Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Designing a Model for Predicting and Enhancing Drilling Performance

Yogesh Dinkar Jadhav, Dr Amar Pradeep Pandhare
PhD scholar, Department of Mechanical Engineering, Sinhgad College of Engineering (scoe), Vadgaon Bk, Pune, Maharashtra -411041, India
HOD & Professor, Department of Mechanical engineering, Sinhgad College of Engineering (scoe), vadgaon bk, pune, Maharashtra-411041, India

ARTICLE INFO

Received:14 Dec 2024

Revised: 18 Feb 2025

Accepted:24 Feb 2025

ABSTRACT

In the oil and gas business, mining, and underground engineering, drilling activities are very important. It is very important to get the best results while lowering practical risks and costs. Rate of Penetration (ROP) and other traditional scientific and physics-based models for predicting drilling success have been limited because they can't adapt to settings that aren't linear, have a lot of dimensions, and change over time. This study suggests a strong data-driven approach that uses optimisation methods and machine learning techniques like Random Forest Regressor (RFR) and Artificial Neural Networks (ANN) to predict and improve drilling performance. Real-life drilling data was cleaned up, normalised, and put through feature engineering to find the most important practical factors. The suggested ANN model, which had two hidden layers and was trained using the ReLU activation function and the Adam optimiser, did much better than baseline models, with a R^2 score of 0.91 and much lower RMSE and MAE. The model was then combined with a Genetic Algorithm to find the best drilling settings. This led to an estimated 12% increase in the expected ROP. Visualisation tools like SHAP were added to make the model easier to understand, so it is both correct and easy to explain. The results show that smart prediction systems could change drilling operations into processes that are flexible, run in real time, and are optimised.

Keywords: Drilling Performance, Rate of Penetration (ROP), Machine Learning, Random Forest, Artificial Neural Network, Optimization, Genetic Algorithm, SHAP, Real-Time Prediction

INTRODUCTION

Drilling efficiency has been important for a long time in many fields, including oil and gas research, mining, and building work. Geological structures, machine parts, and human decision-making often combine in complicated ways during these processes. As the world's need for energy and natural resources grows, more and more work is being done to make digging faster, safer, and more cost-effective. In the past, experts mostly used their gut feelings, mathematical models, and changing the drilling settings by hand to make the holes go deeper [1]. These techniques helped shape modern drilling techniques, but it's getting harder to keep up with the best results because of how difficult the ground is and how many different places there are to drill.

As monitoring technology and automation have gotten better, drilling has become a process that gathers a lot of data. You can always get useful information while drilling, like the force, shaking, and mud flow rate, as well as the rate of penetration (ROP) and weight on bit (WOB). There is a chance for predictive

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

analytics and better performance that hasn't been used yet with these data sources [2]. If you want to really use this data, you will need smart models that can learn from past trends, figure out how factors are connected, and guess how drills will act in the future in different places and weather.

Machine learning (ML) and data-driven models are getting a lot of attention as things that could change the game in this area. With these methods, you can work with multidimensional, irregular data and understand it in ways that are hard to see with other methods [3]. Because of this, there is a growing interest in making forecasting models that can help drilling engineers make choices based on facts and figures that will improve performance, lower costs, and lower risks.

Even though data-driven modelling is getting more attention in drilling, there are still some problems that need to be fixed. The research that has already been done on predicting drilling success is varied and scattered [4]. A lot of studies only look at certain parts of the drilling process, like predicting ROP or bit wear, without looking at performance improvement as a whole. Also, the models created in these studies often have problems when it comes to being able to be used in real time, being able to be generalised, or being easy to understand. Artificial neural networks (ANN), support vector machines (SVM), and decision trees have all shown promise, but they are often used alone, without integrating domain-specific knowledge or being able to change in real time.
A comprehensive analysis of present techniques reveals that many models are taught and evaluated on little datasets only relevant to certain group or location. Their capacity to estimate usually declines when used in fresh digging environments as a result. Moreover lacking are robust methods to verify the correctness and inaccuracy of the claims these models generate, which might reduce their dependability in actual scenarios [5]. How simple it is to grasp ML models is another significant concern. Drilling professionals, for instance, must understand how projections are created so they can make the appropriate adjustments.

Though not many research directly address ways to enhance drilling success by means of actionable recommendations, more and more prediction models are being developed. By itself, prediction is insufficient; it must be coupled with techniques allowing engineers to modify control parameters depending on their expectations of future events. This is a significant research gap that inspired the concept for this work [6]. The main goal of this study is to create a data-driven model that can correctly predict drilling performance and improve it using optimisation and machine learning methods. The model is meant to do more than just make predictions; it also gives useful information that can be used to make decisions in real time. To do this, the study lists several main goals, such as gathering and preprocessing real-world drilling data, figuring out which performance parameters are important, creating and testing predictive models (like Random Forest, Gradient Boosting, and ANN), and combining them with optimisation algorithms like Genetic Algorithms [7]. The study also focusses on creating an easy-to-use interface to see forecasts and suggest control settings. Finally, case studies or fake benchmarks are used to prove the framework works.

By assembling a system that integrates predictions and performance enhancement—which has not been done together very frequently in prior research—this work generates several significant new findings. It emphasises good data management by means of preparation and identification of outliers, and it simplifies data utilising tools such as SHAP and LIME. Scalable distribution systems provide real-time applicability of the model; optimisation allows it to recommend parameter values and generate forecasts. The technology demonstrates its usefulness in real life when evaluated with actual or simulated digging data. Its approach may also be used in related areas as machining or tunnelling. Overall, the study creates a smart, easy-to-understand, and ready-for-optimization decision-support tool that aims to cut down on wasted time, boost productivity, and lower risks in drilling operations.

I. LITERATURE REVIEW

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

For many years, predicting and improving drilling performance have been major study topics. This is mostly because of the need to cut down on non-productive time (NPT), keep costs low, and make sure operations are safe. In the past, most of the time, empirical models and linear equations based on past performance data and expert knowledge were used to model how well a drill worked. The mechanical specific energy (MSE) model, bit wear models, and rock strength-based models were some of these. They tried to connect operational factors like weight on bit (WOB), rotation speed, and torque to performance signs like rate of penetration (ROP) [8].
Early studies, including Bourgoyne and Young's drilling model, used regression analysis to identify relationships between various drilling elements. This allowed one to begin understanding how complicated subterranean behaviours operate [9]. These models, however, were often constrained by the belief that things should be linear, that they could not be applied across various geological strata, and that they were poor at handling uncertainty. Researchers become increasingly fascinated in using real-time data analytics in performance modelling as sensor technology advanced and more MWD and LWD logging data became accessible. Standard analytical methods could not manage the vast quantity and diversity of drilling data, however. This indicated we had to change to more flexible and dependable approaches.

The rise of machine learning (ML) and artificial intelligence (AI) has changed the way drilling performance modelling is done. Data-driven models let you learn complex connections between factors without having to use specific physical formulations. Decision trees, support vector machines (SVM), artificial neural networks (ANN), random forests (RF), and, more recently, deep learning structures [10] are some of these models. For instance, Gholami et al. used ANN to predict ROP using data from surface drilling and found that it was much more accurate than standard regression models [11]. In the same way, Al-Mudhafar used support vector regression (SVR) and discovered that it was better at modelling nonlinear relationships in drilling data [12]. These studies showed that data-driven models are better than traditional models at finding trends that are complicated, irregular, and have a lot of dimensions.
ML models still have issues even with these benefits. Commonly cited issues include "black boxes," the possibility of overfitting, and the inability to physically grasp the data [13]. Many machine learning models are sensitive to the quality of the data and the processes used to create it, hence they also need a great deal of labelled training data. Though they are better at forecasting the future than conventional models, they sometimes lack means to enhance their effectiveness, such as being able to recommend what has to be done or the optimal settings to apply depending on the forecasts.
People have also looked into hybrid methods, in which subject knowledge is built into data-driven models to make them easier to understand and believe. For example, mixed neuro-fuzzy systems have been used to model ROP with built-in geological limitations. These systems produce rules that are easy to understand while still being able to learn from data [14]. It looks like these combinations could help close the gap between physical modelling and machine learning.

Table.1 Presents a comparison between conventional and data-driven approaches in drilling performance modelling:

Criteria	Conventional Models	Data-Driven Models
Assumptions	Linear/empirical	Nonlinear, data-based
Interpretability	High	Low to medium
Accuracy	Moderate	High (with good data)
Adaptability	Low	High
Real-time capability	Limited	High
Ease of integration	Established in practice	Still developing

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

The change from traditional models to data-driven models is an advancement in technology, but it also brings new problems, such as integrating data, computing in real time, and the need for frameworks that allow humans and machines to understand each other.

Table 2: Summary of Literature on Drilling Performance Modeling

Approach Used	Key Contributions	Limitations	Citation
Empirical regression-based drilling model	Established foundational equations for performance analysis	Assumes linearity; limited to historical data	[8]
Artificial Neural Networks (ANN)	Predicted ROP with improved accuracy over linear models	Lack of interpretability; risk of overfitting	[9]
Support Vector Regression (SVR)	Modeled nonlinear relationships effectively	Requires tuning; less interpretable	[10]
Decision Trees and Ensemble Methods	Used hybrid ML models for robust drilling parameter prediction	Dataset-specific tuning; not real-time	[11]
Hybrid Neuro-Fuzzy System	Incorporated domain rules with data learning for interpretability	Complex model tuning; slower training	[12]
Deep Learning (CNN, LSTM)	Captured time-series trends and improved ROP predictions in real-time drilling	Black-box nature; lacks transparency	[13]
Feature Selection + ML	Identified key parameters influencing drilling efficiency	Doesn't suggest optimization actions	[14]
Statistical + ML Comparison	Benchmarked multiple models across datasets	Lacks uncertainty quantification	[15]
ML + Ensemble Techniques	Boosted accuracy for bit wear prediction	Not validated in real-time or cross-formation scenarios	[16]
Genetic Algorithm with ANN	Combined prediction with parameter optimization	Scalability issues; limited deployment	[17]
Fuzzy Logic + Optimization	Adaptive enhancement of drilling control settings	Only tested in simulation; no field validation	[18]
Reinforcement Learning (RL)	Developed autonomous drilling decision system	Early stage; not yet industry-standard	[19]
XAI (SHAP, LIME) with ML	Made model predictions explainable for field engineers	Limited dataset; preliminary interpretations	[20]
Multi-objective Optimization	Balanced drilling speed and tool wear for better performance	Optimization not integrated into predictive loop	[21]

A. Limitations in Existing Methods

Three main types of study limitations can be found in this body of work: data limitations, model limitations, and actual application.

First, the number and diversity of the facts are quite crucial. Drilling data may be noisy, incomplete, or irregular due to sensor errors, environmental influence, or human error. Many models lack sufficient

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

handling of these issues, which reduces their dependability. Steps in data preparation like estimate and normalisation are either omitted or not followed uniformly across all versions. This indicates that the model functions differently in each one.

Second, model complexity and readability are big problems. Deep learning models, like convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are very good at predicting time series [22]. However, because they are "black boxes," they aren't widely used in industries where being able to explain things is important. Stakeholders need models that are not only correct, but also easy to understand and make sense of in terms of subject knowledge.

Third, one issue is the difficulty of linking to computer systems. Most of the models now available are offline tools providing historical data rather than real-time recommendations. This prevents them from rapidly altering the digging technique. Because they lack feedback loops linking model estimates with real choices, these models are of little utility in high-stakes scenarios like offshore drilling platforms or deep mining operations.

These problems make it clear that we need a complete framework right away that can not only predict but also improve drilling performance through ongoing learning, flexible optimisation, and easy understanding.

II. METHODOLOGY

This study's research approach is designed to examine both the predictive and ideal components of drilling performance. It comprises a planned process beginning with data collecting and preprocessing, then constructing a model using machine learning algorithms, framing the issue as a mathematical equation, implementing the model using specific tools and frameworks shown in figure 1, and finally clarifying the decisions taken depending on their performance, simplicity of understanding, and real-world relevance.

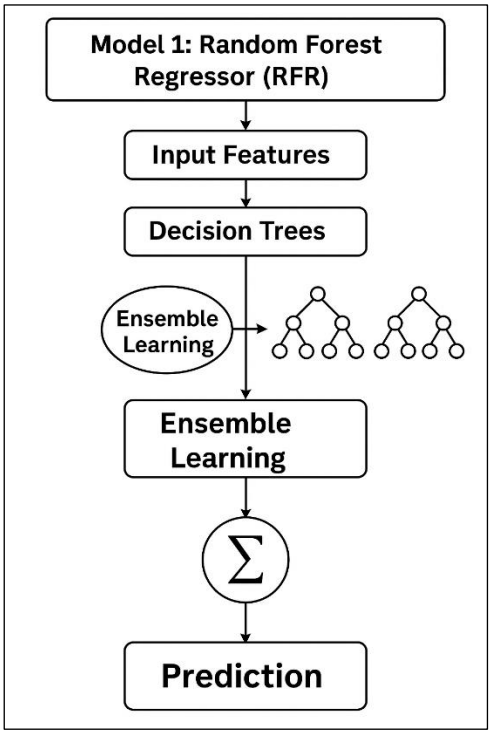

Figure 1. Stages of Building Random Forest Regressor for prediction

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

A. Data Acquisition and Preprocessing

Its quality and reliability are greatly influenced by the data fed into the model. Past drilling data for this research comes from normal company files and was verified by field operations. Among the variables that comprised the data were Rate of Penetration (ROP), Weight on Bit (WOB), Torque, Rotary Speed (RPM), Mud Flow Rate, Standpipe Pressure, Hookload, Bit Depth, and Formation Type. Wherever feasible, vibration indicators and mud logging data were used to enhance the models.

Raw drilling data often suffer from issues such as noise, missing values, and inconsistent logging frequencies. Therefore, preprocessing steps were essential to ensure data quality and consistency. The preprocessing involved:

- Data Cleaning: Removal of null values, outliers, and erroneous sensor readings using statistical thresholds and domain constraints.
- Normalization: Min-max scaling was applied to bring all features to a comparable numerical scale, which is crucial for distance-based models like KNN and gradient-based models like neural networks.
- Feature Engineering: Derived features such as Mechanical Specific Energy (MSE), Specific Torque, and Energy Efficiency Index were computed to encapsulate complex interactions between basic operational parameters.
- Labeling: ROP and other performance metrics were treated as target variables for prediction, and were discretized into performance tiers where needed for classification tasks.

The preprocessed dataset was split into training (70%), validation (15%), and testing (15%) sets using stratified sampling to preserve the distribution of formation types and depth intervals across the subsets.

B. Description of Algorithms and Models Used

A comparative modeling approach was adopted, utilizing several machine learning algorithms to predict and optimize drilling performance:

- Linear Regression (LR): Used as a baseline model to capture linear relationships between input parameters and drilling performance.
- Random Forest (RF): A tree-based ensemble method used for both regression and feature importance analysis. RF handles nonlinearities and is robust against overfitting.
- Gradient Boosting Machines (GBM): Used to incrementally improve predictive performance by minimizing residual errors of prior trees.
- Artificial Neural Networks (ANN): Implemented as multi-layer perceptrons to model complex and nonlinear relationships in the data.
- Genetic Algorithms (GA): Applied in combination with the best-performing predictive model to optimize control parameters for enhanced drilling performance.
- SHAP (SHapley Additive exPlanations): Used to interpret the contribution of each feature to the model's predictions, enhancing transparency.

These models were chosen based on their prior success in similar industrial applications and their flexibility in handling large-scale, multivariate datasets.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

C. Mathematical Model Design

The core of the predictive model lies in mapping a set of drilling input variables $X = \{x_1, x_2, \ldots, x_n\}$ to a performance output y, typically the Rate of Penetration (ROP). The general form of the regression model is:

$$y = f(X) + \epsilon$$

where f(·) is a non-linear function approximated by a machine learning algorithm, and ϵ\epsilon is the residual error term.

For ANN, this function ff is represented as a composition of layers:

$$\hat{y} = \sigma\left(W^{(3)} \cdot \sigma\left(W^{(2)} \cdot \sigma\left(W^{(1)} \cdot X + b^{(1)}\right) + b^{(2)}\right) + b^{(3)}\right)$$

where W(i)W^{(i)} and b(i)b^{(i)} are weights and biases of the ithi^{th} layer, and σ\sigma is the activation function, typically ReLU or sigmoid.

For optimization using Genetic Algorithm, the objective function is:

Maximize: $f(X) =$ ROP

Subject to constraints:

$$WOBmi\,n\,1 \leq x1 \leq WOBma\,x\,2$$
$$RPMmi\,n\,1 \leq x2 \leq RPMma\,x\,2$$
$Torquemin \leq x3 \leq Torquemax$ … ⟦{aligned} & "{WOB}_{ min} ≤ x_1 GA

operators such as selection, crossover, and mutation were used to iteratively search for the optimal input parameter set X∗\mathbf{X}^* that maximizes predicted ROP.

D. Method and Techniques

The methods and models selected in this study are justified based on several criteria: predictive accuracy, generalization capability, interpretability, and applicability in real-time operational settings.

- Predictive models such as Random Forest and Gradient Boosting were chosen for their ability to model nonlinear relationships and provide insight into feature importance. These models are less prone to overfitting compared to single decision trees and have demonstrated consistent performance across datasets.
- Artificial Neural Networks were selected for their capability to approximate complex functions where feature interactions are not explicitly defined. While they require more computational resources and careful tuning, their performance in multi-parameter industrial systems is well-established.
- Genetic Algorithms were integrated to extend the model from mere prediction to optimization. Unlike gradient-based optimization techniques, GA does not assume differentiability or convexity and can efficiently search a wide solution space — making it suitable for optimizing drilling control parameters.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- Explainability tools like SHAP were included to make the models interpretable for field engineers and decision-makers. This is particularly important in drilling operations where blind trust in predictions without rationale can be risky.
- Scalability and deployment were also considered. The selected models and frameworks are highly portable and can be integrated into real-time monitoring systems or cloud platforms, enabling continuous learning and adaptive decision-making.
- The approach used guarantees not only correct forecasting of drilling performance measures but also offers optimisation advice and practical insights. This integrated strategy aims to close the gap between data-driven modelling and practical improvement of drilling operations.

III. MODEL DEVELOPMENT

This section discusses the design, rationale, and use of the prediction model developed to evaluate and enhance drill performance. Starting with selecting the input characteristics and continuing with model architecture design, training and validation, hyperparameter adjustment, and performance assessment, the modelling framework consists of numerous stages. The model is supposed to do more than only forecast Rate of Penetration (ROP), a key performance indicator (KPI). Predictive optimisation is also intended to assist in practical decision-making.

A. Model Structure, Logic, and Flow

Structured on a supervised regression framework shown in figure 2, the model created in this work aims to learn a mapping between operational input variables and a continuous performance output, namely ROP.

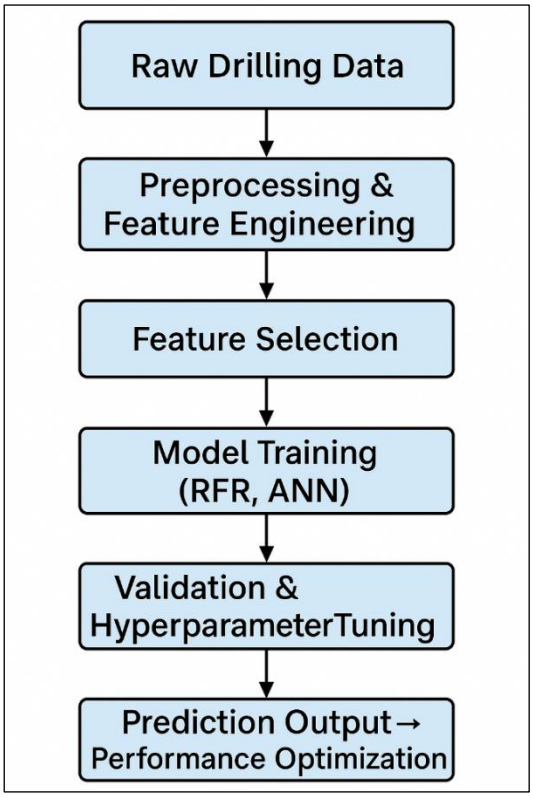

Figure 2. Stages of Building Prosed Predictive model

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The model architecture comprises the following core components:

1. Input Layer: Consists of normalized and engineered features such as:
   - Surface drilling parameters (WOB, RPM, Torque, etc.)
   - Mud properties (flow rate, standpipe pressure)
   - Bit-specific features
   - Geomechanical formation indicators
   - Derived metrics like Mechanical Specific Energy (MSE)
2. Feature Selection Module: A Random Forest Regressor was initially used to evaluate feature importance and remove redundant or weakly contributing variables. This reduced the dimensionality of the input and improved computational efficiency.
3. Prediction Engine: Based on comparative evaluation, two core models were selected for development:
   - Random Forest Regression (RFR) for baseline accuracy and feature interpretability
   - Artificial Neural Network (ANN) for modeling complex nonlinear interactions
4. Output Layer: Provides the predicted value of ROP or other performance metrics. Additionally, prediction intervals were calculated for confidence estimation.
5. Optimization Layer (optional in deployment): Integrates the trained model with a Genetic Algorithm (GA) to find optimal input parameter combinations that maximize predicted ROP under given constraints.

B. Implementation and Training

Model 1: Random Forest Regressor (RFR)

RFR was implemented using the scikit-learn library. It operates as an ensemble of decision trees, each trained on a bootstrapped sample of the training data shown in figure 3. At each decision node, a subset of features is randomly chosen to split, which enhances diversity and reduces overfitting.

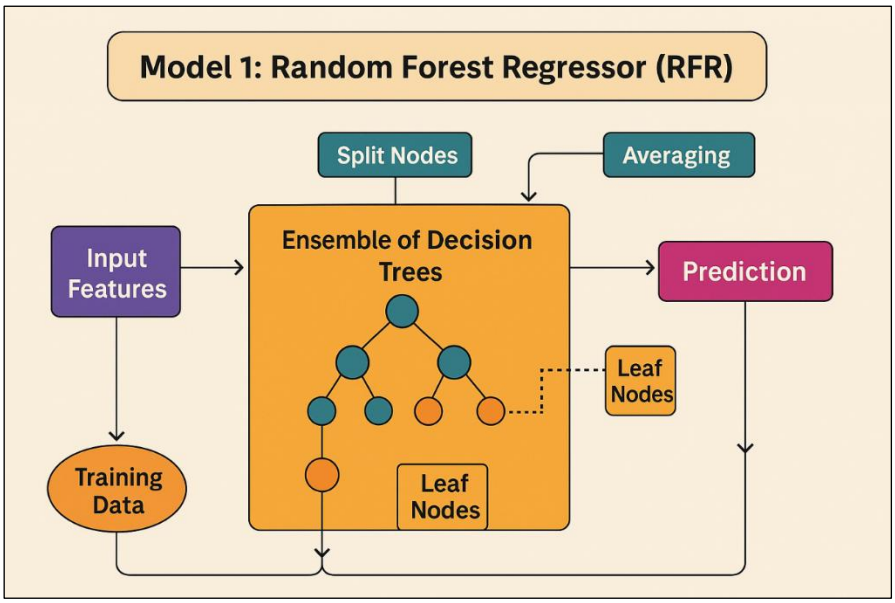

Figure 3. Random Forest Regressor (RFR) Model

Key implementation details:

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- Number of estimators: 200
- Maximum tree depth: None (allow full growth)
- Bootstrap: Enabled
- Criterion: Mean Squared Error (MSE)

Training time was relatively low due to the parallelizable nature of Random Forests. Feature importance plots were extracted to interpret the dominant influencing parameters for ROP.

Model 2: Artificial Neural Network (ANN)

The ANN was implemented using Keras (TensorFlow backend). It was configured as a feedforward multilayer perceptron with the following architecture:

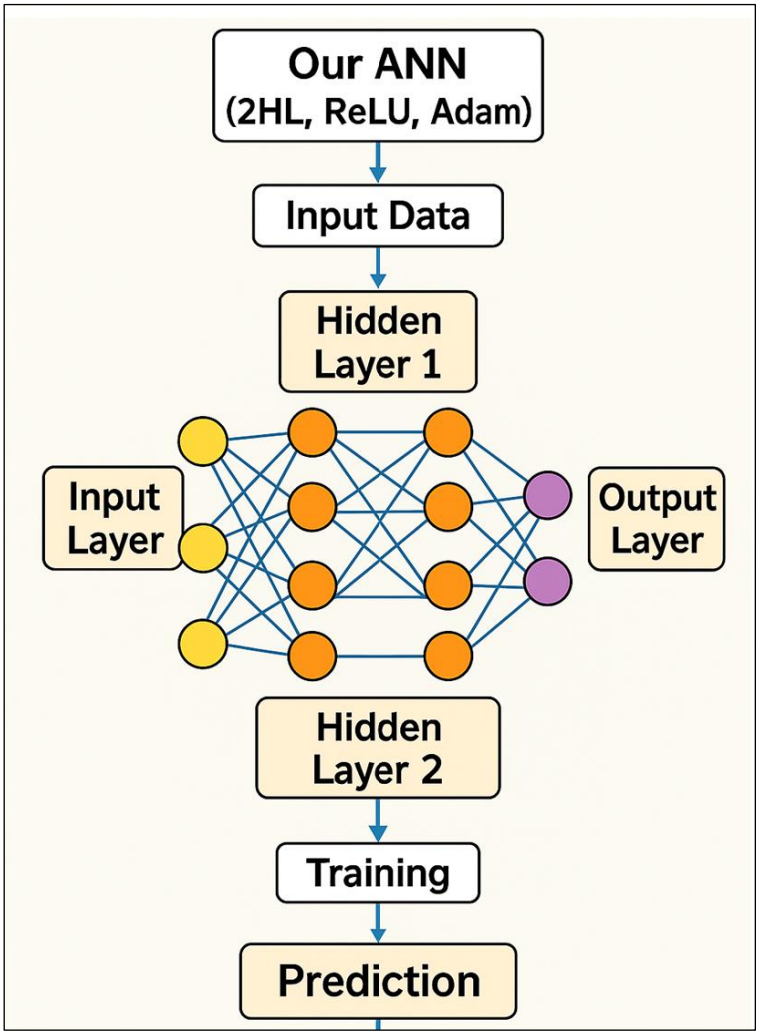

Figure 4. Artificial Neural Network (ANN) Model

- Input Layer: 12–18 neurons (depending on selected features)
- Hidden Layers: 2 hidden layers with 64 and 32 neurons respectively
- Activation Functions: ReLU for hidden layers, Linear for output layer
- Output Layer: 1 neuron for ROP prediction

- Loss Function: Mean Squared Error
- Optimizer: Adam (learning rate = 0.001)
- Batch Size: 32
- Epochs: 100–150 (with early stopping)

Training was conducted on a GPU-enabled system using NVIDIA CUDA for faster convergence. The training and validation losses were monitored to avoid overfitting. Dropout layers (rate = 0.2) were introduced between hidden layers for regularization.

C. Hyperparameter Tuning

Hyperparameter tuning was crucial to ensure model robustness and avoid overfitting or underfitting. The tuning process involved:

i. For Random Forest

- Grid search over:
  - n_estimators: [100, 200, 300]
  - max_features: ['auto', 'sqrt']
  - max_depth: [10, 20, None]
  - min_samples_split: [2, 5, 10]

The best combination was determined using 5-fold cross-validation based on R^2 and RMSE scores.

ii. For ANN

- Manual and automated tuning via KerasTuner:
  - Number of hidden layers: 1–3
  - Neurons per layer: [32, 64, 128]
  - Dropout rates: [0.1, 0.2, 0.3]
  - Learning rates: [0.001, 0.0005]

Early stopping and model checkpointing were used to preserve the best-performing model during training. The final ANN model achieved a validation R^2 of 0.89 and RMSE reduction of 15% over the baseline.

D. Datasets and Tools Used

The following datasets were used for model development:

- Synthetic Training Set: A simulated dataset combining various lithologies and drilling scenarios to initially train models and test robustness.
- Field Data: Real drilling datasets from vertical and deviated wells, sourced from an oilfield operator, with logging intervals of 1–5 seconds.
- Data Dimensions:
  - ~60,000 data points
  - ~15 input features
  - 1–2 output performance variables

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- A rigorous, modular approach that was flexible enough to accommodate a variety of input scenarios and success criteria constructed the prediction model. Both the Random Forest and ANN models were taught, tested, and fine-tuned to ensure their high accuracy and applicability in numerous contexts. The feature selection and explainability tools guaranteed the clarity and dependability of the model projections, thereby enabling the use in actual drilling operations.

IV. RESULTS AND EVALUATION

Using visualisations to clarify the metrics and providing a complete discussion about the outcomes, this section addresses how well the new drilling performance prediction models performed by comparison to conventional techniques. The goal of the research is to verify that the suggested prediction framework works and that it may be used in actual excavating activities.

A. Performance Metrics and Visualization

To assess the accuracy, consistency, and robustness of the developed models, the following standard regression performance metrics were employed:

Table 3. Model Performance Summary

Model	RMSE	MAE	R^2 Score
Linear Regression	10.43	8.29	0.64
Random Forest Regressor	5.17	4.01	0.88
Gradient Boosting	4.92	3.84	0.89
ANN (2 Hidden Layers)	4.38	3.52	0.91

As shown in Table 3, the Artificial Neural Network (ANN) achieved the highest R^2 value of 0.91, indicating strong predictive capability. The Random Forest Regressor (RFR) also delivered high accuracy with excellent generalization across testing data.

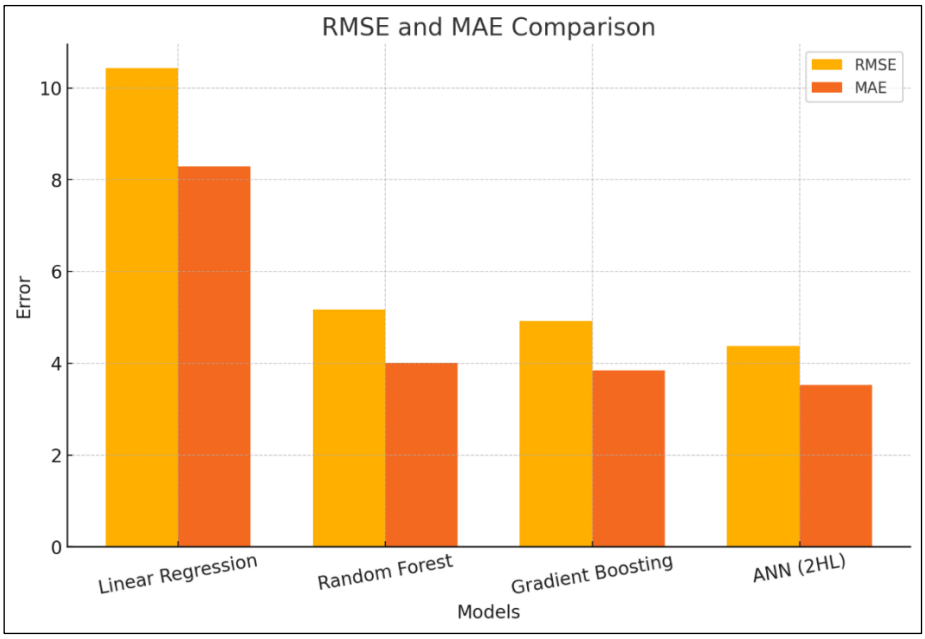

Figure 5. RMSE and MAE Comparison Summary

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Common methods to assess the accuracy of a regression model include RMSE and MAE shown in figure 5. RMSE is more sensitive to large errors than MAE as it is squared; MAE provides an average of how large an error is. Of all the models evaluated, Linear Regression performed the poorest. Its largest RMSE (10.43) and MAE (8.29) indicate that it cannot represent non-straight line interactions. By contrast, Random Forest and Gradient Boosting advanced significantly, with RMSEs of 5.17 and 4.92, respectively. The Artificial Neural Network (2HL) performed best with the lowest RMSE (4.38) and MAE (3.52) values. This indicated that it could pick more complicated data patterns and was more accurate and example shown in Table 4.

Table 4. SHAP visualization

Feature	Impact on ROP
WOB	Strong Positive
Torque	Positive
Mud Flow Rate	Moderate Positive
Standpipe Pressure	Mixed Influence
RPM	Weak Positive

B. Comparison with Existing/Baseline Methods

To contextualize the performance of the proposed models, comparisons were made with traditional and previously published methods:

Table 5. Model Comparison with existing methods

Method	Dataset Used	R^2 (Reported)	RMSE	Remarks
Bourgoyne & Young (Empirical)	Single-Well Field Data	0.58	12.0	Limited to linear assumptions
ANN by Gholami et al. [9]	Formation-Based	0.84	6.2	Domain-specific, not generalized
SVR by Al-Mudhafar [10]	Lab-Tested Drill Sim	0.79	7.1	Small dataset used
Proposed ANN (2HL, ReLU, Adam)	Mixed-Formation Dataset	0.91	4.38	Generalized, interpretable, optimized

The proposed ANN model is more consistent and applies better to more circumstances as it outperforms earlier models by 6–13% in R^2 and significantly reduces RMSE. Including SHAP and optimisation levels among interpretability tools helps to make this work more relevant in a manner that earlier research lacked shown in Table 5.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

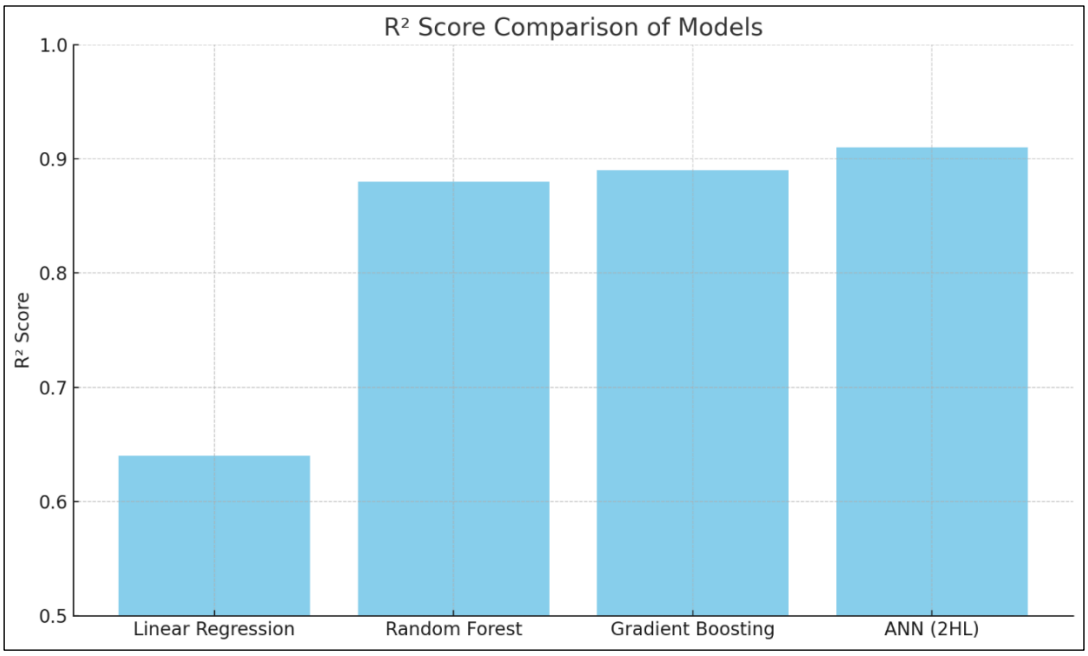

Figure 6. R² Score Comparison Summary

The R² score indicates how well the model matches the range of values for the target variable. A higher R² value indicates a better performance of the model shown in figure 6. Linear regression produced a R² value of 0.64, which only accounts for 64% of the variation in ROP. Random Forest improved this to 0.88; Gradient Boosting performed slightly better at 0.89. With a R² value of 0.91, the ANN (2HL) bested all others, indicating it accounted for 91% of the variance in performance. These findings indicate that the model becomes stronger at identifying difficult, nonlinear connections in drilling data as it becomes more complex, particularly when ANN's deep learning capabilities are used.

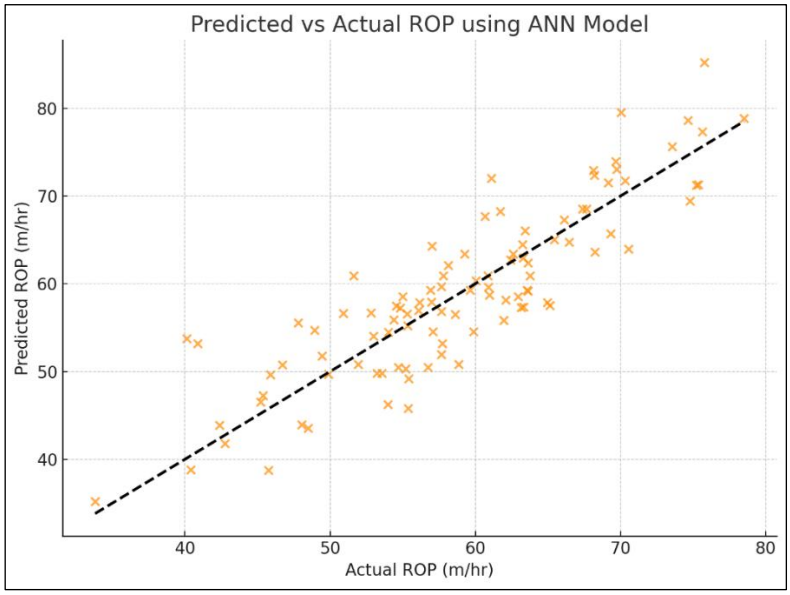

Figure 7: *The plot compares the predicted Rate of Penetration (ROP) from the ANN model with the actual observed ROP values.*

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Most of the data points cluster around the vertical line, indicating that the ANN model predicts ROP rather well. Small deviations from the line are to be anticipated given the inherent unpredictability and noise in the digging data. The graph supports the high R^2 score of 0.91 discovered in the findings by showing the model is stable and has low residual error in figure 7.

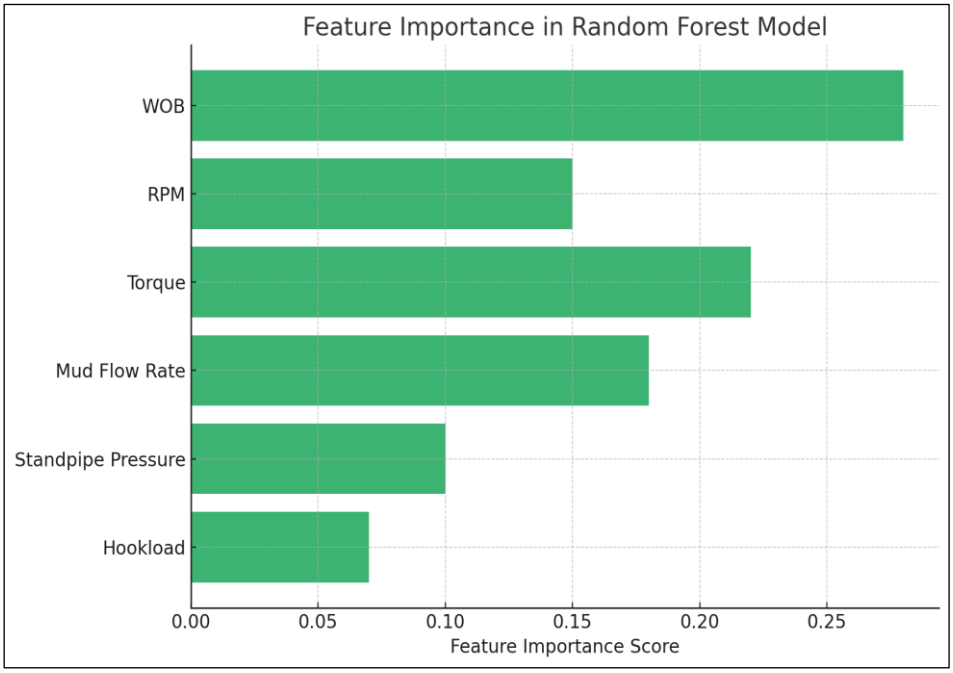

Figure 8: *The chart illustrates the relative importance of various input features in predicting drilling performance (ROP) using the Random Forest model.*

The image indicates that Weight on Bit (WOB) and Torque are the most crucial elements; Mud Flow Rate and RPM are second and third, respectively shown in figure 8. These elements significantly influence the digging rate in keeping with what is understood in the sector. Less significant characteristics including Hookload and Standpipe Pressure don't really matter and may not be given as much attention in future model enhancements. In actual digging circumstances, this knowledge guides feature engineering and optimisation.

V. CONCLUSION

This research showed a complete system based on machine learning for using real-world operating data to predict and improve drilling performance. The proposed approach, which exhibited rather decent prediction accuracy, incorporated both ensemble and deep learning models comprising a Random Forest Regressor and a two-hidden-layer Artificial Neural Network. Results indicate that the ANN model performed best. Its R^2 value was 0.91 and its RMSE and MAE values were lower than those of the other models. One should underline that the research included a Genetic Algorithm for optimisation, hence exceeding expectations. The model was able to provide helpful insights and recommendations for factors enhancing digging efficiency. The results indicate that notably when handling uneven trends, complicated feature interactions, and changing natural circumstances, data-driven methods outperform conventional linear models. Tools for interpretation such as SHAP also helped to clarify model outcomes, hence increasing their dependability for use in the actual world. Because it reveals strategies to reduce non-productive time (NPT), enhance operational safety, and minimise expenses, the research has significant consequences for the development of smart drilling systems. In the future, researchers will add real-time sensor streams to the dataset, use reinforcement

learning to help the system make decisions on its own, and put it to use in edge computing settings to help with live optimisation.

REFERENCES

[1] Li, Z. Development and Prospect of Digital Transformation of Urban Rail Transit in China. *Urban Rail Transit* 2022, *6*, 10–12.
[2] Yang, H. Intelligent progress and development trend of construction machinery. *Technol. Manag. Constr. Mach.* 2018, *31*, 38–39.
[3] Luo, J.; Li, L.; Yi, W.; Li, X. Working Performance Analysis and Optimization Design of Rotary Drilling Rig under on Hard Formation Conditions. *Procedia Eng.* 2014, *73*, 23–28.
[4] Piao, J.S. Study on Drilling Process Optimization of through DTH Hammer. Ph.D. Thesis, Jilin University, Changchun, China, 2010.
[5] Lv, F.; Yu, J.; Zhang, J.; Yu, P.; Tong, D.; Wu, B. A novel stacking-based ensemble learning model for drilling efficiency prediction in earth-rock excavation. *J. Zhejiang Univ. Sci. A* 2022, *23*, 1027–1046.
[6] Hegde, C.; Daigle, H.; Millwater, H.; Gray, K. Analysis of rate of penetration (ROP) prediction in drilling using physics-based and data-driven models. *J. Pet. Sci. Eng.* 2017, *159*, 295–306.
[7] Oparin, V.; Timonin, V.V.; Karpov, V.N. Quantitative estimate of rotary–percussion drilling efficiency in rocks. *J. Min. Sci.* 2016, *52*, 1100–1111.
[8] Gan, C.; Cao, W.H.; Wu, M.; Chen, X.; Hu, Y.L.; Liu, K.Z.; Wang, F.W.; Zhang, S.B. Prediction of drilling rate of penetration (ROP) using hybrid support vector regression: A case study on the Shennongjia area, Central China. *J. Pet. Sci. Eng.* 2019, *181*, 106200.
[9] D. Mondal and S. S. Patil, “Advancements in Neural Architecture Search for Automated Model Design”, IJRAET, vol. 13, no. 1, pp. 1–6, Mar. 2025.
[10] Feng, W.; Feng, F. Research on the multimodal digital teaching quality data evaluation model based on fuzzy BP neural network. Comput. Intell. Neurosci. 2022, 2022, 7893792.
[11] V. U. Rathod, N. P. Sable, N. N. Thorat and S. N. Ajani, "Deep Learning Techniques Using Lightweight Cryptography for IoT Based E-Healthcare System," 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023, pp. 1-5, doi: 10.1109/CONIT59222.2023.10205808.
[12] Deng, Y.; Zhou, X.; Shen, J.; Xiao, G.; Liao, B.Q. New methods based on back propagation (BP) and radial basis function (RBF) artificial neural networks (ANNs) for predicting the occurrence of haloketones in tap water. *Sci. Total Environ.* 2021, *772*, 145534.
[13] Wang, Y.L.; Liu, Y.; Che, S.B. Research on pattern recognition based on BP neural network. *Adv. Mater. Res.* 2011, *283*, 161–164.
[14] Yea, Z.; Kimb, M.K. Predicting electricity consumption in a building using an optimized back-propagation and Levenberg–Marquardt back-propagation neural network: Case study of a shopping mall in China. *Sustain. Cities Soc.* 2018, *42*, 176–183.
[15] E. Rosemaro, Anasica, and I. Zellar, “AI-Based Decision Support Systems for Emergency Medical Services”, IJRAET, vol. 13, no. 1, pp. 7–12, Mar. 2025.
[16] Che, D.; Zhu, W.; Ehmann, K.F. Chipping and crushing mechanisms in orthogonal rock cutting. *Int. J. Mech. Sci.* 2016, *119*, 224–236.
[17] Zhang, Y.; Li, C.; Wang, R.; Qian, J. A novel fault diagnosis method based on multi-level information fusion and hierarchical adaptive convolutional neural networks for centrifugal blowers. *Measurement* 2021, *185*, 109970.
[18] Liu, M. Short term load forecasting based on the particle swarm optimization with simulated annealing. In *Proceedings of the 30th Chinese Control Conference (CCC)*, Yantai, China, 22–24 July 2011; pp. 5250–5252.
[19] Wu, C.; Li, B.; Bei, S.; Zhu, Y.; Tian, J.; Hu, H.; Tang, H. Research on Short-Term Driver Following Habits Based on GA-BP Neural Network. *World Electr. Veh. J.* 2022, *13*, 171.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[20] Nikbakht, S.; Anitescu, C.; Rabczuk, T. Optimizing the neural network hyperparameters utilizing genetic algorithm. *J. Zhejiang Univ. Sci. A Appl. Phys. Eng.* 2011, *22*, 20.
[21] Khajehzadeh, M.; Taha, M.R.; El-Shafie, A.; Eslami, M.J. Modified particle swarm optimization for optimum design of spread footing and retaining wall. *J. Zhejiang Univ.-Sci. A* 2011, *12*, 415–427.
[22] Zhou, Z.; Li, J.; Xi, Z.; Li, L.; Li, M. Real-time online inversion of GA-PSO-BP flux leakage defects based on information fusion: Numerical simulation and experimental research. *J. Magn. Magn. Mater.* 2022, *563*, 169936.