

1 Introduction

With the continuous growth of global energy demand, the exploration and development of oil and gas resources are increasingly extending toward deep, ultra-deep, and complex geological formations (Mohn and Sørensen, 2019). These challenging environments significantly increase the technical difficulty and operational risk of drilling activities (Bourgoyne et al., 2018). Drilling operations usually account for a large proportion of the total cost of oil and gas field development, and therefore improving drilling efficiency is of critical importance for reducing operational costs and enhancing economic performance (Aadnøy and Looyeh, 2019).

Among the various indicators used to evaluate drilling performance, the rate of penetration (ROP) is widely regarded as one of the most important parameters (Bourgoyne and Young, 1974). As shown in Figure 1-1 (Gan et al., 2019), ROP mainly reflects the interaction between drill bit and the formation. ROP represents the drilled depth per unit time—in simple terms, how fast the drill bit moves through the rock. ROP directly determines the duration of drilling operations and thus strongly influences drilling cost and overall project efficiency (Aadnøy and Looyeh, 2019). A higher ROP can shorten drilling time and reduce costs. However, ROP is influenced by many factors. These include the combination of drilling parameters, the drilling process itself, as well as organization and management. However, the drilling process is a highly complex physical system involving multiple interacting factors, including drilling parameters, formation properties, drilling fluid characteristics, and bit–rock interaction mechanisms. These factors interact in nonlinear and dynamic ways, making accurate prediction of ROP a challenging task (Hegde and Gray, 2017).

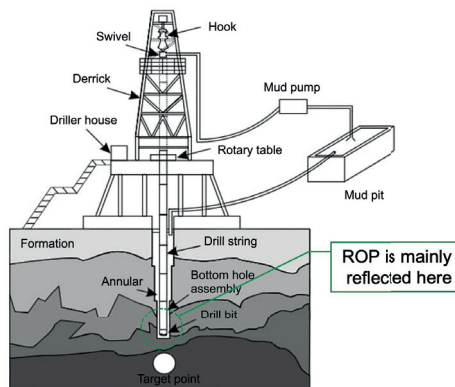


Figure 1-1 Drilling process schematic



Early attempts to model drilling performance relied primarily on empirical or semi-empirical approaches derived from drilling mechanics and field experience (Bourgoyne and Young, 1974). One of the most influential models is the Bourgoyne–Young model, which applies multiple regression techniques to relate drilling parameters and formation properties to penetration rate. In addition, the concept of mechanical specific energy (MSE) was introduced as an indicator of drilling efficiency and has been widely used in drilling optimization studies (Teale, 1965). Although these models provide valuable insights into the physical mechanisms of drilling processes, their predictive capability is often limited in complex geological environments or when applied to different drilling conditions (Dupriest and Koederitz, 2005).

With the rapid development of sensor technology and digital drilling systems, modern drilling rigs are capable of collecting large volumes of real-time operational data (Hegde and Gray, 2017). Various sensors continuously record parameters such as weight on bit (WOB), rotary speed (RPM), torque, mud flow rate, standpipe pressure, and drilling depth. These data streams provide unprecedented opportunities for analyzing drilling behavior using data-driven methods (Agwu and Akpabio, 2018).

Machine learning, as an important branch of artificial intelligence, has shown great potential for modeling complex nonlinear relationships in large datasets (Jordan and Mitchell, 2015). Over the past decade, machine learning techniques have been increasingly applied in drilling engineering for tasks such as ROP prediction, drilling parameter optimization, drilling dysfunction detection, and formation evaluation (Shadravan et al., 2019). Compared with traditional empirical models, machine learning methods are capable of automatically extracting patterns from historical data and thus often achieve higher prediction accuracy (Agwu and Akpabio, 2018).

Artificial neural networks (ANNs) are among the most widely used machine learning techniques for ROP prediction due to their strong nonlinear modeling capability (Bilgesu et al., 1997; Hegde and Gray, 2017). Other algorithms such as support vector machines (SVM), random forests (RF), gradient boosting models, and deep learning methods have also been applied in recent studies (Cortes and Vapnik, 1995; Breiman, 2001; Friedman, 2001). These models have demonstrated promising results in predicting drilling performance under various geological conditions.

Despite these advances, several challenges remain in the application of machine learning to drilling engineering, shown in Figure 1-2.

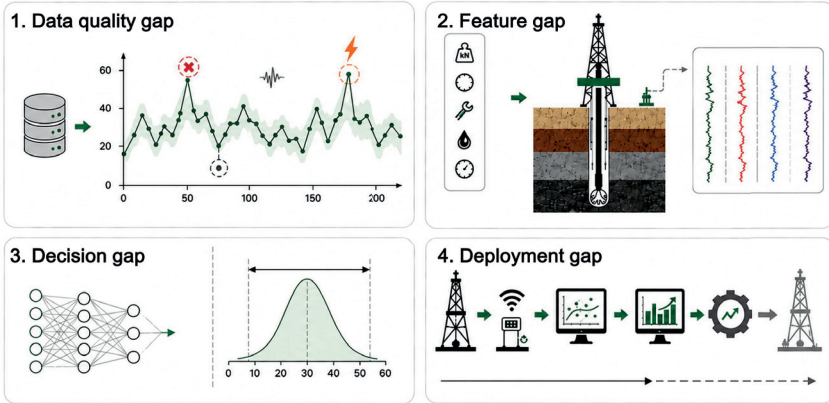


Figure 1-2 Research Gaps in Existing ROP Studies

First, drilling data collected from field sensors often contain noise, missing values, and abnormal measurements caused by sensor malfunction, signal interference, or operational instability (Aggarwal, 2017; Chandola et al., 2009; Hodge and Austin, 2004; García et al., 2016). Poor data quality can significantly degrade the performance of machine learning models. Therefore, effective data preprocessing and anomaly detection methods are essential for improving model reliability.

Second, many existing studies rely primarily on drilling operational parameters while neglecting geological information and energy-related indicators that may better represent formation characteristics (Teale, 1965; Bourgoyne and Young, 1974). The integration of geological and physical features can significantly improve the interpretability and predictive performance of ROP models (Dupriest and Koederitz, 2005; Hegde and Gray, 2017).

Third, most machine learning studies focus mainly on improving prediction accuracy. However, uncertainty associated with model predictions is rarely quantified. In real engineering applications, drilling decisions should not rely solely on point predictions but must also consider the reliability and uncertainty of model outputs (Gal and Ghahramani, 2016; Kendall and Gal, 2017; Lakshminarayanan et al., 2017; Gawlikowski et al., 2021).

Finally, many existing studies focus on algorithm development without considering the integration of predictive models into practical engineering systems. In drilling operations, engineers require decision-support tools capable of providing actionable recommendations for adjusting drilling parameters in real time. Therefore, it is necessary to develop a comprehensive



framework that integrates data preprocessing, machine learning prediction, uncertainty quantification, and parameter optimization (Bishop, 2006; Murphy, 2012; Russell and Norvig, 2021; Hegde and Gray, 2017; Dupriest and Koederitz, 2005).

Motivated by these challenges, this research aims to develop an integrated machine-learning-based framework for predicting ROP and optimizing drilling parameters, ultimately supporting intelligent drilling decision-making in practical engineering environments.

1.1 Objectives

The overall objective of this research is to develop a machine-learning-based framework capable of accurately predicting the rate of penetration and providing intelligent recommendations for drilling parameter optimization.

To achieve this goal, the following research tasks are carried out. The complete framework of the dissertation can be shown in Figure 1-2.

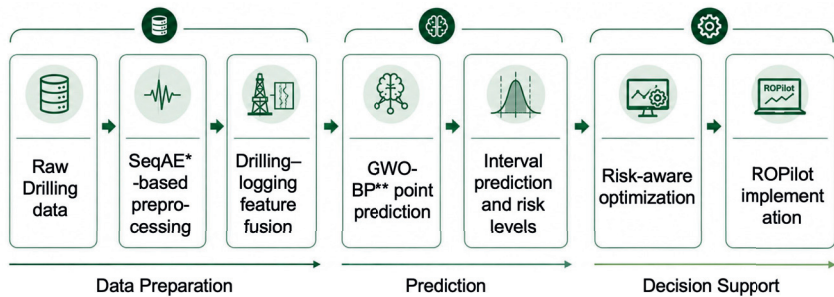


Figure 1-3 End-to-End Research Framework

1. Drilling data preprocessing

A systematic data preprocessing framework is developed to address the problem of abnormal measurements in drilling datasets. Considering the limitations of traditional statistical methods such as the three-sigma rule, MAD, and IQR, an improved SAE-based anomaly detection approach is proposed to enhance data reliability.

2. Feature integration for ROP prediction

A comprehensive feature system integrating drilling parameters with logging data is constructed to better represent formation characteristics and drilling efficiency.

3. Machine learning-based ROP prediction



Artificial neural networks are used to build predictive models for ROP estimation. To improve model performance and avoid local optima during training, swarm intelligence optimization algorithms are applied to optimize network parameters.

4. Uncertainty quantification of ROP prediction

Uncertainty quantification methods such as Monte Carlo sampling and bootstrap techniques are introduced to evaluate the reliability of model predictions and construct prediction intervals.

5. Risk-aware drilling parameter optimization

Based on predicted ROP distributions and uncertainty information, a risk-aware optimization framework is developed to balance drilling efficiency and operational risk.

6. Development of an intelligent drilling decision support system

An integrated system combining prediction models, uncertainty analysis, and parameter optimization is developed to provide real-time decision support for drilling engineers.

This thesis moves ROP research from a single predicted value to a complete workflow for prediction, interpretation, optimization, and intelligent drilling decision support.

1.2 State-of-the-art

Existing studies on ROP prediction and drilling optimization can generally be categorized into four main research directions:

- Traditional empirical and physics-based models
- Data-driven machine learning models
- Optimization algorithms for drilling parameter tuning
- Uncertainty quantification and decision support systems

1.2.1 Traditional ROP Prediction Models

Early research on drilling performance mainly relied on empirical and semi-empirical models derived from drilling mechanics (Teale, 1965; Maurer, 1962). The Bourgoyne–Young model remains one of the most influential approaches for predicting penetration rate (Bourgoyne and Young, 1974). Other studies introduced the concept of mechanical specific energy to evaluate drilling efficiency and optimize operational parameters (Dupriest and Koederitz, 2005; Pessier and Fear, 1992).



Although these models provide useful engineering insights, their applicability is often limited when drilling conditions vary significantly across formations or well sections.

1.2.2 Machine Learning Methods for ROP Prediction

With the rapid development of artificial intelligence, machine learning techniques have been widely adopted for ROP prediction(Shadravan et al., 2015; Hegde et al., 2019).

Artificial neural networks have been extensively used to model nonlinear relationships between drilling parameters and penetration rate(Ahmadi et al., 2010; Al-AbdulJabbar et al., 2020). Support vector machines and ensemble learning methods such as random forests and gradient boosting have also been applied to improve prediction performance(Friedman, 2001; Rasmussen and Williams, 2006).

These methods typically outperform traditional empirical models due to their ability to learn complex patterns from historical data.

1.2.3 Optimization Algorithms in Drilling Engineering

Optimization algorithms have been widely used to improve drilling efficiency by identifying optimal combinations of drilling parameters (Dupriest and Koederitz, 2005; Aadnøy and Looyeh, 2011).

Swarm intelligence algorithms such as particle swarm optimization (PSO), grey wolf optimization (GWO), and whale optimization algorithm (WOA) have demonstrated strong performance in solving complex optimization problems (Kennedy and Eberhart, 1995; Mirjalili et al., 2014; Mirjalili and Lewis, 2016; Yang, 2014). These algorithms are capable of exploring large search spaces and avoiding local optima.

1.2.4 Uncertainty Quantification and Intelligent Drilling Systems

In recent years, uncertainty quantification has attracted increasing attention in machine learning-based engineering applications. For drilling engineering, deterministic point predictions alone are often insufficient because drilling decisions are affected not only by the expected rate of penetration, but also by the reliability, variability, and potential risk associated with model outputs. Methods such as Monte Carlo dropout, bootstrap aggregation, Bayesian neural networks, and ensemble learning have been introduced to estimate predictive uncertainty and construct prediction intervals (Gal and Ghahramani, 2016; Kendall and Gal, 2017; Lakshminarayanan et al., 2017). These methods enable engineers to evaluate whether a predicted ROP value is sufficiently reliable for operational decision-making.



In parallel with academic developments, intelligent drilling systems have also been increasingly adopted in industrial practice. These systems integrate real-time data acquisition, operational state recognition, performance evaluation, predictive analytics, and drilling optimization into digital platforms. For example, the digital provides a series of drilling-related digital solutions covering data management and visualization, performance evaluation, and operations optimization. Its solutions include tools such as the data gateway, the ingest, the data store, and the trace for data aggregation, quality enhancement, storage, and real-time visualization, as well as the state, the pronova, the advise, the supervise, and the prognose for operational recognition, benchmarking, reporting, advisory support, and predictive analysis (TDE Group Limited, n.d.). According to the company, the state automatically detects operational states from sensor data streams using a combination of physical, rule-based models and machine learning algorithms, while the prognose is designed to identify drilling dysfunctions, support fast decision-making, reduce operational risk, and suggest corrective actions (TDE Group Limited, n.d.).

1.3 Thesis Outline

The remainder of this thesis is organized as follows.

Chapter 2 introduces the concept of drilling big data and the general research process of machine learning. It first discusses the acquisition, transmission, and storage of drilling data, emphasizing the role of field sensors, logging systems, and digital drilling infrastructure. Then, the chapter outlines the standard workflow of machine learning research, including problem definition, data preprocessing, model training, evaluation, deployment, and maintenance. This chapter provides the data and methodological foundation for the subsequent development of ROP prediction and optimization models.

Chapter 3 presents the drilling data processing and analysis framework used in this research. The chapter first describes the characteristics of the selected drilling dataset, defines the input and output variables, and analyzes the main quality issues in raw drilling data, such as missing values, abnormal points, and non-stationary fluctuations. Conventional outlier detection methods, including the three-sigma criterion, MAD, and IQR, are introduced and compared. Based on their limitations, a sequence autoencoder-based anomaly detection method is proposed and evaluated to improve data quality for subsequent ROP modeling.

Chapter 4 develops optimized machine learning models for ROP prediction by integrating drilling parameters with logging information. The chapter first explains the motivation and framework for drilling–logging data fusion and constructs a multi-source feature system to



represent both operational conditions and formation properties. Several swarm-intelligence optimization algorithms, including PSO, GWO, and WOA, are then applied to improve the performance of BP neural network models. The optimized models are compared in terms of prediction accuracy, statistical significance, convergence behavior, robustness, and interpretability using SHAP analysis.

Chapter 5 focuses on uncertainty quantification and engineering analysis of ROP prediction. It first evaluates baseline uncertainty estimation methods, including MC Dropout and Bootstrap, and analyzes their advantages and limitations in constructing prediction intervals. To improve interval reliability and engineering usability, an interval optimization method based on conformal calibration and risk stratification is proposed. The chapter further evaluates the prediction intervals using coverage, interval width, Winkler score, and risk-level distribution, thereby extending ROP prediction from deterministic point estimation to uncertainty-aware decision support.

Chapter 6 proposes a risk-aware drilling parameter optimization strategy for improving ROP. The optimization problem is formulated by considering controllable drilling parameters, engineering feasibility constraints, predicted ROP improvement, and predictive uncertainty. Different risk metrics and optimization strategies are introduced to balance drilling efficiency and operational reliability. The optimization results are analyzed from the perspectives of ROP improvement, parameter adjustment patterns, risk-level variation, and sensitivity to risk weights, providing practical guidance for drilling parameter recommendation.

Chapter 7 describes the design and implementation of an intelligent decision support system for data-driven ROP prediction and risk-aware drilling optimization. The chapter introduces the overall system architecture, design objectives, and core functional modules, including data preprocessing, drilling–logging data fusion, ROP prediction, uncertainty and risk analysis, and drilling parameter optimization. A dual-mode operating mechanism is developed to support both model evaluation and engineering application. Through a typical workflow, the chapter demonstrates how the proposed methods can be integrated into a practical software tool for intelligent drilling decision support.

Chapter 8 summarizes the main conclusions and scientific contributions of this thesis. It reviews the major findings obtained from drilling data preprocessing, optimized ROP prediction, uncertainty quantification, risk-aware parameter optimization, and system implementation. The limitations of the present research are also discussed, particularly with respect to data



availability, model generalization, and field deployment conditions. Finally, future research directions are proposed to further enhance the applicability, robustness, and real-time decision-making capability of the developed framework.



2 Drilling Big Data and Process of Machine Learning

In machine learning research, it is widely recognized that data preprocessing fundamentally determines the upper performance limit of a model, whereas the learning algorithm primarily functions to approximate the optimal solution within the constraints imposed by data quality and representation (Kotsiantis et al., 2006; García et al., 2016). Consequently, data preprocessing is not merely a preliminary step in model development but a decisive factor governing model robustness, generalization capability, and predictive accuracy. High-quality data processing enhances signal extraction, reduces noise interference, mitigates bias, and ensures consistency, thereby establishing a solid foundation for reliable modeling results.

With the rapid advancement of information technology and digital oilfield construction, petroleum companies have extensively deployed intelligent monitoring systems across drilling sites and surface equipment (Bai and Guo, 2016; Aadnøy and Looyeh, 2019). These systems continuously record multi-source, high-frequency operational data, including mechanical, hydraulic, geological, and environmental parameters. As a result, drilling operations have entered the era of big data, characterized by high volume, high velocity, high variety, and varying data quality (Mayer-Schönberger and Cukier, 2013; Chen et al., 2014). These real-time monitoring datasets constitute the core data source of the drilling big data platform and provide essential support for intelligent drilling optimization and predictive modeling of the rate of penetration (ROP) (Hegde and Gray, 2017; Shadravan et al., 2015).

This chapter provides a systematic overview of the drilling big data framework from three perspectives: data acquisition, transmission, and storage. In addition, this chapter introduces the general workflow of machine learning for engineering problem-solving. By integrating systematic data management with a rigorous machine learning pipeline, this framework establishes a methodological foundation for subsequent ROP prediction and optimization research.

2.1 Drilling Big Data Acquisition

The drilling site constitutes the fundamental source of drilling big data, encompassing a wide range of information including engineering parameters, geological measurements, and on-site audio–visual records (Aadnøy and Looyeh, 2019; Bourgoyne et al., 1991). These heterogeneous data types collectively form the foundation of data-driven drilling analysis and intelligent optimization. According to the acquisition mechanism, drilling big data can be broadly classified into manual-entry data and equipment-based automatically collected data.