Energy Reports 9 (2023) 5863–5876

Contents lists available at ScienceDirect

Energy Reports

journal homepage: www.elsevier.com/locate/egyr

Review article

Predictive big data analytics for drilling downhole problems: A review

Aslam Abdullah M. *, Aseel A., Rithul Roy, Pranav Sunil

Petroleum Engineering Research Group, School of Chemical Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India

ARTICLE INFO

Article history:
Received 8 November 2022
Received in revised form 28 April 2023
Accepted 10 May 2023
Available online 23 May 2023

Keywords:
Data mining
Oil and gas industry
Predictive modeling

ABSTRACT

With the recent introduction of data recording sensors in exploration, drilling and production processes, the oil and gas industry has transformed into a massively data-intensive industry. Big data analytics has acquired a great deal of interest from researchers to extract and use all the possible information. This paper presents an outline for predictive big data analytics to forecast and analyze some downhole problems such as pipe sticking, dog leg and pipe failure depending on several variables. Different methodologies were studied under big data, enabling the identification of the paradigm change in data storage and processing while handling vast diversified data generated in a short span of life. The evaluated data pattern sets are fed into different established predictive models and risk prediction windows to highlight future irregularities for the prevention of accidents. Finally, the game theory is used to evaluate the best predictive model to discover the optimal model for the identification of downhole problems.

Contents

* Correspondence to: School of Chemical Engineering, Vellore Institute of Technology, Vellore, 632014, Tamil Nadu, India.
E-mail address: aslamabdullah.m@vit.ac.in (Aslam Abdullah M.).

https://doi.org/10.1016/j.egyr.2023.05.028

Abbreviations

ANN	Artificial Neural Network
AVP	Average Validity Percent
GTE	Game Theory Explorer
HPC	High-Performance Computing
IC	Integrated Circuits
IoT	Internet Of Things
JSON	JavaScript Object Notation
LWD	Logging While Drilling
MAE	Mean Absolute Error
MPa	Megapascal
MPP	Massive Parallel Processing
MWD	Measurement While Drilling
NE	Nash Equilibrium
PBs	Petabytes
RE	Residual Error
REs	Residual Errors
RMSE	Root Mean Squared Error
RTDs	Resistance Temperature Detectors
SAW	Surface Acoustic Wave
SVM	Support Vector Machine
TBs	Terabytes

Nomenclature

°C	Degree Celsius
^{0}	Degree
g	Gram
M	Sample mode (XP, Yp)
P	Training sample
net_{pj}	Total of the node j's inputs
W_{ji}	Weight value of the network
(D, Y)	Sample space, D for *m* characteristics, and Y for *n* category
C1, C2,….., Cn	Potential values
P(C1), P(C2),….., P(Cn)	Occurrence probability
$H(C)$	System's entropy
C_i	Category
X	Fixed the sample's Characteristic
$H(C\|X = x_i)$	Conditional information entropy
$H(C\|X)$	Conditional entropy
P_{actual}	Actual data utilized in creating the prediction model
$P_{Predicted}$	Equivalent predictive data that the prediction model generates

1. Introduction

Data are currently torrenting into every sector of the global economy (Economist, 2010). Big data has become an increasingly essential topic as an abundance of information is produced and explored. The use of new tools and procedures to digitalize information on a large scale far beyond what was feasible with conventional approaches termed "big data". Big data is a very ambiguous term (Loh, 2011), usually requiring sophisticated data storage, administration, analysis, and visualization tools due to the size and complexity of the data sets (Chen et al., 2012). Relatively few businesses were fully utilizing big data for their strategic objectives (Ross et al., 2013). Big data analytics' basic tenet is that by analyzing massive amounts of unstructured data from numerous sources, useful insights may be produced that can assist businesses in transforming their operations and gaining an advantage over their rivals (Chen et al., 2012). Data scientists uncover more information on big data (Bile Hassan and Liu, 2020; Leung and Jiang, 2014) using various techniques like data mining algorithms (Alam et al., 2021), data analytics methodologies (Jiang and Leung, 2015; Leung and Carmichael, 2010; Leung and Jiang, 2015), machine learning (Ahn et al., 2019; Cuzzocrea et al., 2020a,b), data extraction (Cuzzocrea et al., 2020a,b), and/or computational and statistical modeling (Leung, 2018). Big data analysis has the potential to benefit society.

One of the key phases of the oil and gas industry's digitization seems to be big data (BD) analytics. The oil and gas exploration and production industries now generate enormous datasets on a daily basis as a result of recent technological advancements. According to reports, oil, and gas corporations are quite concerned about maintaining these datasets (Mohammadpoor and Torabi, 2020). Out of the data gathered in the oil and gas sector, only 5% is believed to be utilized, but as the petroleum and natural gas businesses pursue their technological transition, this percentage will rise dramatically. The upstream oil and gas business is indisputably swamped with digital data. Several big data technologies have been implemented for retail and marketing, which provides greater space for growth in this sector (Zhu et al., 2022). Big data processing techniques or technology collects and handles a large amount of data (Mohammadpoor and Torabi, 2020) since it analyzes data quickly and precisely helps to solve complex problems. Through real-time recording, big data technology discovers, examines, and extracts important information from the bulk data (Hsu and Robinson, 2017), this aids in identifying prospective trends and aid engineers in foreseeing potential issues (Xue, 2020; Qilong, 2020). Big data technologies are equally applicable in the oil and gas drilling industry. The term "drilling" refers to the extensive set of procedures required to build circular wells using excavation techniques. Numerous measurement systems and sensors are used in the process of drilling, as well as the ongoing advancement of the technology of drilling (Xie et al., 2018). The analysis of the seismic, mud logging and other important data generated during drilling can be done using big data technologies, which forecast reservoir characteristics, simulate drilling operations, and improve safety

during drilling (Xue, 2020; Qilong, 2020). Optimization of drilling operations is gaining importance in the oil and gas industry (Liu et al., 2022a,b). Many Downhole problems are commonly encountered during drilling operations.

The term "downhole" problems refers to the difficulties in drilling operations that create interruptions in soil tension around the borehole. The common downhole problems encountered are pipe sticking, pipe failure, dog legs, telescoping holes, key seats in holes, shale problems, and lost circulation problems. The contact between the drilling mud, the formation, and the actual excavation of the hole causes downhole problems (Hassan, 2018) what keeps a hole open is the equilibrium between well bore mud pressure and chemical composition on one side and pore pressure and soil pressures on the other (or stable). If this equilibrium is disturbed, well-bore complications are likely to arise.

The utilization of big data analytics and associated methodologies to address downhole problems will be investigated. Big data helps to reach the target by understanding the causes and planning remedies under controlled well cost (Lake, 2006). A better outcome from big data will be achieved using a prediction model and algorithm (Pornaroontham et al., 2023). To achieve a better outcome, three steps will be followed: Data mining involves gathering and processing data to identify patterns; algorithms will be used to incorporate these patterns based on predictive modeling (Ren and Du, 2023), and game theory will be used to find out the best appropriate model.

The first stage in creating an algorithm for detecting downhole problems is gathering enough incident-free actual drilling data. Then the data are analyzed to create a pattern for the data processing model. This data will be used to create a set of prediction models (after being cleaned and all outliers have been eliminated). The topmost and bottommost borders of the secure operating windows (risk predictive windows) for such authorized actual operating parameters are then established using the Residual Errors (REs) that were obtained when developing the predictive analytics. During the safe operating window, it is simple to monitor the actual diverging behavior of the streamed borehole data in real-time. An anomaly will be deemed for each parameter that falls beyond a predetermined safe operational window. The alarm can nonetheless go off if more consecutive data points over the predetermined threshold are identified as outliers. A method called game theory is applied to choose the optimum algorithm model (Elmgerbi and Thonhauser, 2022; Wu et al., 2022; Wang et al., 2022).

2. Big data and its characteristics

Big data consists of unstructured (not ordered and text-heavy) and multi-structured data (containing many data formats arising from interactions between humans and machines) (Trifu and Ivan, 2016). The phrase big data (also known as big data analytics or business analytics) identifies the magnitude of the accessible data collection. There are further properties of the data that make it suitable for big data applications. IBM appropriately identifies these features as three V's. These three V's refer to volume, variety, and velocity (Pence, 2014). However, recent publications have added two additional V's to provide a detailed explanation of big data. The other Vs consist of veracity and value (Ishwarappa and Anuradha, 2015).

Globally, every sector has been profoundly transformed by big data (Baaziz and Quoniam, 2013). This technology can be used in both the upstream and downstream sectors of the oil and gas industries. And under downstream, big data can be used in refining, oil, and gas transportation, and upstream, it can be used in exploration, drilling, reservoir engineering, and production engineering. It is a significant data source for the implementation of data analytical methodologies and techniques, like business analytics, machine learning, and pattern matching, for a wide variety of dependable big data analytical problems. Among these concerns are the evaluation of tool dependability, modeling, prediction, failure detection, dependability in tool geometry, manufacturing, evaluation, predictive maintenance, as well as lifetime cost analysis (Khalili-Garakani et al., 2022).

The role of big data analytics is to manage and evaluate massive volumes of data, retrieve substantial knowledge and relevant information, and acquire important ideas to allow efficient decision-making through vast databases (datasets) that are growing rapidly (Fraden, 2010). In contrast to the usual enterprise data warehouse setting, the big data environment needs magnetic, dynamic, and deep analytical capabilities (Fraden, 2010). The principal goal of this technology is to efficiently manage huge datasets in a short period of time (Tiwari et al., 2018). These advancements have facilitated the operational and strategic efficiencies of enterprises.

Social networks such as Facebook and Twitter are the primary generators of big data. Social network topology has a significant influence on physical technical networks since the majority of traffic is generated by social networking sites and linked sites (Nair et al., 2014). Due to the fact that big data involves large data sets and, in certain circumstances, complex problems, it is highly essential to have access to creative and efficient technology (Mohammadpoor and Torabi, 2020).

Big data technologies may be utilized to evaluate seismic activity, mud logging, and logging data, to predict reservoir features, simulate and decrease the drilling duration, and improve drilling safety. Mathematically, in this technology, one-to-one relationships are determined by fitting discrete data. If the amount of information gathered is small, the outcome of the fitting might be linear. Furthermore, the more and more data collected, the more precise the resulting statistics. Within the era of big data, static data are transferred into dynamic ones (Xue, 2020). As a result of recent technological advancements, the Internet of Things (IoT), fog computing, and cloud computing are now accessible to solve data storage and calculation problems (Beckwith, 2013; Mounir et al., 2018). Today, big data analytics is an indispensable tool for organizations of all sizes and across a variety of sectors. By embracing the power of big data, companies may get previously impossible insights about their customers, their businesses, and the world around them (Chen et al., 2012).

Big data is not the outcome of a single silver-bullet technology, but rather the highly complementary combination of several technologies and creative concepts (Perrons and Jensen, 2015). Despite the fact that this type of analytics depends on solid data science foundations, there are a number of key considerations for putting these approaches into effect (Kezunovic et al., 2020). The storage and processing of large data sets, as well as the transformation of large data, sets into knowledge, are the primary challenges connected with big data. It is often believed that the massive amount of big data means that useful information is hidden and must be unearthed, but analysts cannot simply intuit the data's value content (Shull, 2013). In any industry, big data analytics can give new perspectives. It may result in the accurate identification or forecasting of new scientific hypotheses, consumer behavior, societal phenomena, weather patterns, and economic situations (Jayalath et al., 2014). Table 1 compares traditional data and big data analytics.

3. Big data in oil and gas

The recent technological advancements have caused the oil and gas exploration and production industries to create massive datasets every day. According to reports, handling these datasets

Table 1
Traditional vs Big Data Analytics.

Parameters/ Characteristics	Traditional Data Analytics	Big Data Analytics	Reference
Data Storage	Isolated proprietary servers	Private or public or hybrid cloud	
Velocity	low to moderate depending on business volume	Velocity is high	Lukić, (2017)
Structured data	Data is structured	Data may be structured, semi-structured, or unstructured	Rajendran et al. (2016)
Analytics	Provide previously collected data and status reports	Need for real-time, direct feedback from the consumer, sentiment analysis	Shukla et al. (2015)
Data sources	Trusted homogeneous sources providing organizational and trading pattern data	Heterogeneous sources providing data such as Google searches, social media, sensor data, streaming data	Almeida, (2018)
Data processing	Centralized	Distributed	Lukić, (2017)
Object of analysis	Sample from the known population	From entire population	Rajendran et al. (2016)
Data volume	Business volume, and extent of digitization	Very high	Almeida, (2018)
Database technology	Rational data sources	NoSQL data sources	Lukić, (2017)
Data size	Very small	Large size	Xu et al. (2016)
Data usage	Easy to handle	Difficult to handle	Xu et al. (2016)
System configuration	Normal system configuration is sufficient	High system configuration is required	Shukla et al. (2015)
Database tools	Traditional tools are enough	special kind is required	Lukić, (2017)
Data functions	Normal functions are enough	Special kinds of functions are required	Rajendran et al. (2016)
Data time	Traditional data is produced every hour or day.	However, big data is produced more often, mostly per second.	Shukla et al. (2015)
Stability of data	Stable as well as interrelated	Uncertain and unstable relationship	Xu et al. (2016)

is a significant challenge for oil and gas businesses. In several upstream and downstream oil and gas sector activities, these datasets are captured in numerous formats and produced in vast quantities (Belhadi et al., 2019; Kamble and Gunasekaran, 2020; Tiwari et al., 2018). This aspect of big data is evident in several oil and gas industry sectors, including exploration, drilling, and production. During oil and gas exploration, seismological reports produce a vast quantity of information that is utilized to create 3D and 2D representations of the underlying strata (Mohammadpoor and Torabi, 2020). Tools like Measurement While Drilling (MWD) and Logging While Drilling (LWD) send data to the surface in real-time (Mohammadpoor and Torabi, 2020). The analysis of decisions in the petroleum and natural gas industries is enhanced by the processing and analysis outcomes of such data. Oil and gas reservoir exploration and development, the development of oil production, and capital budgeting are reliant on the modeling and processing of massive data sets. It is necessary to implement a High-Performance Computing (HPC) system in order to reduce the time of data processing and give results of efficient real-time data processing (Hsu and Robinson, 2017). The oil and gas industry mainly depends on emerging technology to enhance and improve its capabilities and aid in the discovery of more energy. Here are some significant applications of big data analytics in the oil and gas industry.

3.1. Upstream sector

3.1.1. Manage seismic data

In order to identify potential petroleum sources, upstream analysis begins with the collection of seismic data (using sensors) across a potential area of concern. Once the data has been collected, it is processed and evaluated to establish a drilling site. Combining seismic data with other data sets (such as a company's previous data on prior drilling operations and research data) may help estimate the amount of oil and gas in oil reservoirs (Panes et al., 2022).

3.1.2. Optimize drilling processes

Customizing predictive models that predict future equipment failures is one method to enhance drilling operations. The machine is initially equipped with detectors to gather information. This data and equipment information (model, operational parameters) are analyzed using machine learning techniques to identify usage patterns that are probable to cause equipment failure (Baaziz and Quoniam, 2013).

3.1.3. Improve reservoir engineering

The data needed to improve reservoir production may be gathered using a variety of downhole sensors (sound sensors, pressure sensors, and temperature sensors). For instance, businesses may develop reservoir management software employing big data analytics to get quick and actionable data on changes in the pressure of reservoir, temperature, flow, and acoustics in order to improve reservoir performance and profitability.

3.2. Downstream sector

Big data predictive analytics may help oil and gas companies improve their asset management by lowering the amount of time. Comparing the equipment's previous operating data with its present operating data is the first stage in conducting an analysis of the equipment's performance. The performance forecast is then further adjusted by taking into account the device's final needs and failure scenarios. After that, maintenance professionals are given the expected performance of the device so that they can make decisions about, for instance, the replacement of this asset (Khalili-Garakani et al., 2022).

3.3. Midstream sector

The petroleum sector is highly complicated when it comes to logistics, and the primary objective is to move oil and gas with as much danger as is humanly feasible (White et al., 2019). Sensor analytics are used by businesses to verify that their energy products are transported in a secure manner. The analysis of sensor data from pipelines and tankers by software designed for predictive maintenance may identify irregularities (such as fatigue cracks, stress corrosion, seismic ground movements, and so on), allowing for the prevention of accidents (Kadry, 2020).

4. Downhole problems

As a result of improvements in drilling techniques, exploration for oil and gas operations is currently being conducted more regularly than ever before at rock depths of hundreds of meters. Downhole circumstances can be challenging (Ren et al., 2019). Some of the challenging downhole circumstances include strong abrasive development, pressure exceeding 207 megapascals (MPa), temperatures above 200 °C, shocks and vibrating levels above 15 g, horizontal paths instead of a normal vertical borehole, and others (Carter-Journet et al., 2014a,b). This phenomenon has a detrimental effect on the uptime of systems, productivity, the costs of maintaining such systems, and the operational dependability of service providers and drilling operators (Carter-Journet et al., 2014a,b; Beckwith, 2013). The most typical downhole issues are pipe sticking, pipe failure, dog legs and telescopic holes (crookedness of hole), key seats in holes, shale problems, and lost circulation problems.

4.1. Pipe sticking

When excavating deep, high-pressure wells in geologically difficult locations, pipe sticking is one of the most common challenges that might arise, and it can be both expensive and hazardous. Basically, there are three main types of pipe sticking.

4.1.1. Mechanical sticking

The pipe gets stuck because of mechanical factors such as distorted wellbore geometry, an undergauge hole, poor hole cleaning, key seating, junk in the hole, cement related problems, and collapsed casing (Elmgerbi and Thonhauser, 2022). When drilling through particularly major formations, pipe sticking may sometimes be an undesired consequence. An unstable wellbore has a greater risk of caving in, collapsing, or sloughing (flowing inward). By reinforcing the wellbore and utilizing an appropriate mud system that is compatible with the formation physically and chemically, mechanical sticking caused by wellbore instabilities can be prevented. Lake (2006). In directional drilling, a small hole will be made into the side of the wall by the drill string rotating with a side force (lateral force) acting on it (groove). These grooves can be found on undetected ledges along washouts or at doglegs (Lake, 2006).

4.1.2. Differential sticking

A portion of the drill pipe becomes stuck in the mud cake as a result of this circumstance. This takes place when the pressures in the wellbore and those in the formation are extremely different from one another. Hydrostatic pressure, formation pore pressure, sand content, well bore diameter, the radius of drill collar, mud pressure, contact area (drill collar and mud cake), coefficient of friction, depth of well, mud weight, cake thickness, and specific gravity of the mud are the contributing factors (Lake, 2006). If the angle of a well is less than 5^0, it is deemed straight. Even in a straight, deep hole, deviations of 6^0 to 8^0 are allowed (Kayode and Lami, 2020). For example, in a 1000 meter well, if the angle is10°, drift or displacement is 1000 x Sin 10 = 1000 x.01745 = 17.4 m; for 20, drift per 1000 m = 35 m; for 30, the said drift is 52.5 m and so on increasing with angle. As such contact with hole wall is likely and unavoidable.

4.2. Dog leg

Dog leggings are a huge problem in drilling. No hole is precisely vertical, and every hole has a tendency to spiral, according to recent surveying methodologies. A dogleg is a problematic circumstance that occurs from a sudden shift in hole deviation (inclination and/or azimuth). The factors contributing are dog leg angle, dog leg severity (degree per 10 m), length of string below key set, and change in azimuth over the interval (degree $\Delta\phi$). Drill pipe or drill collars experience bending forces at the dog leg point. As a result, there are large variations in axial stress in the pipe cross-section. It is maximum at the outside point and minimum at the inner point. However, string rotation quickly moves these locations, and high-degree stress reversal occurs at every cross-sectional location. When the stress exceeds its limiting value, fatigue failure of the drill pipe or drill collar occurs. A thrust force on the formation in contact with the pipe enters play at the dog leg point as a result of the pipe bending and string tension attempting to straighten the same. A keyhole can be dug using this thrust force on drill pipe tool joints with rotating energy.

4.3. Pipe failure

Drill string pipes are made up of two separate components: the pipe's body and its connector. However, for drill collars, the connection is weaker than the body's strength. A new drill pipe's connection (tool joint) is stronger than its body. The drill pipe fails at the tool joint, while the drill collar fails at the connection. The most frequent failures are due to mud clogging and tool joint failure (which accounts for 50% of fishing jobs).

5. Analysis of downhole problems using big data

Big data predictive analysis can be implemented to find the best method for detecting downhole problems, which will help oil and gas firms in identifying trends and predicting occurrences throughout their operations to swiftly respond to disturbances and enhance operational efficiencies. By using big data techniques, the focus will be on analyzing existing data to make accurate forecasts of future events. The general steps involved in the prediction technique are categorized into five groups.

(a) Collection of downhole drilling data using sensors.
(b) Extraction, processing, loading, cleansing, and storage of the data, in addition to other data-collecting pre-procedures.
(c) Additional calculation, analysis, and data mining on the acquired data
(d) Intelligent modeling using big data technologies to forecast future downhole problems
(e) Use of risk predictive windows to implement early safety warnings and identification of the downhole problem based on association relationships.

5.1. Collection of downhole drilling data using sensors

The sensors mounted on drill pipes and equipment collect a large volume of data, which is crucial for the predictive model. The oil and gas industry is known for its large purchases of various types of sensors, such as acoustic sensors, temperature sensors, and pressure sensors (Xie et al., 2018). These sophisticated sensors are used to capture, measure, and collect a vast assortment of downhole drilling operations, environmental data, and dynamic data.

The temperature sensor measurement can be air temperature, liquid temperature, or the temperature of solid matter. The most widely utilized temperature sensors in contemporary electronics are thermocouples, thermistors, Resistance Temperature Detectors (RTDs), and semiconductor-based Integrated Circuits (IC) (Li et al., 2022). Acoustic sensors have a broader function than just the detection of sound waves. The role of acoustic sensors extends

Table 2
Downhole Problems, factors, and impacts.

Downhole problems	Factors	Impacts	Reference
Pipe Sticking	Hydrostatic Pressure, Formation Pore Pressure, Sand Content, Well Bore Diameter, Radius of Drill Collar, Mud Pressure, Contact Area (drill collar and mud cake), Coefficient of Friction, Depth of Well, Mud Weight, Cake Thickness, Specific Gravity of Mud	Equipment failure Stress Occurs Property damage, including clean-up and recovery costs, lost product value, and operator property damage.	Haghshenas et al. (2008)
Dog Leg	Dog Leg Angle, Dog Leg Severity, Degree per 10 m, Length of string below key set, Change in azimuth over the interval, (Degree $\Delta\ \phi$)	Harm to local water and air pollution	Lyons et al., (2012)
Pipe Failure	Torque, Diameter of the tool joints, Drill Collar Size	Vibration of Equipment Stress Occurs Death of any person Explosion or fire not intentionally set by the operator. Operator, product, and property harm. Equipment failure	Rezaei et al. (2015)

beyond the detection of sound waves. Specifically, their use for sensing mechanical vibrations in solids for the manufacture of microbalances and Surface Acoustic Wave (SAW) devices grew in popularity (Fraden, 2010). In oil and gas applications, pressure sensors are used extensively, with the measurement procedure involving almost all pressure categories, including gauge, absolute, differential pressure, high pressure, and micro differential pressure. Pressure sensors can be categorized according to the method by which they can sense pressure changes, strain gauge, piezoelectric, capacitive, manometers, and bourdon tube.

There are thousands of sensors strategically placed to monitor the site. A significant number of sensors are required by the digitalized oilfield process to collect data through modeling, monitoring systems, real-time data optimization, and control drilling. Monitoring is done for a number of characteristics, including vibrations, pressure, fluid flow rate, temperature, and fluid stuck up. Real-time performance is evaluated by pipe modeling and visualization, and parameters are logged.

Several characteristics are monitored, and data is captured at regular intervals, providing frequent assessments of the downhole environment, drilling output, and downhole characteristics. The operator will be provided with real-time, discontinuous, and discrete downhole drilling data, from which the engineer will be able to extract the temperature, vibrations, pressure, fluid flow rate, and other measured variables that may contribute to issues downhole such as pipe sticking, dog legs, and pipe failure. The downhole failure, parameters, and their effects are given in Table 2.

5.2. Big data storage and management

The downhole drilling data obtained is an enormous, unstructured, and complicated data collection that is challenging for conventional data processing technologies to manage (Chen et al., 2014). The information gathered by the sensors is multidimensional, and due to the ever-increasing amount of data being created, quicker and more effective methods of data analysis have become necessary. Along with the necessary infrastructures for storing and managing enormous data, there are also specific tools and methodologies for big data analytics that are essential for making successful judgments at the proper time (Elgendy and Elragal, 2014). There are techniques and tools which can analyze (process, decode, and interpret) the operation status and the change in parameters simultaneously (Kale et al., 2015; Pritchard et al., 2016). The data are gathered by drilling operators or service providers, and downhole data acquired from global geographic drilling operations will continually amass and grow in quantity, ultimately becoming a dataset that exceeds the storage and processing capacity of a single server (Chen et al., 2014). Multiple distributed servers are used to store, transmit, and process the collected data before extracting, transforming, and loading it into different databases for advanced analytics. These data are very large, ranging in size from Terabytes (TBs) to Petabytes (PBs) (Meeker and Hong, 2014). Moreover, large data sets may have considerable variability, hindering the data processing and administration, and varying integrity as a result of data inconsistency, incompleteness, complexity, delay, deceit, assumptions, and horizontal scalability to merge dissimilar information (Chen et al., 2014; Elgendy and Elragal, 2014; Hu et al., 2014). Relational databases, data marts, and data warehouses are classic techniques for storing and retrieving structured data.

Several solutions, such as distributed databases and Massive Parallel Processing (MPP) databases for delivering high query productivity and platform stability, as well as non-relational databases, were utilized for big data. Non-relational databases, such as NoSQL, are created to store and manage non-relational data. NoSQL seeks huge scalability, a flexible data format used for streamlined creation and deployment of applications. Compared to relational database systems, NoSQL decouples data storage and management. These databases emphasize scalable, high-performance data storage and enable data administration operations to also be done at the application level as opposed to database-specific languages (Bakshi, 2012).

After storing the data, it has to be analyzed using big data tools and techniques. According to (He et al., 2011), there are four essential needs for processing large amounts of data. The foremost prerequisite is rapid data processing. Due to the fact that disk and internet traffic conflicts with request performance during performing data, it is vital to minimize the time necessary for performing data. The next criterion is the speed of query execution. Many queries are response-time essential due to the demands of high workloads and real-time requests. As a result, the data placement structure needs to be able to maintain high query processing rates as the number of inquiries grows quickly. Consequently, the prerequisite for large-scale data collection is the efficient use of storage capacity.

Due to limited disk space, it is essential that data storage must be carefully handled throughout processing and that challenges regarding the storage of the data should be minimalized. The quick expansion in user behavior might need extensible storage space and processing speed. The ability to adjust well to workload patterns that are extremely dynamic is the final need. Massive

datasets are processed by a variety of applications and consumers, for a variety of purposes in a diverse range of ways, requiring the operating system to be highly adaptive to unforeseen processing dynamics and not specific to each of these workload patterns (He et al., 2011).

5.3. Big data analysis

5.3.1. Challenges in big data analysis

Data mining is the extraction of relevant information and insights from large datasets using statistical and computational methods. Data mining is an integral part of big data analytics, which entails processing, analyzing, and interpreting large and complex datasets to discover patterns, trends, and insights that can assist organizations in making informed decisions. Nonetheless, data extraction for big data analytics is not error-free. During the data collection process, these errors can influence the quality and dependability of the insights generated from the data (Amirian et al., 2015). Examples of common data collection errors include:

- Sampling errors: When the sample data used for analysis is not representative of the population as a whole.
- Measurement errors: When the data collected is not accurate or trustworthy.
- Data entry errors: Errors in data entry occur when information is erroneously recorded.
- Processing errors: Errors in processing occur when data is improperly processed.

Data cleansing, data validation, data normalization, and data transformation are some of the methods used by data mining and analytics practitioners to reduce the likelihood of these errors. Data cleansing is the process of identifying and rectifying data errors and inconsistencies, such as absent values, outliers, and duplicate entries. Validating data involves validating its accuracy and completeness, such as by ensuring that all required fields are populated. The normalization of data entails transforming the data into a standard format, such as converting all dates to a common format. For example, consider a scenario in which a drilling company wishes to improve the precision of its drilling operations by analyzing data collected from downhole sensors. The collected data include measurements of temperature, pressure, and other parameters that help the company determine the characteristics of the drilled rock formation. Nevertheless, errors may occur during the data collection procedure, such as faulty sensors or incomplete data due to technical issues. These errors can result in erroneous analysis and may lead to drilling in the incorrect location, resulting in increased costs and decreased efficiency (Liu et al. 2022).

5.3.2. Tools and technologies

The gathered downhole drilling data must be processed quickly and precisely. Apache Hadoop, Map Reduce, MongoDB, and Cassandra are widely used big data analysis tools, with Cassandra being utilized extensively in the oil and gas sectors. Through real-time recording, big data technology discovers, examines, and extracts important information from a mass of data. MapReduce and Hadoop implement parallel processing models and frameworks for doing large data analytics with effectiveness, dependability, scalability, and manageability (Hu et al., 2014), and mostly their technological architecture comes from Apache Hadoop.

5.3.2.1. Apache Hadoop. Hadoop is known for its open-source, extension-based platform used for distributed big data analysis programs that derive from the Apache open-source project. It is also a sort of software architecture that distributes processing on massive amounts of data (Le, 2022). Linear expansion is the main feature of this platform, which is extremely portable and compiled in Java. The MAD system helps to load data files into the disturbed file system and run them in parallel to MapReduce computation of data. Using Hadoop, data can simply be copied and loaded, which is later interpreted by MapReduce at processing time instead of loading time. Herodotos Herodotou et al. (2011) Hadoop can process all types of data sources, also it can adapt to any changes occurring in the data sources (Cuzzocrea et al., 2011). In terms of data warehousing, Hadoop is quite effective. Any node performance concerns lack an impact on the server because the Hadoop data is stored and backed up on a minimum of three nodes. The data is accessible without the user's permission. Hadoop falls short, however, in terms of datasets and real-time analysis. Currently, implementing Hadoop's data warehouse is a prudent solution. The initial release lacks compliance with other high-level programming languages.

5.3.2.2. Map reduce. Map Reduce is a parallel programming approach, derived from "Map" and "Reduce" in functional programming languages that is appropriate for processing large amounts of data. It is Hadoop's central processing unit and is responsible for data processing and analytics (Cuzzocrea et al., 2011). The MapReduce paradigm is built on adding additional computers or resources as opposed to expanding the processing power or storage capacity of a single computer; in other words, scaling out as opposed to scaling up [Data Science and Big Data Analytics (2012)]. In order to minimize the time required to accomplish a work, the core concept of MapReduce is to divide it into subtasks and execute them in parallel (Cuzzocrea et al., 2011).

5.3.2.3. Peer-to-peer networks. This technique employs a decentralized and distributed network architecture that interconnects millions of devices. As each system node is capable of storing data, the storage capacity is almost endless. Message-passing interfaces allow each node to communicate and share data. This interface supports the hierarchical master/slave paradigm, which enables the effective utilization of the slave's computing resources through dynamic resource allocation (Datta et al., 2006; Nguyen et al., 2020; Singh and Reddy, 2015). The primary disadvantage of this interface is its lack of fault tolerance, as there is no method for addressing errors. Consequently, the fai lure of a single node might result in the shutdown of the entire network. The system also has a connectivity bottleneck, which hinders data synthesis and processing in real-time.

5.3.2.4. MongoDB. This is written in C++, a document-oriented technology using a NoSQL database based on JavaScript Object Notation (JSON). NoSQL database technology has the ability to manage large datasets, like documents. JSON is based on JavaScript, which can process data and build a list of value pairs or a collection of values. MongoDB offers a flexible and adaptable framework that may be altered to meet all the needs of different clients (Trifu and Ivan, 2016).

5.3.2.5. Cassandra. It is particularly effective in situations where spending extra time learning a difficult system can provide you with a lot of power and flexibility. It also uses NoSQL data technology and is column-oriented (Hashem et al., 2015).

5.4. Intelligent modeling using big data technologies to forecast future downhole problems

The fundamental objective of the model is to monitor the anomalies during the observed downhole performance characteristics with utmost precision and time. Therefore, the information that will be used for training and building models must

match particular criteria. The first and foremost requirement is that the data should be free of error. Secondly, the collected data must have a suitable size. At the beginning of each phase, the data sets collected in the first few hours will be utilized to train and build the model. Depending on the data resolution, the time required to collect adequate data sets to be used as inputs will differ depending on the rig (data sampling frequency). The patterns of data sets derived from downhole drilling data utilizing big data analysis are incorporated into the prediction model, which will assess the environment in the well hole. The model will focus on identifying and forecasting three major downhole problems (pipe sticking, dog leg, and pipe failure).

Various approaches have been presented thus far to identify early indications and create warnings for the major prevalent downhole challenges. Even though most modern developments have sought to address the evident shortcomings of the older ones. These are limited, particularly for real-time applications, which was the main impetus for the discussed function. This paper provides a predictive model which can identify and characterize the frequent indicators of unwanted downhole accidents at their earliest phases (Tsuchihashi et al., 2021; Wang et al., 2020). The selection and classification of training data frames have a substantial effect on these models' performances.

Wang et al. (2020) introduced the work on recognizing the most prevalent undesired downhole occurrences using a novel machine learning-focused approach. The method utilizes real-time data from downhole drilling to identify anomalous circumstances. The first method was used to classify the drilling process, while the second algorithm was used to find anomalies. The approach was evaluated to determine its ability to discern between typical and irregular downhole drilling circumstances with a negligible rate of false alarms. For discovering underlying incidence early, there must be similarities in the perspective. Minor modifications in the well configuration (setup) or downhole circumstances will result in an entirely different situation. Consequently, a substantial number of false alarms are anticipated (Lee et al., 2021).

Classical models such as linear regression, penalized regression, partial least squares, Support Vector Machines (SVM), and Artificial Neural Network (ANN) for nonlinear regression are used generally as learning techniques (Xaio, 2015; Bishop, 2006; Vapnik, 1995). It has been found that these models are effective in solving a wide variety of problems in a variety of sectors, such as engineering, finance, and healthcare, among others. As an example, linear regression is frequently utilized when attempting to predict the connection between two variables, but SVM are frequently utilized when attempting to classify data (Hastie et al., 2009).

Classification and regression trees are the most popular techniques for developing analytics of explorative data and prediction models (Loh, 2011; Xaio, 2015; Breiman et al., 2017). These models make use of a hierarchical framework that divides the data into subsets, which makes it possible to identify complicated patterns within the data. Primarily, the prediction model used here comprised three learning techniques: ANN, Decision tree algorithm (Quinlan, 1986), and Support vector algorithm to develop a prediction model that can detect downhole problems. These models have been extensively studied in the literature and have exhibited promising results in various applications (Breiman et al., 2017). The purpose of combining multiple learning techniques was to create a reliable and precise prediction model capable of detecting anomalies in downhole data.

5.4.1. Learning algorithm using ANN

The ANN model is designed and studied using Neuroshell 2V4.0, an accessible commercial neural network analysis, and modeling application program (El-Abbasy et al., 2014). Established data for a selection of parameters are utilized for training the ANN throughout to provide quality prediction models based on the ANN. The collected parameter data sets are divided into different aspects, such as 60% for training and 20% each for testing and validation. Fig. 1 depicts the predictive model (flow chart) to identify the downhole problems using big data.

The data set is used according to the above, 60% of the training data sets are used to train the network, while the testing data sets are modified or trained continuously and rectified by modifying the weight of the network links. Error-back propagation neural network, or BP neural network, is the main ANN model. This comprises three parts: an input layer, a hidden layer, and an output layer. $f(x) = \mathrm{f(x)} = 1/(1+e\text{^-x})$ commonly adopted, the learning set has M sample mode (XP, Yp). For the P training sample ($P = 1, 2, \ldots \mathrm{M}$), the total of the node j's inputs is recorded as net_{pj}, while its outputs are recorded as

$$O_{pj}, net_{pj} = \sum_{j=0}^{N} W_{ji} O_{pj} = f(net_{pj}), \tag{1}$$

Unless the initial weight value of the network is chosen arbitrarily, the error between the network output and anticipated output d_{pi} for each input sample P may be represented as

$$= \sum E_P = \left(\sum_j (d_{pi} - O_{pj})^2 \right) /2 \tag{2}$$

Correction formula of weight value of the network

$$W_{ji} = W_{ji}(t) + \eta \delta_{pi} O_{pj} \tag{3}$$

$$\delta_{pj} = \begin{cases} f\left(net_{pj}\right)\left(d_{pi} - O_{pj}\right), & output \\ f\left(net_{pj}\right) \sum_k \delta_{pk} W_{kj}, & input \end{cases} \tag{4}$$

5.4.2. Learning algorithm using decision tree algorithm

A decision tree is a non-parametric, guided learning approach that is used for regression and classification applications. It features a tree-like, hierarchical structure with a root node, branches, internal nodes, and leaf nodes. The information theory-based decision tree algorithm (Kamble and Gunasekaran, 2020) incorporates the ID3, CART, and C4.5 algorithms. The ID3 algorithm is the ancestor of the C4.5 and CART. The ID3 algorithm works as follows:

- Entropy of categorization system information

Consider the sample space (D,Y) of a classification system, where D represents for sample (m characteristic) and Y refers for n category, and where *C1, C2, ….., Cn* are the potential values. Each category's occurrence probability is denoted by *P(C1), P(C2), ….., P(Cn).*This categorization system's entropy is:

$$H(C) = -\sum_{i}^{n} P(C_i) .log_2 P(C_i) \tag{5}$$

In discrete distributions, the occurrence probability $P(C_i)$ of category C_i may be calculated by dividing the occurrence time of that category by the entire sample size. For continuous distribution, zone-by-zone discretization is often necessary.

- Conditions-dependent Entropy

Based upon the definition of conditional entropy, the conditional entropy in the categorization is the informational entropy

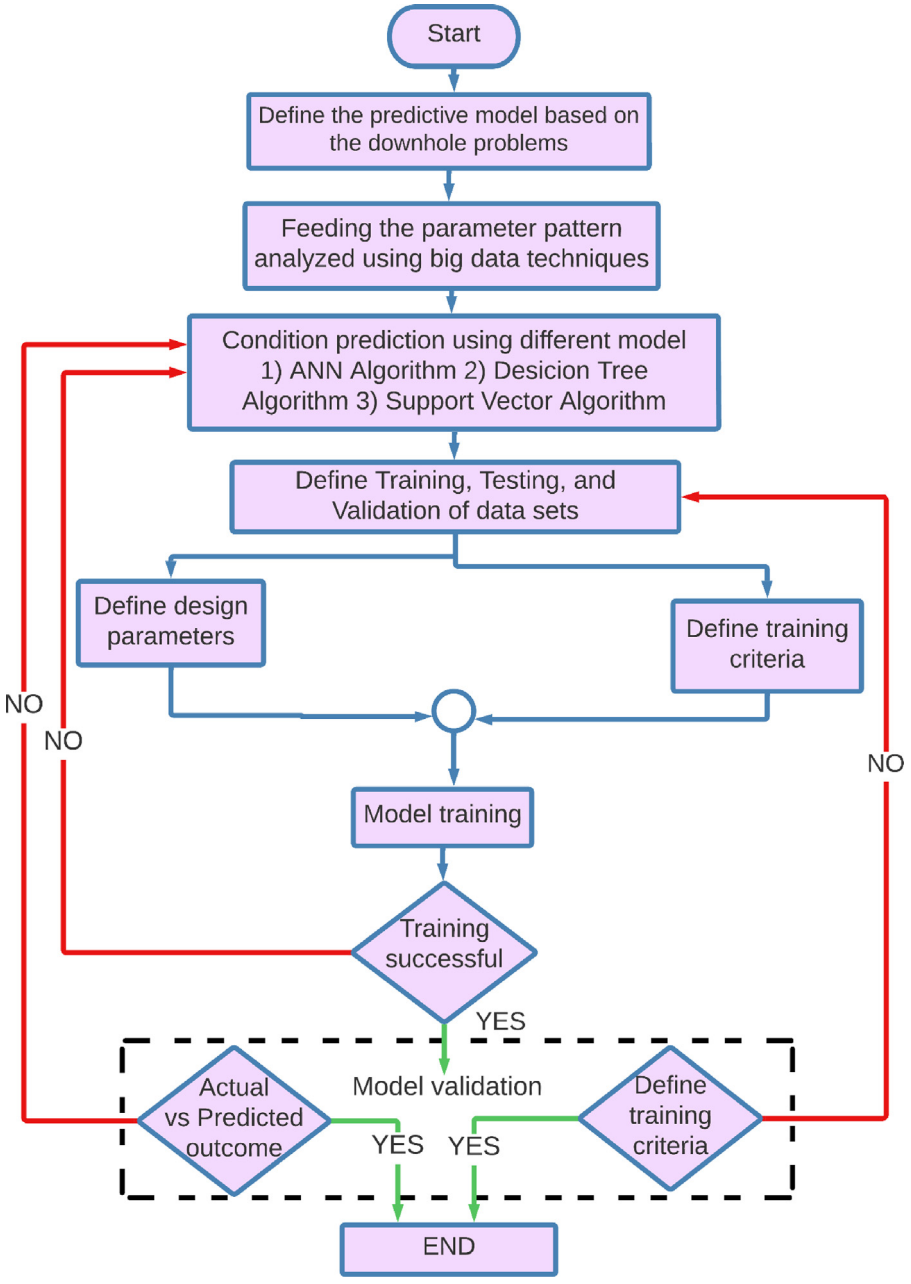

Fig. 1. Predictive model (flowchart) to identify the Downhole Problems Using Big Data.

when a fixed sample's characteristic X is considered. Since the potential values of this Characteristic X include $x_1, x_2, \ldots\ldots\ldots, x_n$, it is necessary to fix it in the computation of conditional entropy, every possible value must be fixed. Following this, the statistical expectation is calculated. Hence, the possibility of sampling Characteristic X value $x_i isP_i$, the conditional information entropy when this characteristics is fixed as value $x_i isH(C|X = x_i)$, and $H(C|X)$ is the conditional entropy when Characteristic X is set in the classification method $(X = (x_1, x_2, \ldots\ldots, x_n))$

$$H(C|X) = P_1H(C|X = x_1) + P_2H(C|X = x_2) + \ldots\ldots.P_nH(C|X = x_n) = \sum_{i=1}^{n} P_iH(C|X = x_i) \tag{6}$$

The provided algorithm is intended to be autonomous in identifying and categorizing the downhole event, since the drilling sector is moving towards to the new digital age to limit social intervention, especially for key choices connected to security and the sustainability.

5.4.3. Learning algorithm using support vector algorithm (SVM)

The SVM model (Hansen, 2020) successfully overcomes many of the flaws of the ANN optimizing approach. On the basis of the idea of structural risk reduction, the SVM model ensures the teaching approach's generalizability. In addition, the SVM optimization procedure may also be transformed into a problem of nonlinear optimization. Thus, the algorithm's worldwide efficiency can be ensured, and the neural network's locally optimal problem can be solved. In addition, the support vector machine has a rigorous conceptual and analytical basis that eliminates the experience aspect of neural networks. Furthermore, the SVM model can be extended to multi-class classification problems by applying either the one-vs-one or one-vs-all methods. The SVM builds a classifier for each pair of classes in the

one-vs-one method, and the final prediction is reached by casting a vote between these classifiers. The SVM builds a classifier for each class in the one-vs-all method, and the prediction is then made by choosing the class with the highest score.

5.5. Model setup stage

After developing all the prediction models, they are assessed by applying the same training and development datasets. Utilizing the validation dataset, all created model's performances are evaluated. Each model's evaluation is based on specific mathematical validation criteria (Al-Barqawi and Zayed, 2006; Zayed and Halpin, 2005). The accuracy and error of the different algorithm models are then verified then using the validation data set. The model with the minimum error, best accuracy, and best model performance is selected.

$$AIP = \left\{ \sum_{i=1}^{n} \left| 1 - \left(\frac{E_i}{C_i} \right) \right| \right\} * \frac{100}{n} \tag{7}$$

$$AVP = 100 - AIP \tag{8}$$

$$MAE = \frac{\sum_{i=1}^{n} |C_i - E_i|}{n} \tag{9}$$

$$f_i = \frac{1000}{1 + MAE} RMSE = \sqrt{\sum_{i=1}^{n} \frac{(C_i - E_i)^2}{n}} \tag{10}$$

$$RE = \sqrt{(P_{actual} - P_{Predicted})^2} \tag{11}$$

where AIP is the Average Invalidity Percent, AVP is the Average Validity Percent, RMSE is the Root Mean Squared Error, MAE is the Mean Absolute Error, RE is the Residual Error, P_{actual} is the Actual data utilized in creating the prediction model and $P_{Predicted}$ is the Equivalent predictive data that the prediction model generates.

All three models are evaluated based on the above equations. To anticipate the mistake, Eqs. (7) and (8) display the average validity/invalidity percentages (i.e., AVP and AIP). The model is said to be sound for AIP values around 0.0 but is not robust for AIP values near 100. Likewise, the Root Mean Square Error (RMSE) is computed by the Eq. (11). RMSE value should be near zero if the model has high efficiency. The Mean Absolute Error (MAE) is given by equation (Chen et al., 2012) and has a range of zero to infinity. Reliable yields should be around zero (Trifu and Ivan, 2016). MAE can also determine the fitness function fi of the given models (Trifu and Ivan, 2016), given in Eq. (10). The equation suggests if a model's fi value is near 1000, considered to be valid. Else it is taken as invalid. In the next stage, using equation (Chen et al., 2012), the Residue Error (RE) is calculated for every single data point independently in all three models.

5.6. Risk-predictive windows

This step begins by running two required sub-functions, one of which determines the threshold value that may be used as a benchmark to track the real-time performance of prediction models. While the second function's primary responsibility will be to determine the configurations of the risk-predictive window (safe operational window sizes) for the continuous drilling process when the model identifies the concerned downhole problem (pipe sticking, dog leg, pipe failure), which eventually will alert the alarms and propose the user to make the decision. Real-time data fluctuations caused by sensor faults or other causes might trigger a false alert; this can be avoided by removing dubious data points.

5.7. Use of game theory to evaluate the best predictive model

The use of game theory (Wen et al., 2000) can helps to identify the most possible model to predict downhole problems, with Nash Equilibrium (NE) being one of the main concepts in this theory. By solving the game theory model, this NE will indicate the best possible model. Game theory is a useful additional concept that data scientists can apply to predict how rational engineers will make decisions. The major components that help in analyzing a data-driven decision-making problem include the set of options or choices available. The set of outcomes is based on these choices. A decision analysis based on game theory would be promising for finding the optimal risk criteria for downhole problems. Prediction of the downhole problem has been challenging. In general, the workflow chart can be summarized as follows:

1. Identify two cases, i.e., entities with conflicting interests
2. Find the models for each case
3. Obtain the payoff table with case A's model in rows and case B's model in columns. The payoff of case A should be listed in the first table, and that of case B be listed in the second table
4. Can use Game Theory Explorer (GTE) to find the NE
5. Analysis of the NE and locating the model that minimizes downhole problems

Assumptions used in game theory for selecting the best algorithm in downhole drilling:

- Rationality: All agents, such as the drilling operators or software algorithms, are assumed to be rational decision-makers who aim to maximize their own utility or payoffs.
- Complete information: All agents have complete information about the game, including the rules, strategies, and payoffs. In this case, the drilling operators and algorithms have access to all the relevant data and information about the downhole drilling process.
- Non-cooperative behavior: All agents behave non-cooperatively, meaning they do not form alliances or coordinate their actions with others. In this case, the drilling operators and algorithms act independently to maximize their own utility.
- Zero-sum game: The selection of the best algorithm can be viewed as a zero-sum game where the total payoff of one agent is equal to the total loss of the other agent. For example, if the drilling operators choose the wrong algorithm, it may result in increased drilling time and costs.
- Simultaneous moves: The drilling operators and algorithms make their decisions simultaneously. In other words, the selection of the best algorithm is made based on the available data at the time, such as cumulative drilling hours and temperature.

The first step in constructing a game theory analysis is to write down the names of the entities involved. The major aspects are ANN, SVM algorithm, and the decision tree algorithm. The second step is listing the most relevant choices available in each case to solve the problems. For example, in the case of each model, one can consider data performance and economy. The third step involves creating the scenario matrix. The advantage of having a scenario matrix is that it reveals all conceivable decision combinations and allows you to consider each one individually. In many instances, this step proves to be the most important step in the process. The next step is to create a pay-off matrix. A payoff matrix is a mechanism to represent the outcomes of each case decision. The payoff matrix is created by selecting values from the scenario matrix. The last step includes looking for dominant strategies and NE. A dominating strategy is a decision that is

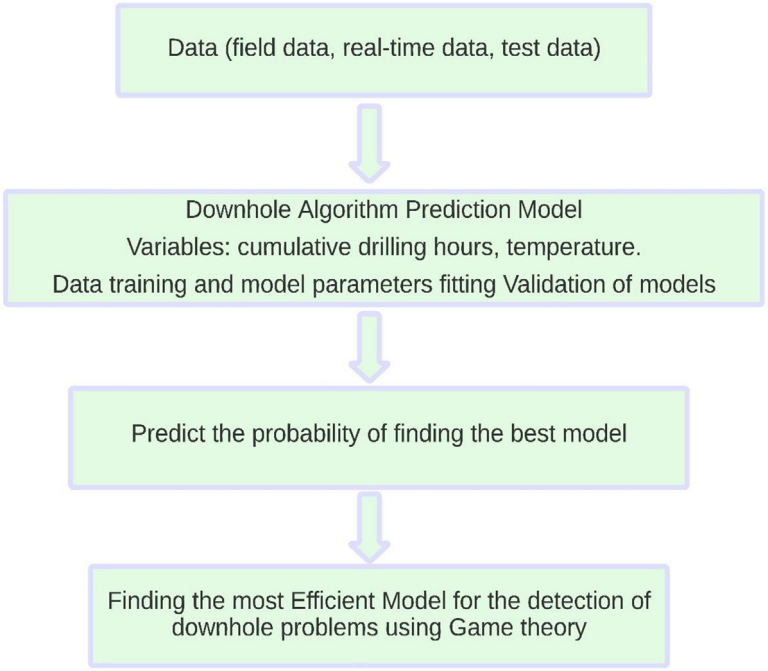

Fig. 2. Methodology for identifying suitable model using game theory.

favorable. A NE depicts the ideal state of a downhole problem in which optimal decisions are made. The methodology for the prediction of downhole problems is depicted in Fig. 2.

5.7.1. Game theory and adversarial learning

In machine learning, adversarial learning and game theory are two distinct approaches. The mathematical framework of game theory is used to analyze and model the strategic decision-making of multiple agents or participants, each with their own preferences and objectives. In contrast, adversarial learning is a machine learning technique in which a model is trained to perform well in the presence of an adversarial attack. While both approaches entail strategic decision-making and can be applied to similar circumstances, their underlying objectives and methodologies are distinct. Game theory is used to analyze and model interactions between agents and to determine the optimal strategies for each participant in a given scenario. In contrast, adversarial learning is used to train models that are resistant to attacks by malignant agents (Zhou et al., 2019).

5.8. Effectiveness of big data analytics and its economic benefits

In recent years, the effectiveness of big data analytics in the oil and gas industry has been widely acknowledged as it has evolved into a crucial instrument for enhancing operational efficiency and lowering costs. With the increase in data generated by sensors and other sources during exploration, drilling, and production, big data analytics has enabled businesses to obtain operational insights and make data-driven decisions.

According to the case study (Santos et al., 2023), an oil company needed to optimize their workover rig scheduling to minimize costs and production losses. The study developed a data-driven optimization methodology using text mining, clustering, and regression modeling to predict workover duration and optimize rig schedules. Computational experiments showed superior performance and improved prediction accuracy, with solutions deviating less than 10% and requiring less rescheduling compared to the company's current methodology.

By implementing best safety practices initiatives detected by an automated drilling condition detection monitoring service, (Duffy et al., 2017) improved the drilling rig's efficacy. In their case study on pad drilling in the Bakken formation, they focused on W2 W connection time during drilling operations. According to their findings, a single cluster of nine wells drilled with the same equipment saved more than 11.75 days. In addition, overall non-drilling time was reduced by 45 percent.

Big Data was used to construct a methodology to investigate the effects of completion parameters on well productivity in a study conducted by (Khvostichenko and Makarychev-Mikhailov, 2018). They collected data from 4,500 wells receiving slick water treatment. They studied the effects of two distinct compounds, linear guar gels and flow back aides based on surfactants. In addition, they extracted the monthly production figures from the IHS Energy databases. The t-test was adopted to evaluate the data statistically.

5.9. Opportunities and critical challenges

There are numerous opportunities for big data analytics to enhance the oil and gas industry, as well as significant obstacles that must be overcome to ensure a successful implementation. One opportunity is the ability to monitor equipment in real time and anticipate potential issues in advance, thereby reducing downtime and boosting productivity. The methodology flowchart for detecting downhole drilling problems is given in Fig. 3. Moreover, big data analytics can enhance the precision of reservoir modeling and optimize production processes. Utilizing big data analytics can also result in cost savings by identifying operational inefficiencies and reducing wasteful expenditures.

6. Conclusions

This paper presents a framework for predictive big data analytics to forecast and investigate downhole problems within the oil and gas industry. The adoption of data recording sensors in exploration, drilling, and production processes has resulted in the

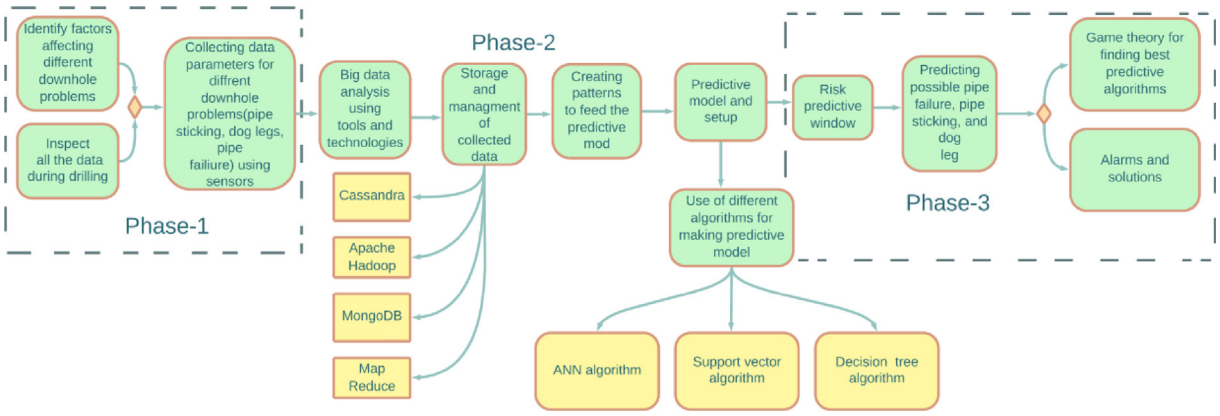

Fig. 3. Methodology flowchart for detecting downhole drilling problems.

transformation of the industry into a data-intensive sector that has rapidly generated vast quantities of diverse data. A variety of big data methodologies were investigated to determine the paradigm shift in data storage and processing, thereby facilitating the identification of downhole challenges such as pipe sticking, dog leg, and pipe failure. Various established predictive models were evaluated using risk prediction windows to identify upcoming irregularities for the prevention of accidents. The best predictive model for identifying downhole problems was determined using game theory. The economic benefits of big data analytics in the oil and gas industry cannot be overstated, as it can help reduce costs, increase efficiency, and enhance safety. However, a number of significant obstacles, including the complexity of data, a scarcity of qualified personnel, and data privacy concerns, must be overcome for successful implementation. Future research could investigate the use of machine learning algorithms and artificial intelligence to enhance the predictive accuracy of models and decision-making in the industry. To realize its maximum potential, additional research could be conducted on the implementation of big data analytics in other oil and gas industry sectors, such as refining and distribution.

CRediT authorship contribution statement

Aslam Abdullah M.: Conceptualization, Investigation, Formal analysis. **Aseel A.:** Drafting of the document. **Rithul Roy:** Drafting of the document. **Pranav Sunil:** Drafting of the document.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article

Acknowledgments

The authors are thankful to Kappa Software for providing Academic Licensed Software to perform the research. We thank the administration of Vellore Institute of Technology for providing all the necessary facilities. All authors read and approved the final manuscript.

References

Ahn, S., Couture, S.V., Cuzzocrea, A., Dam, K., Grasso, G.M., Leung, C.K., McCormick, K.L., Wodi, B.H., 2019. A fuzzy logic based machine learning tool for supporting big data business analytics in complex artificial intelligence environments. FUZZ-IEEE 1–6. http://dx.doi.org/10.1109/FUZZ-IEEE.2019.8858791.

Al-Barqawi, H., Zayed, T., 2006. Condition rating model for underground infrastructure sustainable water mains. J. Perform. Constr. Facil. 20 (2), 126–135. http://dx.doi.org/10.1061/(ASCE)0887-3828(2006)20:2(126).

Alam, Md T., Ahmed, C.F., Samiullah, Md, Leung, C.K., 2021. Mining frequent patterns from hypergraph databases. In: PAKDD, Part II. pp. 3–15. http://dx.doi.org/10.1007/978-3-030-75765-6_1.

Amirian, E., Leung, J.Y., Zanon, S., Dzurman, P., 2015. Integrated cluster analysis and artificial neural network modeling for steam-assisted gravity drainage performance prediction in heterogeneous reservoirs. Expert Syst. Appl. 42 (2), 723–740. http://dx.doi.org/10.1016/j.eswa.2014.08.034.

Baaziz, A., Quoniam, L., 2013. How to use big data technologies to optimize operations in Upstream Petroleum Industry. Int. J. Innov. 1 (1), 19–25. http://dx.doi.org/10.5585/iji.v1i1.4.

Bakshi, K., 2012. Considerations for big data: architecture and approach. In: 2012 IEEE Aerospace Conference. pp. 1–7. http://dx.doi.org/10.1109/AERO.2012.6187357.

Beckwith, R., 2013. Downhole electronic components: Achieving performance reliability. J. Pet. Technol. 65 (08), 42–57. http://dx.doi.org/10.2118/0813-0042-JPT.

Belhadi, A., Zkik, K., Cherrafi, A., Yusof, S.M., el fezazi, S., 2019. Understanding big data analytics for manufacturing processes: Insights from literature review and multiple case studies. Comput. Ind. Eng. 137, 106099. http://dx.doi.org/10.1016/j.cie.2019.106099.

Bile Hassan, I., Liu, J., 2020. A comparative study of the academic programs between informatics/BioInformatics and data science in the U.S.. In: IEEE 44th Annual COMPSAC. pp. 165–171. http://dx.doi.org/10.1109/COMPSAC48688.2020.00030.

Bishop, C.M., 2006. Pattern recognition and machine learning. Springer, p. 4. http://dx.doi.org/10.1007/978-0-387-31073-2.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 2017. Classification And Regression Trees. Routledge, http://dx.doi.org/10.1201/9781315139470.

Carter-Journet, K., Kale, A., Falgout, T., Heuermann-Kuehn, L., 2014a. Drilling optimization: utilizing lifetime prediction to improve drilling performance and reduce downtime. http://dx.doi.org/10.2118/170270-MS.

Carter-Journet, K., Kale, A., Zhang, D., Pradeep, E., Falgout, T., Heuermann-Kuehn, L., 2014b. Estimating probability of failure for drilling tools with life prediction. http://dx.doi.org/10.2118/171517-MS.

Chen, Chiang, Storey, 2012. Business intelligence and analytics: From big data to big impact. MIS QUART 36 (4), 1165. http://dx.doi.org/10.2307/41703503.

Chen, M., Mao, S., Liu, Y., 2014. Big data: A survey. Mob. Netw. Appl. 19 (2), 171–209. http://dx.doi.org/10.1007/s11036-013-0489-0.

Cuzzocrea, A., Leung, C.K., Deng, D., Mai, J.J., Jiang, F., Fadda, E., 2020a. A combined deep-learning and transfer-learning approach for supporting social influence prediction. Procedia Comput. Sci. 177, 170–177. http://dx.doi.org/10.1016/j.procs.2020.10.025.

Cuzzocrea, A., Leung, C.K., Wodi, B.H., Sourav, S., Fadda, E., 2020b. An effective and efficient technique for supporting privacy-preserving keyword-based search over encrypted data in clouds. Procedia Comput. Sci. 177, 509–515. http://dx.doi.org/10.1016/j.procs.2020.10.070.

Cuzzocrea, A., Song, I.-Y., Davis, K.C., 2011. Analytics over large-scale multidimensional data. ACM 14th DOLAP 11 (101), http://dx.doi.org/10.1145/2064676.2064695.

Datta, S., Bhaduri, K., Giannella, C., Wolff, R., Kargupta, H., 2006. Distributed data mining in peer-to-peer networks. IEEE Internet Comput. 10 (4), 18–26. http://dx.doi.org/10.1109/MIC.2006.74.

Duffy, W., Rigg, J., Maidla, E., 2017. Efficiency improvement in the bakken realized through drilling data processing automation and the recognition and standardization of best safe practices. http://dx.doi.org/10.2118/184724-MS.

Economist, A., 2010. Special report on managing information: data, data everywhere.

El-Abbasy, M.S., Senouci, A., Zayed, T., Mirahadi, F., Parvizsedghy, L., 2014. Artificial neural network models for predicting condition of offshore oil and gas pipelines. Autom. Constr. 45, 50–65. http://dx.doi.org/10.1016/j.autcon.2014.05.003.

Elgendy, N., Elragal, A., 2014. Big data analytics: A literature review paper. pp. 214–227. http://dx.doi.org/10.1007/978-3-319-08976-8_16.

Elmgerbi, A., Thonhauser, G., 2022. Holistic autonomous model for early detection of downhole drilling problems in real-time. Process Saf. Environ. Prot. 164, 418–434. http://dx.doi.org/10.1016/j.psep.2022.06.035.

Fraden, J., 2010. Handbook of Modern Sensors. Springer, New York, http://dx.doi.org/10.1007/978-1-4419-6466-3.

Haghshenas, A., Paknejad, A.S., Rehm, B., Consultant, D., Schubert, J., 2008. The why and basic principles of managed well-bore pressure. In: Managed Pressure Drilling. Elsevier, pp. 1–38. http://dx.doi.org/10.1016/B978-1-933762-24-1.50007-3.

Hansen, K.B., 2020. The virtue of simplicity: On machine learning models in algorithmic trading. Big Data and Soc. 7 (1), 205395172092655. http://dx.doi.org/10.1177/2053951720926558.

Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Ullah Khan, S., 2015. The rise of big data on cloud computing: Review and open research issues. Inf. Syst. 47, 98–115. http://dx.doi.org/10.1016/j.is.2014.07.006.

Hassan, Z., 2018. Common drilling well problems (Reasons, indications, mitigation and prevention). Mitigating Drilling Problems Using Nano-Based Drilling Fluid http://dx.doi.org/10.13140/RG.2.2.19138.48327.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer, New York, http://dx.doi.org/10.1007/978-0-387-84858-7.

He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., Xu, Z., 2011. RCFile: A fast and space-efficient data placement structure in MapReduce-based warehouse systems. In: IEEE 27th International Conference on Data Engineering. pp. 1199–1208. http://dx.doi.org/10.1109/ICDE.2011.5767933.

Hsu, C.S., Robinson, P.R., 2017. Springer Handbook of Petroleum Technology. Springer Sci. Rev, http://dx.doi.org/10.1007/978-3-319-49347-3.

Hu, H., Wen, Y., Chua, T.S., Li, X., 2014. Toward scalable systems for big data analytics: A technology tutorial. IEEE Access 2, 652–687. http://dx.doi.org/10.1109/ACCESS.2014.2332453.

Ishwarappa, Anuradha, J., 2015. A brief introduction on big data 5Vs characteristics and hadoop technology. Procedia Comput. Sci. 48, 319–324. http://dx.doi.org/10.1016/j.procs.2015.04.188.

Jayalath, C., Stephen, J., Eugster, P., 2014. From the cloud to the atmosphere: Running MapReduce across data centers. IEEE Trans. Comp. 63 (1), 74–87. http://dx.doi.org/10.1109/TC.2013.121.

Jiang, F., Leung, C., 2015. A data analytic algorithm for managing querying, and processing uncertain big data in cloud environments. Algorithms 8 (4), 1175–1194. http://dx.doi.org/10.3390/a8041175.

Kadry, H., 2020. Blockchain applications in midstream oil and gas industry. Int. Petrol. Technol. Conf. -Abstract http://dx.doi.org/10.2523/IPTC-19937.

Kale, A., Zhang, D., David, A., Heuermann-Kuehn, L., Fanini, O., 2015. Methodology for optimizing operational performance and life management of drilling systems using real time-data and predictive analytics. In: SPE Digital Energy Conference and Exhibition. http://dx.doi.org/10.2118/173419-MS.

Kamble, S.S., Gunasekaran, A., 2020. Big data-driven supply chain performance measurement system: a review and framework for implementation. Int. J. Prod. Res. 58 (1), 65–86. http://dx.doi.org/10.1080/00207543.2019.1630770.

Kayode, E.s., Lami, O., 2020. Evaluation of differential pressure sticking and stuck pipe in oil and gas drilling technology and its production operations. World Acad. J. Eng. Sci. 7 (2), 114–130.

Kezunovic, M., Pinson, P., Obradovic, Z., Grijalva, S., Hong, T., Bessa, R., 2020. Big data analytics for future electricity grids. Electr. Power Syst. Res. 189, 106788. http://dx.doi.org/10.1016/j.epsr.2020.106788.

Khalili-Garakani, A., Nezhadfard, M., Iravaninia, M., 2022. Enviro-economic investigation of various flare gas recovery and utilization technologies in upstream and downstream of oil and gas industries. J. Clean. Prod. 346, 131218. http://dx.doi.org/10.1016/j.jclepro.2022.131218.

Khvostichenko, D., Makarychev-Mikhailov, S., 2018. Effect of fracturing chemicals on well productivity: Avoiding pitfalls in big data analysis. http://dx.doi.org/10.2118/189551-MS.

Lake, Larry, 2006. Drilling problems and solutions. In: Petroleum Engineering Handbook. SPE J.

Le, Y., 2022. Research on data resource management of biomass energy engineering based on data mining. Energy Rep. 8, 1482–1492. http://dx.doi.org/10.1016/j.egyr.2022.02.048.

Lee, D., Lai, C.W., Liao, K.K., Chang, J.W., 2021. Artificial intelligence assisted false alarm detection and diagnosis system development for reducing maintenance cost of chillers at the data centre. J. Build. Eng. 36, 102110. http://dx.doi.org/10.1016/j.jobe.2020.102110.

Leung, C.K., 2018. Mathematical model for propagation of influence in a social network. In: Encyclopedia of Social Network Analysis and Mining. Springer, New York, pp. 1261–1269. http://dx.doi.org/10.1007/978-1-4939-7131-2_110201.

Leung, C.K.S., Carmichael, C.L., 2010. FpVAT A visual analytic tool for supporting frequent pattern mining. ACM SIGKDD Explor. 11 (2), 39–48. http://dx.doi.org/10.1145/1809400.1809407.

Leung, C.K.S., Jiang, F., 2014. A data science solution for mining interesting patterns from uncertain big data. In: IEEE Fourth International Conference on Big Data and Cloud Computing. pp. 235–242. http://dx.doi.org/10.1109/BDCloud.2014.136.

Leung, C.K.S., Jiang, F., 2015. Big data analytics of social networks for the discovery of following patterns. In: DaWaK. pp. 123–135. http://dx.doi.org/10.1007/978-3-319-22729-0_10.

Li, X., Tan, J., Li, W., Yang, C., Tan, Q., Feng, G., 2022. A high-sensitivity optical fiber temperature sensor with composite materials. Opt. Fiber Technol. 68, 102821. http://dx.doi.org/10.1016/j.yofte.2022.102821.

Liu, Y., Cao, J., Zhang, Q., 2022b. The product marketing model of the economic zone by the sensor big data mining algorithm. Sustain. Comput.- Infor. 36, 100820. http://dx.doi.org/10.1016/j.suscom.2022.100820.

Liu, S., Zhou, X., ji wei, Zhang., Xie, Z., 2022a. The optimization algorithm for application in directional drilling trajectories of energy field. Energy Rep. 8, 1212–1217. http://dx.doi.org/10.1016/j.egyr.2022.01.235.

Loh, W., 2011. Classification and regression trees. WIREs Data Min. Knowl. Discov. 1 (1), 14–23. http://dx.doi.org/10.1002/widm.8.

Meeker, W.Q., Hong, Y., 2014. Reliability meets big data: Opportunities and challenges. Qua. Eng. 26 (1), 102–116. http://dx.doi.org/10.1080/08982112.2014.846119.

Mohammadpoor, M., Torabi, F., 2020. Big data analytics in oil and gas industry: An emerging trend. Pet. J. 6 (4), 321–328. http://dx.doi.org/10.1016/j.petlm.2018.11.001.

Mounir, N., Guo, Y., Panchal, Y., Mohamed, I.M., Abou-Sayed, A., Abou-Sayed, O., 2018. Integrating big data: simulation, predictive analytics, real time monitoring, and data warehousing in a single cloud application. SPE J. http://dx.doi.org/10.4043/28910-MS.

Nair, R., L., D., Shetty, S., 2014. Research in big data and analytics: An overview int. J. Comput. Appl. 108 (14), 19–23. http://dx.doi.org/10.5120/18980-0407.

Nguyen, T., Gosine, R.G., Warrian, P., 2020. A systematic review of big data analytics for oil and gas industry 4.0. IEEE Access 8, 61183–61201. http://dx.doi.org/10.1109/ACCESS.2020.2979678.

Panes, P., Macariola, M.A., Niervo, C., Maghanoy, A.G., Garcia, K.P., Ignacio, J.J., 2022. A bibliometric approach for analyzing the potential role of waste-derived nanoparticles in the upstream oil and gas industry. Clean. Eng. Technol. 8, 100468. http://dx.doi.org/10.1016/j.clet.2022.100468.

Pence, H.E., 2014. What is big data and why is it important?. J. Educ. Technol. Syst. 43 (2), 159–171. http://dx.doi.org/10.2190/ET.43.2.d.

Perrons, R.K., Jensen, J.W., 2015. Data as an asset: What the oil and gas sector can learn from other industries about big data. Energy Policy 81, 117–121. http://dx.doi.org/10.1016/j.enpol.2015.02.020.

Pornaroontham, P., Kim, K., Kulprathipanja, S., Rangsunvigit, P., 2023. Water-soluble organic former selection for methane hydrates by supervised machine learning. Energy Rep. 9, 2935–2946. http://dx.doi.org/10.1016/j.egyr.2023.01.118.

Pritchard, D.M., York, P., Roye, J., 2016. Achieving savings through reliability using real time data. In: Offshore Technology Conference. http://dx.doi.org/10.4043/26935-MS.

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn. 1 (1), 81–106. http://dx.doi.org/10.1007/BF00116251.

Ren, Z., Du, C., 2023. A review of machine learning state-of-charge and state-of-health estimation algorithms for lithium-ion batteries. Energy Rep. 9, 2993–3021. http://dx.doi.org/10.1016/j.egyr.2023.01.108.

Ren, Y., Wang, N., Jiang, J., Zhu, J., Song, G., Chen, X., 2019. The application of downhole vibration factor in drilling tool reliability big data analytics—A review. ASCE-ASME J. Risk Uncertain. Eng. B 5 (1), http://dx.doi.org/10.1115/1.4040407.

Rezaei, H., Ryan, B., Stoianov, I., 2015. Pipe failure analysis and impact of dynamic hydraulic conditions in water supply networks. Procedia Eng. 119 (1), 253–262. http://dx.doi.org/10.1016/j.proeng.2015.08.883.

Ross, J.W., Beath, C., Quaadgras, A., 2013. You may not need big data after all. Harv. Bus. Rev. 91 (12), 90.

Santos, I.M., Hamacher, S., Oliveira, F., 2023. A data-driven optimization model for the workover rig scheduling problem: Case study in an oil company. Comput. Chem. Eng. 170, 108088. http://dx.doi.org/10.1016/j.compchemeng.2022.108088.

Shukla, P., Radadiya, B., Atkotiya, K., 2015. An emerging trend of big data for high volume and varieties of data to search of agricultural data. Peer Rev. 8, 164–169.

Shull, F., 2013. Getting an intuition for big data. IEEE Softw. 30 (4), 3–6. http://dx.doi.org/10.1109/MS.2013.76.

Singh, D., Reddy, C.K., 2015. A survey on platforms for big data analytics. J. Big Data 2 (1), 8. http://dx.doi.org/10.1186/s40537-014-0008-6.

Tiwari, S., Wee, H.M., Daryanto, Y., 2018. Big data analytics in supply chain management between 2010 and 2016: Insights to industries. CAIE 115, 319–330. http://dx.doi.org/10.1016/j.cie.2017.11.017.

Trifu, M.R., Ivan, M.L., 2016. Big data components for business process optimization. Inform. Econ. 20 (1), 72–78. http://dx.doi.org/10.12948/issn14531305/20.1.2016.07.

Tsuchihashi, N., Wada, R., Ozaki, M., Inoue, T., Mopuri, K.R., Bilen, H., Nishiyama, T., Fujita, K., Kusanagi, K., 2021. Early stuck pipe sign detection with depth-domain 3D convolutional neural network using actual drilling data. SPE J 26 (02), 551–562. http://dx.doi.org/10.2118/204462-PA.

Vapnik, V.N., 1995. The nature of statistical learning theory. Springer Sci. Rev. New York http://dx.doi.org/10.1007/978-1-4757-2440-0.

Wang, G., Chao, Y., Cao, Y., Jiang, T., Han, W., Chen, Z., 2022. A comprehensive review of research works based on evolutionary game theory for sustainable energy development. Energy Rep. 8, 114–136. http://dx.doi.org/10.1016/j.egyr.2021.11.231.

Wang, C., Liu, G., Yang, Z., Li, J., Zhang, T., Jiang, H., Cao, C., 2020. Downhole working conditions analysis and drilling complications detection method based on deep learning. J. Nat. Gas Eng. 81, 103485. http://dx.doi.org/10.1016/j.jngse.2020.103485.

Wen, J., Li, Z.J., Wei, L.S., Zhen, H., 2000. The improvements of BP neural network learning algorithm. In: WCC 2000 – ICSP 2000. 2000 5th International Conference on Signal Processing Proceedings. 16th World Computer Congress 2000. pp. 1647–1649. http://dx.doi.org/10.1109/ICOSP.2000.893417.

White, B., Kreuz, T., Simons, S., 2019. Midstream. In: Brun, K., Kurz, R. (Eds.), Compression Machinery for Oil and Gas. Elsevier, pp. 387–400. http://dx.doi.org/10.1016/B978-0-12-814683-5.00009-2.

Wu, Q., Xie, Z., Li, Q., Ren, H., Yang, Y., 2022. Economic optimization method of multi-stakeholder in a multi-microgrid system based on Stackelberg game theory. Energy Rep. 8, 345–351. http://dx.doi.org/10.1016/j.egyr.2021.11.148.

Xaio, C., 2015. Using machine learning for exploratory data analysis and predictive models on large datasets (M.S. Thesis). University of Stavanger, Norway.

Xie, H., Shanmugam, A.K., Issa, R.R.A., 2018. Big data analysis for monitoring of kick formation in complex underwater drilling projects. J. Comput. Civ. Eng. 32 (5), http://dx.doi.org/10.1061/(ASCE)CP.1943-5487.0000773.

Xue, Q., 2020. Prospect of big data application in drilling engineering. pp. 279–312. http://dx.doi.org/10.1007/978-3-030-34035-3_8.

Zayed, T.M., Halpin, D.W., 2005. Deterministic models for assessing productivity and cost of bored piles. Constr. Manag. Econ. 23 (5), 531–543. http://dx.doi.org/10.1080/01446190500039911.

Zhou, Y., Kantarcioglu, M., Xi, B., 2019. A survey of game theoretic approach for adversarial machine learning. WIREs Data Min Know. 9 (3), http://dx.doi.org/10.1002/widm.1259.

Zhu, L., Wei, J., Wu, S., Zhou, X., Sun, J., 2022. Application of unlabelled big data and deep semi-supervised learning to significantly improve the logging interpretation accuracy for deep-sea gas hydrate-bearing sediment reservoirs. Energy Rep. 8, 2947–2963. http://dx.doi.org/10.1016/j.egyr.2022.01.139.