead with a training time of 50 seconds, followed by K-nearest neighbor (75 seconds). In contrast, the training time of support vector machine and random forest is 125 seconds and 95 seconds, respectively. Although the training time is longer, their classification accuracy is higher, showing their advantages in processing complex data. In terms of F1-Score, support vector machine takes the lead again; reaching 0.81, indicating that the model has a good balance between precision and recall. Random forest (0.78) and decision tree (0.77) follow closely, while K nearest neighbor has a lower F1-Score of 0.74, showing its shortcomings in processing petroleum engineering big data. The performance of the AUC-ROC curve is consistent with the trend of classification accuracy. The AUC-ROC values of support vector machine and random forest are 0.9 and 0.88, respectively, indicating that they have high stability and strong classification ability in distinguishing different categories. The AUC-ROC values of decision tree and K nearest neighbor are 0.87 and 0.84, respectively, which are relatively low, indicating that they are more unstable in classification performance.

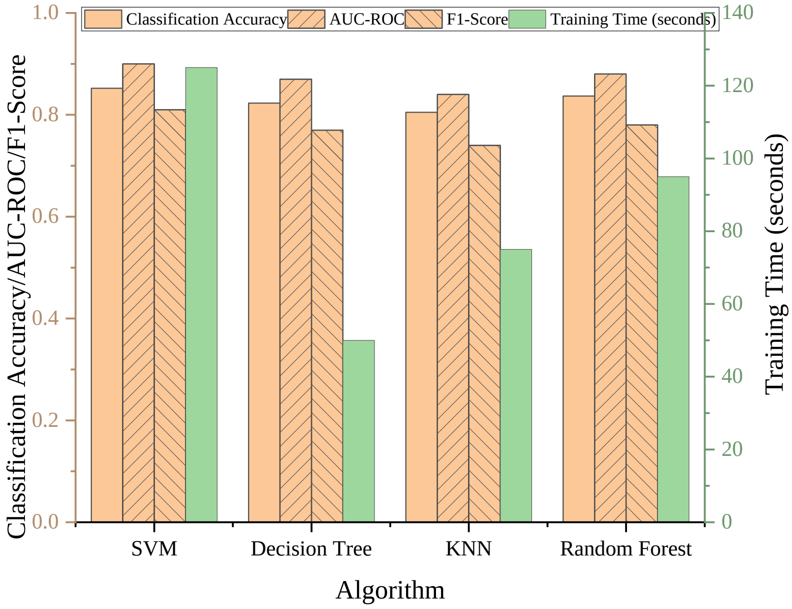

Figure 1. Performance of traditional machine learning algorithms

Table 1. Model stability analysis (standard deviation of multiple trainings)

| Algorithm | Classification Accuracy Std. Dev. (%) | Training Time Std. Dev. (seconds) | F1-Score Std. Dev. | AUC-ROC Std. Dev. |
|---|---|---|---|---|
| SVM | 1.2 | 8.5 | 0.02 | 0.01 |
| Decision Tree | 1.5 | 5.2 | 0.03 | 0.02 |
| KNN | 1.8 | 7 | 0.04 | 0.03 |
| Random Forest | 1.3 | 6.3 | 0.02 | 0.01 |

The SVM is the most stable in terms of the standard deviation of various indicators. Its classification accuracy standard deviation is 1.2%, the training time standard deviation is 8.5 seconds, and the F1-Score standard deviation and the standard deviations for AUC-ROC are 0.02 and 0.01. respectively, showing that its model has high consistency in multiple experiments and good robustness. In contrast, although the decision tree is relatively stable in terms of classification accuracy and training time, its F1-Score standard deviation is 0.03 and AUC-ROC standard deviation is 0.02, which is slightly higher than SVM, indicating that its performance in some experiments fluctuates greatly, especially in the balance between precision and recall, showing a certain instability. The stability of the KNN algorithm is poor, with a standard deviation of 1.8% for classification accuracy, 7 seconds for training time, 0.04 for F1-Score, and 0.03 for AUC-ROC, all of which show that the model fluctuates greatly in different experiments and exhibits low reliability, especially in the F1-Score and AUC-ROC indicators. The stability of the random forest is between SVM and decision tree, with a standard deviation of 1.3 for classification accuracy, 6.3 seconds for training time, 0.02 for F1-Score and 0.01 for AUC-ROC, respectively, indicating that it exhibits good stability and low fluctuations in classification tasks, especially in AUC-RO and F1-Score, which are similar to SVM, as shown in Table 1.

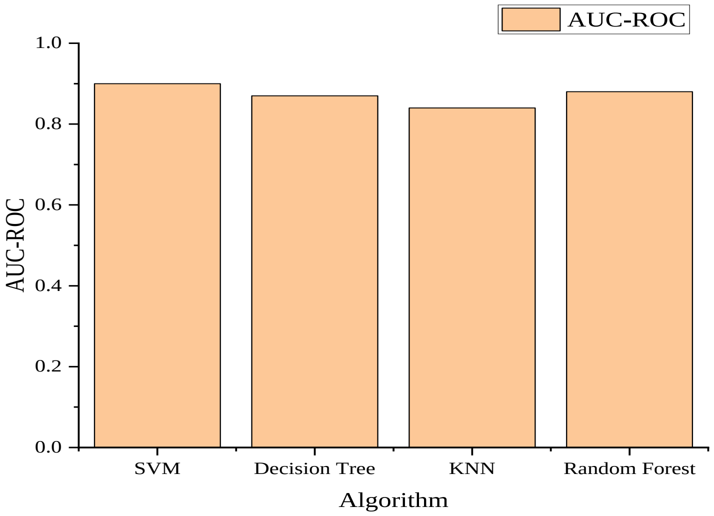

Figure 2. Comparison of AUC-ROC curves

In the comparison of AUC-ROC in Figure 2, the SVM performs best with an AUC-ROC score of 0.9, indicating that it has a strong ability to distinguish different categories. This shows that SVM is very effective in classification tasks, especially for scenarios that require high-precision classification. The decision tree algorithm has an AUC-ROC score of 0.87, which is slightly lower than SVM, but still performs well. This shows that despite its excellent performance, it may not be as good as SVM in some cases, probably because decision trees are prone to overfitting problems in complex data. The KNN algorithm has an AUC-ROC score of 0.84, the lowest of the four algorithms. This suggests that KNN may have some difficulties in distinguishing categories, especially when there is noise in the data or when the distribution between classes overlaps greatly. Its lower score indicates that KNN is not as robust as other algorithms in this particular task. Random forest has an AUC-ROC score of 0.88, which is relatively stable and ranks second only to SVM. Although lower than SVM, random forest is still highly competitive on complex data sets because it can combine the advantages of multiple decision trees and can effectively handle the complexity in the data.

Experiment 2: In the improved deep learning model, different activation functions (such as Sigmoid, Tanh, etc.) are used for comparison to observe their effects on the model effect.

The experiment investigates how deep learning model performance is affected by several activation functions (Sigmoid, Tanh, and ReLU). The best activate function for the classification job of petroleum industry large data is determined by comparing the performance of different activation functions in terms of model classification accuracy, training duration, F1-Score, model peace of mind, etc.

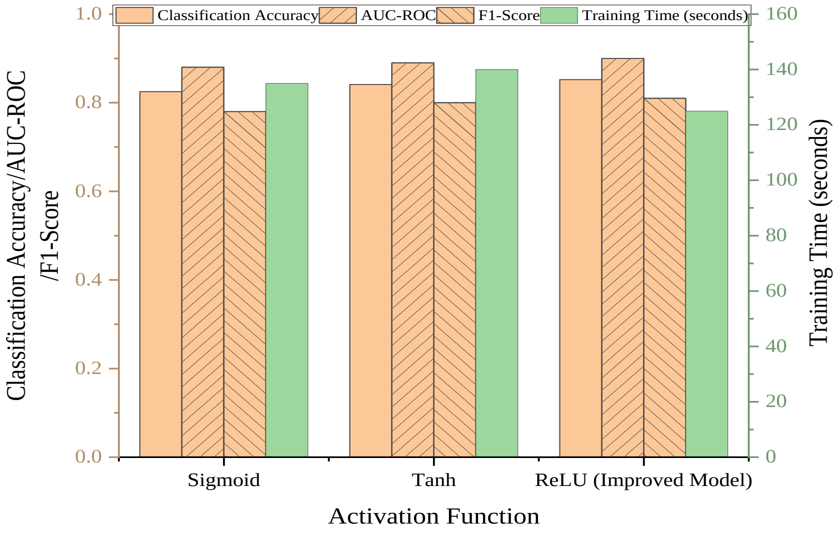

Figure 3. Performance of different activation functions

In the comparison of different activation functions, the ReLU activation function (improved model) performs best, with a classification accuracy of 0.852, a training time of 125 seconds, an F1-Score of 0.81, and an AUC-ROC of 0.9. Compared with Sigmoid and Tanh, the ReLU activation function showed higher accuracy and faster training speed in the classification task. This shows that ReLU can effectively alleviate the gradient vanishing problem and improve the training efficiency and stability of the model. The classification accuracy of the Tanh activation function is 0.841, the training time is 140 seconds, the F1-Score is 0.8, and the AUC-ROC is 0.89. Although its classification accuracy is higher than Sigmoid, its training time is longer, and its F1-Score and AUC-ROC are slightly lower than ReLU, indicating that Tanh may not be as efficient as ReLU in some cases. The Sigmoid activation function has a classification accuracy of 0.825, a training time of 135 seconds, an F1-Score of 0.78, and an AUC-ROC of 0.88, which is the worst performance (see Figure 3). Although Sigmoid may be effective in some tasks, its lower accuracy and higher training time indicate that ReLU and Tanh may be more appropriate in this task.

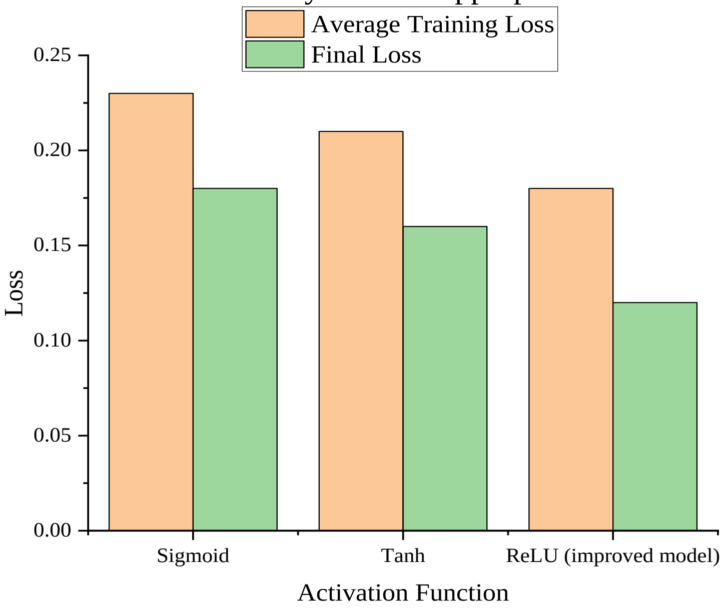

Figure 4. Error analysis (loss value) under the influence of activation function

In the comparison of loss values, the ReLU activation function (improved model) performs best, with an average training loss of 0.18 and a final loss of 0.12, showing lower training loss and better model convergence. This shows that the ReLU activation function can optimize the model more effectively, thereby accelerating the training process and improving the final performance of the model. The average training loss of the Tanh activation function is 0.21, and the final loss is 0.16. Compared with ReLU, Tanh's loss value is slightly higher, although it still performs well. Tanh may encounter certain gradient vanishing problems during training, resulting in relatively high loss values. The average training loss of the Sigmoid activation function is 0.23, and the final loss is 0.18, which is the highest among the three activation functions, as shown in Figure 4. This shows that Sigmoid converges slower than ReLU and Tanh during training, and fails to achieve good results in the final loss.

## 5. Conclusion

This paper proposes a neural network classification algorithm based on deep learning, aiming to improve the efficiency and accuracy of the classification task of petroleum engineering big data by application field. By introducing the ReLU activation function and the Softmax output layer, combined with the Adam optimizer, we successfully improve the performance of the model in multi-category classification and significantly improved the classification accuracy and application efficiency. According to the study findings, the neural network framework based on deep learning outperforms more conventional machine learning algorithms (like support vector algorithms, decision trees, K nearest neighbors, and random forests) in terms of precision of classification, educating time, and model durability. In particular, in the selection of activation functions, the ReLU activation function shows better performance than Sigmoid and Tanh, which is manifested in lower training loss, faster convergence speed, and higher classification accuracy. In addition, for different activation functions, the experiment further analyzes their impact on training loss and final loss. The experimental results show that the ReLU activation function can effectively reduce training loss, ensure that the model converges quickly during training, and ultimately achieve better classification results. Future research can further explore the combination of different deep learning architectures and optimization algorithms to further improve classification accuracy and model generalization capabilities, and provide support for big data processing and intelligent development in the oil industry.

## References

*[1] Sadat Lavasani M, Raeisi Ardali N, Sotudeh-Gharebagh R, et al. Big data analytics opportunities for applications in process engineering[J]. Reviews in Chemical Engineering, 2023, 39(3): 479-511.*

*[2] Xie C, Du S, Wang J, et al. Intelligent modeling with physics-informed machine learning for petroleum engineering problems[J]. Advances in Geo-Energy Research, 2023, 8(2): 71-75.*

*[3] Bahaloo S, Mehrizadeh M, Najafi-Marghmaleki A. Review of application of artificial intelligence techniques in petroleum operations[J]. Petroleum Research, 2023, 8(2): 167-182.*

*[4] Ozowe W, Daramola G O, Ekemezie I O. Petroleum engineering innovations: Evaluating the impact of advanced gas injection techniques on reservoir management[J]. Magna Sci. Adv. Res. Rev, 2024, 11(1): 299-310.*

*[5] Mohammed M A, Akawee M M, Saleh Z H, et al. The effectiveness of big data classification control based on principal component analysis[J]. Bulletin of Electrical Engineering and Informatics, 2023, 12(1): 427-434.*

*[6] Ali A H, Mohammed M A, Hasan R A, et al. Big data classification based on improved parallel k-nearest neighbor[J]. TELKOMNIKA (Telecommunication Computing Electronics and Control), 2023, 21(1): 235-246.*

*[7] Wang H, Adenutsi C D, Wang C, et al. Construction and Application of a Big Data System for Regional Lakes in Coalbed Methane Development[J]. ACS omega, 2023, 8(20): 18323-18331.*

*[8] James G G, Oise G P, Chukwu E G, et al. Optimizing business intelligence system using big data and machine learning[J]. Journal of Information Systems and Informatics, 2024, 6(2): 1215-1236.*

*[9] Ohalete N C, Aderibigbe A O, Ani E C, et al. Advancements in predictive maintenance in the oil and gas industry: A*

*review of AI and data science applications[J]. World J. Adv. Res. Rev, 2023, 20(3): 167-181.*

*[10] Melberg K, Gressgård L J. Digitalization and changes to work organization and management in the Norwegian petroleum industry[J]. Cognition, Technology & Work, 2023, 25(4): 447-460.*

*[11] Arinze C A, Izionworu V O, Isong D, et al. Integrating artificial intelligence into engineering processes for improved efficiency and safety in oil and gas operations[J]. Open Access Research Journal of Engineering and Technology, 2024, 6(1): 39-51.*

*[12] Weijermars R, Waheed U, Suleymanli K. Will ChatGPT and related AI-tools alter the future of the geosciences and petroleum engineering?[J]. First Break, 2023, 41(6): 53-61.*