Assessment of Machine Learning Models for Reservoir Characterization and Drilling Automation

Thesis submitted in partial fulfillment

for the Award of Degree of

Doctor of Philosophy

By

Saurabh Tewari

RAJIV GANDHI INSTITUTE OF PETROLEUM TECHNOLOGY

JAIS, 229304

PPE15001 ***2021***

CERTIFICATE

It is certified that the work contained in the thesis titled *Assessment of Machine Learning Models For Reservoir Characterization And Drilling Automation* by *Saurabh Tewari* has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all the requirements of Comprehensive, Candidacy, and Open Seminar.

Dr. U. D. Dwivedi

(Thesis Supervisor)

DECLARATION BY THE CANDIDATE

I, *Saurabh Tewari*, certify that the work embodied in this thesis is my own bonafide work and carried out by me under the supervision of *Dr. U. D. Dwivedi* from *July 2015* to *July 2020,* at the *Department of Petroleum Engineering & Geoengineering,* Rajiv Gandhi Institute of Petroleum Technology, Jais. The matter embodied in this thesis has not been submitted for the award of any other degree. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, etc., reported in journals, books, magazines, reports dissertations, theses, etc., or available at websites and have not included them in this thesis and have not cited as my own work.

Date:

Place: Jais, Amethi **Saurabh Tewari**

CERTIFICATE BY THE SUPERVISOR

It is certified that the above statement made by the student is correct to the best of my knowledge.

Dr. U. D. Dwivedi

(Thesis Supervisor)

Head of the Department

CERTIFICATE

CERTIFIED that the work contained in the thesis titled *Assessment of Machine Learning Models for Reservoir Characterization and Drilling Automation* by *Mr. Saurabh Tewari* has been carried out under my supervision. It is also certified that he fulfilled the mandatory requirement of TWO quality publications that arose out of his thesis work.

It is further certified that the two publications (copies enclosed) of the aforesaid *Mr. Saurabh Tewari* have been published in the Journals indexed by –

(a) SCI

(b) SCI Extended

(c) SCOPUS

Dr. U. D. Dwivedi **Dr. Shivanjali Sharma**

(Thesis Supervisor) (Convener, DPGC)

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: *Assessment of Machine Learning Models for Reservoir Characterization and Drilling Automation*

Name of the Student: *Saurabh Tewari*

Copyright Transfer

The undersigned hereby assigns to the Rajiv Gandhi Institute of Petroleum Technology Jais all rights under copyright that may exist in and for the above thesis submitted for the award of the “Doctor of Philosophy".

Date:
Place: Jais, Amethi **Saurabh Tewari**

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for the author's personal use provided that the source and the Institute's copyright notice are indicated.

ACKNOWLEDGEMENT

I would like to express my gratitude, respect, and appreciation to all those who have been associated with the successful completion of my Doctoral Program. I would like to express my gratitude and respect to my doctoral advisor Dr. U. D. Dwivedi, for his excellent advice, encouragement, esteemed guidance, and complete freedom to test the innovative ideas throughout my research work. Mostly, his honesty, discipline, perfection, and excellent technical writing skills were a constant source of inspiration for me.

I am thankful to Dr. Susham Biswas for his constant motivation to achieve my research goals. I would like to extend my thanks to all the faculty members of the Department of Petroleum Engineering and Geoengineering for their help and support.

Finally, yet importantly, I convey my heartfelt gratitude to my parents, family, and friends for their blessings, unconditional love, sacrifice, and continuous support is given to me to pursue my dreams.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS/NOTATIONS

DT	Measured Depth	MT	Milled Tooth
TVD	True Vertical Depth	MSE	Mean Square Error
ROP	Rate of Penetration	MAE	Mean Absolute Error
WOB	Weight on bit	RMSE	Root Mean Square Error
RPM	Rounds per minutes	LM	Levenberg-Marquardt
TQ	Torque	BR	Bayesian Regularization
SPP	Standpipe pressure	SCG	Scaled Conjugate Gradient
MW	Mud weight	IADC	International Association of Drilling Contractors
FR	Flow Rate	UNDP	United Nation Development Programme
TG	Total Gas	CCD	Centre Composite Design
IN	Inclination	BBD	Box Behnken Design
AZ	Azimuth	CCRD	Central Composite Rotatable Design
BT	Bit type	CL	Clay
BS	Bit Size	CM	Carbonate mudstone
RSM	Response surface methodology	DM	Dolomitic mudstone
PDC	Polycrystalline Diamond Cutter	DP	Dolomitic packstone
SS	Siltstone	DS	Dolomitic sandstone
MLP	Multilinear perceptron	DW	Dolomitic wackestone
SVM	Support vector machine	PS	Packstone
GB	Gradient boosting	RF	Random forest
HEM	Heterogeneous ensemble method	HoEM	Homogeneous ensemble methods
SG	Savitzky–Golay	10-FCV	10-Fold cross-validation
TT1	Acoustic transit time 2 log	SP	Spontaneous potential log
ADT	Acoustic transit time 1 log	RLL3	Deep laterolog resistivity log
MCAL	Caliper 2 log	RILM	Medium induction log
DCAL	Caliper 1 log	RILD	Deep induction log
NPOR	Neutron log	GR	Gamma log
DPOR	Density log	SPOR	Sonic log
GA	Genetic algorithm	HEM	Homogeneous ensemble methods

PREFACE

In the oil and gas industry, a huge amount of data is generated through sensory measurements during exploration to production phases of the reservoir. Uncertainties and inexactness are present in all the reservoir measurements due to heterogeneity and stochastic distribution of reservoir characteristics. Conventionally, field data are interpreted by experienced experts to extract useful information. However, with the advent of measurements-while-drilling and smart-well technologies, there is a significant increase in the volume of data generated and to be analyzed. Therefore, processing and analysis of this huge data pose a significant challenge to the prevailing technologies used in the oil and gas industry. In this study, the intelligent modeling approach has been investigated to provide cost-effective solutions for three major problems of petroleum domain viz. (a) Lithofacies identification (b) Drilling optimization (c) Production rate estimation.

Initially, ensemble-based big data analytics have been proposed for quantitative lithofacies modeling of the unconventional mudstone reservoir. The performance testing of five standard ensemble methods (viz. Bagging, AdaBoost, Rotation forest, Random subspace, and DECORATE) has been done to identify the prevailing subsurface lithofacies. The results have been generated using single well data existing in the Kansas region of the U.S.A. Random subspace–SVM combination has given the highest classification accuracy as compared to all base combinations experimented. Further, the application of heterogeneous ensemble models (HEMs) has been explored to generate more generalized results for lithofacies modeling using multiple-well data. Additionally, stability analysis of HEMs has also been performed to ensure the reliability of the proposed methodology. The performance of HEMs depends upon the selection of efficient base

classifiers for the quantitative lithofacies modeling. A validation curve has been found as an efficient measure for identifying the stable search range for machine learning parameters. The stacking ensemble has shown great potential to extract lithofacies information from well logs data. The training and testing classification accuracies of HEMs are the highest among the other classifiers used in this study.

A novel data-driven drill bit selection technique has been developed for oil and gas field applications. To overcome the problems associated with existing methods, Response surface methodology (RSM) and Artificial Bee Colony (ABC) have been used to develop an intelligent data-driven approach for the selection of suitable bit types. RSM has been implemented to generate the objective function for ROP. Moreover, the developed function is optimized using ABC to achieve optimum values of ROP and drilling parameters including bit types. The results are also compared with the existing artificial neural network (ANNs) model for drill bit selection. The combination of RSM and ABC provides a more reliable bit selection modeling approach as compared to ANN-based on cost-per-foot estimations. The ROP objective function developed through RSM is less complex than the ANN-based objective function due to the absence of an exponential function. ANN requires more computational cost for the development of the ROP function and its optimization. This study provides an alternate intelligent approach to bit selection based on optimum values of ROP. However, these models are case-specific as well as data-dependent models and require calibration for other field data.

Advancements in intelligent predictive models have also enabled the automated selection of drill bit types using previously drilled offset wells data. Data-driven machine learning algorithms can be utilized for suitable drill bit selection. However, real-field data

typically involves an unequal distribution of data samples resulting in a complex imbalance multi-class classification problem for drill bit selection. To overcome this problem, architectural adaptations along with the data re-sampling technique, have been incorporated in the existing ensemble methods for handling the imbalanced data problem in drill bit selection. Random forest with bootstrap class weighting has provided the best overall accuracy for the selection of drill bit types using previously drilled Norwegian offset wells data. It is observed that drill bit selection becomes difficult in the lower formations due to uncertainty in subsurface conditions. Data imbalance condition exists due to the drilling of thin lithofacies that harm the performance of classifiers. Conventional classifiers can't be trusted for the drill bit selection, especially for critical drilling zones.

In another research work, data-driven models are applied for correlating the two-phase flow behavior of oil and gas with setting variables of surface installed chokes. Surface chokes are widely installed equipment on wellheads to control the hydrocarbon flow rate and to maintain the bottom-hole pressure. The design of these chokes and production strategies can be improved with the estimation of future production rates of hydrocarbon. An extensive literature review has been performed to address the issues related to existing intelligent models for daily hydrocarbon production through surface chokes. 'Random forest' and 'Extremely randomized trees' paradigms have been proposed to predict the oil and gas production rates through surface installed chokes. A comparative study has been performed to investigate the efficacy of the proposed and other popular machine learning models (viz. ANNs, SVR, etc.) for the prediction of the two-phase flow rate of oil and gas. Random forest and ExtraTree ensembles have outperformed the popular estimation models, viz. ANNs, SVR and LSSVR, for production forecasting.

Chapter 1

Introduction

1.1 Introduction

The oil and gas industry plays a vital role in meeting the ever-growing energy demand of the human race needed for its sustainable existence. Newer unconventional wells are drilled for the extraction of hydrocarbons that requires advanced innovations to encounter the challenges associated with various exploration and production operations. To remain competitive in the globalized energy market, the oil and gas industry requires newer innovative technologies that can facilitate uninterrupted, cost-effective, high-quality sustainable productions. With the ever-increasing use of advanced instrumentation and control in the oil and gas industry, a huge amount of data is generated through installed sensors in various real field operations, from exploration to production of hydrocarbon. Due to heterogeneity and stochastic distribution of reservoir properties, uncertainty and inexactness are present in all the measurements of reservoir properties [1]. Also, petroleum data suffer from several complexities such as nonlinearity, high dimensionality, noise, and imbalanced data conditions. Petroleum data processing is required in several data-related tasks such as modeling and optimization, reservoir simulation, classification, clustering, forecasting, and monitoring of various petroleum events and operations [1]. These data demand advanced computational tools to be employed for their processing and analysis. Therefore, data-driven machine learning models have been designed for handling complex problems related to classification and estimation tasks [1, 2]. These models have great potential in processing petroleum domain data.

In the petroleum industry, the sensor-based measurements are acquired in various forms such as well logs, seismic logs, mud logs, reservoir data, well test data, production data, etc. [1,2] These data are further analyzed and interpreted by geologists, geophysicists, and petroleum engineers to extract useful information in decision making for various field operations. Traditionally, these data are interpreted by manual analysis by a domain expert. However, it requires great human efforts, time, and cost involving high chances of errors [3]. Conventional interpretation methods provide only limited operational knowledge and often miss recognizing several useful information hidden in the data. The sensory data contain serious problems of noise, nonlinearity, high dimensionality, etc. which decrease the accuracy of the conventional interpretation methods sometimes below 50 % in the oil and gas industry [3].

The areas of big data analytics and machine learning have become very promising research fields for a wide range of applications due to the development of high computational systems [4,5]. These methods can process huge amounts of data, extract useful information from raw data, and can easily identify the hidden patterns in the given data. These advanced techniques can easily filter out noise, reduce dimensionality, model nonlinear relationships, and sometimes helpful in handling reservoir uncertainties [3,4]. The machine learning modeling approach has several advantages such as economical solution provider, quick mitigation of real-field problems, real-time deployment, facilitate automation to real-field operations, found to be more robust and reliable. However, data dependency, data availability underdeveloped systems, and multidisciplinary expertise are the important issues associated with machine learning models [3,4].

Machine learning models can provide practical solutions to complex petroleum problems. Therefore, hybrid computational models, such as ensemble methods, a

committee of machines, etc., were suggested for processing complex petroleum data [5]. These techniques are of significant importance, especially, when a high classification or estimation accuracy is targeted, as they can increase the generalization capabilities of an ML model by enhancing its modeling strategy [5]. Such hybridized models enable physically meaningful computation for time-demanding applications. In this research work, different supervised models, hybridization strategies, structural designs of hybrid models, learner screening criteria, and hybrid computational modeling approaches are tested and validated using real field data with a primary focus on their applications in petroleum systems and operations. Primarily, the research work carried out in this thesis focuses on the applications of machine learning models for lithofacies identification, suitable drill bit selection, optimization of drilling rate of penetration, and estimation of hydrocarbon production rate using diverse petroleum data.

The first application reported in this thesis work is on the identification of lithofacies for unconventional mudstone reservoirs. These geological formations are producing a huge amount of hydrocarbon, especially natural gas. The mudstone lithofacies are difficult to identify due to their overlapping properties [6]. Well-logs are mainly utilized for the identification of the subsurface lithofacies along with the depth through human experience [7]. The manual interpretation of well-logs is a time-consuming, and costly task [7]. It also requires human expertise for its interpretation [8]. However, the utilization of advance computational machine learning techniques automatize the analysis of well-logs and integrate results from other sources such as core analysis and seismic analysis to provide more authentic results [8,9]. This study has investigated popular supervised classifier models along with the development of several higher hybrid computational models for reducing uncertainties associated with reservoir modeling and enhancement of reservoir modeling accuracy.

The second area of research work carried out in this thesis is on the utilization of modified ensemble methods for handling the imbalanced data problem in drill bit selection. With the advancement of intelligent predictive models, the automated selection of drill bit type is possible using previously drilled offset wells' data. Data-driven machine learning algorithms can be utilized for suitable drill bit selection. However, real-field data typically involves an unequal distribution of data samples resulting in a complex imbalance multi-class classification problem for drill bit selection. In this analysis, two methods, namely AdaBoost and Random forest, have been combined with the data resampling technique for the complex drill bit selection procedure. The four other popular machine learning techniques namely, K-nearest neighbor, Navies Bayes, Multilayer perceptron, and Support vector machine, were also evaluated individually to understand the degrading effects of imbalanced data. These models are trained and tested on the drilling data obtained from Norwegian Volve oil and gas wells. Proper pre-processing of input drilling data has been done before the training of machine learning models. Diverse data-driven experimental scenarios have been simulated to analyze the performance of data-driven models for drill bit selection.

Several researchers have also suggested a newer approach for drill bit selection based on optimum values of drilling rate of penetration (ROP). A separate study has been conducted for finding suitable drill bit types for drilling target formation based on the optimum ROP values. Instead of deploying the conventional method of offset well logs, Response surface methodology (RSM) and Artificial bee colony (ABC) have been combined to develop an intelligent data-driven modeling approach for the selection of optimum bit type for drilling operations. This approach also makes the optimal use of operational control parameters and is found more efficient than conventional methods. RSM has been implemented to generate the objective function for ROP. The developed

ROP function is further optimized through ABC to obtain the optimum values of drilling variables and drill bit types for target geological formations. The results were also compared with the existing Artificial neural network (ANN) and Genetic algorithm (GA) model for suitable drill bit selection.

The third area of research work carried out in this thesis is on hydrocarbon production through surface installed chokes. Surface chokes are widely installed equipment in the oil and gas industry. The multi-phase flow behavior of hydrocarbon near the surface orifical is always a matter of concern because it influences the overall hydrocarbon production rate. Several theories have been proposed to explain these phenomena, however, none of them are fully able to justify the flow and its flow regime behavior. Therefore, data-driven models have been investigated for the estimation of the hydrocarbon flow rate. Random forest and ExtraTree ensemble have been applied to correlate the surface measured production variables with oil and gas flow rates. It has been found that a data-driven model performs much better than analytical or any other existing empirical models.

1.2 Literature review

1.2.1 Intelligent lithofacies recognition

Reservoir characterization is defined as the act of developing a reservoir model that has a resemblance with a real-world reservoir with similar properties and behaviors for storing and producing hydrocarbons. These models are utilized to simulate the flow behavior of fluids in various conditions to optimize the production strategies of the actual reservoir. To cultivate such a reservoir model, accurate information of reservoir properties such as porosity, permeability, reservoir pressure, temperature, recognition of lithofacies, etc. is essentially required [1].

Quantitative lithofacies modeling is one of the most challenging parts of reservoir characterization that involves the identification of subsurface lithofacies [1]. These subsurface layers contain information about the depositional environment along with the sedimentary process of rock formation [2]. Several types of qualitative analysis of rock formations are performed to understand their geometries, grain size, sedimentary structure, etc. [2]. These analyses usually involve well logs, core analysis, advanced geochemical, Rock-Eval pyrolysis, etc. [3] However, it is time taking and expensive affair. It also requires human proficiency and huge efforts for its analysis and accurate interpretation. The quantitative approach has been preferred over qualitative analysis to achieve fast, accurate, and more economical modeling of the reservoir. The information about the subsurface lithofacies is extracted from the conventional well logs that are recorded throughout the depth of the reservoir formation. Well-logs can record the physical properties of rock that change logs response with depth [4]. Conventionally, experts manually examine well logs to identify different layers of lithofacies through their experience. However, there is always a high chance of human error with complex well-logs. The manual analysis is never considered as a good practice for the interpretation of well-logs.

Well-logs are sensor-based measurements of lithofacies such as spontaneous potential logs, resistivity logs, neutron density logs, spontaneous potential logs, etc. [5]. Several logs are used for the accurate identification of lithofacies along with the depth of the formation. However, these sensory recorded data are complex in nature and contain issues namely, nonlinearity, high dimensionality, imbalance, and noise along with uncertainties in measurement due to heterogeneous behavior of the reservoir [1]. The excavation of valuable information from the raw sensory logging data through manual technique becomes a difficult task with high risk. Hence, intelligent

mechanization is necessary to examine and extract useful information from the huge amount of logging data for quantitative lithofacies modeling.

Several intelligent machine learning models were applied for the automatic identification of lithofacies through computational processing of well logs data. Some of the important applied intelligent models are as follows as (a) Clustering technique [6]. (b) Principle component analysis (PCA) with Artificial neural networks (ANNs) [7, 8]. (c) Support vector machine (SVM) [4]. (d) Self-organizing map (SOM) and Multi-resolution graph-based clustering (MRGC) [3] and (e) Random forest (RF) [9]. It can be seen that the application of conventional classifiers is mainly reported for lithofacies modeling. The utilization of a hybrid computational approach has been rarely done to extract facies information.

The hybrid computational approach is a popular data mining methodology that helps to extract useful information for the highly complex data structure. Several hybrid computational approaches have been in applied diverse engineering fields such as the committee of machines, ensemble methods, etc. These are multiple classifier systems that combine several supervised classifiers to enhance the efficacy of conventional classifiers. It has been also mathematically proven that the performance of supervised classifiers can be improved through multiple classifier systems [10,11]. Ensemble methods are found to be capable of handling nonlinear, high dimensional, noise, and imbalanced data.

1.2.2 A comparative study of heterogeneous ensemble methods for the identification of geological lithofacies

Mudstone lithology is a kind of sedimentary rock that contains most of the unconventional hydrocarbon reservoirs. It can also play multiple roles from hydrocarbon generation to storage [12]. Hidden sweet spots are present inside mudstone

lithofacies that support hydrocarbon production. These unconventional reservoir systems are quite difficult even for conventional interpretation techniques for lithofacies identification. Understanding of petrophysical properties of rocks and their spatial distribution in association with lithofacies are essential for the development of a reservoir model to produce hydrocarbon [13]. Several conventional techniques are performed for the recognition of lithofacies such as rock-eval pyrolysis, geomechanical spectroscopy logs, etc. [3]. However, these are found to be costly, time-taking, and demands expertise. It has been found that ensemble methods are more efficient than single supervised classifiers or learners. However, there exists another category of ensemble methods that show greater potential than the previously defined ensemble classifier for lithofacies recognition. These can dig deep for more information from raw well-log data. It also integrates the outcomes of various classifiers for pattern recognition tasks. Ensemble methods can be characterized into two kinds: (a) homogeneous ensemble methods (HoEMs) such as Bagging, Rotational forest, Random subspace, etc., and (b) heterogeneous ensemble methods (HEMs) such as Stacking, Voting, StackingC, etc. HoEMs generate several hypotheses in the feature space using similar classifiers (e.g., a group of five hundred ANNs) and combine them to achieve maximum accurate classification results. However, HEMs combine various dissimilar classifiers (e.g., A cluster containing n SVMs, n ANNs, n Naïve Bayes, etc.) in the feature space. It has been proved mathematically that these models are more efficient than HoEM due to the heterogeneity in the base classifiers and provide more generalized results with reduced prediction errors [14]. However, these methods are not much investigated for their possible application in the petroleum domain.

1.2.3 Intelligent drill bit selection

Rocks formations are crushed into chips through drill bits that are installed at the bottom of the drill string. Several methods have been applied for finding the optimum bit type mainly based on measured data taken from offset well logs with limited applicability. Out of these methods, the most popular method in use is cost per foot (C_{PF}) estimation for drilled intervals [15]. The method is popular as it is based on the operating cost of the drilling operation. C_{PF} can be measured using a formula.

$$C_{PF} = \frac{Bit_C + Rig_C(Bit_{RT} + CN_T + Trip_T)}{DF} \tag{1.1}$$

where Bit_C is the bit cost in dollars, Rig_C is the cost of rig per hour, Bit_{RT} is the bit running time in hours, CN_T is connection time in hours, $Trip_T$ is the trip time in hours, and DF is the sectional length of wellbore drilled in feet. C_{PF} is used in combination with other methods as it doesn't depend on the operational parameters but the drilling economy is highly affected by them. C_{PF} also has one more disadvantage that it can't be used in the case of directional and horizontal wells. C_{PF} has been proven efficient in the analysis of historic drilling data obtained from the offset wells and current supervision of bit run [15,16]. Back in the 1960s, Teale formulated a notion of Specific Energy (S_E) by establishing a relationship between bit energy requirement and its performance [15,16]. In rock drilling, SE is the energy spent by the machine to eradicate unit rock volume [17]. The S_E formula is given below.

$$S_E = \frac{RPM * WOB}{ROP * BD} \tag{1.2}$$

where RPM is the rounds per minute in rpm, WOB is the weight on the bit in pounds, ROP is the drilling rate of penetration in feet per hour, and BD is the bit diameter in inch. As it is clear from the formula, SE is governed by only three parameters mainly,

ROP, WOB (weight on bit), and RPM, which proves to be less advantageous. It also can't distinguish between various formations based on mechanical properties and vibrations effects on the dulling of a bit [17].

The international association of drilling contractors (IADC) has adopted a new dull grading system for selecting bits on their degree of dullness for roller cone as well as fixed cutter bits. In this regard, IADC used to report various parameters of drill bit such as teeth wear, bearing conditions, etc. on a scale from 1 to 8, where 1 is excellent and 8 being the poor condition. The dullness of the drill bit is a crucial factor as if the bit wears fast, it adds to drilling cost and more time consumption which indirectly affects the economics of drilling operation. So, one needs to carefully select the bit by evaluating data as it will lead to extra cost to projects [18].

In 1964, Hightower selected drill bits from geological information and logs from offset wells [19]. Sonic logs were used to find the right bit for drilling operations by estimating the formation strength to define the drillability of the formation. The rock strength utilized in this method was not measured directly from sonic logs but estimated indirectly from the theory of elasticity. Bourgoyne and Young [20] utilized a multiple regression approach for ROP modeling and considered drill bit types as an important factor influencing the drilling operation. Rabia et al. [21] proposed the selection of a bit based on mechanical specific energy. Fear et al. [22] selected drill bits based on the geology and rock properties of the formation. Perrin et al. [16] proposed a novel drilling index for the evaluation and selection of drill bit types for drilling operations. Uboldi et al. [23] utilized rock strength measurements as criteria for the choice of drill bits. Bahari and Seyed [15] applied mathematical correlations as objective functions for the optimization of various drilling variables and operational costs.

Recently, data-driven intelligent models have been utilized to find suitable types of drill bits. These models are reported to be more accurate as they learn from previous well data, defying traditional methods for selecting the appropriate drill bit [24]. Bilgesu et al. [25] used ANN for the prediction of drill bit types for drilling target geological formations. Yilmaz et al. [26] trained the ANN model using previously drilled wells offset data and predicted the drill bits types for the development wells required to be drilled internal and external of the same field. They also tested the trained ANN model for the prediction of drill bit types for the development wells that were required to be drilled in an adjoining field. Bahari et al. [15] utilized a Genetic algorithm (GA) for the accurate computation of constants for the Bourgoyne-Young ROP model. Edalatkhah et al. [27] also selected the suitable drill bit types using ANN and GA for South Pars Field wells. Momeni et al. [28] applied ANN for the estimation of drilling ROP and bit types [14]. Momeni et al. [29] combined ANN and GA for drill bit selection based on optimal ROP. They selected the drill bit types based on the optimum values of ROP and drilling variables. Abbas et al. [18] also supported the notion of drill bit selection depending upon optimum values of ROP using ANN and GA. Here ANN was primarily utilized for the development of the objective function and GA for optimization of the ROP objective function for the drill bit selection.

Several researchers have suggested that the selection of drill bits should be performed based on the optimum values of ROP. This condition results in the development of an unconstrained bounded optimization problem, where a function of ROP is required to be defined using drilling variables. However, the exact relationship between ROP and drilling variables is unknown and undefined which makes optimization of ROP a difficult task. According to Kolmogorov's theorem, multilayer feedforward perceptron (MLP) ANN architecture can be utilized to define any

continuous function in its approximation form [30]. The approximation function (objective function) requires an activation function and input variables that are predefined during the training of the MLP neural network. Three-layered MLP architecture can be expanded in a mathematical form with connection weights and bias of neurons that will act as coefficients of approximation function. This technique helps to solve real-field complex optimization problems, especially where the association between input and target variables is unknown such as bit selection based on optimum ROP values. In the case of complex approximation function, paradigms such as ant colony, swarm optimization, GA, etc. can be implemented to retort the optimization problem as stated in the literature [31]. However, researchers reported several issues with ANN such as overfitting, underfitting, stuck up in local minima/maxima, lack of proper guidelines for the selected network architecture, [32]. This also opens the opportunity to investigate other techniques that can generate approximation functions to optimize ROP values for drill bit selection. More research work is required to automate the drill bit selection procedure for better reliability and efficacy.

1.2.4 Intelligent forecasting of hydrocarbon production

Hydrocarbon production forecasting is an important task for reservoir engineers to measure the performance of installed production systems. It also plays a vital role in the estimation of remaining hydrocarbon inside the producing reservoir formations, optimization of production operations, reservoir management, and business planning [35]. Continuous recording and monitoring of daily hydrocarbon production data are usually done to forecast future well production. However, it is a challenging task due to the reservoir's heterogeneity and complex interactions of the reservoir with hydrocarbon production systems [36]. The production of multiphase fluid through surface choke is also influenced by the behavior of the producing reservoir formation in static and

dynamic conditions [36]. Accurate assessment of reservoir properties, itself, is a problematic issue and its heterogeneity adds uncertainties in all types of reservoir measurements [37,38]. In the petroleum domain, production forecasting has always been considered a thought-provoking and popular problematic task due to the complexity of acquired production data [39].

Several empirical models and correlations have been proposed to predict the multi-phase flow through surface installed chokes. Wellhead chokes are widely installed to control the oil and gas flow rates on the surface, to maintain downhole pressure, and also to produce back pressure that protects the reservoir from formation damage [40-43]. To regulate the flow rate for meeting various regulations, chokes are installed to minimize various problems owing to varying production rates which may be slugging of surface equipment, avoid excess sand production caused due to high drawdown, and water/gas coning [44,45]. A significant part of the production optimization relies heavily on the study of the flow behavior through the chokes i.e., whether the flow is subsonic or sonic [46]. The critical pressure ratio is calculated to distinguish between sub-sonic and sonic flow conditions. It is approximately 0.55 for natural oil and gas above which subsonic condition prevails. When the fluid velocity in a choke matches the traveling velocity of sound in the fluid under in-situ conditions then such type of flow is termed sonic flow [47]. Under sonic flow conditions, the pressure wave downstream of the choke cannot go upstream through the choke because the medium is traveling in the reverse direction at a similar velocity [42].

Several correlations have been developed theoretically or empirically using experimental or field production data to study the simultaneous oil and gas flow behavior in sonic and sub-sonic conditions through chokes. Tangren et al. [48] contributed the first study on wellhead chokes and their effects on the production rate of

hydrocarbons for the continuous liquid phase. Gilbert [49] correlated oil production rate with wellhead surface choke size, gas oil ration, and wellhead pressure. Ros [50] reported that a correlation existed between upstream pressure, restriction size of choke, and flow rate of hydrocarbon. Several other researchers also proposed similar correlations for the oil production rate using diverse field data [51-55]. Al-Attar and Abdul Majeed [56] tested several proposed correlations to provide the best fitting for East Baghdad Oil field production data. They found that the revised correlation was similar to the Gilbert equation with different constants values. Mirzaei-Paiaman and Salavati [42] proposed a newer correlation for the flow of oil through wellhead chokes using Persian oil field data. All the correlations have been developed either theoretically or empirically for wellhead chokes and multiphase hydrocarbon production based on experimental data or field data [42][49-54]. Theoretical correlations developed using field data require a large number of parameters collection from fields which is a time-consuming and costly affair. On the other side, experimentally developed empirical correlations lack generalizability due to the limited range of experimental data. Therefore, advanced machine learning techniques have been employed to model the production rate of hydrocarbon with the surface installed chokes.

Researchers have widely utilized machine learning techniques to correlate variables that have complexities in their relationships. Machine learning techniques have achieved more reliable and generalized prediction models for several engineering domains such as reservoir characterization, drilling automation, etc. Morzaei-paiaman and Salavati, [42] applied the ANN model for the estimation of the oil production flow rate. He also compared the prediction results of ANNs with the correlations proposed by Mirzaei-Paiaman [43], Gilbert [49], Ros [50], Achong [51], and Baxendell [52] to prove that ANNs results are more accurate than empirical and theoretical correlations. Elhaj et

al. [53] studied ANN along with Fuzzy logic, Functional networks, Decision tree, and Support vector machines for single gas flow rate forecasting. Choubineh et al. [35] applied hybrid ANN training-based optimization for modeling the hydrocarbon flow rate. Figure 1.1 shows data diversity existing in Petroleum domain.

Figure 1.1 Data diversity existing in Petroleum domain [57].

ANNs are one of the most popular intelligent modeling techniques for correlating complex variables together. However, ANN has certain limitations such as stuck up in local minima/maxima, over-fitting/ under-fitting, etc. [54]. These shortcomings challenge the reliability and generalizability of the results provided by ANN. Other machine learning models are not much studied for modeling hydrocarbon production rate from surface installed chokes. Moreover, some intelligent models can handle the limitations and shortcomings of ANN such as Support vector regression (SVR), Least-square support vector regression (LSSVR), RF, Regression tree, etc. Nejatian et al. [55] estimated the choke flow coefficient for sub-sonic flow conditions of natural gas using LSSVR. Gorjaei et al. [56] utilized PSO-least square support vector regression (PSO-

LSSVR) to predict the two-phase flow of oil and gas through the surface installed choke. These studies have shown that other machine learning techniques can also be adopted for modeling hydrocarbon production rates through surface installed chokes.

1.3 Motivation

Technical advancements in data acquisition technologies have changed the conventional workflow of the oil and gas industry [3]. Advance data acquisition technologies such as mud pulse telemetry, logging systems, seismic tools, etc. provide more accurate information and lead to the automation of various field operations [1,3]. However, these technological advancements have also resulted in newer challenges in data processing and analysis for the extraction of useful information from raw data. Further, the captured petroleum field data are naturally complex with high uncertainties in their measurements due to the heterogeneity of hydrocarbon reservoirs [1]. Thus, oil and gas field data require advanced computational techniques for their processing and analyses.

Petroleum researchers have suggested intelligent machine learning models for handling the problems related to the petroleum domain. Supervised and unsupervised machine learning models are widely applied for solving field-related issues such as drill pipe stuck up [57], lithofacies identification [3], reservoir properties estimation [3], drilling parameter estimation [15], drill bit selection [18], hydrocarbon production forecasting [38-56], etc. These intelligent paradigms have given impressive accurate outcomes for various petroleum applications. However, these models have their limitations and shortcomings as explained in the literature review section of this chapter. These models are highly influenced by bias and variance associated with the training data which increases the chances of large errors in the prediction outcomes and compromised generalization performance. To overcome the abovementioned data-

related issues, ensemble methods are proposed in this study. Ensemble methods are less explored intelligent paradigms that have the potential to solve the problems of the oil and gas industry. The ensemble approach is particularly investigated in this thesis to handle issues related to lithofacies identification, drill bit selection, and production forecasting.

This thesis approaches oil and gas problems at two levels i.e. data level as well as algorithm level to enhance the efficacy of existing machine learning models. Initially, acquired field data have been studied to eliminate data-related issues such as abnormal data samples, noise, high dimensionality, imbalance, etc. A proper preprocessing stage has been developed to solve any data-related issue before training and testing an intelligent paradigm at the data level. Secondly, machine learning algorithms have been investigated at the algorithmic level, compared, and modified to achieve the best possible outcomes. The primary research goals of this thesis are to investigate the applicability, reliability, and effectiveness of potential machine learning models for providing data-based solutions for various issues in the upstream industry using diverse petroleum data. The focus of this study is to perform a comparative and comprehensive investigation, on various conventional and advanced intelligent models for their suitability and efficacy, in solving important issues in the petroleum industry. Also, investigations have been carried out to provide proper guidelines for the development of the computational framework and preprocessing of the raw field data before training and testing of machine learning models.

1.4 Research questions

- How HoEMs can be applied for the modeling of lithofacies of mudstone reservoir?

- Which is the most suitable HoEMs and base classifier combination for the identification of geological mudstone lithofacies?
- How do better generalization can be achieved on multiple well data using HEM approach?
- How does an alternative approach can be developed to challenge existing ANN model for drill bit selection based on optimum values of drilling ROP?
- What can be done to handle the imbalanced data problem during drill bit selection?
- What are the performance efficiency of earlier applied machine learning models for forecasting oil and gas production through a surface installed choke ?
- Is there proper guideline and workflow available in the petroleum domain for machine learning applications?

1.5 Research gap

The petroleum data have several issues such as high dimensionality, noise, imbalance, nonlinearity, diversity in data format, large storage space requirements, computational expense, etc. that are needed to be addressed adequately to make the results more authentic. There are no proper guidelines, for the implementation of machine learning models to solve the problems of the oil and gas industry, in the existing literature. Most of the existing literature lacks a comprehensive and comparative study for machine learning models for similar data and problems. Hybrid computational models are not much explored in the petroleum domain that may provide more generalized results. The overfitting and underfitting conditions also hamper the reliability of prevailing intelligent models for real field applications. There is also lack of understanding in oil and gas problems at two levels i.e. data level as well as algorithm level for existing

machine learning models. Acquired field data contain a lot of data-related issues such as abnormal data samples, noise, high dimensionality, imbalance, etc. A proper preprocessing stage has to be developed for solving any data-related issue before training and testing an intelligent paradigm at the data level.

1.6 Research objectives

Considering the limitations of the existing work as discussed above sections, the main objectives behind the research work carried out in this thesis are as following:

- To propose homogeneous ensemble methods for automatic identification and recognition of geological lithofacies for unconventional mudstone reservoirs.
- To investigate and apply heterogeneous ensemble methods for geological lithofacies modeling.
- To propose Response surface analysis and Artificial bee colony for drill bit selection based on optimum values of drilling penetration rate.
- To apply the combination of ensemble methods and data resampling techniques for handling the imbalanced data problem during drill bit selection.
- To compare the performance of existing machine learning models for forecasting oil and gas production through a surface installed choke.
- To propose proper working frameworks and guidelines for the application of various data-driven models in the petroleum domain to handle data-related issues.

1.7 Hypothesis of work

- Modeling of lithofacies of mudstone reservoir is possible through HoEM.

- Diverse performance behaviors can be measured for ensemble-base combinations.
- HEM provides better generalization on multiple well data.
- Existing ANN based bit selection model can be outperformed with an RSM and ABC combination based on optimum values of drilling ROP.
- Imbalanced data problem can be handled through ensemble and data resampling techniques.
- The estimation performance of machine learning models varies for oil and gas production through a surface installed choke.
- Machine learning applications need proper guideline and framework for handling data related issues.

1.8 Problem Statement

The processing and analysis of this huge data pose a significant challenge to the prevailing technologies used in the oil and gas industry. With the advancement of smart sensors, and IIoT, technologies, huge amount of data are generated by upstream industry. This requires hybrid machine learning models for their processing to extract useful information. Hybrid ensemble models can be utilized for sloving various issues at data level as well as algorithm level also. However, these methods are not much investigated in the literature for their possible application in the petroleum domain. In this thesis, the intelligent modeling approach has been investigated to provide cost-effective solutions for three major problems of petroleum domain viz. (a) Lithofacies identification (b) Drilling optimization (c) Production rate estimation.

- Modeling of lithofacies is a challenging task due to the prevailing uncertainties in all the reservoir related measurements and less generalization capability of single supervised models.
- The selection of suitable drill bit types is a challenging problem in drilling operations. Earlier models have their own limitations and shortcomings which are needed to be thoroughly studied and required development of an alternative data driven solution.
- Ensemble classifiers are required to be investigated for drill bit types selection problem which formulates as multiclass imbalanced classification challenge and need adjustments at data and algorithm levels.
- Several empirical models and correlations have been proposed to predict the multi-phase flow through surface installed chokes. However, no proper guidelines and performance comparison are available in the literature for ML implementations.

1.9 Conceptual framework for processing petroleum data

A diverse variety of data exists in the petroleum domain that is difficult to process together even if they are describing the same reservoir properties such as core data, well logs, and seismic data, etc. All the before said data types are helpful for the identification of lithofacies, however, they require separate processing techniques for their interpretation. Petroleum data also suffer from several complications such as nonlinearity, high dimensionality, noise, and imbalanced data conditions as mentioned earlier. These impurities adversely affect the pattern recognition capabilities of machine learning models and increase the prediction errors significantly due to bias and variance of data. Additional processing steps have been added before the training and testing of

different machine learning models. Intelligent modeling proposed in this research work is broadly classified into three stages for rectifying the issues related to petroleum data namely, the preprocessing stage, model development stage, and post-processing stage. These stages are further explained in detail in forthcoming chapters. Figure 1.2 shows a conceptual framework for the identification of lithofacies using well log data.

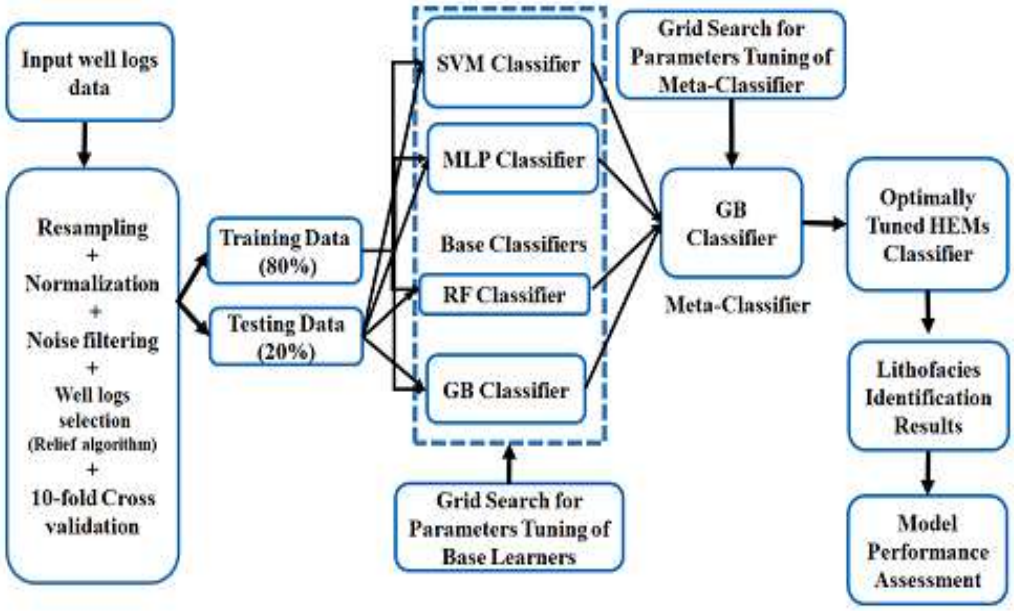


Figure 1.2 Conceptual framework for the identification of lithofacies using well logs data [54].

1.10 Dataset description

The datasets used for training, validation, and testing of different machine learning models, used in the thesis, primarily belong to Norwegians Volve and Kansas, U.S.A. Both of these oil and gas fields have published their data in the public domain for academic and research purposes. A general introduction about both the fields has been described as given below.

1.10.1 Kansas field

Kansas region is mainly composed of sedimentary rocks with a maximum width of 2850 m. A large number of unconformities occur in the Kansas region with the sedimentary strata having 15–50% of the post-Precambrian period [58]. Northeastern Kansas is enclosed by Pleistocene glacial deposits. A thick layer of Mesozoic rock is present in the western Kansas region. Mesozoic rock layers are mainly made up of limestones, chalks, sandstone, marine shales, and nonmarine shale contents. Panoma field and Hugoton field, existing in western Kansas, comprise large natural gas-producing reservoirs. Pennsylvanian and Permian systems are the broadest structures of rock containing bedded rock salts in several layers. The pre-Pennsylvanian system existing in Kansas contains dolomites, marine, limestones layered alternatively with sandstones and shales. The Precambrian basement is composed mainly of quartzite, granite, and schist. Permian strata contain the carbonate reservoirs that produce the majority of natural gas. In 1992, Mississippian strata produced 43% of cumulative hydrocarbon production of the Kansas field out of which 19% contributed to cumulative oil production [59]. The numerous unconformities available in the Kansas region help trapping and migration of petroleum. Basal Pennsylvanian in Kansas has a huge deposition of hydrocarbon along with its length. A detailed description of the petroleum geology of the Kansas region can be found in Newell [59,60], Merriam [58], Adler et al. [61], and Jewett and Merriam [62]. Manual interpretation of wells logs data of such a huge hydrocarbon producing region is time-consuming and costly. Therefore, automatic detection and identification of subsurface lithofacies using machine learning algorithms are highly desirable to minimize cost and time. Figure 1.3 shows maps of Kansas oil and gas fields.

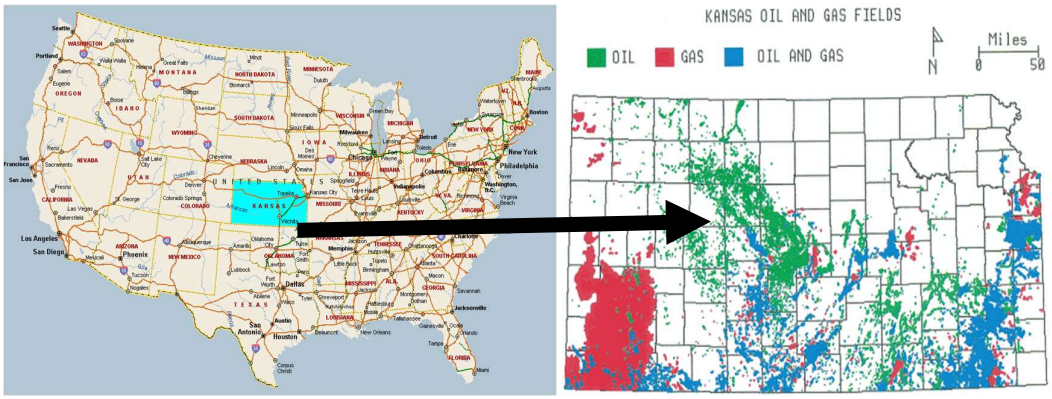


Figure 1.3 Maps of Kansas oil and gas fields.

1.10.2 Volve Field

The Volve field was explored in the central North Sea with 80 m below water depth in 1993. Its hydrocarbon-producing reservoir consists of sandstone belonging to the Jurassic age at the depth of 2750 m-3120 m. Drilling operations started on the Volve field in May 2007 followed by hydrocarbon production in the same year with lifetime anticipation of three to five years. However, its production continued three years longer than the expected number of years. Total oil production of the Volve field was expected to be nine and a half million barrels with a recovery of 54%. Later, it was shut down in October 2016 by a joint decision of the investors viz. Statoil, Bayen Gas, and ExxonMobil. The wells data were made public for research and academic purposes on the website of Equinor company. The three issues stated during the drilling of the abovesaid field were wellbore stability, rock mechanics, and loss of circulation problems. Pore pressure (PP) recorded during drilling varied along with the formation depth. In well 15/9-F-4, the PP gradient fell in the Shetland group whereas increased in Draupne shale. The initial PP gradient recorded in well 15/9-F-4 was 1.14 sg or 335 bar that dropped up to 0.95 sg or 280 bar in different reservoir formations. The value of maximum collapse pressure estimated in the lower Hordaland Group was 1.40 sg and

1.38 sg in Draupne shales. Hordaland shale was unstable for high-angle wells with higher pore pressure. Balder formation in the Rogaland group had a low fracture gradient because of its tuffaceous and friable nature. It also acted as a loss zone with a high vulnerability to washouts. Cromer Knoll and Sola formations were the most unstable formation zones. Ty formation and Balder formation were recognized as potential loss zones before drilling operations. Figure 1.4 shows geographical location of the Norwegian Volve field.

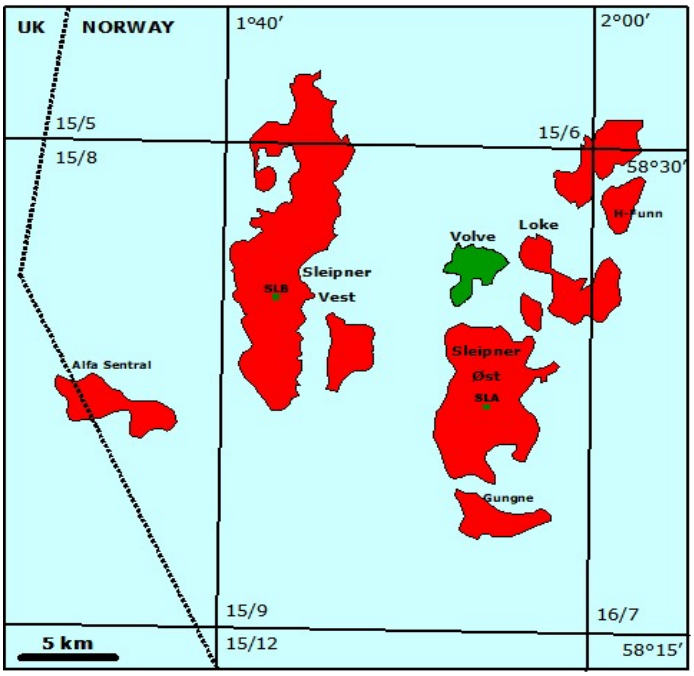


Figure 1.4 Geographical location of the Norwegian Volve field.

1.11 Dissertation outline

Chapter 2 investigates the feasibility of the application of the ensemble approach for the development of lithofacies recognition models. It also describes the comparative performance testing of five potential ensemble methods for quantitative lithofacies modeling. Additionally, it discusses the effect of imbalanced well-logs data on the

performance of supervised learners. The comparative results have been generated using single well data existing in the Kansas region of the U.S.A.

Chapter 3 explains the utilization of heterogeneous ensemble models (HEMs) for lithofacies identification using multiple wells data for generating more generalized results for the Kansas region. This study also elaborates on the benefits of HEMs over other techniques for quantitative lithofacies modeling.

Chapter 4 describes the application of Response surface methodology (RSM) for the development of the objective function of ROP for its optimization. Further, drill bits are selected based on optimum ROP values for the target formation.

Chapter 5 describes possible solutions for complex imbalanced multi-class classification problem that occurs during data-driven drill bit selection. Architectural modifications in ensemble methods are proposed with the data re-sampling techniques to provide a new and efficient approach for handling the complex drill bit selection process. Additionally, four popular machine learning techniques are also evaluated to understand the performance degrading effects of imbalanced drilling data obtained from Norwegian wells. The issue of data imbalance has been discussed in detail with possible remedies for the selection of suitable drill bits.

Chapter 6 illustrates the modeling of hydrocarbon production through surface installed chokes using machine learning models. The multi-phase flow behavior influences the overall hydrocarbon production rate. Therefore, data-driven models have been proposed and investigated for the estimation of the hydrocarbon flow rate.

In **Chapter 7**, the main contributions and results of this thesis work are summarized. The potential future implications of these studies along with the future research scope are also highlighted in this chapter.

Chapter 2

Study of Homogeneous Ensemble Methods for the Identification of Geological Lithofacies

2.1 Introduction

Quantitative lithofacies modeling is an essential part of reservoir characterization to identify different subsurface layers. Well-logs are primarily utilized to recognize reservoir rock layers underlying along with the depth of the formation. Conventionally, interpretation of logs is performed manually which requires cost, time, and domain expertise. Quantitative lithofacies modeling is a difficult task because of several factors as listed below.

(a) Limited availability of core samples due to economic and logistic constraints;

(b) Overlapping lithofacies facies;

(c) Uncertainties in subsurface measurements;

(d) Hard to identify thinner geological layers;

(e) Variations within thick lithofacies;

(f) Difficult to detect facies boundaries; and

(g) Big data problem.

With the advancement in computational capability, intelligent paradigms are applied for the automatic detection and classification of well logs data [5]. Mainly, single supervised classifiers have been utilized for the recognition of lithofacies [3,5]. However, they have their limitations and shortcomings [5]. Therefore, advanced hybrid computational models have been investigated in this study for quantitative lithofacies modeling.

Several intelligent machine learning models were applied for the automatic identification of lithofacies using computational processing of well logs data. Some of the important applied intelligent models are as follows: (a) Clustering technique [6]; (b) Principal component analysis (PCA) with Artificial neural networks (ANN) [7, 8]; (c) Support vector machine (SVM) [4]; (d) Self-organizing Map (SOM) and Multi-resolution graph-based clustering (MRGC) [3]; and (e) Random forest [9]. It can be seen that the application of conventional supervised classifiers is mainly reported for lithofacies modeling. The utilization of a hybrid computational approach has been rarely explored to extract facies information.

This chapter presents the application and comparison of five ensemble methods namely, Bagging, AdaBoost, Rotation forest, Random subspace, and DECORATE for quantitative lithofacies modeling. It also considers seven popular classifiers (namely Naïve Bayes (NBC), Logistic regression (LogR), Multilayer perceptron (MLP), Radial basis function (RBF), SVM, Classification & regression tree (CART), and C4.5 Decision trees) as base classifiers in the proposed ensemble-based lithofacies modeling. The primary objective of implementing ensemble methods for lithofacies modeling is to enhance the efficacy of the base classifiers.

2.2 Background

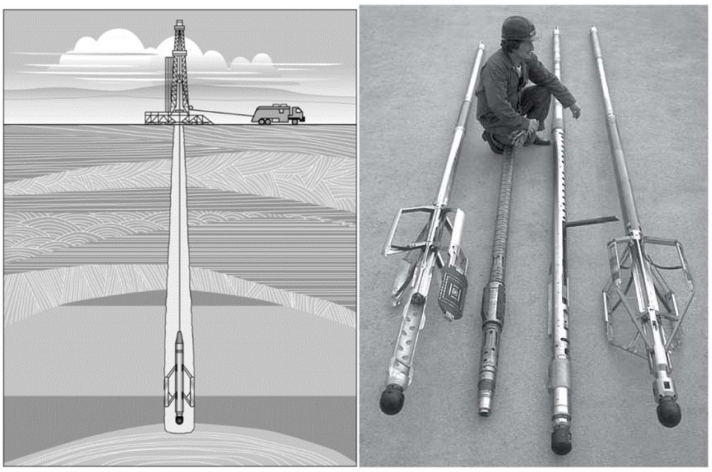
The exploration of oil and gas initiates with the identification of potential geological formations, seismic survey, and drilling of wild cats. After drilling of wild cat wells, coring is initially performed to collect the rock samples. Collected samples are tested in the laboratory to measure the formation's rock and fluid properties. However, the process of coring is found to be costly which limits the availability of core samples. Thus, well logging techniques are implemented. The wireline tools are lowered in the drilled formations to assess the presence of sufficient hydrocarbon inside the

lithological formations. The logging started on 5th, September 1927, when H. Doll and Schlumberger brothers at Pechelbronn first recorded semi-continuous resistivity measurement in the Alsace oil field [63]. A sonde is a measurement device that is lowered inside the wells during wireline operation to record the characteristic of the subsurface geological formations inside the wellbore. These logs serve varied purposes for geologists, production geologists, and petro physicists. The geologist is primarily concerned with the depositional environment, lithology, and stratigraphy of subsurface formations prevailing inside wellbores. Exploration geologists develop a large-scale image of underlying geology using different logs that are helpful in reservoir modeling and to determine the location for drilling a new well. The production geologist also performs a similar task with more information to produce a detailed geological reservoir model which plays an important role in reservoir management.

The petrophysicist utilizes all the available information to study the physical, chemical, mineralogical properties of reservoir rocks along with the distribution of fluid inside the reservoir formations. The petrophysicist mainly identifies sweet spots, estimates reservoir properties, predicts the volume of hydrocarbon present inside the reservoir formation, and designs strategies for reservoir management for the long-term recovery of oil and gas. Conventionally, experts manually examine well-logs to identify different layers of lithofacies through their experience. However, there is always a high chance of human error during the manual interpretation of complex well-logs. The manual analysis is not considered good practice for the interpretation of well-logs.

Nowadays, measurements while logging has been utilized to record multiple data from the downhole wellbore. This system typically consists of a downhole sensor unit, a telemetry system, a power source, and a surface display device. A huge amount of data are generated with measurement-while-drilling (MWD) or logging-while-drilling

(LWD) that are manually interpreted by experts to identify subsurface information. Table 2.1 shows different types of logging usually performed on the oil and gas fields. The information collected through MWD or LWD data analyses is mainly utilized for reservoir characterization so that an accurate reserve model can be developed to estimate the static and dynamic behavior of the hydrocarbon reservoir.



(a) (b)

Figure 2.1 Wireline logging tools (a) depicts elements of logging tool viz. measurement sonde, wireline, and mobile laboratory. (b) The four well logging sonde tools: (left to right) dipmeter, sonic log, density logging tool, and dipmeter with multiple electrodes [63].

2.3 Homogeneous ensemble methods

Ensemble methods are multiple classifier systems that help to identify hidden patterns inside the complex data. It combines the decisions of several supervised classifiers for classification and estimation tasks to achieve more accurate results. Ensemble methods improve the performance of conventional classifiers known as base classifiers [64,65]. Ensemble techniques integrate diverse hypotheses to obtain the highest possible classification accuracy for a specific base classifier in feature space [66]. Ensemble

methods have distinct characteristics for handling the issues of complex petroleum data. Therefore, ensemble classifiers seem to be a better alternative for the extraction of useful information from sensory logging data.

To combine outcomes of base classifiers, the ensemble uses either weighted or unweighted voting rules [64-66]. The architecture of the ensemble approach can be generated in numerous ways such as random choice of training data, a random selection of feature space, increasing diversity of training data, manipulation of the error function, etc. [64-66]. The decisions of base classifiers are combined through fusion methods such as majority voting, Borda count, algebraic combiners, etc. [67]. Base classifiers are also termed weak classifiers, unstable classifiers, classifiers of low complexity, and badly performing classifiers [67]. It has been reported that the performance of a weak or unstable classifier can be enhanced in three ways namely, regularization [68], noise injection [69], and a system of multiple classifiers [64-66]. In this research work, five ensemble methods have been tested for the recognition of lithofacies as given below.

2.3.1 Bagging

Breiman et al. [70] united notions of bootstrapping and aggregating techniques together for the training of base classifiers. A random independent sampling of training data with replacement is done to generate bootstrap samples [71]. These bootstrap samples are utilized for the training of base classifiers simultaneously. Bootstrapping generates m bootstraps replicate samples $X_m = (X_1, X_2, ...X_M)$ where (m=1,2,3,….M)) from training set X with replacement. Base classifiers are trained simultaneously on X_m random data samples. Aggregating combines the outcomes of base classifiers together for the final classification decision. During the testing phase, class labels of test data are decided by majority votes acquired through aggregating the decisions of trained base/weak classifiers. Base classifiers are trained on unlike set of samples so

they are different from each other. The classification of test data samples is decided by majority voting acquired from base classifiers. Figure 2.2 shows a generalized workflow of the 'Bagging ensemble' classifier used in QLM. Bagging technique is implemented in following ways.

- Generate bootstrap replicate samples X_m, $X_m = (X_1, X_2, ...X_M)$ from the training data X (m=1,2,3,..M).
- Train the base classifiers $C_m(x)$ on X_m datasets. (m=1,2,3,4….M).
- Combine classifiers outcomes using simple majority voting to a final decision rule as given below. $\alpha_i = \arg\max_{y\in\{1,-1\}} \sum_i \delta_{\text{sgn}}(C_i(x)), yl$ where $\delta_{i,j} = \{^{1\, i=j}_{0\, i\neq j}$ is the Kronecker symbol, yl is the class label of the classifier.

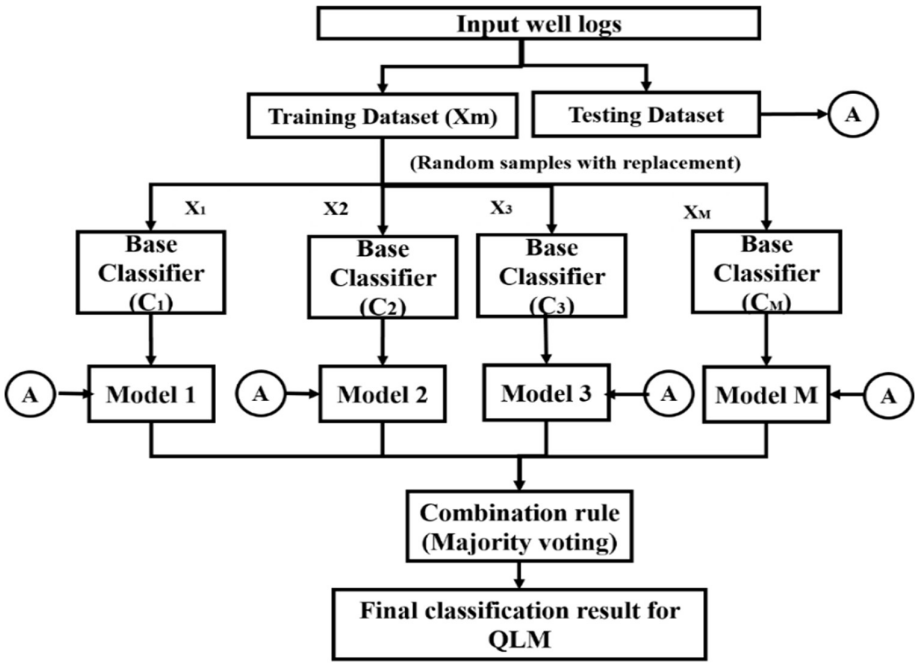


Figure 2.2 A generalized workflow of the 'Bagging ensemble' classifier used in QLM.

2.3.2 AdaBoost

Freund and Schapire, modified the boosting algorithm to develop the AdaBoost ensemble [71]. AdaBoost trains ensemble classifiers by resampling the original data and

combine their decisions through weighted majority voting rule. AdaBoost generates consecutive bootstrap X_m samples and initially assigns equal weights to all training data samples. AdaBoost trains the preliminary base classifier $C_1(x)$ on the initial bootstrap data samples X1. Later, weights are adjusted according to the misclassifications made by the initial base classifier $C_1(x)$. The weights of incorrectly classified training samples are increased in a next modified training set. Hence, the chance of recurrence of the misclassified samples in next training samples X_2 increases for $C_2(x)$. Classification results are produced by the combination of the weighted votes or decisions of the base classifiers $C_m(x)$. The weights of base classifiers are decided based on their classification performance. AdaBoost can be implemented in the following ways.

- Generates bootstrap training samples $X_m = (X_1, X_2, ...X_M)$ and combines with these sets of weights W_m, $W_m = (W_1, W_2, ...W_M)$. (m=1,2,3,4…M)
- Train the base classifiers $C_m(x)$ by using weighted samples $X_m W_m$, $\left(X_m W_m = (X_1 W_1, X_2 W_2, ...X_M W_M\right)$ and calculate the probability estimates of error as given below. $err_m = \frac{1}{M}\sum_{j=1}^{M} W_j \zeta_j$ where $\zeta_j = \begin{cases} 1 & otherwise \\ 0 & if\ X_j is\ classified\ correctly \end{cases}$ and combining weight is calculated as follows $C_m = \frac{1}{2}\log\left(\frac{1-err_m}{err_m}\right)$.
- Set $w_i^{m+1} = w_i^m \exp(C_w \zeta_i^m)$ if the error is between 0 to 0.5, (i=1,2,3,..M) and renormalize so that $\sum_{i=1}^{M} W_i^{m+1} = M$, else if set all the weights to 1 and repeat the step again.

- Combine classifiers outcomes using simple majority voting to a final decision rule as given below. $\alpha_i = \arg\max_{y\in\{1,-1\}} \sum_i \delta_{\text{sgn}}(C_i(x)), yl$ where $\delta_{i,j} = \{^{1\, i=j}_{0\, i\neq j}$ is the Kronecker symbol, yl is the class label of the classifier.

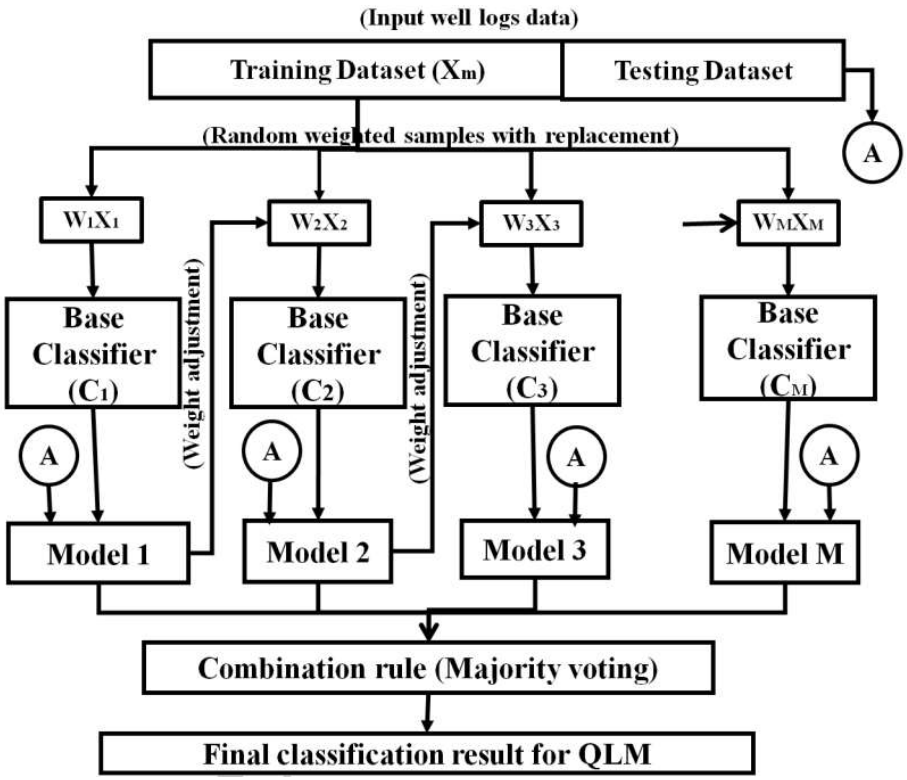


Figure 2.3 A generalized workflow of the 'Adaboost ensemble' classifier used in QLM.

2.3.3 Rotation Forest

Rotation forest trains its base classifiers using extracted features of the subsets generated from the training dataset. Initially, training data are partitioned into subsets, and then the principal component analysis is applied to extract the features from these subsets. Let $X_m = (X_1, X_2, ...X_M)$ be the original training data samples containing H features and dimensionality $N \times H$. Let Y be the vector containing class label $Y_n = (Y_1, Y_2, ...Y_N)$. Initially, features H are randomly split into M subsets: $H_{i,j}$ $(for - j = 1,2,..M)$. Let $X_{i,j}$' is the bootstrap samples data samples generated from

75% of $X_{i,j}$. Apply principal component analysis on $X_{i,j}$'to obtain the principal coefficients. The coefficients are arranged in a rotation matrix $Ro_{i,j}$ is rearranged into $Ro_{i,j}$' to develop the same order of features as in H. Build a classifier S_i using $(XRo_{i,j}'Y)$ as a training set. For a given sample, classifier S_i assigns the confidences to the hypothesis that the sample belongs to a particular class (CL_i). The confidences for each class CL_i are calculated by the average combination methods. Classification of a given sample is done according to the class having the largest assigned confidence. Confidence for each class is given by the average combination method. Confidence for each class is given by the average combination methods as shown below.

$$Co_i = \frac{1}{l}\sum_{j=1}^{l} P_{i,j}(XRo_{i,j}), i = 1,2,3,...C \tag{2.1}$$

Where, $P_{i,j}(XRo_{i,j})$ is the probability of dataset belonging to the ith class. The sample will be assigned to the class having the highest confidence Co_i.

2.3.4 Random subspace

Ho [72] proposed the concept of random subspace to utilize the benefit of feature extraction for high dimensional data. In the Random subspace, base classifiers are trained with random sets of features extracted from training data. The final decision about the class label is taken by the majority voting rule. It is reported that the performance of Random forest is higher especially with the data having redundant features [72]. Random subspace paradigm can be implemented as given below.

- Initially, extract random n-dimensional subspace X_i, from original m-dimensional feature space X.
- Train classifier $C_i(x)$ in X_i.(i=1,2,3,4…m)

- Combine classifiers outcomes using simple majority voting to a final decision rule as given below. $\alpha_i = \arg\max_{y\in\{1,-1\}} \sum_i \delta_{\text{sgn}}(C_i(x)), yl$ where $\delta_{i,j} = \{^{1\, i=j}_{0\, i\neq j}$ is the Kronecker symbol, yl is the class label of the classifier.

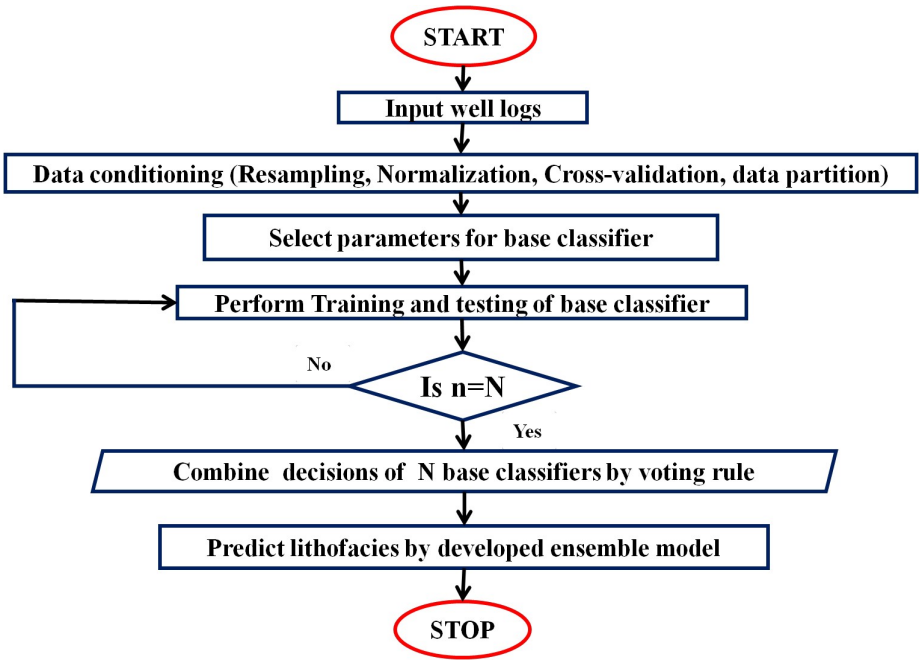


Figure 2.4 The generalized flowchart of the proposed ensemble methods for quantitative lithofacies modeling.

2.3.5 Diverse ensemble creation by oppositional relabeling of artificial training examples (DECORATE)

Melville et al. [73] developed a new ensemble architecture based on the maximization of the data diversity concept. It has been verified that overall generalization error reduces with an increase in the training data diversity. In this ensemble approach, base classifiers are trained in iteration one after the other on the combination of training and diversity data. Diversity data samples are artificially generated at every iteration based on the data distribution of original data. The class labels are assigned to these artificial data samples to maximize diversity in the final results. At every iteration, a new base

classifier is trained with the combination of training and artificial data, however, the addition of this classifier depends upon a reduction in training error with an increase in classification accuracy of the ensemble. This ensemble continues its iterations until it reaches its stop criteria such as maximum ensemble size, the maximum number of iteration, etc. [73].

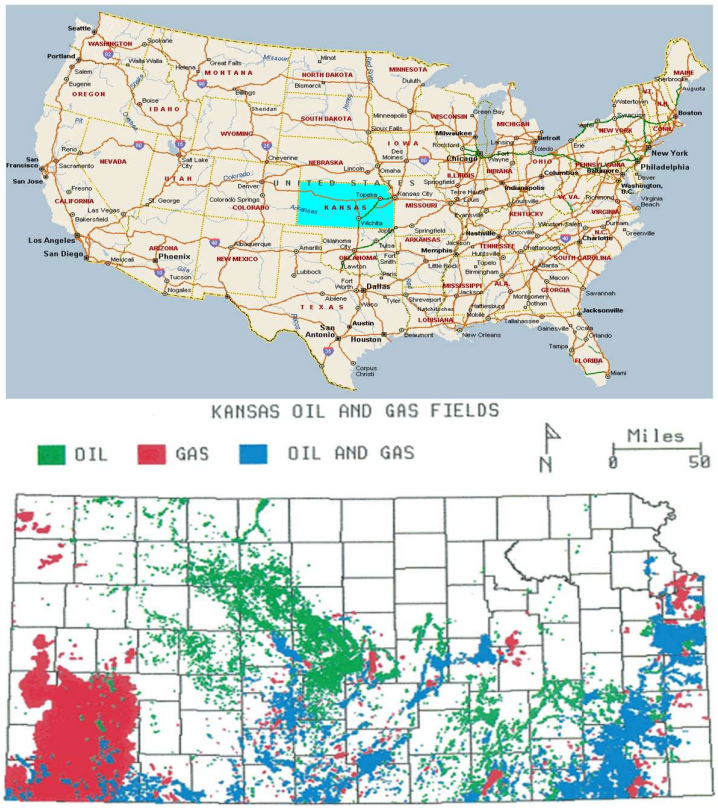


Figure 2.5 Political maps of the U.S.A. with the Kansas region and distribution of oil and gas wells [58].

2.4 *Data description*

2.4.1 Kansas oil and gas field

The well-logs data utilized for the training and testing of ensemble classifiers is downloaded from the Kansas geological survey (KGS) website which is one of the largest data repositories provided for research purposes. Geological well-logs

downloaded from this website belong to single Paradise A well data (API:15-163-24133) situated in the Kansas region of the U.S.A [58]. This well-logs data contains several samples having missing or garbage or null values that are removed through resampling operation. Nine lithofacies are considered in the training and testing datasets for the evaluation of ensemble classifiers. Total 2281 data samples were extracted from the downloaded las-file with the following lithofacies: clay, limestone, packstone, dolomite, dolomite mudstone, dolomite packstone, dolomite wackestone, silt, argillaceous clay [58]. These were acknowledged as class labels during the training and testing of well logs data. Figure 2.4 shows political maps of the U.S.A. with the Kansas region and distribution of oil and gas wells.

Table 2.1 The statistical details of various well-logs data available in Paradise well [58]

S. No.	Input Parameters	Minimum	Maximum	Units
1.	Depth (DT)	3200	3771	ft
2.	Caliper 1 (DCAL)	7.468	8.982	inch
3.	Density Porosity (DPOR)	0.652	32.118	pu
4.	Gamma Ray (GR)	16.781	414.152	API
5.	Neutron Porosity (NPOR)	0	44.143	pu
6.	Bulk Density Correction (RHOC)	0.01	0.296	gm/cc
7.	Deep Induction Resistivity (RILD)	1.905	65.158	Ohm.m
8.	Medium Induction Resistivity (RILM)	2.031	140.706	Ohm-m
9.	Deep Laterolog Resistivity (RLL3)	2.866	266.481	Ohm-m
10.	Spontaneous Potential (SP)	-1.797	71.61	MV
11.	Acoustic Transit Time 1 (DT)	51.522	235.961	Usec/ft
12.	Micro Inverse Resistivity (MI)	0.609	41.818	Ohm-m
13.	Micro Normal Resistivity (MN)	0.138	44.097	Ohm-m
14.	Sonic Porosity (SPOR)	2.774	133.212	Pu
15.	Caliper 2 (MCAL)	6.947	8.911	In
16.	Acoustic Transit Time 2 (TT1)	54.36	662.885	Usec/ft

2.4.2 Performance evaluation indicators

The performance of ensemble classifier is assessed through popular statistical indicators such as classification accuracy, true positive rate (TPR) or sensitivity, true negative rate (TNR) or specificity, Type-1 error, Type-2 error, and area under the receiver operating characteristic curve (AUROC) as described in Table 2.2 given below.

Table 2.2 The statistical parameters utilized for the analysis of the results of lithofacies identification.

S. No.	Performance indicator	Descriptions
1	$\text{Accuracy}=\frac{TP+TN}{TP+FP+FN+TN}$	where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.
2	$\text{Sensitivity(Recall)}=\frac{TP}{TP+FN}$	where TP is true positive, and FN is false negative
3	$\text{Specificity}=\frac{TN}{TN+FN}$	where TN is true negative and FN is false negative
4	$\text{Precision}=\frac{TP}{TP+FP}$	where TP is true positive, and FP is false positive
5	$\text{Type-1 error}=\frac{FN}{FN+TN}$	Type-1 error (miss) is the probability of wrong samples being classified to a specific class,
6	$\text{Type-2 error}=\frac{FP}{TP+FP}$	whereas Type-2 error (false-alarm) is the probability of samples belonging to a specific class being categorized to the wrong class
7	Area under ROC (AUROC) curve is calculated to measure the classification performance of a classifier.	The Receiver Operating Curve (ROC) curve is a plot of True Positive Rate (TPR) versus False Positive Rate (FPR).

2.5 Results and discussion

The performances of five ensemble methods were evaluated using Kansas well logs data along with six supervised classifiers namely, MLP, C4.5, NBC, Log R, CART, RBF, and SVM. Table 2.3 to 2.8 illustrates the classification results for ensemble methods in combination with seven base classifiers. Table 2.3 contains the individual performance of every base classifier considered in this study. The individual classification accuracy of SVM is higher than all the other base classifiers. These base classifiers can be arranged in the decreasing according to their performance as follow as SVM, C4.5, CART, LogR, MLP, NBC, and RBF as shown in Table 2.3.

Table 2.3 Classification performance of the conventional classifiers for Kansas well-logs data.

S.No.	Classifier	Accuracy	AUROC	Sensitivity	Specificity	Precision
1.	NBC	71.2407	0.935	71.2	95.8	0.731
2.	MLP	73.0381	0.950	73.0	94.8	0.698
3.	LogR	86.1026	0.985	86.1	97.8	0.859
4.	RBF	56.5103	0.876	56.5	89.7	0.442
5.	CART	88.8207	0.972	88.8	98.4	0.889
6.	C4.5	89.6537	0.959	89.7	98.2	0.896
7.	SVM	90.925	0.947	90.9	98.5	0.909

Table 2.4 contains the average classification results of the Bagging ensemble in a combination with each base classifier considered in this study. C4.5 and CART have produced equivalent outcomes as shown in Table 2.4. Therefore, C4.5 and CART are equally efficient bases in the Bagging approach for quantitative lithofacies modeling. It is also inferred from Table 2.4 that SVM is the second effective base classifier with a slight decrease of 0.3946 % in its performance as compared to Table 2.3. LogR emerges

as the third effective base classifier, however, its AUROC performance is marginally superior to SVM. There is a major reduction in the enactment whereas the RBF classifier has produced the worst outcomes as compared to all the remaining base classifiers as shown in Table 2.4.

Table 2.4 Average classification accuracy of 'Bagging ensemble' for Kansas well-log data.

Base	Accuracy	AUROC	Sensitivity/ recall	Specificity	Precision	Type-I error	Type-2 error
MLP	75.916	0.969	75.9	95.3	0.748	0.047	0.241
NBC	71.3284	0.938	71.3	95.8	0.731	0.042	0.287
LogR	85.6203	0.986	85.6	97.8	0.854	0.022	0.144
CART	90.3113	0.988	90.3	98.5	0.904	0.015	0.097
C4.5	90.3113	0.988	90.3	98.5	0.904	0.015	0.097
RBF	57.431	0.880	57.4	90	0.555	0.100	0.426
SVM	89.9167	0.978	89.9	98.2	0.899	0.018	0.101

Table 2.5 contains the performance of the Adaboost ensemble in the combination of base classifiers for the quantitative lithofacies modeling. The C4.5 is the frontrunner base classifier in the AdaBoost ensemble architecture for quantitative lithofacies modeling. It has provided the best performance in the AdaBoost ensemble as shown in Table 2.5. In AdaBoost, CART and SVM have given similar and second-best performance outcomes. MLP and NBC remain relatively mediocre bases with AdaBoost ensemble for the classification of lithofacies, while RBF is performed as the inferior base classifier.

Table 2.5 The performance of 'AdaBoost ensemble' for Kansas well-logs.

Base	Accuracy	AUROC	Sensitivity/ recall	Specificity	Precision	Type-I error	Type-2 error
MLP	77.7291	0.954	77.7	95.7	0.740	0.043	0.223
NBC	71.1092	0.876	71.1	95.8	0.727	0.042	0.289
LogR	85.9711	0.927	86.0	97.8	0.858	0.022	0.14
CART	90.5305	0.985	90.5	98.6	0.906	0.014	0.095
C4.5	91.4073	0.987	91.4	98.6	0.914	0.014	0.086
RBF	56.861	0.787	56.9	90	0.469	0.100	0.431
SVM	90.5305	0.983	90.5	98.3	0.905	0.017	0.095

Table 2.6 The performance of 'Rotation' for the classification of Kansas field well-logs

Base	Accuracy	AUROC	Sensitivity/ recall	Specificity	Precision	Type-I error	Type-2 error
MLP	76.1947	0.966	76.2	95.4	0.743	0.046	0.238
NBC	62.3849	0.916	624	94.5	0.646	0.055	0.376
LogR	86.0149	0.985	86.0	97.8	0.858	0.022	0.14
CART	88.9522	0.993	89.0	98.1	0.891	0.019	0.11
C4.5	90.7497	0.966	90.7	85	0.908	0.15	0.093
RBF	58.3516	0.900	58.4	90.3	0.516	0.097	0.416
SVM	91.0566	0.975	91.1	98.4	0.911	0.016	0.089

Table 2.6 demonstrates the performance of the Rotation forest ensemble in combination with six base classifiers considered in this study for lithofacies modeling. SVM has appeared as the frontrunner base classifier. C4.5 has attained second place in

terms of its performance and average AUROC though CART moved to third place. The MLP and NBC remain mediocre performing classifiers with Rotation forest too, however, RBF is still the worse base classifier for the recognition of lithofacies. Table 2.7 summarizes the performance of the Random subspace ensemble for geological lithofacies modeling of the Kansas oil-field well-logs data. Again, SVM has emerged as an effective base classifier with the Random subspace ensemble with the highest accuracy of 92.28% as compared to all base combinations experimented in this study. C4.5 has acquired the second performance metrics with Random subspace. CART acquired the third place in terms of classification performance shown in Table 2.7. LogR has given mediocre performance with Random subspace and can be placed at fourth place in the list of the suitable base classifier for lithofacies modeling. MLP, NBC, and RBF remain weak performers for the identification of lithofacies.

Table 2.7 The performance of 'Random subspace ensemble' for Kansas oil field well-log data.

Base	Accuracy	AUROC	Sensitivity/ recall	Specificity	Precision	Type-I error	Type-2 error
MLP	69.9693	0.967	70.0	37.0	0.755	0.63	0.3
NBC	70.1447	0.929	70.1	95.3	0.729	0.047	0.299
LogR	80.3157	0.973	80.3	96.5	0.804	0.035	0.197
CART	89.6098	0.990	89.6	98.2	0.896	0.018	0.104
C4.5	89.9605	0.992	90.0	98.2	0.899	0.018	0.1
RBF	59.6668	0.909	59.7	90.5	0.543	0.095	0.403
SVM	92.2841	0.990	92.3	98.7	0.923	0.013	0.077

Table 2.8 illustrates the performance of the DECORATE ensemble with its six different base classifiers. C4.5 has provided the best results in terms of overall

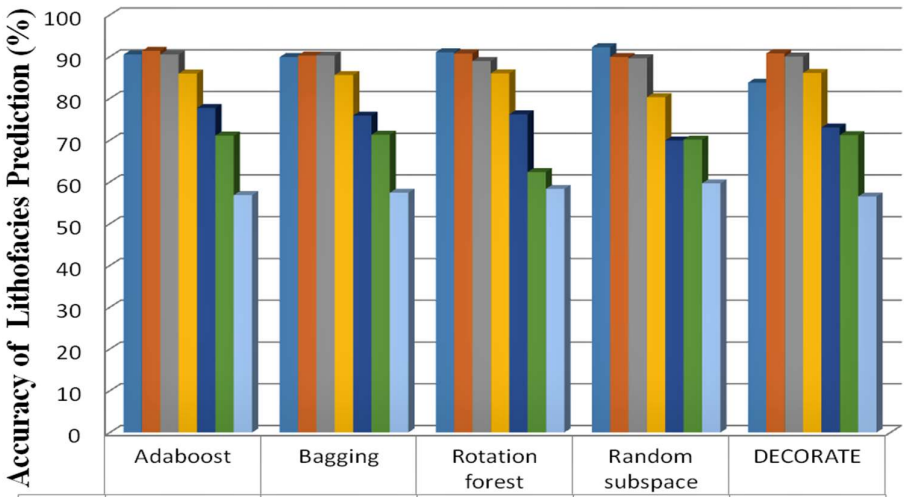
performance metrics. CART has provided the second-best results followed by LogR. The performance of SVM has been declined with DECORATE compared to its performance with all other ensemble methods. MLP, NBC, and RBF have emerged as poor performance base members with DECORATE for geological lithofacies modeling of the Kansas oil-field well-logs data. Figure 2.5 contains a confusion matrix for Random subspace classifiers which has achieved the highest accuracy in combination with SVM. It can also be observed that the overall classification accuracies of CART, C4.5, and SVM are in close competition with each other when used as a base classifier with ensembles, as shown in Tables 2.3-2.8. Figure 2.6 contains summary of results for ensemble methods to predict the geological lithofacies.

Table 2.8 The performance of 'DECORATE' for the facies recognition of Kansas wells.

Base	**Accuracy**	**AUROC**	**Sensitivity/ recall**	**Specificity**	**Precision**	**Type-I error**	**Type-2 error**
MLP	73.0381	0.950	73.0	94.8	0.698	0.052	0.27
NBC	71.2407	0.935	71.2	95.8	0.731	0.042	0.288
LogR	86.1026	0.985	86.1	97.8	0.859	0.022	0.139
CART	90.0482	0.987	90.0	98.5	0.901	0.015	0.1
C4.5	90.7935	0.988	90.8	98.6	0.908	0.014	0.092
RBF	56.5103	0.876	56.5	89.7	0.442	0.103	0.435
SVM	83.779	0.904	83.8	97	0.837	0.030	0.162

		Actual Lithofacies								
Predicted Lithofacies		Clay	Carbonate Mudstone	Packstone	Wackstone	Dolomite Packstone	Dolomite Wackstone	Dolomite Mudstone	Dolomitic Sandstone	Silt
	Clay	183	6	7	0	1	1	0	7	0
	Carbonate Mudstone	4	351	3	0	4	18	0	0	0
	Packstone	5	8	221	1	1	2	0	1	5
	Wackstone	0	0	6	31	1	0	0	0	0
	Dolomite Packstone	0	9	2	0	132	0	0	6	1
	Dolomite Wackstone	6	15	0	0	0	671	0	0	0
	Dolomite Mudstone	0	1	2	0	1	0	166	8	1
	Dolomitic Sandstone	11	5	1	0	7	0	13	302	0
	Silt	0	0	4	0	1	0	1	0	48
	Absolute Accuracy	87.56%	88.86%	89.34%	96.88%	89.19%	96.97%	92.2%	93.21%	87.27%

Figure 2.6 Confusion matrix depicting overall classification accuracy of 'Random subspace ensemble' with SVM as a base classifier.



	Adaboost	Bagging	Rotation forest	Random subspace	DECORATE
SVM	90.5305	89.9167	91.0566	92.2841	83.779
C4.5	91.4073	90.3113	90.7497	89.9605	90.7935
CART	90.5305	90.3113	88.9522	89.6098	90.0482
LogR	85.9711	85.6203	86.0149	80.3157	86.1026
MLP	77.7291	75.916	76.1947	69.9693	73.0381
NBC	71.1092	71.3284	62.3849	70.1447	71.2407
RBF	56.861	57.431	58.3516	59.6668	56.5103

Figure 2.7 Summary of results for ensemble methods to predict the geological lithofacies.

2.6 Summary

Five ensemble methods have been trained and tested with six different base classifier combinations for quantitative lithofacies modeling. It has been observed that the performance of ensemble methods depends upon the behavior of the base classifiers. Base classifiers have shown improvement from their performance when applied in ensemble architecture. Most suitable ensemble-base combination found after comprehensive investigation are as follows: (a) Bagging-CART/C4.5, (b) AdaBoost-C4.5, (c) Rotation forest-SVM, (d) Random subspace-SVM, and (e) DECORATE-C4.5. Random subspace –SVM combination has given the highest classification accuracy of 92.28% as compared to all base combinations experimented. The analysis of results indicates that Ensemble methodology has a huge potential for lithofacies modeling. This study has motivated us to investigate other hybrid computational approaches such as the Random committee, Stacking ensembles, Voting ensemble, Bucket of models, cascading, Committee of machines, Clustering ensemble, etc. for their potential application in petroleum engineering.

Chapter 3

A Comparative Study of Heterogeneous Ensemble Methods for the Identification of Geological Lithofacies

3.1 Introduction

Most of the unconventional reservoirs are located in mudstone lithology that forms a very peculiar hydrocarbon generation and storage system. These are the type of sedimentary rock that acts as a source, cap, and storage reservoir for hydrocarbon generation to accumulation. Mudstone lithology contains sweet spots that support hydrocarbon production due to the presence of rich source rocks [12]. The textual and mineralogical contents of mudstones have heterogeneous distributions that are not apparent [12]. Well-logs are mainly utilized for the identification of the subsurface lithofacies along with the depth through human experience. However, the manual interpretation of well-logs is a time-consuming, and costly task. It also requires domain expertise for its interpretation. Further, mudstone reservoirs have overlapping and complex lithology along with the depth of the reservoir formation which is difficult to interpret using conventional techniques (Figure 3.1). Therefore, more advanced techniques are required for automatic lithofacies modeling.

Most of the machine learning models, described for lithofacies modeling in the literature, are based on single supervised or unsupervised classifiers. However, it has been proved that the performance of single classifiers can be improved using hybrid computational models such as a multiple-classifier system, a committee of machines, composite systems, etc. [11] Multiple classifier systems, like ensemble methods, can excavate more valuable information from raw sensory data. It combines the decisions of several classifiers for classification and regression tasks. The ensemble approach can be

categorized into two types: (a) homogeneous ensemble methods (HoEMs) such as Bagging, Random forest, etc., and (b) heterogeneous ensemble methods (HEMs) such as Voting, Stacking, etc. HoEMs in feature space combine several hypotheses generated by the identical type of supervised classifiers which are utilized as base classifiers (e.g., a cluster of hundreds of SVMs). In the case of HEMs, different classifiers are utilized to generate and combine diverse hypotheses to achieve maximum possible prediction accuracy for the existing feature space. It has been proved that heterogeneity in base classifiers helps to develop more reliable, robust, and generalized classifier models [14]. Therefore, heterogeneous ensemble methods are found to be more efficient in handling complex, nonlinear, multidimensional, and imbalanced data as compared HoEMs [5,14]. However, these methods are not much explored for the processing of petroleum data.

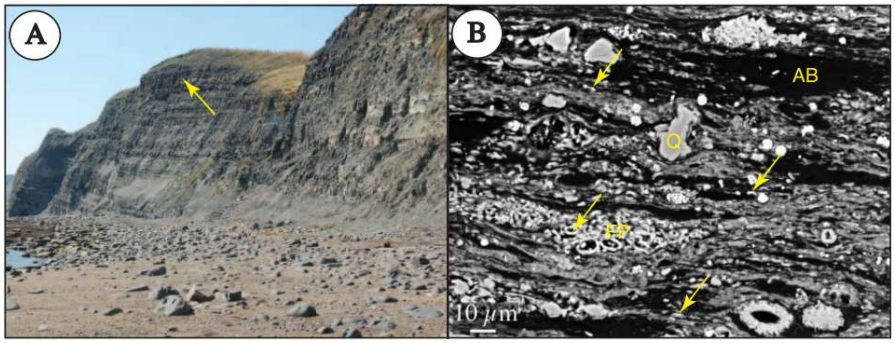


Figure 3.1 Variability of mudstone (a) Kimmeridge clay formation of Upper Jurassic in Dorset, England (b) Backscattered electron image of siliciclastic mudstone samples collected from the tip of the arrow shown in A [12].

In this chapter, two HEMs, namely Voting and Stacking, ensembles have been applied for the quantitative modeling of mudstone lithofacies using Kansas oil-field data. RF, gradient boosting (GB), SVM, and MLP have been incorporated as base classifiers in the applied HEMs architecture. A comprehensive comparison has also been performed among these classifiers for lithofacies identification. Multiple wells

data have been considered to achieve better generalized results for lithofacies modeling. Overall, the coming sections evaluate the pattern recognition ability of HEMs for multifarious mudstone lithofacies using multiple wells logs data. Figure 3.2 shows a conceptual architecture of Stacking ensemble for the identification of lithofacies.

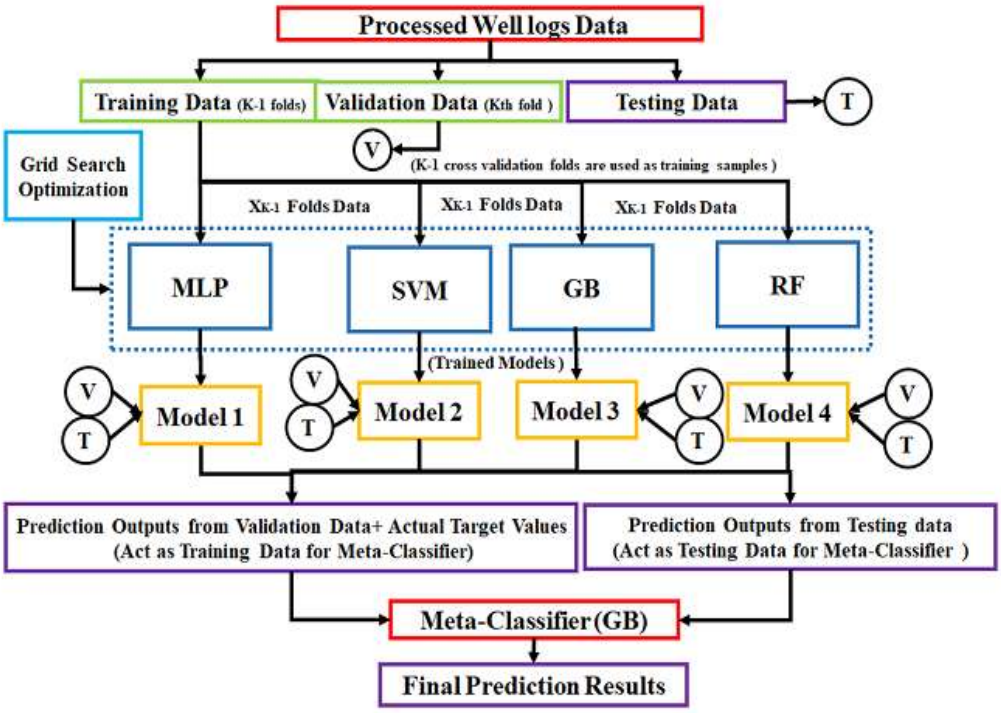


Figure 3.2 A conceptual architecture of Stacking ensemble for the identification of lithofacies.

3.2 Stacked generalization ensemble

Wolpert proposed the stacked generalization ensemble which is popularly known as Stacking [74]. It integrates the results of diverse supervised classifiers in its architecture as a base classifier. Dissimilar base classifiers quest the feature space with their varied viewpoints to discover the best possible hypotheses for a particular classification job [75]. It combines the results of base classifiers with a meta-classifier to deliver the ultimate classification outcome. The Stacking ensemble can also be generated by the

unification of the same base classifiers having varied parametric values. The choice of base and meta-classifier is always a matter of concern. The design of a suitable arrangement of classifiers in a large feature space is quite difficult. It has been proved that the Stacking ensemble helps to minimize the generalization error.

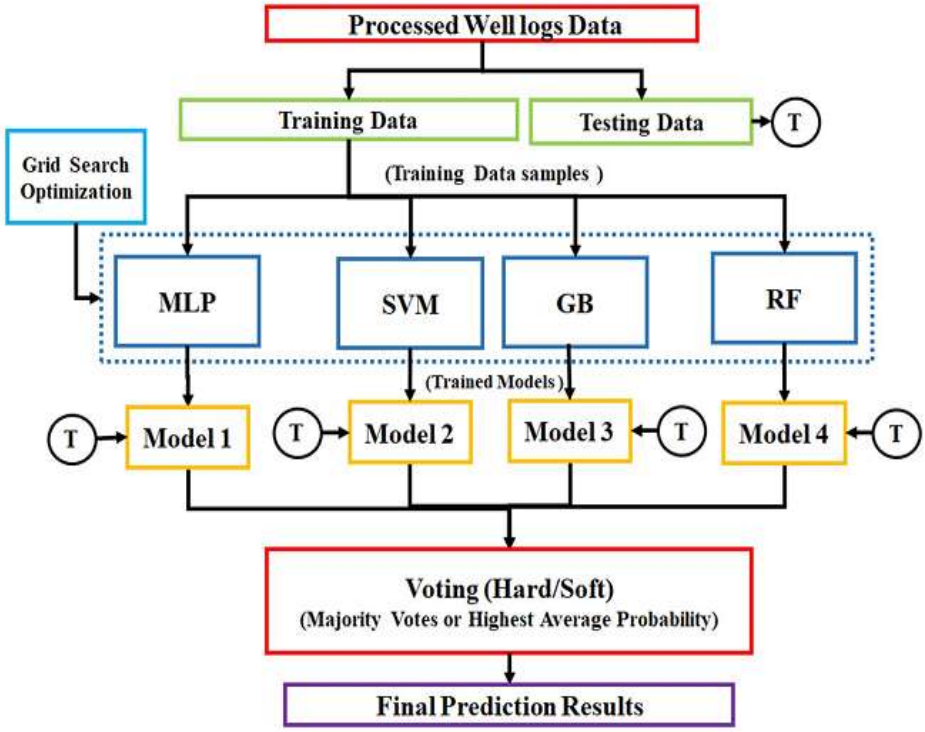


Figure 3.3 A theoretical framework of Voting ensemble utilized for the lithofacies recognition task.

Stacking ensemble can also be created by merging the decision of similar base classifiers having different parametric values. The selection of base and meta-classifier combination is always a matter of concern during the design of stacking ensemble architecture. It is also difficult to design the most suitable configuration of classifiers in large feature space. Wolpert (1992) proved that the stacking ensemble is good in reducing the generalization error by decreasing bias and variance error associated with data. Initially, input data are split into training and testing datasets. Further, the training

dataset is again split into K identical subsets similar to K-fold cross-validation technique. Base classifiers are trained on (K - 1) subsets, while the Kth subset is retained as a validation set. After training with (K - 1) subsets, base classifiers are individually tested with the Kth validation subset and also with the testing data. The outcomes of each base classifier with validation and test datasets will act as new training and testing data for meta-classifier. Moreover, the meta-classifier will be trained with the prediction outcomes of the validation set and the actual values of the target variable.

Algorithm

Let us assume input: $X = \{(x_i, y_i) \mid x_i \in \chi, y_i \in Y\}$

Output: Trained Stacking classifier

Step1: Learn first level classifier

For $t \leftarrow 1$ to T do

Learn a base classifier C_t based on X

Step 2: Construct new dataset from X

For $t \leftarrow 1$ to m do

Construct a new data set that contains $\{x_i^{new}, y_i\}$, where

$x_i^{new} = \{C_j(x_i) for\ j = 1\ to\ T\}$

Step 3: Learn the second level classifier

Learn a new classifier C^{new} based on newly constructed dataset.

Return $C(x) = C^{new}(c_1(x), c_2(x) ... c_T(x))$

3.3 Voting ensemble

Voting ensemble also combines the diverse supervised classifiers for the pattern recognition task. It offers flexibility in the combination strategies for combining the results of base classifiers. However, voting never uses any paradigm for the combination of outcomes from base classifiers as in the case of Stacking. Two combination schemes are applied for the unification of the judgments of base classifiers, namely majority vote rule (hard voting) and average predicted confidence probabilities (soft voting) to forecast the class of test data [64]. In hard voting, class labels of test samples are decided by the majority voting rule. The final classification of the test data sample is decided by the maximum number of votes received by each class. The soft

voting strategy allocates weights to the individual base classifier. It produces prediction probabilities during the testing phase for every test sample belonging to a particular class. Further, these probabilities are multiplied with the weights then averaged. Test data samples are finally allocated into that class that attains the highest average confidence probability. This strategy allocates data samples as argmax (argument of maxima) of the sum of assigned probabilities [64-66]. Figure 3.3 shows the theoretical framework of the Voting ensemble used for lithofacies recognition.

Algorithm

Let us assume input: $X = \{(x_i, y_i) \mid x_i \in \chi, y_i \in Y\}$
Output: Trained Voting classifier
Step1: Learn first level classifier
For $t \leftarrow 1$ to T do
Learn a base classifier C_t based on X
Step 2: Use same training dataset X
For $t \leftarrow 1$ to m do
Train new classifiers
Step 3: Combine the classifier's decision
Use combination strategies to final decision
Return $C(x) = C^{new}(c_1(x), c_2(x)...c_T(x))$

3.4 Data description

The well logs data were downloaded from the Kansas geological survey (KGS) website [76]. The digital "Las" files format contain 13,000 data from which 3425 data samples are mined related to nine lithofacies, namely dolomitic wackestone (DW) (1015), clay (CL) (320), dolomitic mudstone (DM) (240), dolomitic sandstone (DS) (455), siltstone (SS) (85), dolomitic packstone (DP) (265), carbonate mudstone (CM) (520), packstone (PS) (465) and wackestone (WS) (60). The above-said lithofacies are acknowledged as class labels for the classification of well logs data into their respective lithofacies. The downloaded "Las" files belong to Paradise A, Deforest, and Strahm wells existing in the Kansas field (Table 3.1 and Figure 3.4).

Table 3.1 The statistical description of input variables from well logs data.

S. No.	Well logs	Minimum	Maximum	Units
1.	Depth (DT)	434	3771	ft
2.	Sonic Porosity (SPOR)	2.774	133.212	Pu
3.	Density Porosity (DPOR)	0.652	32.118	pu
4.	Gamma Ray (GR)	16.781	414.152	API
5.	Neutron Porosity (NPOR)	0	44.143	pu
6.	Bulk Density Correction (RHOC)	0.01	0.296	gm/cc
7.	Deep Induction Resistivity (RILD)	1.905	65.158	Ohm.m
8.	Spontaneous Potential (SP)	-1.797	71.61	MV
9.	Deep Laterolog Resistivity (RLL3)	2.866	266.481	Ohm-m
10.	Caliper 1 (DCAL)	7.468	8.982	inch
11.	Micro Inverse Resistivity (MI)	0.609	41.818	Ohm-m
12.	Micro Normal Resistivity (MN)	0.138	44.097	Ohm-m
13	Caliper 2 (MCAL)	6.947	8.911	In
14.	Acoustic Transit Time 2 (TT1)	54.36	662.885	Usec/ft
15.	Medium Induction Resistivity (RILM)	2.031	140.706	Ohm-m
16.	Acoustic Transit Time 1 (DT)	51.522	235.961	Usec/ft

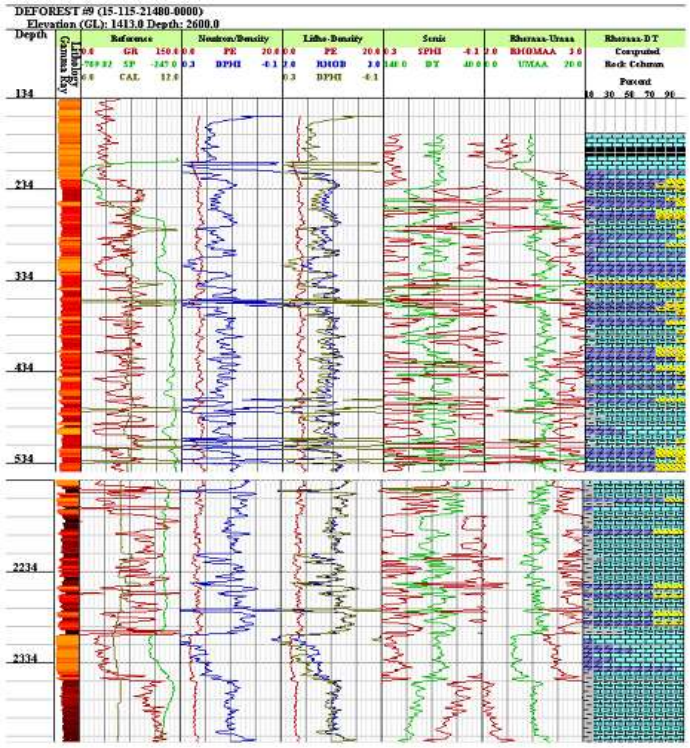


Figure 3.4 Well logs data of Paradise well existing in Kansas region of U.S.A.

3.5 Data-driven workflow for HEMs

Additional preprocessing steps have been added before training and testing of different machine learning models for the lithofacies recognition. Intelligent modeling proposed in this research work is broadly classified into three stages viz. (a) Preprocessing stage (b) Model development stage and (c) Post-processing stage. Figure 3.5 contains a graphical framework of HEMs used for lithofacies.

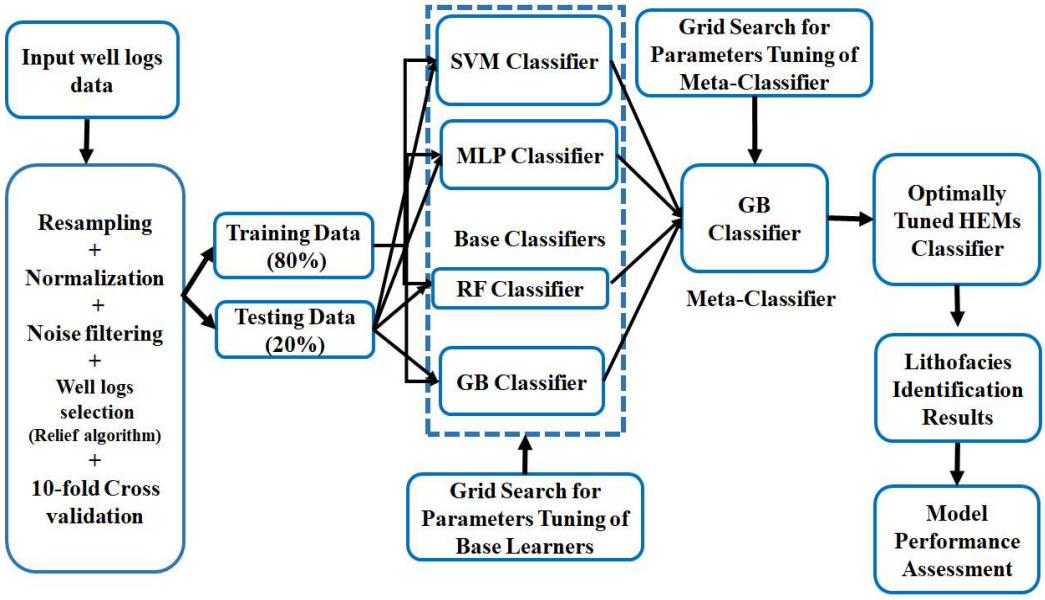


Figure 3.5 A generalized framework of HEMs for the identification of lithofacies.

In the preprocessing stage, the resampling of petroleum data was done to eliminate samples containing null, garbage, and missing values. After resampling, the input data were normalized to reduce the impact of larger values on the smaller values of predictor variables. Later, noise filtering of input well logs was done to minimize the effects of noise during the pattern recognition of lithofacies. Tewari et al. [77] studied the influence of noise levels on the classification performance of supervised classifiers and reported its damaging effects on the classification performance. Diaz et al. [78] also suggested that the preprocessing of petroleum data, viz. noise filtering, feature extraction, etc., before the pattern recognition task helps to improve the classification or

estimation accuracy of intelligent algorithms. Several denoising techniques are available in the petroleum and geophysics literature such as low pass filter, high pass filter, Savitzky-Golay filter, wavelets denoising, moving average, Gaussian, etc. High peaks of well logs data are considered as noise components that are generally eliminated using noise filters. Figure 3.6 shows denoising of four well logs using the SG filter technique.

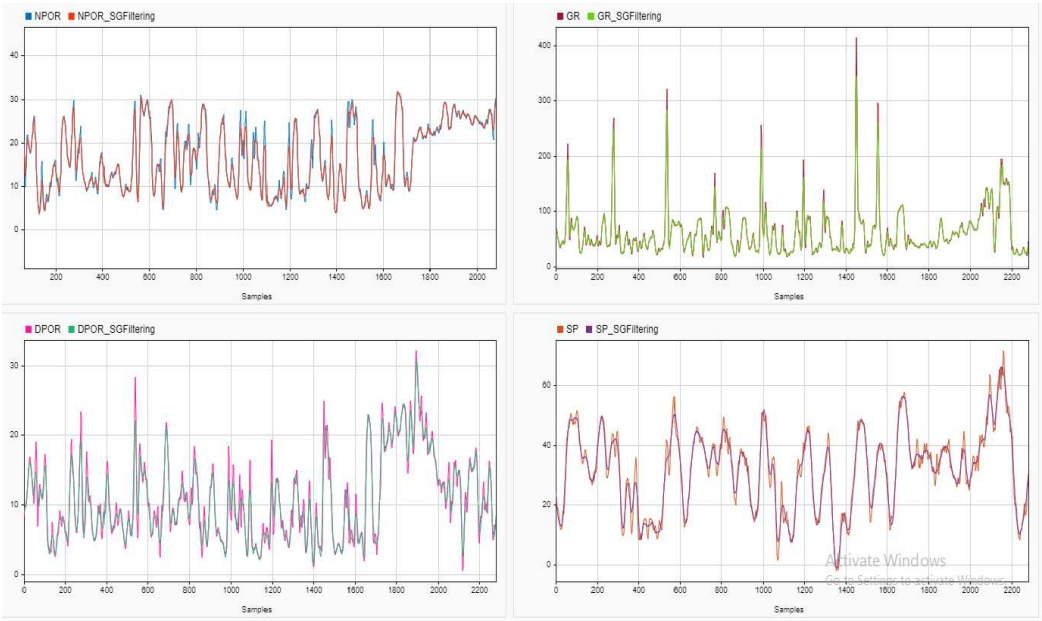


Figure 3.6 Denoising of four well logs using the SG filter technique.

After noise filtering, important data attributes were selected to decrease the dimensionality of data and eliminated redundant well logs. The high dimensionality of input logs data increases computational cost and time during pattern recognition of lithofacies. This can be reduced by the selection of important attributes from input well logs data and removal of the redundant ones. Several attribute selection paradigms are available in the literature such as a forest of tree-based attributes selection, Univariate feature selection, Relief algorithm, etc. Relief algorithm has been primarily applied to select the attributes. Figure 3.7 contains available well logs arranged according to their predictor important weights assigned by Relief algorithm for pattern recognition of lithofacies.

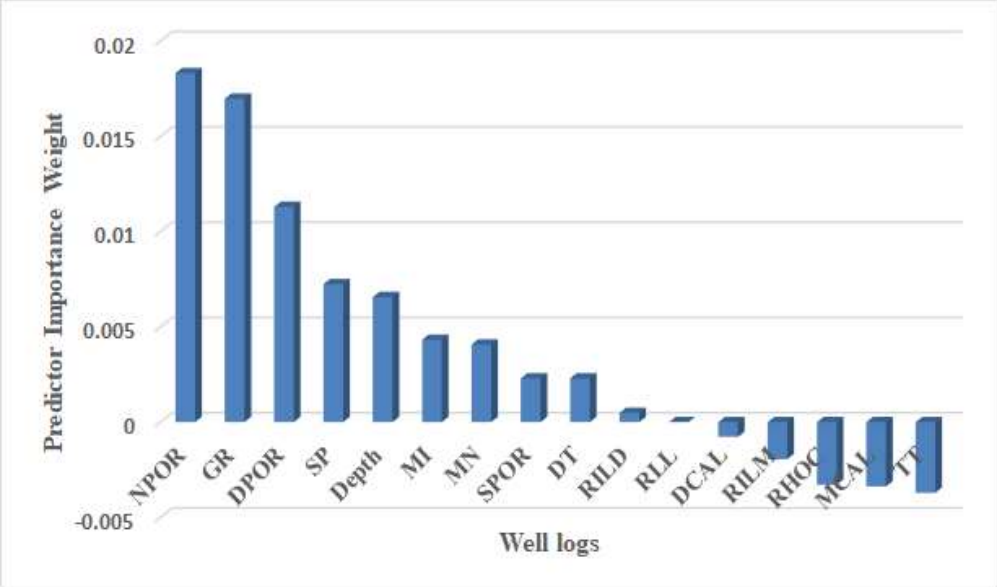


Figure 3.7 Available well logs arranged according to their predictor important weights assigned by Relief algorithm for pattern recognition of lithofacies.

The processed input petroleum data was further divided into training sets and testing sets using a cross-validation technique. There are three cross-validation techniques viz. k-fold, leave-one-out, and hold-out that are popular in the machine learning domain for the generation of training and testing datasets from input data. K fold cross-validation technique has been utilized in this research work for splitting the processed input data for training and testing of intelligent models (K=10). The 10-fold cross-validation (10-FCV) technique has been reported to have minimum variance error as compared to other cross-validation techniques [79]. Cross-validation helps to minimize the chances of overfitting and underfitting of models [79]. The input well logs data were randomly partitioned into ten random sets of training and testing datasets during 10-FCV. Intelligent models are trained and tested on these ten datasets for learning the patterns hidden within the well logs and to predict the respective lithofacies. All the ten test outcomes of the learners are averaged to deliver overall final classification accuracy.

During the training phase, the model parameters are determined and training errors are minimized with each iteration to obtain the best possible performance of intelligent models. The optimum value of the model's parameters is essential to be determined during the training phase so that these models can be generalized for unseen data also. The model parameters were optimized using several tuning algorithms such as Grid search, Random search, Trust region, Simplex methods, Genetic algorithm, Particle swarm optimization, etc. These tuning algorithms have their advantages and limitations.

Machine learning models always have the possibility of getting overfitted or under-fitted during pattern recognition. A separate validation score test was conducted to examine the overfitting and underfitting tendency of intelligent models. A validation curve was utilized to shrink the search range for various parameters. It clearly illustrates the overfitting and underfitting regions of the respective classifiers with a specific parameter variation. In an underfitting state of the intelligent model, training and validation scores are normally recorded to be low, whereas overfitting states result in high training and low validation scores. The parameter search range is primarily comprised of upper and lower constraints of a stable region. In a stable region, no dramatic variation in training and validation scores takes place. However, the model still needs an optimization algorithm that explores within the stable search range to find the best possible value of the model parameters. The search range and optimum values for various model parameters are depicted in Table 3.2. Figures 3.8 and 3.9 show the validation curves of GB and RF classifiers for four important parameters, namely Estimators, Min_samples_split, Max_depth, and Min_samples_leaf.

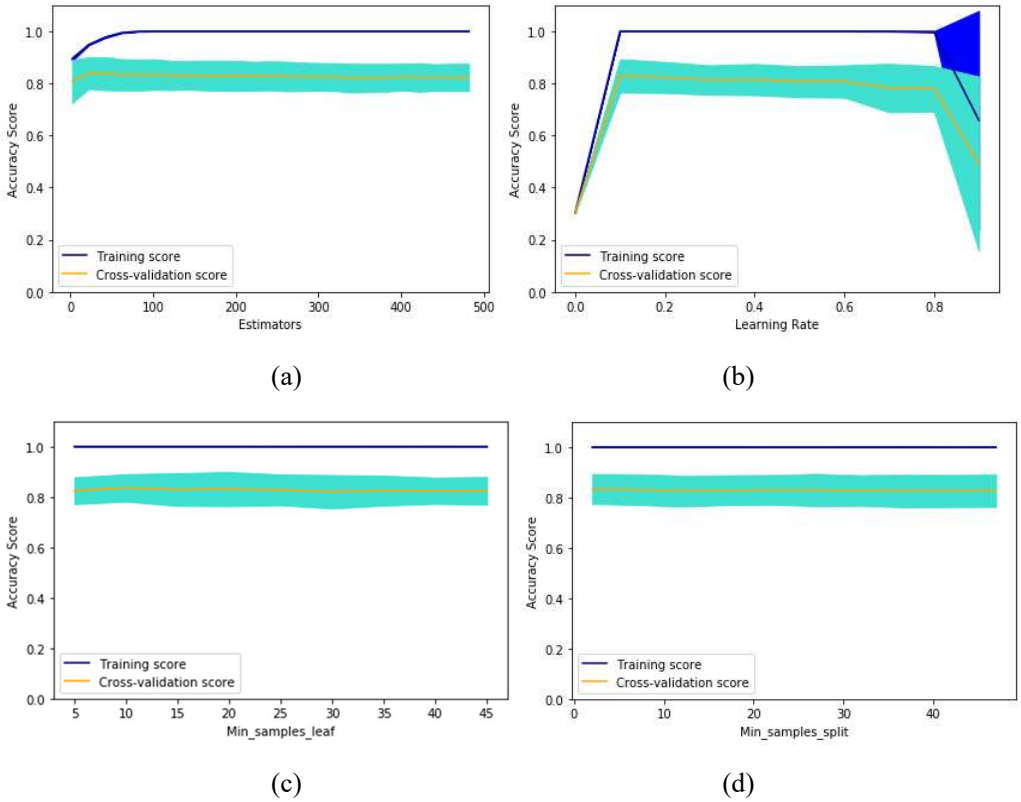


Figure 3.8 Validation curve of GB classifier to identify stable search range for four primary model variables (a) number of estimators (b) learning rate (c) minimum samples required at leaf node and (d) minimum samples required for splitting the internal node.

Figure 3.10 a, b shows the validation curves of SVM for regularization constant (C) and gamma (ϒ) versus accuracy score. The model with optimum parameters' values is saved to classify unseen new data samples. The optimally tuned machine learning model was also tested on unseen data samples to evaluate its generalizability. The performance of optimally tuned intelligent models is evaluated on testing data using statistical parameters viz. coefficient of correlation, root mean square error, mean absolute error, recall, precision, F1 score, P-value, T-value, etc. Figure 3.5 shows a generalized conceptual workflow for the heterogeneous ensemble methods to recognize the subsurface lithofacies.

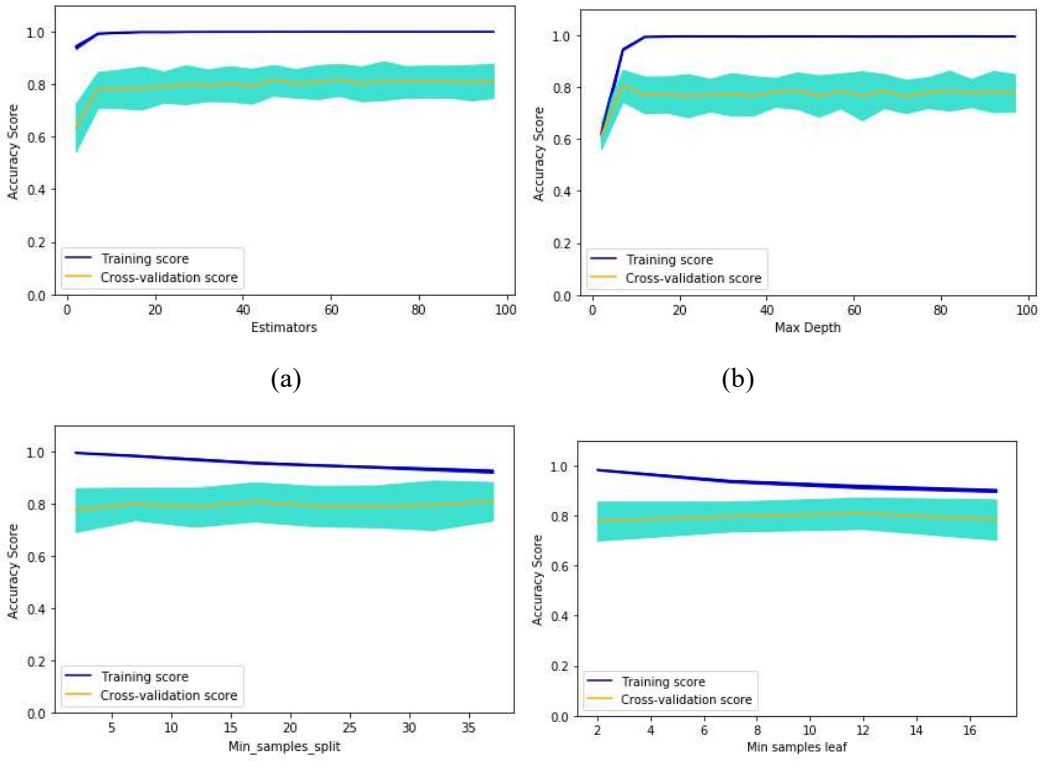


Fig. 3.9 Validation curve for RF classifier to identify stable search range for four primary model parameters (a) number of estimators (b) maximum depth of tree (c) minimum samples required for splitting the internal node. (d) minimum samples are required at the leaf node.

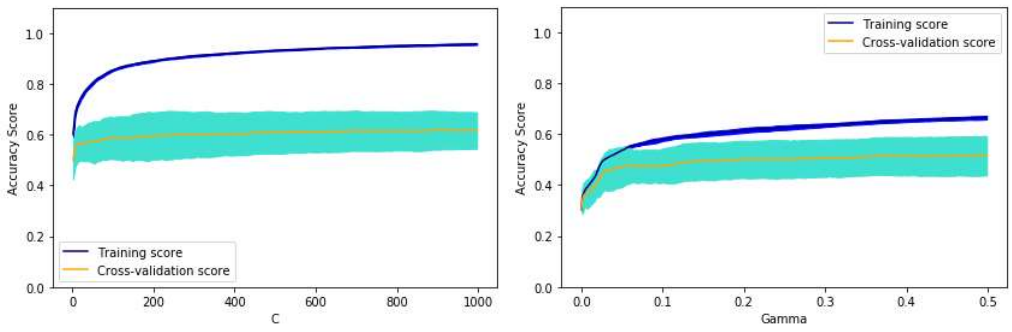


Fig. 3.10 Validation curve for SVM classifier to identify stable search range for two primary model parameters (a) penalty cost parameters for misclassified error samples (b) kernel coefficient of RBF.

Table 3.2 The search range and optimized values of model parameters obtained through Grid search algorithm on input well logs data.

Classifiers	Model Parameters	Search Range	Settings
MLP	Activation function	Identity, logistic, tanh, relu	relu
	Solver	Lbfgs, Sgd, Adam,	Adam
	alpha	0.00001-0.1	0.0001
	Learning_rate_init	0.0001-0.1	0.001
	Learning _rate	Constant, invscaling, and adaptive	constant
	Max_iteration	10-400	200
SVM	C	1-1000	200
	gamma	0.001-1	4
	Kernel	RBF, polynomial, linear	RBF
RF	Estimators	10-1000	100
	Max_depth	0-infinity	10
	Min_samples_split	2-10	2
	Min_samples_leaf	1-10	1
GB	Estimators	10-400	250
	Max_depth	0-infinity	None
	Min_samples_split	2-10	2
	Min_samples_leaf	1-10	1
Voting	Base classifiers	Any supervised classifiers and HoEMs	RF, GB, MLP, SVM
	Weights	0.1-1	1,1,0.5,0.5,1
	Voting	Soft/hard	hard
	Combining strategies	Simple/weighted majority voting	Weighted majority voting rule
Stacking	Base classifiers	Any supervised classifiers & HoEMs	RF, GB, MLP,& SVM
	Meta Classifier	RF, GB, MLP & SVM	GB

3.6 Results and discussion

This section discusses the experimental results obtained during the recognition of nine mudstone lithofacies belonging to Kansas oil and gas fields. The performance of Stacking and Voting ensembles was compared with four popular classifiers, namely GB [80], RF [81], SVM [82], and MLP [30, 31]. Stacking and Voting are two HEMs that were implemented to predict complex lithofacies. Figure 3.5 depicts a generalized conceptual workflow for HEMs to predict the lithofacies of the formations. The performance of HEMs was tested by two separate data-driven experiments for the prediction of lithofacies. In the first experiment, 10-FCV was performed to split the input data samples into training and testing subsets so that generalized prediction outcomes can be obtained. The performance of each classifier has been reported in the form of precision, recall, and F1-score for individual lithofacies. Tables 3.3, 3.4, and 3.5 show precision, recall, and F1-score acquired by HEMs and base classifiers for each lithofacies during 10-FCV. Overall, the classification performance of Stacking has been found higher than all the other classifiers considered in this study. The voting ensemble has secured second place in terms of overall classification performance as shown in Tables 3.3, 3.4, and 3.5. GB and RF classifiers have given similar performance scores for the identification of mudstone lithofacies as shown in Table 3.4. SVM classifier has also maintained good classification performance during 10-FCV for all the lithofacies. MLP becomes the worst performing classifier in terms of evaluation metrics, viz. average precision, average recall, and average F1-score, as shown in Table 3.6. It is also found that Voting, GB, RF, and MLP have fluctuations in their performances for smaller classes, namely SS and WS. However, Stacking and SVM classifiers are successful in maintaining their performances even for smaller classes as shown in Tables 3.4 and 3.6. Smaller classes have contributed a lesser number of data samples

during the training and testing of machine learning models. These classes also represent facies having thin layers that are difficult to identify using conventional well logs interpretation techniques. WS and SS facies are intentionally included with lesser data samples to magnify data imbalance conditions that make classification harder even for strong classifiers such as GB, RF, Voting, etc. Voting and Stacking ensembles have utilized the same base classifiers for the classification of facies; however, Stacking performed better than Voting due to the presence of a meta-classifier for combining the out-comes of base classifiers.

In the second experiment, a separate test was also performed with randomly selected training and testing data samples without 10-FCV. Table 3.6 depicts the overall performance of every classifier utilized in this study with processed input data split into (80%) training subset and (20%) testing subset. The testing accuracy for individual lithofacies is depicted diagonally in confusion matrices. Training and testing classification accuracies of HEMs are found higher than all other machine learning models utilized in this study as shown in Table 3.6. Naturally, subsurface layers exist inside the formations with uneven thickness and patterns. Therefore, uneven data distribution has been considered to represent real-field conditions. This also provides us an opportunity to understand the worst to best possible performance of machine learning classifiers for individual layers during imbalanced data conditions. The uneven data distribution is in particular chosen for this study to understand the effect of data imbalance conditions. Facies having lesser data points such as WS, SS, etc., are designed for magnifying data imbalance effects. Stacking ensemble has shown great potential to extract lithofacies information from well logs data even for smaller classes due to the presence of meta-classifier in its architecture. The Stacking ensemble has scored 83% accuracy for WS and 94% for SS which are challenging smaller lithofacies.

This research work is specially designed to evaluate worst- to best-case scenarios for lithofacies modeling. Layer wise classification accuracy of HEMs along with its base classifiers can be summarized as follows: (a) Stacking (67.9–95.8%), (b) Voting (58.3–94.1%), (c) GB (58.3–94.1%), (d) RF (41.7–94.6%), (e) SVM (58.3–94.1%) and (f) MLP (0.0–88.7%).

In the case of data with high imbalance conditions, performance indicators (viz. accuracy, precision, recall, and F1-score) may give misleading results. Therefore, the testing performance of each classification model is also evaluated using the MCC parameter which is unaffected by data imbalance issues as shown in Table 3.6. It is found that MCC scores of applied models also justify their performance as shown in Table 3.6. DP has emerged as one of the most challenging subsurface rock layers during the testing phase. In this study, most of the time, all the classifiers have identified data samples related to DP as CM. It may be possible that the presence of calcareous mud inside DP has confused base classifiers with CM. This uncertainty may be removed by increasing the number of training data samples that will help in learning discriminatory features between similar layers (Figure 3.11).

Table 3.3 The performance of HEMs after 10 fold cross-validation for lithofacies classification.

Staking ensemble				**Voting ensemble**			
Facies	**Precision**	**Recall**	**F1-score**	**Facies**	**Precision**	**Recall**	**F1-score**
CL	0.77	0.86	0.81	CL	0.80	0.80	0.80
CM	0.78	0.89	0.83	CM	0.76	0.87	0.81
DM	0.92	0.96	0.94	DM	0.98	0.88	0.92
DP	0.80	0.68	0.73	DP	0.86	0.60	0.71
DS	0.87	0.85	0.86	DS	0.74	0.92	0.82
DW	0.96	0.93	0.95	DW	0.95	0.94	0.95
PS	0.93	0.83	0.87	PS	0.94	0.83	0.88
SS	0.76	0.94	0.84	SS	0.8	0.94	0.86
WS	1.00	0.83	0.91	WS	1.00	0.58	0.74
Micro avg.	0.87	0.87	0.87	Micro avg.	0.86	0.86	0.86
Macro avg.	0.87	0.86	0.86	Macro avg.	0.87	0.82	0.83
Weighted avg.	0.88	0.87	0.87	Weighted avg.	0.87	0.86	0.86

Table 3.4 The performance of GB classifier and RF ensembles after 10 fold cross-validation for lithofacies classification.

GB classifier				**RF classifier**			
Facies	**Precision**	**Recall**	**F1-score**	**Facies**	**Precision**	**Recall**	**F1-score**
CL	0.81	0.75	0.78	CL	0.79	0.75	0.77
CM	0.74	0.87	0.8	CM	0.76	0.87	0.81
DM	0.94	0.94	0.94	DM	0.98	0.85	0.91
DP	0.85	0.64	0.73	DP	0.87	0.62	0.73
DS	0.75	0.89	0.81	DS	0.73	0.9	0.81
DW	0.93	0.95	0.94	DW	0.94	0.95	0.94
PS	0.95	0.78	0.86	PS	0.89	0.80	0.84
SS	0.78	0.82	0.8	WS	0.8	0.94	0.86
WS	0.88	0.58	0.7	SS	0.83	0.42	0.56
Micro avg.	0.85	0.85	0.85	Micro avg.	0.85	0.85	0.85
Macro avg.	0.85	0.8	0.82	Macro avg.	0.84	0.79	0.8
Weighted avg.	0.86	0.85	0.85	Weighted avg.	0.86	0.85	0.85

Table 3.5 The performance of SVM and MLP classifiers after 10 fold cross-validation for lithofacies classification.

SVM classifier				MLP classifier			
Facies	**Precision**	**Recall**	**F1-score**	**Facies**	**Precision**	**Recall**	**F1-score**
CL	0.79	0.81	0.80	CL	0.38	0.33	0.35
CM	0.80	0.90	0.85	CM	0.57	0.78	0.66
DM	0.94	0.65	0.77	DM	0.70	0.67	0.68
DP	0.83	0.74	0.78	DP	0.56	0.19	0.28
DS	0.70	0.89	0.79	DS	0.52	0.73	0.6
DW	0.97	0.94	0.95	DW	0.94	0.89	0.91
PS	0.97	0.84	0.90	PS	0.78	0.78	0.78
SS	0.8	0.94	0.86	SS	0.67	0.47	0.55
WS	0.98	0.92	0.96	WS	0.0	0.0	0.0
Micro avg.	0.86	0.86	0.86	Micro avg.	0.69	0.69	0.69
Macro avg.	0.87	0.85	0.85	Macro avg.	0.57	0.54	0.54
Weighted avg.	0.86	0.86	0.87	Weighted avg.	0.68	0.69	0.67

Table 3.6 Classification accuracy of six machine learning models depicted on training (80%) and testing (20%) datasets for lithofacies classification.

Classifier	**Base classifier**	**Training Accuracy**	**Testing Accuracy**	**Avg. Precision**	**Avg. Recall**	**F1-score**
MLP	---	0.6892	0.6554	0.67	0.67	0.63
SVM	---	0.9170	0.8403	0.86	0.87	0.87
GB	Decision tree	0.9105	0.8554	0.86	0.86	0.86
RF	Decision tree	0.9012	0.8481	0.86	0.85	0.85
Voting	MLP, SVM, RF, GB	0.9230	0.8657	0.87	0.87	0.87
Stacking	MLP, SVM, RF, GB	0.9278	0.8832	0.89	0.88	0.88

3.7 Summary

A rigorous facies-wise comparison has been made between Stacking and Voting ensembles for the detection and identification of lithofacies. Stacking has shown nearly 4% and 2% improvement in test accuracy as compared to SVM and RF. Four popular machine learning algorithms have been combined in HEMs as base classifiers to provide more accurate and generalized results. In this study, HEMs have combined MLP, SVM, GB, and RF classifiers to achieve better classification accuracy than their individual performances. The individual performance of the abovementioned classifiers has been evaluated using Kansas oil and field data with proper parameter optimization in their stable search ranges. The Stacking ensemble has shown great potential to extract lithofacies information from well logs data. The training and testing classification accuracies of HEMs have been found highest among the other classifiers used in this study. DP layer is found to be the most challenging facies among all the nine target lithofacies. The Stacking ensemble has given the highest individual identification accuracy for all the layers of lithofacies. Prediction accuracy of individual facies ranges from 67.9 to 95.8% (worst to best possible testing accuracy), and maximum overall accuracy is (training = 92.78% and testing = 88.32%) obtained for Stacking ensemble.

Chapter 4

Intelligent Drilling of Oil and Gas Wells using Response Surface Methodology and Artificial Bee Colony

4.1 Introduction

The demand for hydrocarbons has been increased rapidly in the modern era. To meet the ever-growing oil and gas demand, unconventional reservoirs such as tight oil and gas reservoirs, shale gas, ultra-deep reservoirs, etc. are needed to be drilled in more challenging geological lithological conditions. These difficult geological rock formations require newer technological advancements for successful drilling operations. The use of the conventional drilling approach may result in higher overall drilling costs due to human errors. Thus, drilling parameters are needed to be optimized during drilling operations to achieve maximum efficiency.

Drill bits and ROP are the important drilling parameters that are needed to be optimized for the success of drilling operations due to their large impact on operational efficacy and cost. While optimizing the cost of drilling operations, selection of the most suitable drill bit types is one of the main concerns of driller as all other drilling parameters directly or indirectly rely on the drill bit, although the cost of bits only 5% of the total operational cost [83]. Selecting the right bit types for drilling operations is still one of the most challenging tasks due to its dependency on various factors. The performance of drill bits depends on various aspects such as bit design parameters, formation properties, and other operational field parameters [84, 85]. The concept of drilling optimization is built on the usage of earlier drilled well data for optimizing operational variables for drilling the next well with minimum cost and time [83]. The drilling variables are gradually adjusted to achieve their best possible effective optimum

values to decrease operational cost and time. Drill bits are mostly selected based on the knowledge of bit data of previously drilled wells and from the types of bits available to driller from manufactures. Driller selects the drill bits for new well depending upon his experience while drilling the earlier wells [83]. Drilling operations are also affected by various controllable and uncountable factors which involves a high risk of human error that may increase the cost of overall drilling operations [83]. Therefore, various empirical and data-driven models have been developed based on the known relationships between drilling variables to select the most suitable bit types.

Recently, data-driven intelligent models have been utilized to find suitable types of drill bits. These models are reported to be more accurate as they learn from previous well data, defying traditional methods for selecting the appropriate drill bit [24]. Bilgesu et al. [25] used Artificial neural networks (ANN) for the prediction of drill bit types for drilling target geological formations. Yilmaz et al. [26] trained the ANN model using previously drilled wells offset data and predicted the drill bits types for the development wells required to be drilled internally and externally of the same field. They also tested the trained ANN model for the prediction of drill bit types for the development wells that were required to be drilled in an adjoining field. Bahari et al. [15] utilized a Genetic algorithm (GA) for the accurate computation of constants for the Bourgoyne-Young ROP model. Edalatkhah et al. [27] also selected the suitable drill bit types using ANN and GA for South Pars Field wells. Momeni et al. [28] applied ANN for the estimation of drilling ROP and bit types [14]. Momeni et al. [29] combined ANN and Genetic algorithm (GA) for drill bit selection based on optimal ROP. They selected the drill bit types based on the optimum values of ROP and drilling variables. Abbas et al. [18] also supported the notion of drill bit selection depending upon optimum values of ROP using

ANN and GA. Here ANN was primarily utilized for the development of the objective function and GA for optimization of the ROP objective function to select the drill bit.

Various researchers have suggested the selection of drill bits should be performed based on the optimum values of ROP. This condition results in the development of an unconstrained bounded optimization problem where a function of ROP is required to be defined using drilling variables. However, the exact relationship between ROP and drilling variables is unknown and undefined that makes optimization of ROP a difficult task. According to Kolmogorov's theorem, multilayer feedforward perceptron (MLP) neural architecture can be utilized to define any continuous function in its approximation form [86]. The approximation function (objective function) requires an activation function and input variables that are predefined during the training of the MLP neural network. Three-layered MLP architecture can be expanded in a mathematical form with connection weights and bias of neurons that will act as coefficients of approximation function. This technique helps to solve real-field complex optimization problems, especially where the association between input and target variables is unknown such as bit selection based on optimum ROP values. In the case of complex approximation function, paradigms such as Ant colony, swarm optimization, GA, etc. can be implemented to retort the optimization problem as stated in the literature [87]. However, researchers reported several issues with ANN such as overfitting, underfitting, stuck up in local minima/maxima, lack of proper guidelines for the selected network architecture, [88]. This also opens the opportunity to investigate other techniques that can generate approximation functions to optimize ROP values for drill bit selection.

In this chapter, an alternative solution has been proposed for drill bit selection utilizing Response surface methodology (RSM) and Artificial bee colony (ABC) combination. RSM has been implemented to generate the objective function for ROP due to its strong data fitting characteristic. Further, the generated ROP function is optimized through ABC to acquire the optimal drilling variables and drill bit types for target geological formations. ABC is strategically designed to locate the global optimum value of any given objective function more efficiently in the high dimensional data space. The suggested approach has been compared with the earlier drill bit selection model based on ANN and GA combination. ANN tends to get stuck up in local minima consequently GA sometimes fails to converge when data become too much complex [87,88]. Therefore, the reliability of the ANN and GA combination is a major concern for drill bit selection. RSM has been reported to be a reliable and popular technique for solving optimization problems in various engineering domains [90,91]. Researchers have also reported that ABC is a superior evolutionary optimization paradigm that has outperformed GA in certain applications with faster convergence and lesser iterations [92,93]. This research work investigates the performance of the RSM and ABC combination as an alternative solution to substitute the ANN and GA combination model reported for optimum drill bit selection.

4.2 Materials and methods

In this study, RSM and ABC have been utilized to develop an alternative intelligent data-driven approach for the selection of suitable bit types. The performance of the existing ANN-based drill bit selection model has also been compared with the proposed approach to understand its pros and cons. The intelligent paradigms applied in this study are briefly explained below.

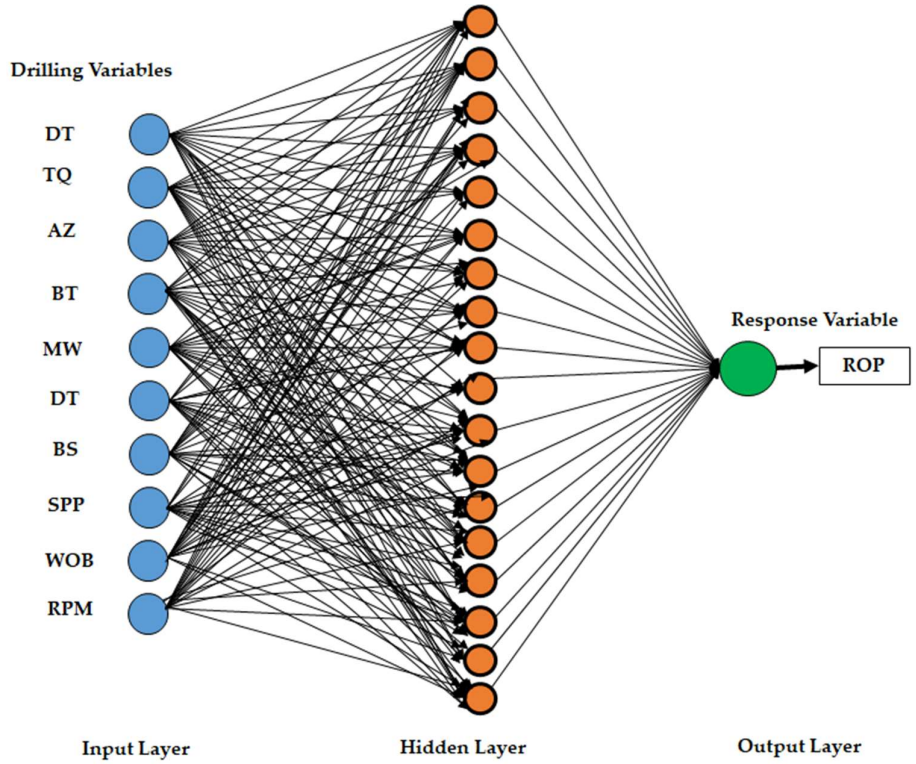


Figure 4.1 The architecture of the ANN investigated in this study.

4.2.1 *Artificial Neural Networks*

ANNs are a nature-inspired intelligent paradigm that is designed based on human brain cells. It is constituted of multiple information processing units known as nodes. The interconnected nodes combine to form a layer and layers combine to form neural networks. Neural nodes are also recognized as neuron units. Each neuron connection has been associated with weights that are attuned during the training to generate approximation function to minimize error for classification or estimation tasks. Generally, ANN comprises multilayers structure viz., input, hidden, and output layers [86,87]. However, hidden layers may vary in number depending on the complexity of training data. Initially, the field data are provided to the input layer which further transmits the raw data to hidden layers for their processing. The results acquired after processing in hidden layers are directed to the output layer where predicted results are

compared with actual target values. The deviation of prediction values from actual targets is provided as feedback to the model for updating associated weights and biases. The number of neurons and hidden layers may vary according to the complexity of problems and data types. ANN is primarily developed for handling classification and regression tasks, however, they are also applied for solving optimization problems. Figure 4.1 depicts the architecture of the ANN utilized in this study.

Kolmogorov's theorem stated that multilayer feedforward perceptron (MLP) neural architecture can be utilized to define any continuous function in the form of an approximation function [86, 87,94]. There exist two stages in the MLP network namely, the learning stage and the prediction stage. Several neural network parameters are predetermined to define neural networks such as the number of neurons, number of layers, propagation rules, connections between neurons, activation function, learning rate, etc. The propagation rule in MLP is the weighted sum of inputs which are given below.

$$\sum_{m=1}^{M} w_{mn} x_m(t) + \beta_n \tag{4.3}$$

where w_{mn} is the connection weights associated with neuron m in the input layer and neuron n in the hidden layer, x_m is the outcome from neuron m in the input layer where M is input layer neurons, t is the associated patterns and β_n neuron bias. The activation function will be multiplied with equation (4.3) to decide whether neurons should be activated or not. There are several popular activation functions used in neural networks such as sigmoid function, hyperbolic tangent function, Softmax function, Softsign, Rectified linear unit, Exponential linear units, etc. The outcome from the K_{th} neuron with activation function existing in the input layer is given below.

$$Y_n = f^a(\sum_{m=1}^{M} w_{mn} x_m(t) + \beta_n) \tag{4.4}$$

Applying the propagation rule twice due to the three-layer architecture of MLP to transmit the values from the input to the output layer. Neurons in the output layer are considered as N. The result of the K_{th} output neuron is presented below.

$$Y_k = (\sum_{n=1}^{N} w_{nk} Y_m(t) + \beta_k) \tag{4.5}$$

Substituting equation (4.4) in equation (4.5), the outcome of the K_{th} neuron can be rewritten as given below.

$$Y_k = \sum_{n=1}^{N} w_{nk} \left(f^a(\sum_{m=1}^{M} w_{mn} x_m(t) + \beta_n) \right) + \beta_k \tag{4.6}$$

Equation (4.6) of MLP is utilized to approximate the objective function in optimization as given in equation (4.7).

$$f(\mathrm{x}_1, \mathrm{x}_2, \ldots \mathrm{x}_M) = \sum_{n=1}^{N} w_{nk} \left(f^a(\sum_{m=1}^{M} w_{mn} x_m(t) + \beta_n) \right) + \beta_k \tag{4.7}$$

$$\text{Constraints } C_1(\mathrm{x}_1, \mathrm{x}_2, \ldots\ldots \mathrm{x}_M) \leq 0$$

$$C_n(\mathrm{x}_1, \mathrm{x}_2, \ldots\ldots \mathrm{x}_M) \leq 0$$

Equation (4.7) acts as the objective function and constraints for optimization problems where the relationship between input and response variable is undefined. In the case of complex approximation functions, algorithms such as Ant colony, Particle swarm optimization, ABC, GA, etc. can be applied to retort the optimization problem.

4.2.2. *Response surface methodology*

RSM is a set of statistical techniques that are quite helpful in optimizing, developing, and improving processes and useful in analyzing and modeling numerous problems in engineering [95,96]. RSM is particularly useful in real-world situations where various

variables affect the performance, quality, and output of the desired process as in the case of drilling operations [96]. The target variable (Y) is termed as the response variable while input variables are known as independent variables. An appropriate relationship can be identified between input and response variables using RSM. RSM helps to develop the correct approximate mathematical function which satisfies a suitable relationship between the objective function and test factor group [96]. Interaction between the input variables can also be included in the response surface equation. The generalized relationship can be developed as given below.

$$y = f(\mathrm{X_n}) + e\sqrt{2} \tag{4.8}$$

where, f is the unknown exact response function which may be complex. 'e' is the error due to unaccountable factors, such as noise, interaction effects, etc., that influence the response or output but are never included in the equation (4.8). Let 'e' be the statistical error which is distributed normally with variance σ^2, and zero mean. Then, the error equation can be written as given below.

$$Error[Y] = \mu = Error[f(\mathrm{X}_n)] + Error[e] = f(\mathrm{x}_1, \mathrm{x}_2, \ldots \mathrm{x}_n) \tag{4.9}$$

where, $x_1, x_2, \ldots x_n$ are known as natural variables in their natural units. Further, these variables are changed into coded variables that are dimensionless and have zero mean and alike standard deviation. The coded form of equation (4.9) can be written as given below.

$$\mu = f(x_1, x_2, \ldots x_n) \tag{4.10}$$

Here, function f is known and undefined. Thus, a suitable approximation function is required to be generated for modeling purposes. In RSM, the first or second-order polynomial equation is primarily generated as approximation functions in place of the actual response function. The second-order model is popularly applied for modeling

various process operations due to its flexibility, diverse functional form, efficient approximation, and model coefficients that can be easily estimated through the least square estimation technique. A second-order response surface equation, based on Talyor series expansion, can be used as given below.

$$y_T = \beta_0 + i = \sum_{i=1}^{n} \beta_i x_i + \sum_{i}^{n} \beta_{ii} x_i^2 + \sum \sum_{i<j=2}^{n} \beta_{ij} x_i x_j + \varepsilon \tag{4.11}$$

where x_i and x_j represent the input parameters, b_0 is the constant of the regression equation, Y is the predicted ROP response, β_i is the linear coefficients, β_{ij} are interaction coefficients, β_{ii} are the coefficients of the square terms, and ε is the fitting error in the equation. To generate a response surface, Stepwise regression methods are utilized to reduce the computational burden.

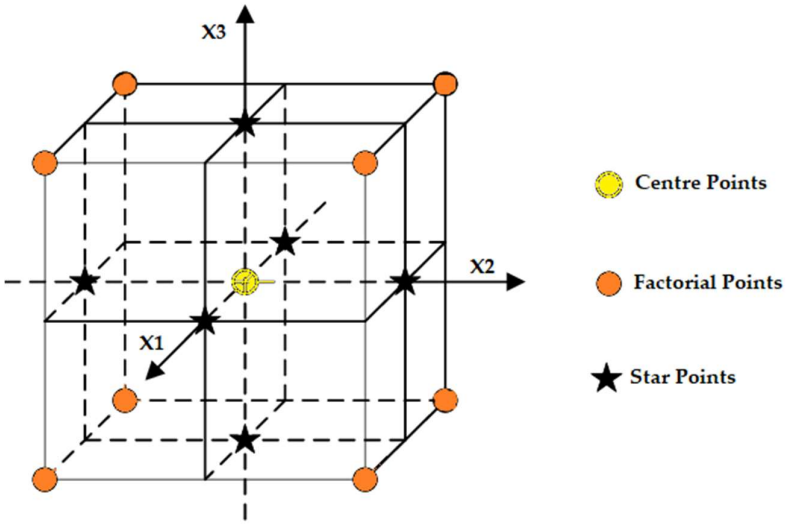


Figure 4.2 The layout of face-centered design (alpha=1) for CCD.

Here, center composite design (CCD) has been considered for developing a second-order approximation function for ROP as shown in equation (11). The input variables are converted into coded variables for fitting the data in CCD between two levels [-1, 1]. CCD contains the factorial point, central point, and axial points which are mostly developed through sequential experimentation as shown in Figure 4.2. Ten factors and a

full factorial design were utilized for the development of the response function. The range of design factors was assigned according to the range of field variables. The interaction, quadratic and linear coefficients were estimated through the least square regression. Further, the importance of each term was determined through an analysis of variance (ANOVA) test whereas redundant terms were eliminated. Equation (4.11) will act as an ROP objective function (approximation function) for the target geological formations containing bit information as an input variable similar to ANN. Further, this equation will be optimized using an Artificial bee colony to find the optimum value of ROP along with its input variables (control variables) including BT. A comprehensive explanation of RSM is available in the cited literature [95,96].

4.2.3. *Artificial bee colony*

Karaboga [97] proposed an Artificial bee colony (ABC) paradigm based on the natural behavior of bees when they search for flowers with nectar. There are three variations of honey bees generally present in natural hives namely, onlookers, employed bees, and scouts. Every bee has an assigned duty that is required to perform. The scout bees perform a random hunt for the flowers having nectar in their nearby environment and remember the location of the flower inside their internal memory [97,93]. This means scouts examine local search feature space for optimal solutions and remember it. After returning to their hive, scouts exchange information about flower locations with other bees using the waggle dance technique [93]. After the waggle dance, employed bees begin their exploration for the nectar having flowers depending upon the information achieved from scouts bees. The employed bees extract the nectar from the target flowers which are known as food sources [93]. Only one employed bee will be assigned to a single flower to exploit their nectar. Thus, each available food source contains an assigned employed bee that creates an initial solution [93]. The value of each solution is

calculated to understand its' significance. A new response is generated for each problem solution using the relationship as given below.

$$B_{i,j} = S_{i,j} + \delta_{i,j}(\mathrm{S}_{i,j} - S_{k,j}) \tag{4.12}$$

$$i \in \{1,2,...IA\}, \quad j \in \{1,2,...\mathrm{O}\}, \; k \in \{1,2,...IA\} \; and \; k \neq i$$

Where, $S_{i,j}$ is parameter j obtained from response *i*, $B_{i,j}$ is the parameter *j* in the new response where *i* represents the number of solutions, $\delta_{i,j}$ is an arbitrary number existing between [-1,1], *k* is a random number from a single answer to the problem, *IA* signifies initial solutions to the given problem, and O is the total number of parameters required during optimization. After calculating a new answer for each solution, they are compared with the previous answer. If the difference is found to be higher between the current and earlier answer, only then it will be accepted otherwise rejected [93]. The step length is adaptively decreased according to the difference between current and previous answers as the search reaches closer to the optimal solution. The waiting onlooker bees in the hive choose the best source depending upon the dance of employed bees. This helps in identifying the global solution existing in the search space along with local solutions [93]. Further, employed bees of an abandoned food site start to behave like scout bees and hunt for newer sources of food. The probability of selection of source through an onlooker bee can be expressed as given below.

$$\mathrm{Pro}_i = \frac{fitness_i}{\sum_{n=1}^{FS} fitness_n} \tag{4.13}$$

where $fitness_i$ represents the fitness of the solution *i* examine via employed bee depending upon nectar quantity at location *i*, and *FS* is the number of the food source. After predefined iterations, there is no improvement in answer value using equation (12), then the employed bees convert into scouts and randomly begin the search for a

newer source of food. ABC algorithm has been utilized for solving various engineering problems such as ROP optimization [98], oil and gas well placement optimization [99], over break prediction in the tunnel [100], etc. A detailed description of ABC optimization can be found in other references [101,102]. Figure 4.3 depicts the flowchart of the Artificial bee colony paradigm that was originally proposed by Karaboga [97]. Figure 4.4 shows the generalized schema of the proposed approach of drill bit selection.

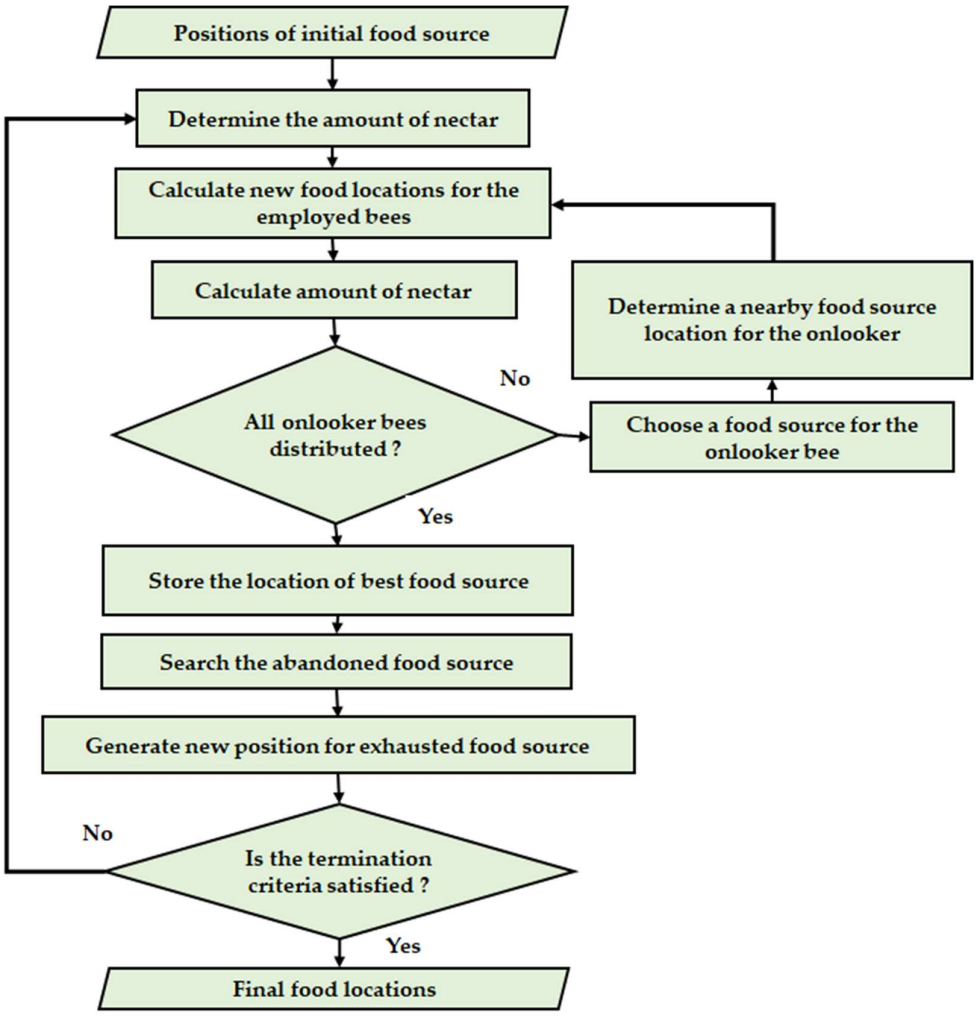


Figure 4.3. Flowchart of the artificial bee colony paradigm [96].

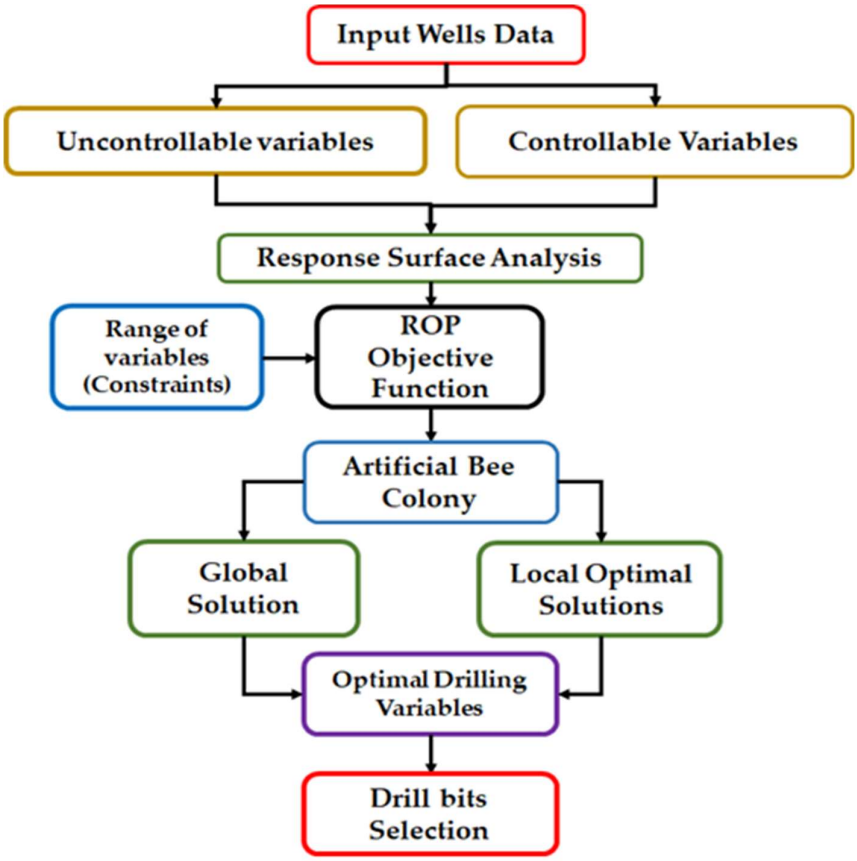


Figure 4.4. A generalized schema of the proposed approach for drill bit selection.

4.2.4. *Data Description*

Volve oil and gas field is situated in the central North Sea near the Norwegian Continental Shelf. It was discovered in 1993 and its production shut down in September 2016 by its investors' companies. The ocean depth near the Volve field is in a range of 85 to 95 meters. This field contains Jurassic sandstone related to the Hugin formation reservoir. The depositional environment of this reservoir is analyzed as tidal to the shallow estuary. The average properties expected from this Hugin reservoir are as follows: porosity (0.2), permeability (910), water saturation (0.23), and shale volume (0.17). Geosteering was particularly utilized to increase the extent of the reservoir linking to various fault blocks. The peak production rate in the Volve field was recorded to be 56,000 barrels per day and produced a cumulative amount of 63 million barrels of

oil with a recovery rate of 54% of the total reservoir estimated over 8 years [103]. Input data used in this research work were obtained from the Equinor company website openly available for research purposes [103]. Table 4.1 contains the geological prognosis of the Volve field. Table 4.2 contains a description of the drilling variables used in the study. Table 4.3 has been provided to inform about different drill bit types (I.A.D.C. code) utilized for drilling the three wells. Wells A (F-4) and B (F-15) were utilized for the training of models and well C (F-12) for testing the developed models. The I.A.D.C. code of drill bits cannot be utilized directly for the training of the ANN model. Thus, they are numbered in the bit-type column of Table 4.3. The location of the Norwegian Volve field situated in the North Sea is depicted in Figure 4.5.

Table 4.1 Geological prognosis of well 15/9-F-12 under study (courtesy: Equinor company) [103].

Group	Formation	Depth (m)	Description
Nordland	Utsira Top	892	Grey claystone, a stringer of sand and siltstone.
	Utsira Base	1084	Well sorted sandstone, minor silt, and limestone stringers.
Hordaland	Skade Top	1259	Claystone, minor limestone/dolomite stringers.
	Skade Base	1347	Medium-grained sorted sandstone.
	Grid Top	2179	Fine-grained sandstone.
	Grid Base	2245	Fine-grained sandstone.
Rogaland	Balder Top	2317	Colored claystone, partly tuffaceous, and limestone stringers.
	Sele Top	2374	Claystone and limestone stringer.
	Lista Top	2445	Non-calcareous claystone and minor limestone stringers.
	Ty Top	2531	Fine to medium sandstone some interbedded claystone, siltstone, and limestone stringers.
Shetland	Ekofisk Top	2698	Limestone with traces of claystone and

			sandstone.
	Tor Top	2715	White limestone with traces of claystone.
	Hod Top	2839	Limestone along with gluconate.
	Blodoeks Top	2944	Marl, argillaceous laminations, and gluconates in parts.
	Hidra Top	2972	Off-white firm limestone.
Cromer Knoll	Roedby Top	2981	Marl along with argillaceous laminations
	Aasgard Top	3001	Interbedded limestone and marl with minor claystone and siltstone.
Viking	Draupne Top	3036	Organic-rich claystone, micaceous, carbonaceous with traces of pyrite.
	Heather Top	3086	Claystone with limestone stringers.
Vestland	Hugin Top	3094	Sandstone and rare claystone stringers.
	Sleipner Top	3266	Sandstone, grey claystone, and layers of coal.

*Table showing different formations for F-12 well along with depth of the reservoir

Table 4.2. Statistical details of drilling data used in this study.

S. No.	Parameters/unit	Range	Units	Code Factor
1.	Measured Depth (DT)	100-3520	m	x_1
2.	Rate of Penetration (ROP)	1.73-201.02	m/hr	--
3.	Weight on bit (WOB)	0.01-19.17	Tons	x_2
4.	Rounds per minutes (RPM)	60-220	rpm	x_4
5.	Torque (TQ)	0-5.24	kN/m	x_3
6.	Standpipe pressure (SPP)	58.6-289.2	Bar	x_5
7.	Mud weight (MW)	1.03-1.42	S.g.	x_6
8.	Inclination (IN)	0.46-54.95	Degree	x_8
9.	Azimuth (AZ)	0.38-334.67	Degree	x_9
10.	Bit type (BT)	1-11	N/A	x_{10}
11.	Bit Size (BS)	8.5-26	Inch	x_7

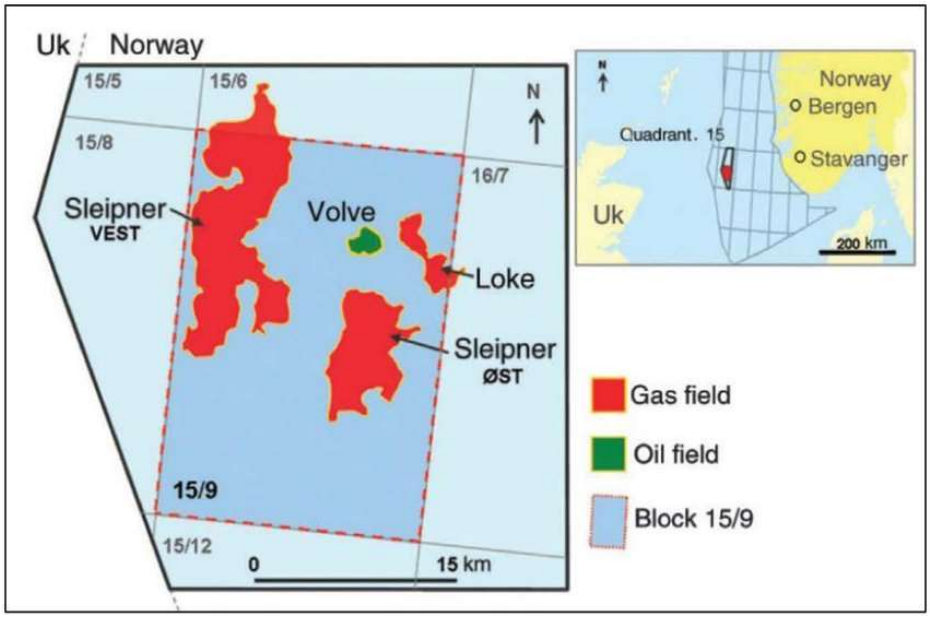


Figure 4.5. Location of Volve oil and gas field in the North Sea [103].

Table 4.3. Drill bit types utilized for drilling of wells at different depths for three Norwegian Volve field wells.

Bit Type	Depth In (m)	Depth out (m)	IADC code	Bit Size (Inch)
WELL A (F-4)				
1	100	310	PDC M415	36”
2	25	1360	MT 115A	17.5”
7	1360	1410	PDC M422	12.25”
8	2770	2993	PDC M222	8.5”
8	2993	3510	PDC M222	8.5”
Well B (F-15)				
10	144	226	MT 115	36”
6	226	1378	PDC M115	26”
3	1378	1381	MT 244	17.5”
4	1381	2536	PDC M332	12.25”
9	2536	3670	PDC M323	8.5”
9	3670	4090	PDC M323	8.5”
5	1378	2591	PDC M322	26”
4	2591	2594	PDC M332	12.25”

5	2591	2596	PDC M322	8.5”
5	2596	3180	PDC M322	8.5”
5	3180	4095	PDC M322	8.5”
5	3185	3498	PDC M322	8.5”
7	2562	2665	PDC M422	12.25”
7	2665	2920	PDC M422	12.25”
WELL C (F-12)				
1	251	1369	PDC M415M	36”
5	1369	2513	PDC M322	17.5”
6	2513	2573	MT 135	17.5”
7	2573	3114	PDC M422	12.25”
8	3114	3520	PDC M222	8.5”

Table 4.4. The reported relationships to calculate the neurons inside the hidden layer of ANN.

Relationships	References	Relationships	References
$\leq 2\times N_i+1$	[107]	$2\times N_i/3$	[110]
$(N_i+N_0)/2$	[108]	$\sqrt{N_i\times N_0}$	[111]
$\frac{2+N_0\times N_i+0.5N_0\times(N_0^2+N_i)-3}{N_i+N_0}$	[109]	$2N_i$	[112,113]

4.3 Results

4.3.1 Development of ROP objective function using ANN

In this study, three-layered MLP architecture was utilized to generate an approximation function for the ROP in terms of operational drilling variables. ANN having a three-layered multilayer perceptron architecture can be utilized to produce the polynomial equation for solving optimization problems [87]. However, parameters of ANN are required to be determined beforehand during the training stage. Three popular training functions were applied to train the ANN namely, (a) Levenberg-Marquardt (LM), (b) scaled conjugate gradient (SCG), and (c) Bayesian regularization (BR). Hornik et al.

(1989) suggested that a singular hidden layer neural network model can be utilized for approximating any nonlinear function [104]. However, the optimum neurons in the hidden layer are required to be estimated before training the ANN model. Table 4.4 contains reported correlations to decide the number of neurons inside the hidden layer. N_i and N_o represents the number of predictor and target variables in Table 4.4. Several ANN models were generated with the neurons ranging from 2-20 which were calculated using the correlations available in Table 4.4. The optimum values of the model parameters are given in Table 4.5. The performance of developed ANN models is compared based on the coefficient of correlation (R^2) and root mean square error (RMSE) as mentioned below.

A. Coefficient of correlation (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(ROP_m - ROP_p\right)^2}{\sum_{i=1}^{n}\left[ROP_m - \frac{1}{n}\sum_{i=1}^{n}\left(ROP_m\right)\right]_i^2} \tag{4.14}$$

B. Root mean square error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(ROP_m - ROP_p\right)^2} \tag{4.15}$$

Zorlu et al. (2008) suggested a ranking method to compare the performance of several neural networks together [105]. Here, an integer (rank) was allocated to every network based upon the goodness of R^2 and RMSE values. Then, ranks allocated to every R^2 and RMSE were added to acquire the total rank for every network configuration separately. The network having the highest total rank was considered as the best model for this research work [105,106]. The results of three-layered neural architectures with three different combinations of training functions are recorded in Tables 4.6, 4.7, and 4.8. The training algorithm LM has acquired the highest rank of 36

through its performance with (10-18-1) configuration. Therefore, the best performing ANN (10-18-1) alignment was considered for the development of the approximation function to obtain optimum values ROP and BT along with other drilling variables.

Table 4.5 Optimum values of model parameters utilized in this research work.

Estimators	**Model Parameters**	**Search Range**	**Optimum value**
ANNs	Configuration	2-20	[10 18 1]
	Learning rate	0.0001-0.5	0.0001
	Maximum number of iteration	100-1000	200
	Activation function hidden layer	Tangential sigmoid function	N/A
	Activation function output layer	purline function	
	Training algorithm	Levenberg-Marquardt	N/A
ABC	Iterations	10-200	100
	Scouts	1-100	70
	Colony size	1-100	100
GA	Iterations	1-5000	3000
	Crossover probability	0.1-1	0.5
	Population size	1-200	100
	Crossover type	-	Uniform
	Elit_ratio	0.001-1	0.001
	Parents portion	0.1-1	0.3

Table 4.6 The outcomes of ANN models trained using the LM function*.

Model No.	No of Neurons	Train R^2	Train RMSE	Test R^2	Test RMSE	Train rating R^2	Train Rating RMSE	Test Rating R^2	Test Rating RMSE	Total Rank
1	2	0.793	11.76	0.597	20.414	3	3	1	1	8
2	4	0.640	16.20	0.634	16.233	1	1	2	2	6
3	6	0.839	10.94	0.720	14.975	4	5	3	4	16
4	8	0.894	9.283	0.850	8.298	6	6	9	10	31
5	10	0.843	11.18	0.801	10.475	5	4	6	8	23
6	12	0.930	6.930	0.831	13.635	9	8	8	5	30
7	14	0.908	8.332	0.761	11.786	8	7	4	7	26
8	16	0.952	6.319	0.825	16.392	10	9	7	3	29
9	18	0.9143	7.99	0.855	10.32	7	10	10	9	36
10	20	0.734	14.59	0.774	12.916	2	2	5	6	15

*Table 4.6 shows the ranking method for selecting the optimum number of neurons for the hidden layer. Here, 18 neurons have achieved the highest total rank of 36 by summing all the training and testing ratings together for the LM training function. The results are also compared with the total rank achieved in Tables 4.7 and 4.8 to decide the optimum neuron configuration in the hidden layer.

Table 4.7 The outcomes of the ANN models trained using SCG function.

No of Neurons	**Train R^2**	**Train RMSE**	**Test R^2**	**Test RMSE**	**Train Rating R^2**	**Train Rating RMSE**	**Test Rating R^2**	**Test Rating RMSE**	**Total Rank**
2	0.723	14.82	0.675	16.430	4	4	4	4	16
4	0.767	12.71	0.804	12.901	8	8	2	9	27
6	0.558	19.32	0.648	16.520	1	2	2	3	8
8	0.710	14.30	0.673	14.263	3	6	3	7	19
10	0.685	16.34	0.600	14.340	2	3	1	6	12
12	0.764	13.88	0.816	11.778	7	7	9	10	33
14	0.814	12.37	0.756	13.538	10	9	7	8	34
16	0.734	14.36	0.745	14.609	5	5	6	5	21
18	0.802	12.12	0.733	16.851	9	10	5	2	26
20	0.757	210.75	0.869	115.82	6	1	10	1	18

Table 4.8. The outcomes of the ANN models trained using the BR function.

Model No.	No of Neurons	Train R^2	Train RMSE	Test R^2	Test RMSE	Train Rating R^2	Train Rating RMSE	Test Rating R^2	Test Rating RMSE	Total Rank
1	2	0.855	10.84	0.714	13.376	1	1	3	8	13
2	4	0.890	8.90	0.819	13.520	2	3	7	6	18
3	6	0.927	7.58	0.751	13.375	3	4	5	9	21
4	8	0.9281	6.94	0.864	13.473	4	5	9	7	25
5	10	0.946	6.492	0.762	14.476	6	6	6	4	22
6	12	0.946	6.359	0.907	10.059	6	7	10	10	33
7	14	0.965	5.131	0.547	19.724	8	9	1	2	20
8	16	0.960	5.37	0.820	14.157	7	8	8	5	28
9	18	0.973	4.416	0.715	19.234	9	10	4	3	26
10	20	0.986	9.907	0.562	28.045	10	2	2	1	15

Figure 4.6 contains the Regression plot, MSE plot, and Error plot of optimal ANN architecture [10-18-1]. Table 4.9 contains weights and bias associated with the selected neural network configuration (10-18-1). The developed ROP approximation function using equation (4.7) is given below.

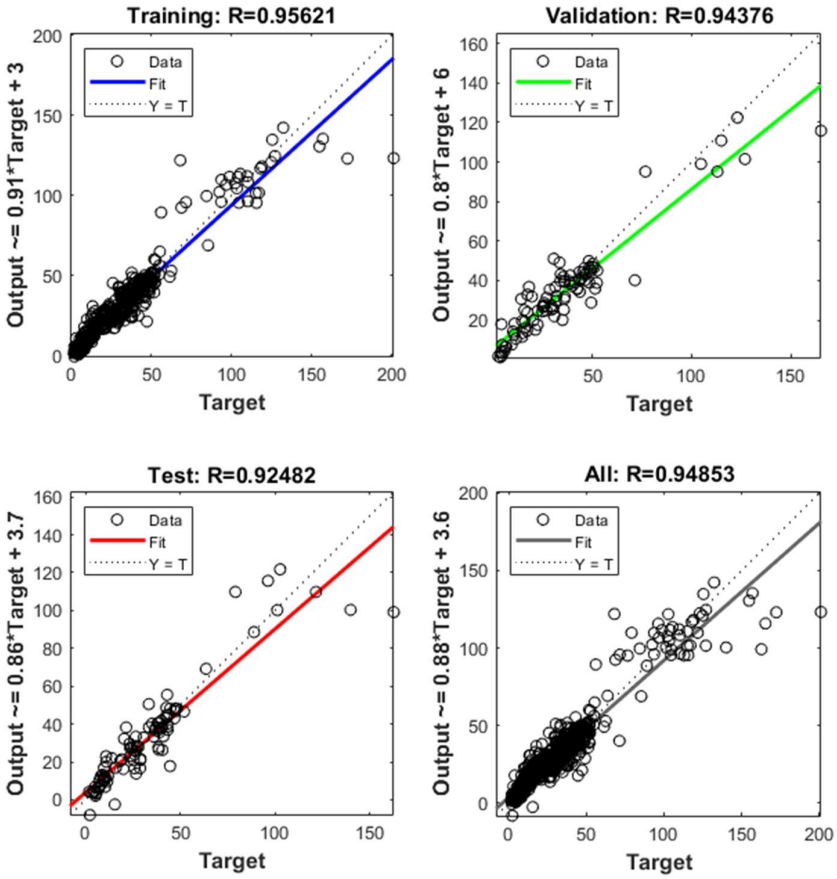
$$\mathrm{f}(x) = \left[\sum_{i=1}^{N} w_{2,\mathrm{i}} \left(\frac{1}{\left[1 + e^{-2Z} \right]} - 1 \right) + b_2 \right] \tag{4.16}$$

$$z = (w_{i,1} \times x_1 + w_{i,2} \times x_2 + w_{1,3} \times x_3 + w_{i,4} \times x_4 + w_{i,5} \times x_5 + w_{i,6} \times x_6 + w_{i,7} \times x_7 + w_{i,8} \times x_8 + w_{i,9} \times x_9 + w_{i,10} \times x_{10})$$

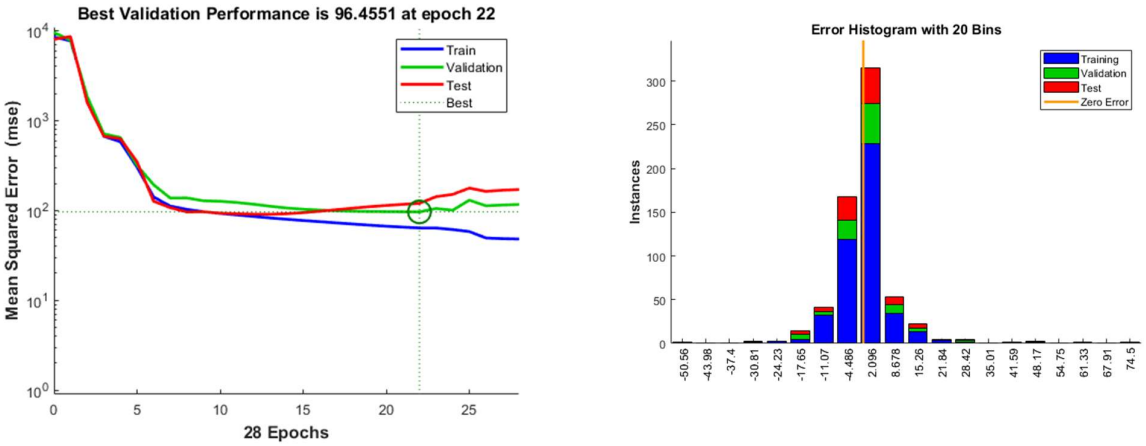
where the tangent sigmoid function is utilized in the hidden layer, while purline is the activation function for the output layer. $w_{i,1}$, and $w_{2,i}$ are the weights of input and hidden

layers. The weights and bias (obtained during the training of 10-18-1 ANN configuration) associated with the equation (4.16) have been provided in Table 4.9. Table 4.10 contains constraint bounds required for the optimization of the ROP equation (4.16) and control variables. Here, TD, IN, and AZ will remain constant because of their predefined nature while optimizing for a particular depth.

Drill bit selection based on optimum ROP values is a bound constrained maximization problem. The developed ROP objective function (equation 4.16) requires an optimization algorithm for determining optimum values of ROP and other operational variables along with BT. GA is an evolutionary paradigm that has been utilized for the optimization of equation (4.16) [106]. During the optimization process, equation (4.16) was maximized using GA with upper and lower bounds as shown in Table 10 according to the following steps. (a) Adjust the model parameters of GA. (maximum no of iterations = 3000, crossover probability = 0.5, population size = 100, parents portion = 0.3, crossover-type = uniform, elite ratio = 0.01, variable type = real). (b) Set the upper and lower bounds for input variables existing in objective equation (16) using Table 10. (c) Randomly generate the initial population for GA. (d) Several combinations of ROP and input variables will be generated during optimization. In the end, GA converges on the best combination of input variables having maximum ROP value. (e) Record the value of ROP and BT in the final solution produced by GA. GA will provide optimum ROP values along with suitable BT and other control variables. The optimization of the ROP objective function (equation 4.16) has been carried out using the Python package Geneticalgorithm 1.0.1 freely available online for research purposes.



(a)



(b) (c)

Figure 4.6. The prediction performance of optimal ANN architecture [10-18-1] for developing ROP objective function (a) Regression plot, (b) MSE plot, and (c) Error plot.

Table 4.9 The weights and bias allocated to the training of optimal ANN configuration [10-18-1].

Connections	Generated Values
Bias in hidden layer	-2.331; -2.775; -0.1308; -0.03132; 1.504; 0.2345; 2.2617; -1.5475; 1.1458; 0.3785; -0.0855; 0.3141; -1.0727; -2.2302; 1.8623; -2.6596; -1.85; 2.721
Connection weights between input and hidden layers	-0.486; -2.186; 0.151; 1.236; 1.964; 0.920; -0.5052; -0.677; 0.008; 0.570; 1.202; -0.577; 0.269; -2.421; 0.145; -0.654; -0.664; -2.189; 1.061; 0.126; 1.048; 1.962; -0.479; -0.984; -1.578; 0.542; -1.061; -0.900; -0.856; 0.401; 0.108; -0.469; -1.358; -0.818; 0.037; 0.413; 0.484; -1.092; -1.283; 0.109; -0.444; -1.698; -3.35; 0.353; -0.194; -0.038; -0.431; -0.647; -0.0049; -0.225; 0.266; 0.517; -0.0142; 1.035; 0.456; -0.271; 0.6465; -0.568; 1.96; -0.423;-0.594; -0.981; 0.441; -0.363; -0.293; -0.0681; -0.313; -0.996; -2.33; 1.21; -1.033; -0.649; 1.68; -0.859; 0.426; 0.880; -0.425; -1.456; -1.231; 0.436; -1.234; -1.301; 1.712; 1.96; 1.55; 0.53; 1.17; -1.79; 0.66; -0.093; -2.81; 1.47; -1.103; -0.346; -3.252; 0.403; -0.52; 0.426; -1.246; 0.8551;-0.721; -3.845; 0.619; 0.902; 1.909; -0.7886; 0.271; 1.015; -3.98; 0.366;-0.101; -1.257; 2.83; 1.017; 1.185; -0.150; -0.0382; -2.389; -0.740; -1.086;-1.868;-0.573; -0.689; -0.337; -1.414; 1.336; 0.797; -0.853; -2.783; -0.484; -0.252; -1.243; 1.548; 1.508; 0.047; -0.109 ; 0.699; 0.56544; -0.317; -0.452; 1.35; -0.1153; -0.661; 1.076; 0.436; 0.986 ;0.55;0.5131; -1.026; -0.5348; -0.1402; 0.156; -0.5217; -0.6648; 1.3074; 0.0939; -0.822;-1.814;-0.092; 1.22;0.242 ;-1.53;1.24;-1.493; 0.112; -0.68; -0.342; 0.62; 0.451; -0.179; -1.14; 3.38; 0.66; -0.351; -0.017; 0.098; -0.433; -1.99; -1.40; -0.112
Bais in the output layer	-0.7816
Connection weights between hidden and output layers	-0.264; -1.214; 0.0348; -0.701; 0.319; 0.410; 0.1114; 0.421; 0.5133; -0.4028; 0.1325; -0.5300; 0.785; -0.475; -0.765; 0.434; 0.376; -0.133

Table 4.10. Range of predictor variables utilized during optimization of ROP as upper and lower bounds.

Depth Interval (m)	Predictor variables utilized during optimization
251-1369	DT = constant, BT = [1, 10], BS = constant, WOB = [0.01, 16.29], RPM = [83, 220], TQ = [-0.15, 15.26]. MW=[1.03,1.36], IN=constant, AZ=constant, SPP=[59,153]
1369-2513	DT = constant, BT = [1, 10], BS = constant, WOB = [0.01, 19.17], RPM = [90, 186], TQ= [7.58, 27.03]. MW=[1.16,1.4], IN=constant, AZ=constant, SPP=[140,252]
2513-2573	DT = constant, BS = constant, BT = [1, 10], WOB = [11.78, 19.07], RPM = [150, 180], TQ = [11.35, 19.07]. MW=[1.39,1.4], IN=constant, AZ=constant, SPP=[226,282.4]
2573-3114	DT = constant, BT = [1, 10], BS = constant, WOB = [1.35, 10.96], RPM = [137, 180], TQ = [1.07, 9.84]. MW=[1.39,1.42], IN=constant, AZ=constant, SPP=[220,290]
3114-3520	DT = constant, BS = constant, BT = [1, 10], WOB = [2.04, 6.87], RPM = [60, 140], TQ = [2.11, 6.15]. MW=[1.39, 1.44], IN=constant, AZ=constant, SPP=[174.6, 233.9]

4.3.2 *Development of the ROP function using RSM*

The CCD design was utilized for the generation of fitting equation (4.17) with a face-centered configuration where alpha was one. The ten input variables were considered as factors that were coded in two levels [1,-1]. There were 128 cube points, 10 center points, and 20 axial points present in the developed CCD design with a total of 158 base runs. The quadratic equation of RSM contains controllable input factors and uncontrollable factors. These uncontrollable or predefined factors such as DT, AZ, etc. were held as constant as they cannot be altered. The objective function of ROP developed using RSM is given below.

$$f(x) = 21.61 \times x_2 + 12.337 \times x_4 - 15.34 \times x_5 + 37.57 \times x_9 - 0.1554 \times x_{10} - 17.67 \times x_1^2 + 2.55 \times x_2^2 - 0.1699 \times x_3^2 - 2.325 \times x_4^2 + 3.35 \times x_5^2 - 2.6398 \times x_9^2 - 11.72 \times x_1 x_2 - 7.04 \times x_1 x_2 - 7.04 \times x_1 x_3 - 12.437 \times x_1 x_9 - 16.11 \times x_2 x_7 + 1.84 \times x_2 x_9 - 8.67 \times x_4 x_6 - 3.767 \times x_4 x_8 + 0.0564 \times x_4 x_{10} - 7.391 \times x_5 x_9 + 9.156 \times x_5 x_8 + 0.0723 \times x_5 x_{10} - 59.62 \quad (4.17)$$

This ROP equation (4.17) shows predictor variables in the quadratic fitting equation. The R^2, adjusted $R^{2,}$ and predicted R^2 for the above-mentioned objective function are 84.41%, 82.68%, and 81.23% respectively as given in Table 4.6. Adjusted R^2 shows the goodness of fit for the developed regression model. Less difference between R^2 and adjusted R^2 shows that important predictors were selected for fitting a quadratic polynomial of RSM. Contours and 3D Surface plots were also generated to have a better visualization of the effects of various predictor variables with ROP as shown in Figures 4.7 and 4.8. These plots were helpful for the manual search of optimal points in the case of a lower number of input variables. However, the manual search technique is not recommended in our case due to the complexity of the developed ROP objective function. The significance level of 5% was used while developing an objective function for ROP (equation 4.17). It has been reported the 5 % significance level balances Type 1 and Type 2 error during hypothesis testing of any regression coefficients [114,115]. The correctness of the developed ROP objective function was validated by analysis of variance (ANOVA) test as shown in Table 4.11. This test demonstrates that coefficients satisfy 5% significance level criteria. All the terms having P values higher than the significance level were eligible for the null hypothesis which resulted in zero value of coefficient terms and was eliminated from the ROP regression equation. Table 4.11 shows all the coefficients having lower P values. A T-test was also performed to validate the significance of the regression coefficient of the ROP function. There are several missing interaction terms in the ROP objective function such as BT*BT, BT*BS, etc. due to their higher P values. The developed equation (4.17) has

been optimized using the ABC algorithm available in the python beecolpy 2.1 packages based on Karaboga and Basturk, [116]. Figure 4.9 shows Residual plots of errors resulted from the developed ROP objective function using RSM.

Table 4.11. Results of the ANOVA test for significant terms in ROP equation (4.17) utilized in this study.

Source	DF	Sum of Squares	Mean Square	F-Value	P-Value	T-Value
Model	63	10.29	0.1633	48.74	0.00	0.000
x_2	1	0.98	0.1022	30.51	0.00	5.524
x_4	1	0.013	0.013	3.94	0.047	1.986
x_5	1	0.0158	0.0157	4.71	0.030	−2.169
x_9	1	0.02	0.02	5.97	0.015	2.444
x_{10}	1	0.0020	0.00204	0.61	0.036	−1.779
$x_1 \times x_1$	1	0.022	0.020	5.98	0.015	−2.445
$x_2 \times x_2$	1	0.1132	0.1132	33.77	0.000	5.811
$x_3 \times x_3$	1	0.0001	0.00005	0.02	0.897	−0.129
$x_4 \times x_4$	1	0.0139	0.0139	4.15	0.042	−2.038
$x_5 \times x_5$	1	0.0243	0.0243	7.24	0.007	2.691
$x_9 \times x_9$	1	0.0343	0.3433	10.24	0.001	−3.200
$x_1 \times x_2$	1	0.1493	0.1493	44.57	0.00	−6.676
$x_1 \times x_3$	1	0.0154	0.0154	4.61	0.032	−2.146
$x_1 \times x_9$	1	0.0027	0.00271	8.09	0.005	−2.844
$x_2 \times x_7$	1	0.1153	0.1153	34.4	0.000	−5.865
$x_2 \times x_9$	1	0.0206	0.0206	6.15	0.013	2.480
$x_4 \times x_6$	1	0.0157	0.01565	4.67	0.031	−2.161
$x_4 \times x_8$	1	0.0134	0.0133	4.01	0.046	−2.002
$x_4 \times x_{10}$	1	0.0206	0.0206	6.15	0.013	2.480
$x_5 \times x_9$	1	0.1299	0.1298	38.75	0.000	−6.225
$x_5 \times x_9$	1	0.0534	0.05303	15.82	0.000	3.978
$x_5 \times x_{10}$	1	0.0137	0.01366	4.08	0.044	2.019

*Table 4.11 contains the ANOVA analysis of significant terms only while insignificant terms that fail in the significance test of 5% are eliminated to reduce the redundancy in equation (4.17).

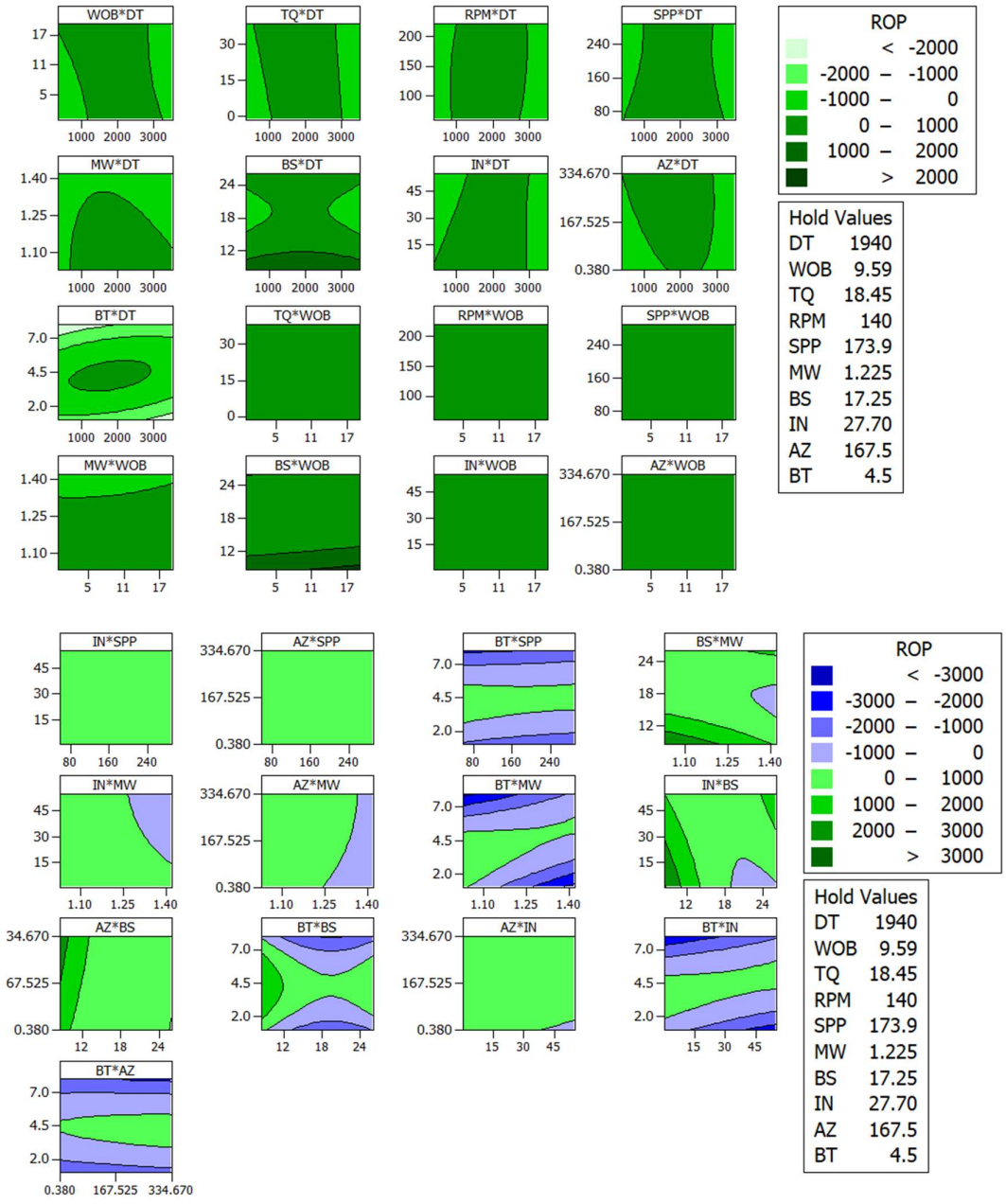


Figure 4.7 Contour plots show interactions of different input variables visualized in 2 D plane for equation (4.17). (Example: In WOB*DT subplot, WOB is on the y-axis whereas DT on the x-axis.).

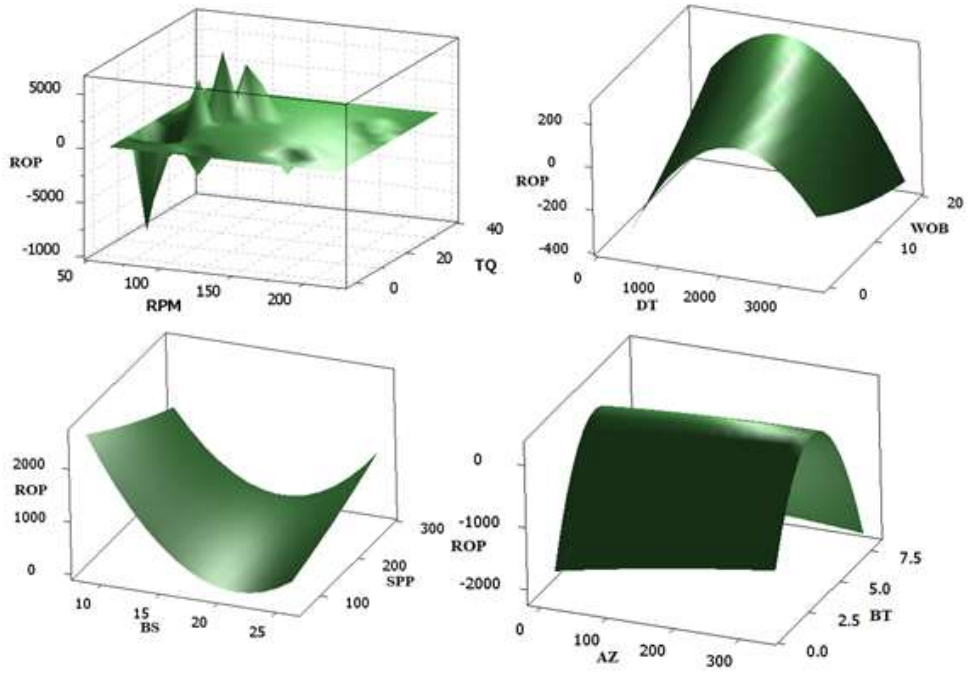


Figure 4.8 Surface plots for ROP objective function generated using the RSM technique. These plots help to visualize the interactions between various input variables. These are graphical visualization of the fitted ROP equation (4.17).

The developed equation (4.17) has been considered as an ROP objective function. During the optimization process, equation (4.17) was maximized using ABC with upper and lower bounds as shown in Table 4.10 according to the following steps. (a) Initialize the search boundaries using the range of parameters provided in Table 4.10 and code the equation (4.17) as an objective function (b) Adjust the other parameters of ABC. (colony size=50, scouts=0.5, iterations=100, min_max='max', nan_protection=True). Here the size of the colony determines bees in the algorithm. Half of its values represent food sources, employed bees, and onlooker bees. (c) The algorithm returns a global optimal solution for the ROP objective function along with the locations of food sources or possible solutions (local maxima). (d) Record the values of ROP and other variables including BT. Table 4.12 and 4.14 contain the optimum values of drilling variables and BT for target formations.

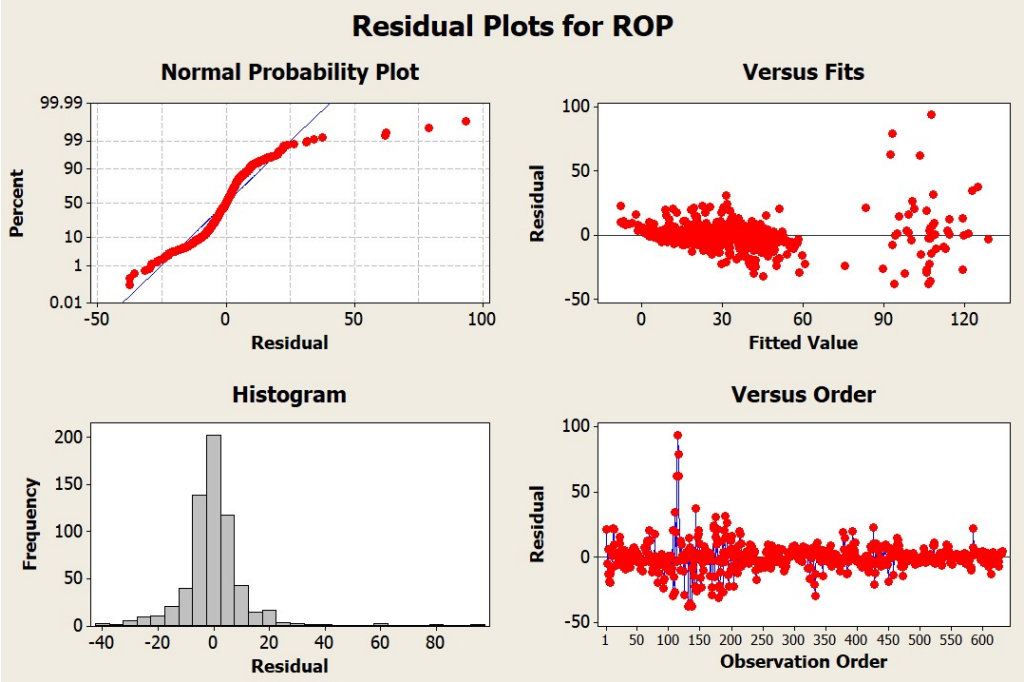


Figure 4.9 Residual plots of errors from developed ROP objective function using RSM.(a) Normal probability plot is used to verify normal distribution of residual data (b) Histogram of residuals provide details about data skewness or outliers presence. (c) Residual vs fits confirm the constant variance of residual. (d) Residual vs order plot check whether residual are uncorrelated or not. These graphs are generated to inspect the goodness of fit of fitting equation (4.17) and ANOVA test.

4.4 Discussion

The selection of suitable drill bits is essential for a successful drilling operation to minimize the overall wellbore cost and increase the efficiency of the drilling operations. In this study, an alternative approach has been investigated for drill bit selection using RSM and ABC combination. RSM has been utilized to develop an objective function for ROP and to determine optimum values of drilling control variables using ABC. Ten drilling variables were considered as input variables for the development of the ROP objective function namely, DT, BT, BS, WOB, RPM, TQ, MW, IN, AZ, and SPP. Three nearby Norwegian wells' data have been considered for testing the proposed

approach of drill bit selection. The five geological zones of well C were utilized for the testing of data-driven drill bit selection techniques as shown in Tables 4.12 and 4.13. Figure 4.4 shows the generalized schema of the proposed approach for drill bit selection. The developed objective equations (4.16 and 4.17) for ROP were optimized with the upper and lower bounds provided in Table 4.10 to obtain the optimum value of BT and ROP. Table 4.12 contains the bit types selected based on optimum ROP values using different data-driven approaches. ANN has wrongly predicted the drill bit types for zones 1, 3, and 4, however, when combined with GA its drill bit selection error reduces to zones 1 only.

ANN has been reported to have a tendency for stuck up in local minima which is why it failed to predict the correct bit type for certain target formations [86, 87]. However, when combined with GA, the optimization task is handled by GA which is a strong optimizer and converses to the correct BT except in zone 1 as compared to actual BT. Equations (4.16) and (4.17) are multimodal equations, developed through high-dimensional data, comprise of large local optimal points. Therefore, detecting a globally optimum solution in search space is a difficult task. In the case of GA, sometimes premature convergence may happen due to strong selection pressure imposed by the selection operator and crossover operator if the initial population lacks desirable diversity. However, ABC utilizes a stochastic search technique which is good at maintaining diversity and escaping local optimal stagnation. Table 4.12 shows that the proposed RSM and ABC combination has precisely estimated the bit-types for five target geological zones. Here, information about actual drill bits used in the real field for drilling the wells has been taken as a standard reference for comparison in Table 4.12. Table 4.14 contains the optimum values of drilling variables acquired through RSM and ABC combination at certain given depths. However, the performance of drill bits

selection techniques must be compared based on the drilling cost involved in the drilling operations. It might be possible that the GA suggested BT for zone 1 is more suited to an applied bit. Therefore, the cost-per-foot analysis should be done for each drill bit to check the economic feasibility of suggested drill bit types.

The drilling cost has a direct relationship with ROP and the running life of the drill bit that is needed to be minimized. Polycrystalline diamond compact (PDC) drill bits were primarily utilized for drilling the Norwegian wells, hence, are considered in this study. Nearly, 10-40 % of dryhole well cost is found dependent on PDC drill bit [16]. PDC bit life fluctuates with its design parameters such as cutter distribution, type of gauge protection material, etc. whereas, design parameters such as nozzle placement, cutter shape, PDC type, etc. directly vary the bottom hole ROP drastically [16]. Therefore, the selection of suitable BT is essential for the minimization of associated drilling costs. The selection performance of different data-driven approaches has also been compared based on cost-per-foot analysis. The prices of drill bits were identified from the manufactures catalogs. The totality of trip time and connection time was expected to be 6 hours per 1000 ft. Equation (1.1) has been utilized to perform the cost-per-foot analysis of the predicted drill bit. Table 4.13 shows the computed cost per foot results for five target zones of well C based on drill bits selected through different data-driven approaches. Figure 4.10 shows that the selection of the bit for five target drilling zones by ANN and GA combination and proposed approach has given nearly similar types of drill bits and cost per foot except for zone 1. The proposed approach has given a lesser cost per foot value for zone 1 as compared to ANN and GA combination as shown in Table 4.13. Table 4.14 shows the optimum value of input drilling variables for certain depths of target zones using the RSM and ABC combination. Therefore, the proposed RSM and ABC combination is found more reliable than ANN-based drill bit

selection models and can also be utilized for drill bit selection purposes. Moreover, these models are case-specific, as well as data-dependent in nature, and require calibration for other fields.

Table 4.12 Comparative analysis of drill bit selection results for the test geological zones.

Bit Selection	Models	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Hou, Chien, & Yuan, (2014)	ANNs	3	5	4	3	8
Edalatkhah, Rasoul, & Hashemi, (2010)/ Abbas et al. (2019)	ANNs and GA	3	5	6	7	8
Proposed approach	RSM and ABC	1	5	6	7	8
Actual BT Data		1	5	6	7	8

Mostly, PDC drill bits have been utilized along with roller cone bits for the drilling of the Norwegian wells considered in this study. These bits are widely applied in offshore conditions due to various benefits such as reducing tripping time, drilling in non-hydrating formations, achieving high RPM and ROP in directional drilling, etc. However, PDC bits are sensitive to fragile and soft formations as in the case of Volve wells considered in this study. These wells comprise softer rock formations that contain fine to medium sandstone, some interbedded claystone, siltstone, and limestone stringers, marl, argillaceous laminations, etc. as shown in Table 4.1. Soft, fragile, and fractured rock formations affect the stability of PDC bits with a large reduction of the bit life. Recently, hybrid drill bits (e.g. Kymera) have been developed that combined the properties of conventional PDC bit and roller cone bit types [117]. These hybrid bits

seem to be a good solution for drilling problematic wellbore sections while maintaining the stability of drilling operations. It may be possible that the drilling cost per foot has been reduced more if hybrid drill bits are employed for drilling the Norwegian wells considered in this study.

Table 4.13 Comparison of drill bit selection results based on cost per foot calculation for different approaches.

Target Zones	ANNs predicted	Cumulative C_{CF} $/ft	ANNs and GA	Cumulative C_{CF} $/ft	Proposed Approach	Cumulative C_{CF} $/ft
Zone 1	3	50	3	50	1	24
Zone 2	5	25	5	25	5	25
Zone 3	4	456	6	396	6	396
Zone 4	3	50	7	50	7	50
Zone 5	8	64	8	64	8	64

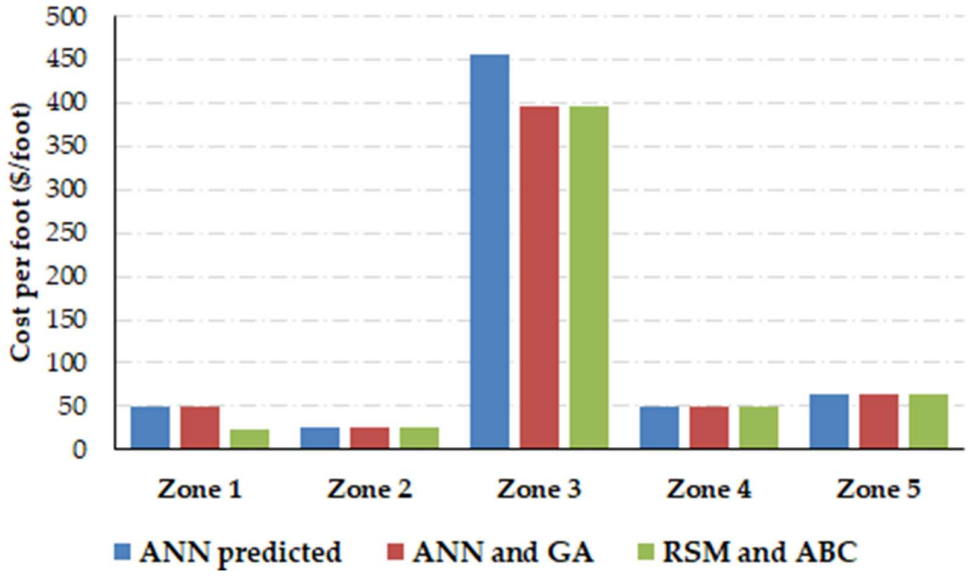


Figure 4.10 The comparison of cost per foot calculated for five target geological zones. ANN and GA combination has given similar cost per foot as compared to the proposed RSM and ABC approach except in zone 1. In zone 1, RSM and ABC have given lower cost per foot results than ANN and GA combination.

Table 4.14 The optimum value of input drilling variables for certain depths of target zones using RSM and ABC combination.

DT	WOB	TQ	RPM	SPP	MW	BT	ROP
Zone 1	3.72	0.01	90	62.6	1.03	1	47.4
Zone 2	3.64	16.67	176	247.3	1.39	5	40.35
Zone 3	16.01	21.8	178	279.3	1.22	6	10.56
Zone 4	6.94	31.25	139	193.8	1.41	7	11.61
Zone 5	6.87	22.28	140	205.9	1.4	8	27.54

4.5 Summary

A comprehensive study has been done to develop a new alternative approach for the selection of drill bit types. RSM and ABC combination has been proposed to select drill bit types based on the optimum values of ROP. The optimum values of operational variables are also determined in this research work for drilling the target formations. The proposed drill bit selection approach is found more accurate than ANN-based prediction of drill bit types. This study provides an alternate intelligent approach for bit selection based on optimum values of ROP. The combination of RSM and ABC provides a more reliable bit selection modeling approach as compared to ANN-based on cost per foot comparison. The prediction correlation coefficient of the RSM objective function is found to be 81.23% while 85.5 % has been found for ANN during the estimation of ROP. The ROP objective function developed through RSM is less complex than the ANN-based objective function due to the absence of an exponential function. ANN requires more computational cost for the development of the ROP function for its optimization. These models are case-specific data-dependent models and require calibration for other field data.

Chapter 5

A Novel Application of Ensemble Methods with Data Resampling Techniques for Drill Bit Selection

5.1 Introduction

The cost of drilling operation for a hydrocarbon well is mainly dependent on various factors namely, the operating cost of the drill rig, the time required for drilling target formations, the number of tripping operations, the life of the drill bit, and drill bit cost [83-85]. Nevertheless, drilling costs can be significantly minimized through the selection of appropriate drill bit designs, which in turn reduces the operating time of the rig with fewer tripping events and more life expectancy of the drill bit. However, the selection of suitable drill bit types for drilling geological formations is a problematic task due to complex interactions between reservoir properties, drill string hardware design. In the previous chapter, drill bits selection is proposed based on optimum values of rate of penetration (ROP) which is one of the most important parameters of the drilling operation. In this chapter, a second approach, based on the prediction capability of classifiers, is proposed for drill bits selection. Recently, data-driven intelligent models have been utilized to find suitable types of drill bits. These models are reported to be more accurate as they learn from previous well data, defying traditional methods for selecting the appropriate drill bit [24]. Bilgesu et al. [25] used Artificial neural networks (ANN) for the prediction of drill bit types for drilling target geological formations. Yilmaz et al. [26] trained the ANN model using previously drilled wells offset data and predicted the drill bits types for the development wells required to be drilled internally and externally of the same field. They also tested the trained ANN model for the prediction of drill bit types for the development wells that were required

to be drilled in an adjoining field. Bahari et al. [15] utilized a genetic algorithm (GA) for the accurate computation of constants for the Bourgoyne-Young ROP model. Edalatkhah et al. [27] also selected the suitable drill bit types using ANN and GA for South Pars field wells. Momeni et al. [28] applied ANN for the estimation of drilling ROP and bit types [14]. Momeni et al. [29] combined ANN and GA for drill bit selection based on optimal ROP. They selected the drill bit types based on the optimum values of ROP and drilling variables. Abbas et al. [18] also supported the notion of drill bit selection depending upon optimum values of ROP using ANN and GA. Here ANN was primarily utilized for the development of the objective function and GA for optimization of ROP objective function for the drill bit selection.

Several researchers have suggested the utilization of supervised classifiers as an alternative approach for the automatic selection of drill bit types based on previously drilled offset wells data. [24-29]. The historical drilling data of previously drilled wells have been provided for the training of Artificial neural networks (ANN) for screening the drill bit which is the first supervised classifier utilized for the selection of drill bits. Several works have supported the utilization of ANN classifier for the drill bit in place of conventional human experience-based drill bit selection. The reported applications never took consideration of various practical and computational aspects of bit selection that will hamper the performance of any supervised classifier in real field conditions such as imbalanced data condition generation, the impact of unstable formations, behavior of supervised classifiers with imbalanced data, etc.

Most of the reported research works are trained on balanced datasets with the fewer drill bit types. The successful applications of ANNs have shown that data-driven models have the potential for the automation of the bit selection process. However, none of them have considered the problem of imbalanced data that will naturally occur due to

the varying thickness of subsurface lithofacies. The actual field data contain the uneven distribution of data samples that result in a complex imbalanced multiclass classification problem during drill bit selection. This uneven distribution of training data samples affects the generalization capability of supervised machine models for unseen data and also makes them unreliable. Therefore, proper investigation of machine learning models is required to evaluate their effectiveness for the screening of drill bit types with complex offset wells data to provide more pragmatic solutions.

Here, the drill bit selection process has been formulated as a multiclass classification problem where diverse drill bit types have acted as class labels. In this work, two ensemble methods namely, AdaBoost and Random forest (RF) have been investigated for handling the complex multiclass imbalanced data problem associated with intelligent drill bit selection. These ensemble paradigms contain boosting techniques in their internal architecture which has been reported useful for solving the imbalanced data issues. They also reduce the bias and variance error associated with training data that provide a better generalization to the prediction results. These ensemble methods are combined with data resampling techniques to enhance their capability of dealing with imbalanced data. Additionally, the behaviour of four popular classifiers namely, K-Nearest neighbors classifier (KNC), Navies Bayes classifier (NBC), Multilayer perceptron (ANN), and Support vector classifier (SVC), have also been studied to select diverse bit types for drilling critically unstable geological formations. The primary motivation of this research work is to explore popular machine learning algorithms in the quest for higher drill bit selection accuracy and better generalization.

Further, it compares the performance of popular machine learning models for drill bit selection; and addresses the problem of imbalanced data which adversely affects the performance of machine learning models for the selection of drill bit. Moreover, the

impact of softer and unstable geological formations that produce critical training data for machine learning models has also been studied to select diverse drill bit types. The behavior of supervised classifiers has been considered for drill bit selection in critical formations. The future implications of automatic drill bit selection are also discussed in this chapter. The comparison of results has been performed to identify the best performing classifier among all the above-mentioned models. All the applied machine learning models have been trained and tested using Norwegian oil and gas field data. The data-related challenges associated with the drill bit selection process have also been discussed. This chapter also discusses issues related to intelligent classifiers and imbalanced petroleum data such as applicability issues, performance difficulties, performance evaluation parameters, and possible data-driven solutions. Overall, a comprehensive study of machine learning models has been performed to assess the challenges associated with the automatic drill bit selection process with practical field datasets.

5.2 Adaptation in ensemble methods for handling imbalanced petroleum data

In this study, oversampling and undersampling approaches have been utilized to generate balanced datasets for the training and testing of ensemble methods. All the data-related techniques used in this research work are briefly explained below.

5.2.1. Data Resampling Techniques

Several real-world problems involve imbalanced data issues where the distribution of samples varies from class to class. It is reported that the majority classes naturally dominate the minority classes during the training of supervised classifiers, which makes them biased and unreliable. However, machine learning models require balanced datasets for their best possible performance [118]. To overcome this imbalanced data

problem, two data sampling techniques were applied for generating balanced datasets for the classifiers to compensate for the ill effects of imbalanced data.

5.2.2. Oversampling

Oversampling technique increases the data samples in the minority class by duplicating the prevailing samples or producing synthetics ones [119]. This approach is widely applied for the generation of the balanced dataset for the training of supervised classifiers. Various oversampling techniques are available in the literature such as random over sampler, focus over sampler, synthetic minority over-sampling technique (SMOTE), etc. SMOTE is a widely applied technique for oversampling. Therefore, in this paper, the SMOTE technique has been applied for balancing the number of data samples for each class. This approach does not produce duplicate copies of existing data samples but synthesizes new ones. It takes the feature space samples for each class and combines them with the features of nearest neighbors [119,120].

5.2.3 Undersampling:

In the undersampling technique, the samples from the majority class are removed to decrease their data samples up to the number of minority class's data samples. This seems to be a straightforward approach for data sampling but is found suitable when the minority class has a sufficient amount of data samples [121]. There are various techniques applied for undersampling of the data samples such as Tomek links, edited nearest neighbors, random under sampler, etc. [120,121]. The random under-sampler technique has been used for the generation of a balanced dataset used in this study. It's a simple and fast method for the generation of a balanced dataset through random sampling from original data. Here, the number of samples in each class of the balanced dataset is predefined by the user. This technique selects bootstrap subsets from the original data for each class based on the user-defined value of samples [120]. It considers each class

independently in case of multiclass imbalance problems which is useful for sampling heterogeneous data having string values in samples [120,121]. Undersampling approach has been recommended only in big data conditions and may result in loss of important information during the removal of data samples from the majority class [120,121]. Both over and under-sampling approaches have limited benefits for handling imbalanced data at the data level, therefore, ensemble methods, having boosting techniques in their internal architecture, are also investigated at the algorithm level to compensate for the effects of imbalance.

5.2.4 Ensemble methods

Ensemble methods are multiple-learner systems that train and combine the outcomes of several supervised learners to produce the final outcomes for pattern recognition tasks [67]. The motivation for the integration of supervised machine learning models is to achieve higher prediction accuracy and improve the generalization ability of ensemble models. The ensemble approach has been reported to be efficient for reducing errors associated with the bias and variance of training data [122]. These methods are also found suitable for handling imbalanced data problems because they integrate boosting techniques within their internal architectures [122]. In this study, two ensemble methods namely, AdaBoost and Random forest are mainly studied for handling complex imbalanced data for the drill bit selection process and are briefly explained below.

5.1.2 AdaBoost

Freund and Schapire (1996) proposed AdaBoost ensemble technique based on the boosting paradigm [71]. It trains the base classifiers using random bootstraps data samples generated from original data and combines their decisions through a weighted majority vote. Initially, it assigns equal weights to all the training data samples. Further, weight adjustments are performed based on the misclassifications made by the initial base

classifier. Weights of misclassified data samples are increased in the next modified training dataset so that the chances of occurrence of misclassified samples will be increased in the next training dataset [71]. AdaBoost is particularly found supportive in handling imbalanced data problems [123]. The assignment of weights to bootstrap subsets is equivalents to resampling data space while combining upper and down sampling [123]. It has accuracy-oriented approach and focuses on the wrongly classified samples while increases the weight until it gets correctly classified. It provides a solution for imbalanced data problem at the data level equivalent to the resampling technique utilized for imbalance reduction. In this, under-sampling of majority classes is performed to produce the balanced dataset and is termed as under-sampled AdaBoost (USA). Figure 5.1. A shows a generalized workflow of the AdaBoost algorithm. The standard AdaBoost ensemble can be applied as given below.

Algorithm

- Produces bootstrap training subsets $(X_n = X_1, X_2 .... X_N)$ from original training data X and is associated with initial equal weights $(W_n = W_1, W_2 .... W_N)$.(n=1,2,3,4…N)
- Base classifiers $C_n(x)$ are trained using weighted training subsets.
- $(W_n X_n = W_1 X_1, W_2 X_2 .... W_N X_N)$ and determine error probability as $Error_n = \frac{1}{N}\sum_{i=1}^{N} W_i \beta_i$
- Where $\beta_i = \{^{1 \quad otherwise}_{0 \quad if\ sample\ correctly\ classified}$ and change in weights is given as $C_n = \frac{1}{2}\log\left(\frac{1 - Error_n}{Error_n}\right)$
- Update weights as $w_i^{n+1} = w_i^n \exp(C_n \beta_i^n)$ if the calculated error is between 0 to 0.5 (i=1,2,3…N) and renormalizes samples weights so that $\sum_{i=1}^{N} W_i^{n+1} = N$ otherwise initialize all the subsets' weights as 1 and repeat the above-given steps.
- Integrate the decisions of all the classifiers $C_n(x)$ by weighted majority voting rule as specified below.

  $\phi(x) = \arg\max \sum_n C_n \theta_{\text{sgn}}(C_n(x)), Y, where\ \theta_{i,j} = \{^{1\ i=j}_{0\ i\neq j}$ is known as the Kronecker symbol and Y is the class label.

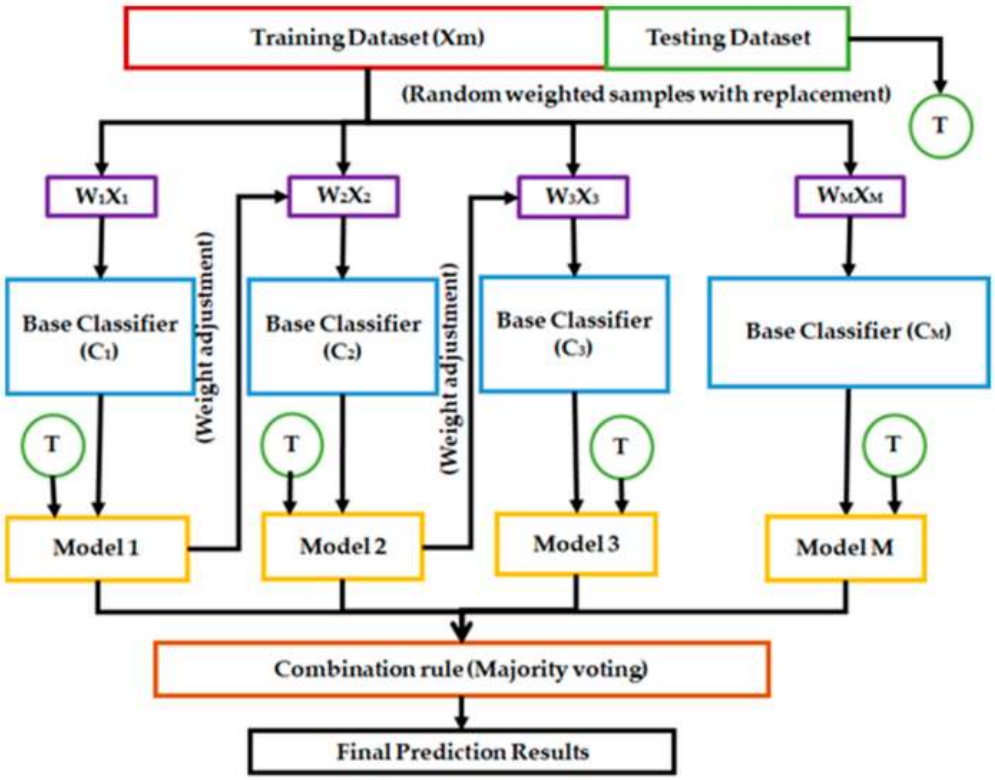


Figure 5.1. A generalized workflow of the AdaBoost algorithm.

5.2.2 *Random forest (RF)*

Breiman (2001) developed an RF algorithm by modifying the Bagging ensemble [124]. RF can be employed for resolving estimation, detection, and recognition-related problems. RF has certain peculiar merits over other classifiers such as computationally fast, few numbers of model parameters for tuning, easier evaluation of generalization error, the capability of handling high dimensionality, can be utilized for attribute selection, etc. [125]. RF is the assembly of decision trees in single ensemble architecture where each decision tree is generated from random training variables [126]. For the training of its decision trees, RF generates random bootstrap data subsets from training data with the replacement of data samples. The final estimation function is in the form of

a loss function that is required to be minimized [126]. All the feature space is available to the root node of the decision tree. Non-splitting nodes in the decision tree are called terminal nodes. The standard RF algorithm can be utilized for imbalanced data classification by adjusting the weight of each class while computing the impurity score for a selected split point [127]. The weights will be adjusted according to the inverse relationship with class frequencies in the training data [127]. This will shift the focus of RF on the minority class samples. This will result in the formation of a weighted class RF technique (WCRF) for the classification of imbalanced data [127].

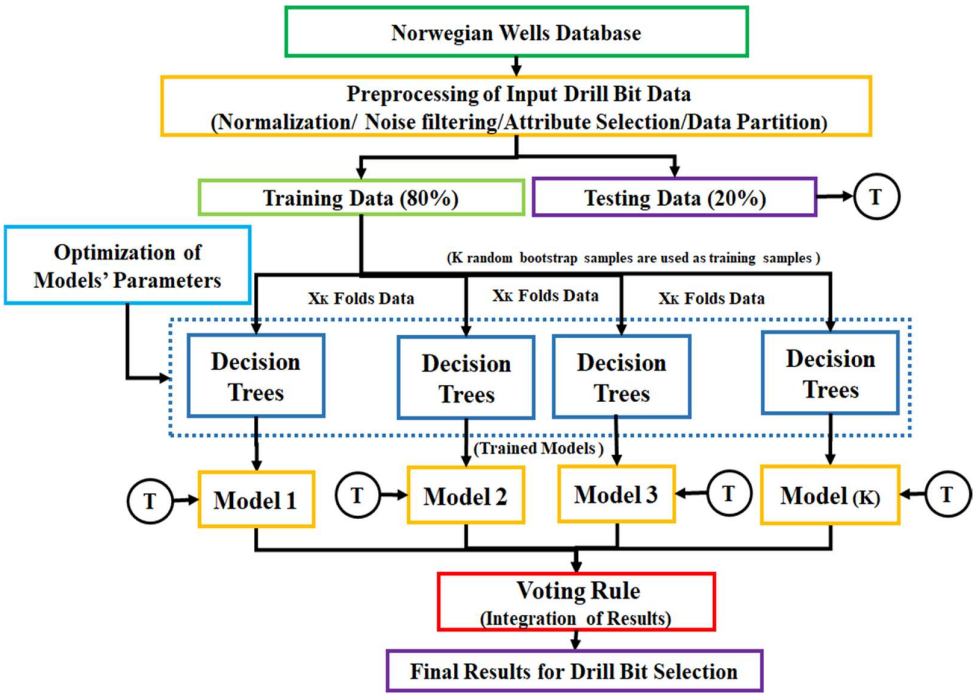


Figure 5.2 A generalized workflow of the drill bit selection process based on the Random forest algorithm.

The second approach that can be applied with RF is the bootstrap weighting approach. Here, the weight adjustment of a class is performed based on its distribution in every bootstrap sample in place of the whole training dataset [127]. Such a configuration of RF is known as RF with bootstrap class weighting (WBCRF). In the third approach, the majority of classes are randomly under-sampled in bootstrap samples to produce

balance datasets (USCRF). This adjustment will explicitly vary the class distribution inside the random bootstrap samples. Figure 5.2 shows a generalized workflow of the drill bit selection process based on the Random forest algorithm.

5.3 Methodology

In this study, drill bit selection has been formulated as a classification problem where diverse bit types have acted as class labels. The performance of four popular classifiers namely, KNC [128], NBC [129-130], MLP [131], and SVC [132,133], have been tested to select drill bits for the given values of operational field variables. AdaBoost and RF are also applied with the resampling technique for screening of drill bit. All the machine learning paradigms have been implemented through the open-source Scikit-Learn python package on the Anaconda platform. Python libraries have several merits over other prevailing platforms such as the capability of handling real-field input wells data, implementation of statistical tests to intelligent paradigms, visualization of results, self-explanatory user guides, community, and forums, etc.

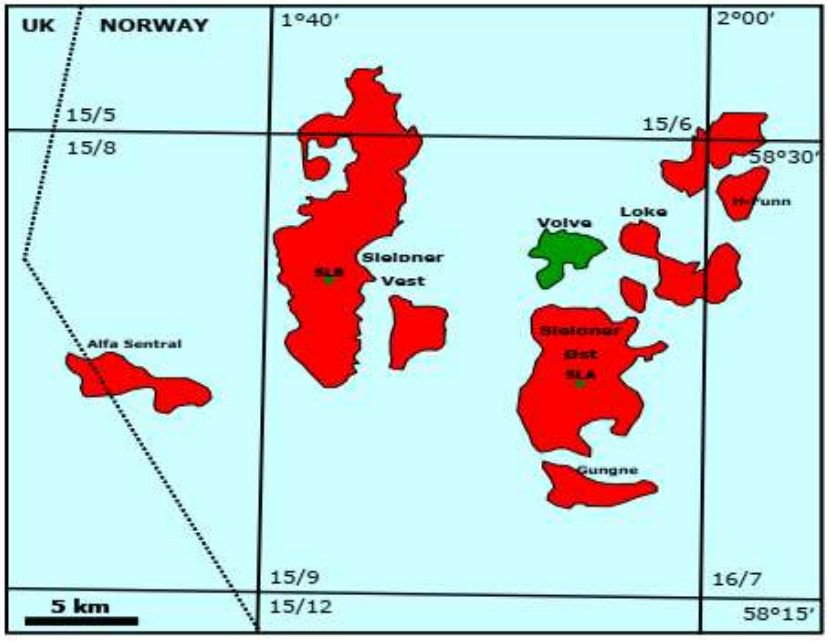


Figure 5.3 Geological location of Volve oil and gas field (Courtesy: Equinor website) [134].

5.3.1 A brief description of Volve field

This field is situated in the central part of the North Sea near the Norwegian Continental Shelf. It was discovered in 1993 and its production shut down in 2016 by its investors' companies. The ocean depth near the Volve field is in a range of 85 to 95 meters. This field contains Jurassic sandstone related to the Hugin formation reservoir. Figure 5.1 shows the location of Volve oil and gas field in the North Sea. The depositional environment of this reservoir is analyzed as tidal to the shallow estuary [134]. The sandstones of the Hugin reservoir contain high contents of quartz and a medium to low range of mica and clay minerals. Various faults can also be found in Hugin formation due to salt and Jurassic extensional tectonics [134]. Draupne formation acts as a worthy spring for oil production due to its organic-rich claystone layer. Smectite contents and argillaceous clay are found in large quantities in Hordaland shales which may be the cause for the higher formation pore pressure and its instability. This formation is not recommended for drilling high angle wellbore due to its easy collapse chances [134]. Balder formation comprises crumbly tuff content which has been reported as the primary reason for mud losses and washouts. The presence of crumbly tuff in Blader formation also decreases its fracture gradient that leads to the instability of formation. Drilling operations in the Sola formation have also suffered from several issues such as a tight hole, collapses, etc. The average properties expected from this Hugin reservoir are as follows: porosity (0.2), permeability (910), water saturation (0.23), and shale volume (0.17) [130]. Geosteering was particularly utilized to increase the extent of the reservoir linking to various fault blocks. Table 5.1 contains the geological prognosis of Well 15/9-F-12 considered under this study.

5.3.2 Data description

The dataset utilized for training and testing of machine learning models belonged to Norwegian Volve oil and gas fields. These data are available online and can be downloaded from the website of the Equinor oil and gas company. The field data of the fourteen Volve oil and gas wells were made public for academic and research purposes in 2018 [134]. Eight wells data were downloaded from the Equinor company website for the testing of machine learning models considered in the study for drill bit selection namely, 15/9-F-4, 15/9-F-5, 15/9-F-7, 15/9-F-9, 15/9-F-10, 15/9-F-11, 15/9-F-14, and 15/9-F-15 [30]. These wells were planned to maximize the production of hydrocarbon from the Hugin formation. Generally, the production wells in the Volve field were multilateral in nature, however, observation and injection wells were in J-shape trajectory [134]. The total number of data points extracted from the final drilling reports of eight wells is shown in Table 5.2. Table 5.3 contains the statistical description of various variables utilized for drill bit selection.

Table 5.1 Geological prognosis of Well 15/9-F-12 under study [134].

Group	Formation	Depth (m)	Description
Nordland	Utsira Top	892	Grey claystone, a stringer of sand, and siltstone.
	Utsira Base	1084	Well sorted sandstone, minor silt, and limestone stringers.
Hordaland	Skade Top	1259	Claystone, minor limestone/dolomite stringers.
	Skade Base	1347	Medium-grained sorted sandstone.
	Grid Top	2179	Fine-grained sandstone.
	Grid Base	2245	Fine-grained sandstone.
Rogaland	Balder Top	2317	Colored claystone, partly tuffaceous, and limestone stringers.
	Sele Top	2374	Claystone and limestone stringer.
	Lista Top	2445	Non-calcareous claystone and minor limestone

			stringers.
	Ty Top	2531	Fine to medium sandstone some interbedded claystone, siltstone, and limestone stringers.
Shetland	Ekofisk Top	2698	Limestone with traces of claystone and sandstone.
	Tor Top	2715	White limestone with traces of claystone.
	Hod Top	2839	Limestone along with gluconate.
	Blodoeks Top	2944	Marl, argillaceous laminations, and gluconates in parts.
	Hidra Top	2972	Off-white firm limestone.
Cromer Knoll	Roedby Top	2981	Marl along with argillaceous laminations
	Aasgard Top	3001	Interbedded limestone and marl with minor claystone and siltstone.
Viking	Draupne Top	3036	Organic-rich claystone, micaceous, carbonaceous with traces of pyrite.
	Heather Top	3086	Claystone with limestone stringers.
Vestland	Hugin Top	3094	Sandstone and rare claystone stringers.
	Sleipner Top	3266	Sandstone, grey claystone, and layers of coal.

Table 5.2 Details of data samples extracted from the Final drilling reports of Norwegian wells.

S. No.	Well No	Data samples Acquired	Classification
1.	F-4	548	Injector well
2.	F-5	721	Injection Well
3.	F-7	187	Production well
4.	F-9	180	Production well
5	F-10	718	Observation/Production well
6.	F-12	631	Production well
7.	F-14	711	Production well
8.	F-15	616	Observation well
	Total	4312	

The input data extracted from the final drilling reports of eight wells contained a variety of sensor-measured variables. The downloaded data are available in pdf format that has been later converted to excel file format for ease of handling. This type of data normally contains issues such as noise, redundant attributes, missing or garbage values, etc. that are required to be cleaned before uploading into machine learning models, otherwise, it will affect models' performance. These input variables will act as predictor variables and unique IADC codes of drill bits will act as class labels. However, IADC bit numbers cannot be directly utilized for class labels instead coded to newer class labels as shown in Table A-1. Loken et al. [135] calculated additional parameters that are based on the natural interactions of conventional drilling variables such as mechanical specific energy (MSE), depth of cut (DC), drill Bit aggressiveness (DBA), and D-exponent (D-EXP) [135]. These interaction drilling variables have been extensively stated in several research works [135-138]. The additional interaction drilling parameters have been calculated as given below.

$$MSE = \frac{WOB}{Area_{bit}} + \frac{120 * \pi * RPM * TQ}{Area_{bit} * ROP} \tag{5.3}$$

$$DC = \frac{ROP}{5 * RPM} \tag{5.4}$$

$$DBA = \frac{36 * TQ}{WOB * BD} \tag{5.5}$$

where WOB is the weight on the bit in tons, RPM is round per minutes in rpm, TQ is torque in kN/m, ROP is the penetration rate of a drill bit in m/hr, $Area_{bit}$ is the area of a drill bit in inch square, and BD is drill bit diameter in inch.

Table 5.3 Statistical details of collected drilling data of eight wells used in this study.

S. No.	Input Variables	Range	Units
1.	Measured Depth (DT)	45-3785	m
2.	True Vertical Depth (TVD)	150-3244.36	m
3.	Rate of Penetration (ROP)	1.62-205.01	m/hr
4.	Weight on bit (WOB)	-7.27-51.57	tons
5.	Rounds per minutes (RPM)	6-311	rpm
6.	Torque (TQ)	-28.53-96.14	kNm
7.	Standpipe pressure (SPP)	3-389.2	bar
8.	Mud weight (MW)	0.99-1.47	s.g.
9.	Flow Rate in (FR)	432-5345	l/min
10.	Total Gas (TG)	0-10.6	%
11.	Bit type (BT)	1-19	--
12.	Bit Size (BS)	8.5-26	inch
13.	D-exponent (DEXP)	0.26-1.55	--
14.	Total flow Area (TFA)	0.663-1.51	$inch^2$
15.	Mechanical Specific Energy (MSE)	2213.0-85127.7	psi
16.	Depth of Cut (DC)	2.5-5.06	m/rev
17.	Drill bit Aggressiveness (DBA)	2.07-6.13	--

5.3.3 Imbalanced data problem

The diverse input variables utilized for the training of machine learning models were obtained from drilling the subsurface lithofacies. These lithofacies have naturally existing subsurface rock layers that occur in a random pattern along with the depth of the geological formations. The thickness of subsurface layers also unevenly varied at different depths of geological reservoir. Different subsurface rock requires diverse drill bits for efficient drilling operations. The thick rock layers generate a large amount of drilling data samples that can be used for classifying associated bit types. However, drilling of thin layer intervals produces a lesser amount of drilling data samples that are available for the training of machine learning models. This results in uneven distribution

of drilling data samples which affects the performance of each supervised classifier. Figure 5.3 shows the number of data samples associated with each bit type available in input drilling data. Imbalanced data samples are difficult to classify and adversely affects the performance of the supervised classifiers algorithm [139]. It can be seen in Figure 5.2 that BT 6 (34), BT 7 (13), BT 11(26), BT 12 (52), and BT 16 (13) contain an extremely lesser number of data samples as compared to other classes. This results in imbalanced data conditions that will automatically jeopardize the whole data-driven bit selection process. A single supervised classifier generally fails to perform adequately with imbalanced data conditions. Popular supervised paradigms such as KNC, ANNs, SVM, etc. become biased for majority classes while ignoring the smaller classes. However, overall classification accuracy will be reported high in case of imbalanced data conditions. To tackle imbalanced data conditions, certain modifications have been suggested by the researcher that is rarely applied in the petroleum domain. This problem can be handled at two levels namely, the data level, algorithm levels. Four major solutions can be applied for handling imbalanced data conditions namely, (a) resampling (b) boosting (c) adaptive algorithm (d) cost-sensitive learning [119]. In this study, boosting, and resampling have been selected for handling the imbalanced data condition occurring during the drill bit selection process. Drill bit selection is a complex multiclass classification problem that requires strong classifier paradigms for its classification. AdaBoost and RF, are the two strong ensemble classifiers that incorporate boosting technique within their internal architectures, can be combined with resampling to handle imbalanced data efficiently.

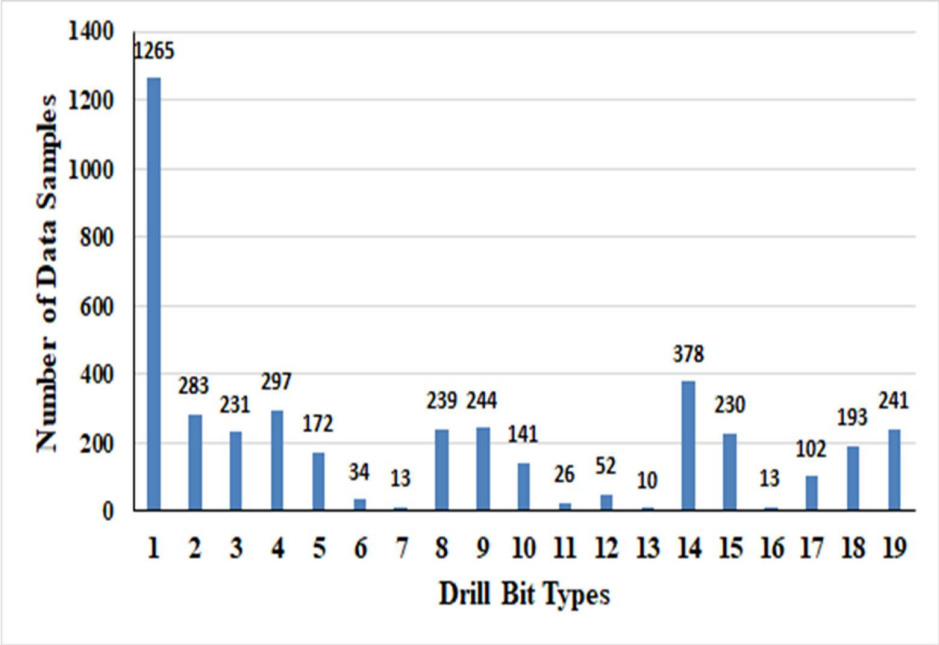


Figure 5.4 Number of data samples available in real field drilling data for each bit type.

5.3.4 Data Preprocessing

Several petroleum researchers have supported the idea of preprocessing real-field wells data before uploading it to machine learning models [140]. Preprocessing of input data helps to enhance the prediction accuracy of machine learning models and reduces the chances of errors. Primarily, data resampling was executed to remove the data samples having null and garbage numeric values. Further, the normalization of input data was performed to diminish the impact of larger values on the smaller ones. Mustaffa and Yusof (2010) compared the normalization techniques and reported that Min-max normalization is particularly suitable for those paradigms which have distance measurement or optimization in their internal design such as K-NN, NBC SVC, etc. [141]. Min-max normalization is also recommended for those input data which don't follow Gaussian distribution which applies to operational field data acquired in this study. The data can be normalized as given below.

$$X_i^{Norm} = \frac{X_i - \min(\mathrm{X})}{\max(\mathrm{X}) - \min(\mathrm{X})} \tag{5.6}$$

where X_{Max} and X_{Min} are maximum and minimum values of operational variables. This technique also ensures that each input variable is uniformly scaled down on the same level.

5.3.5 Noise reduction

The problem of noise in the sensor recorded data has been reported in several research works that affect the performance of machine learning models. Conventional noise filtering techniques such as Fourier transform, moving average, SG filters, etc. are found to be less effective for the removal of noise contents from drilling data [142]. In this study, wavelet filters have been utilized for the denoising of drilling data which is a popular noise filtering technique [143]. In wavelet transform, spare representation of drilling data has been generated to concentrate whole data features into large magnitude wavelet coefficients. The smaller value coefficients are considered as noise components. Later, these smaller coefficients are eliminated during the noise filtering process. The wavelet transform of input data can be given as:

$$W_T(\mathrm{h},\mathrm{k}) = \frac{1}{\sqrt{h}} \int_{-\inf}^{+\inf} \mathrm{T}(t)\psi\left(\frac{t-k}{h}\right) dt \tag{5.7}$$

where k is the scaling factor, h is the factor of expansion and $\psi(\mathrm{t})$ is the wavelet basis function. Further, an inverse wavelet transform can be taken to reconstruct the original waveform of input data. Inverse wavelet transform can be given as:

$$In(t) = \frac{1}{w_\psi} \int_{-\inf}^{+\inf} \int_{-\inf}^{+\inf} W_T(\mathrm{h},\mathrm{k})\psi\left(\frac{t-k}{h}\right) dhdk \tag{5.8}$$

where w_ψ the wavelet factor, h is is the factor of expansion, k is the scaling factor and $\psi(\mathrm{t})$ is the wavelet basis function. The lower wavelet coefficients are removed in the

noise filtering process; however, the properties of the original data are still preserved. In this study, the Haar wavelet has been used for filtering noise components from drilling data. Several research works have also supported the utilization of the Haar wavelet for denoising drilling data [143]. The noise contents are found to be large in drilling data because surface installed sensors have high chances of exposure from surrounding noise. Figure 5.4 shows the denoising of the WOB variable using the 1-D Wavelet filtering technique.

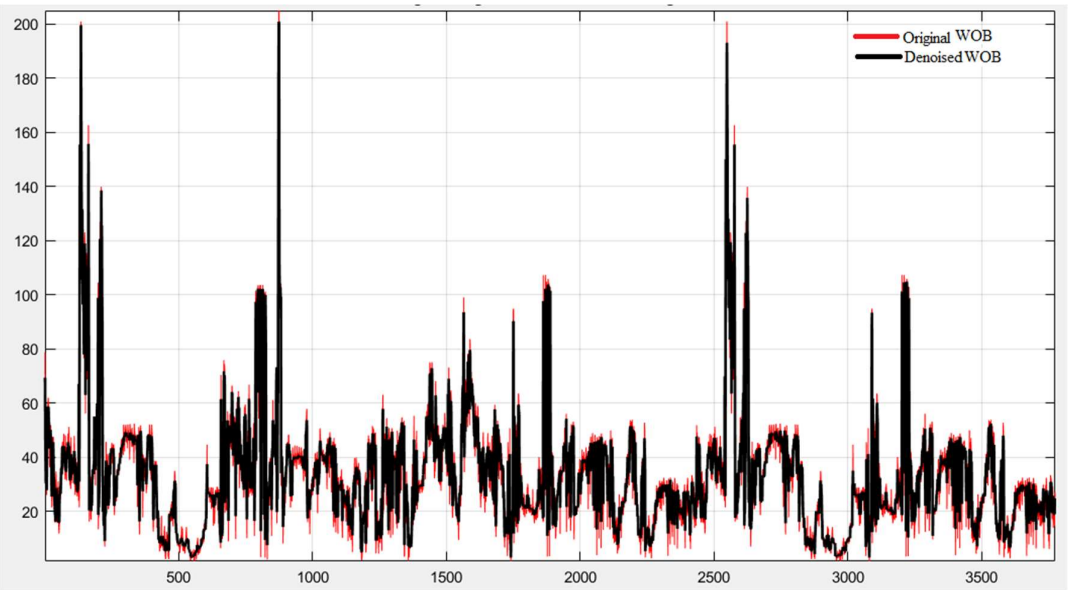


Figure 5.5 The denoising of WOB variable using the 1-D Wavelet filtering technique.

5.3.6 Attribute Selection

Drilling data contain redundant predictor variables that will increase computational cost and time during the training phase of machine learning models. Various feature extraction and attribute selection techniques such as principal component analysis, Fisher discriminant analysis, univariate attribute selection, relief algorithm, correlation heat-map, etc. are available in the literature to eliminate the redundant variables or attributes from input data. The availability of only relevant features or attributes in

training data enhances model accuracy, reduces the influence of noise, and training time of machine learning models. In this work, the forest of decision trees based on feature importance has been calculated for the identification of important drilling variables for the bit selection task. It allocates ranks and weights to input drilling variables depending upon their contribution to the classification task. Figure 5.4 shows predictor variables arranged according to their ranks and weights assigned through a Forest of decision tree-based algorithm. Out of sixteen input (predictor) variables, BS and TFA were recognized as high contributing variables for bit selection whereas TG contribution was the lowest as shown in Figure 5.5. Finally, TG was eliminated from the training datasets due to its redundant nature.

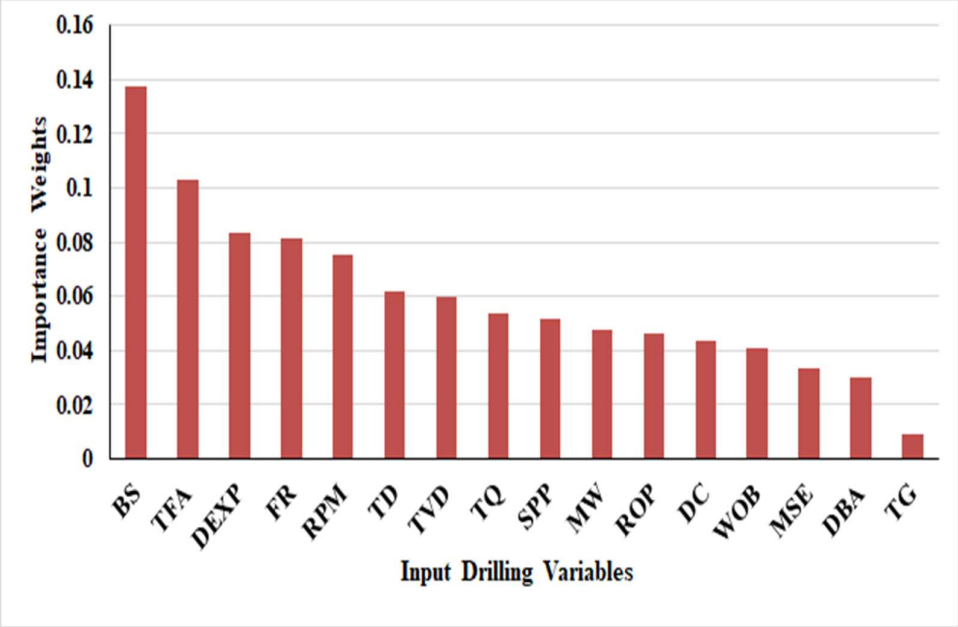


Figure 5.6 Importance of input drilling variables for the selection of drill bit type.

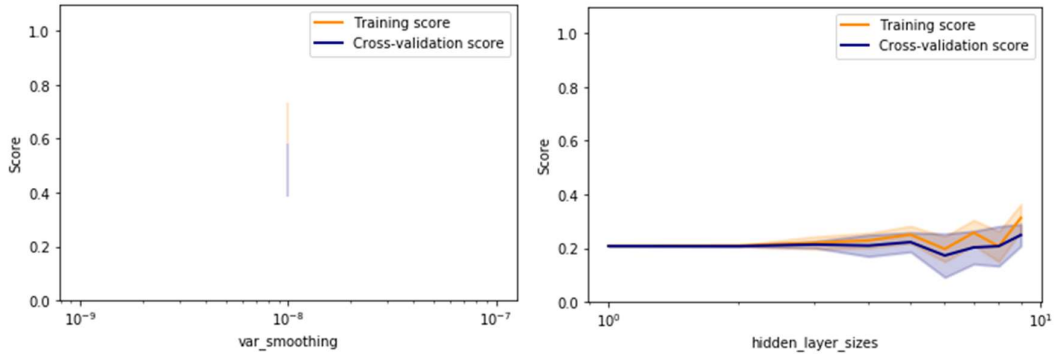
5.3.7 Model training with parameter optimization

The processed drilling data were split into training and testing using a cross-validation technique. Cross-validation of input data is done to avoid the problem of overfitting and underfitting of machine learning models. Several schemes of cross-

validation are available in the literature for data partitions such as k-fold, stratified k-fold, leave one out, leave P-out, hold out, etc. K fold cross-validation was primarily utilized for data partition because it effectively reduces variance error associated with input data [144]. 10-fold cross-validation (10-FCV) splits the training data into K=10 subsets where (K-1) subsets are used for training the machine learning models and Kth for their validation. Iterations will continue until all the subsets have acted at least once as a validation set. The final results of machine learning paradigms are calculated by averaging the accuracies obtained in each iteration. After 10-fold cross-validation, training and testing of machine learning algorithms have been done to evaluate their performances. Generally, machine learning models are prone to overfitting and underfitting conditions. Thus, additional validation curves are generated to identify stable regions existing in search ranges of various models' parameters.

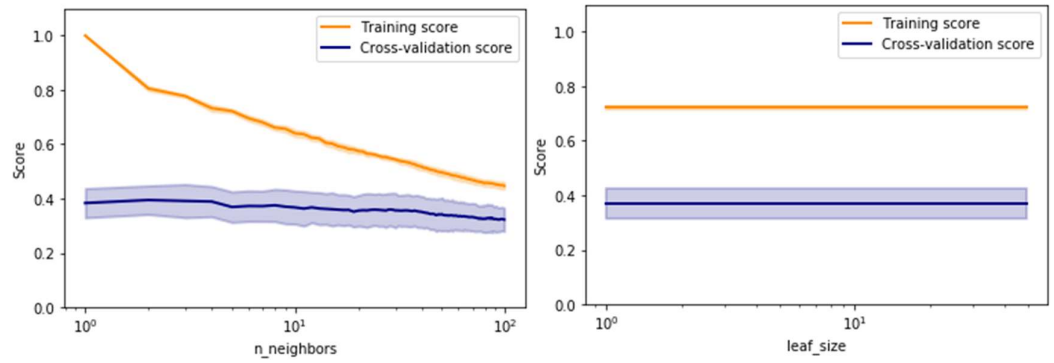
Underfitting conditions, training, and validation scores of machine models will be recorded at lower values. In the case of overfitting, training scores are reported to be high in combination with low validation scores. To avoid overfitting and underfitting conditions, models' parameters are needed to be optimized within the stable regions where no dramatic change of training and validation scores take place as shown in Figures 5.5, 5.6, 5.7, 5.8, and 5.9. The optimization of the model's parameters has been performed using the grid search technique which is a popular parameter tuning algorithm in the petroleum domain. The search ranges and optimum values of models' parameters are shown in Table 5.4. Figure 5.6 depicts the validation curves generated for the Smoothening parameter of NBC and the number of hidden layers of MLP. Figure 5.7 shows validation curves generated for the important parameters of the KNC classifier viz. the number of neighbors and leaf size. Figure 5.8 contains validation curves generated for the regularization and gamma parameters of SVC. Figure 5.9

illustrates validation curves for four important parameters of RF viz. number of estimators, maximum depth of decision tree, minimum samples needed at a leaf node, minimum number of samples needed for splitting the node. Figure 5.10 shows training error minimization versus the number of iterations for SVC classifier.



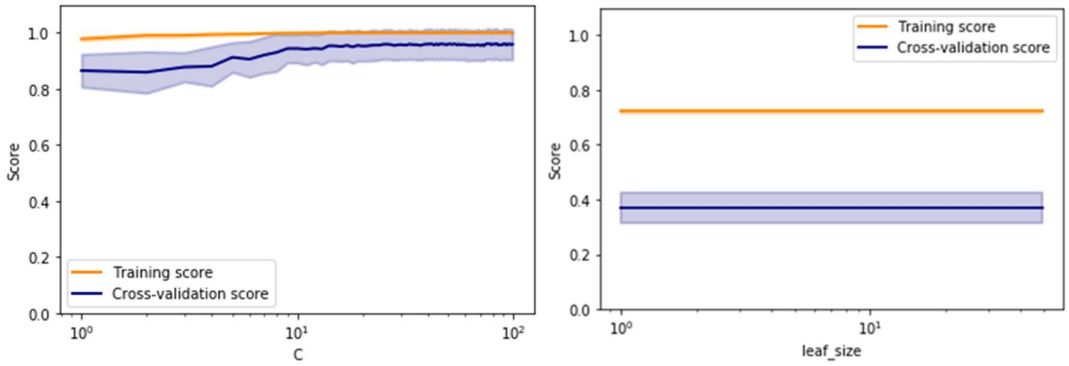
(a) (b)

Figure 5.7 Validation curves generated for NBC and MLP (a) NBC Smoothing, parameter and (b) MLP number of hidden layers.



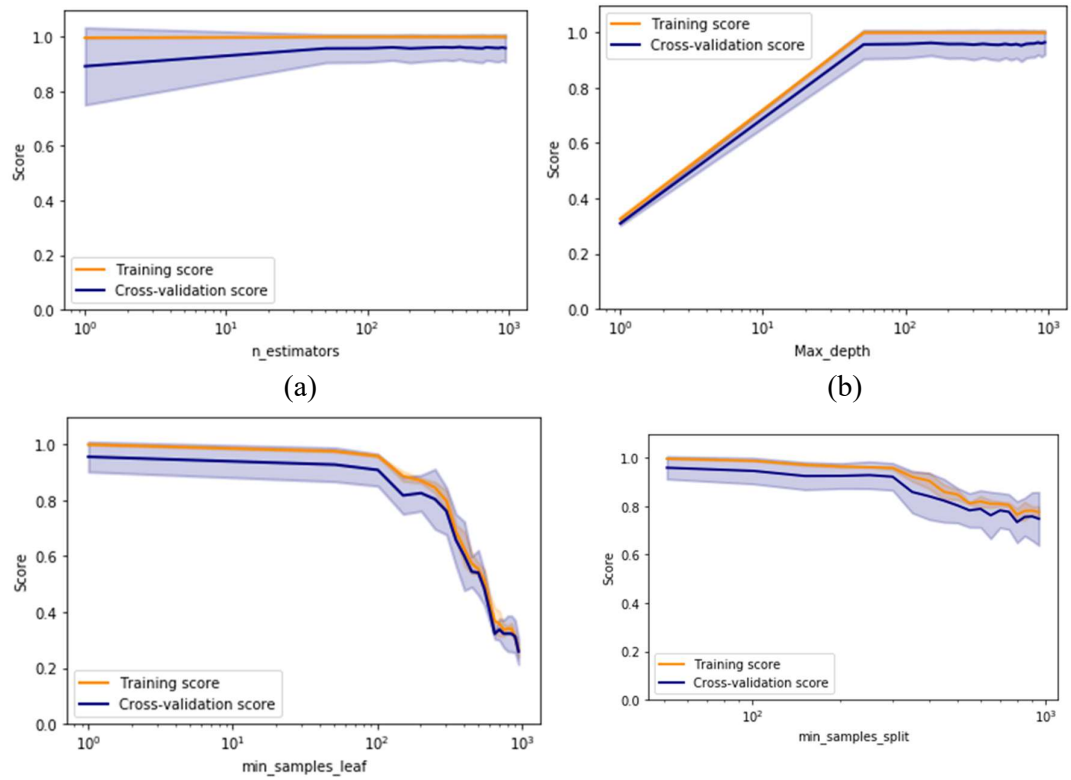
(a) (b)

Figure 5.8 Validation curves generated for the important parameters of KNC classifier (a) Number of neighbors. (b) Leaf size.



(a) (b)

Figure 5.9 Validation curves generated for two important parameters of SVM classifier (a) Regularization parameter C. (b) Gamma parameter.



(a) (b)

(c) (d)

Figure 5.10. Validation curves generated for four important parameters of RF (a) Number of estimators. (b) Maximum depth of decision tree. (c) Minimum samples needed at leaf node. (d) Minimum number of samples needed for splitting.

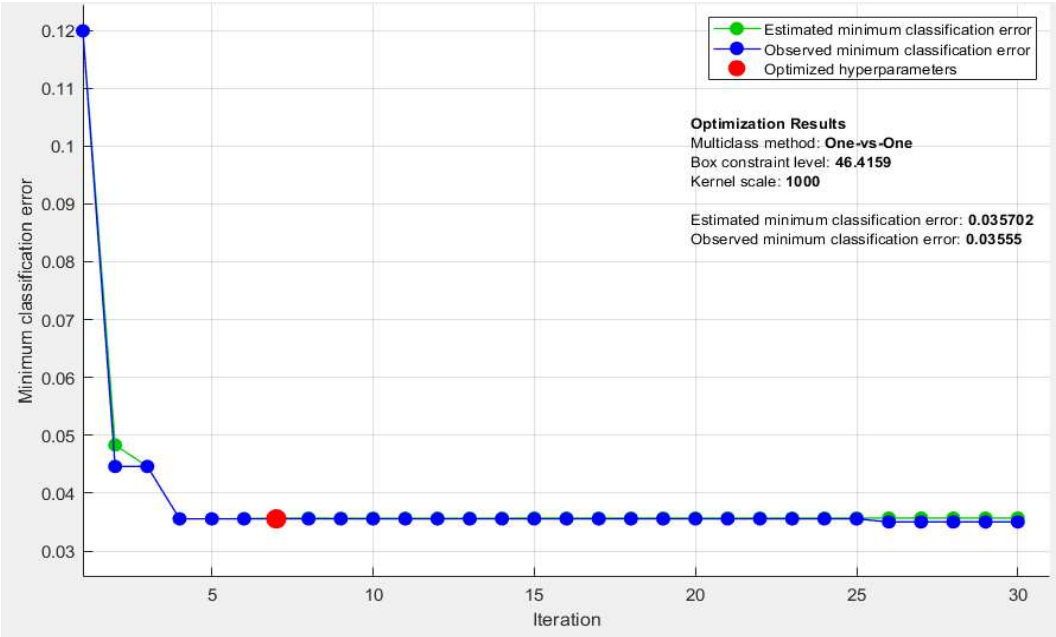


Figure 5.11. Minimization of classification error plot generated during the training phase of SVC using the grid search technique.

Table 5.4 Optimum values of various models' parameters utilized in this study.

Paradigms	Model Parameters	Search Range	Optimum Value
MLP	Learning rate	0.0001-0.5	0.001
	Maximum number of iterations	100-1000	500
	Neurons in the hidden layer	0-1000	20
	Activation function	identity, logistic, tanh and relu	relu
	Solver	adam, lbfgs, and sgd	adam
KNC	Number of Neighbors	1-10	5
	Weight	uniform /distance	Uniform
	Algorithm	Auto, ball_tree, kd_tree,	Auto

		brute	
	Leaf size	1-100	40
NBC	Var_smoothing	1E-9 to 1E-1	1E-8
SVR	Penalty parameter (C)	0.1-10000	100
	Kernel type	Linear, polynomial, Gaussian	Gaussian
	Gamma parameter (y)	0.01- 10	5
RF	Number of estimators	1-1000	100
	Maximum number of iterations	10-1000	1000
	Minimum samples for split an internal node	1-20	2
	Maximum depth of the tree	1-1000	'None'
	Minimum leaf samples	0-25	1
AdaBoost	Number of estimators	1-1000	100
	Base estimator	Any supervised paradigm	Decision tree
	Learning rate	0.1-1	0.1
	Boosting algorithm	SAMME/SAMME.R	SAMME.R

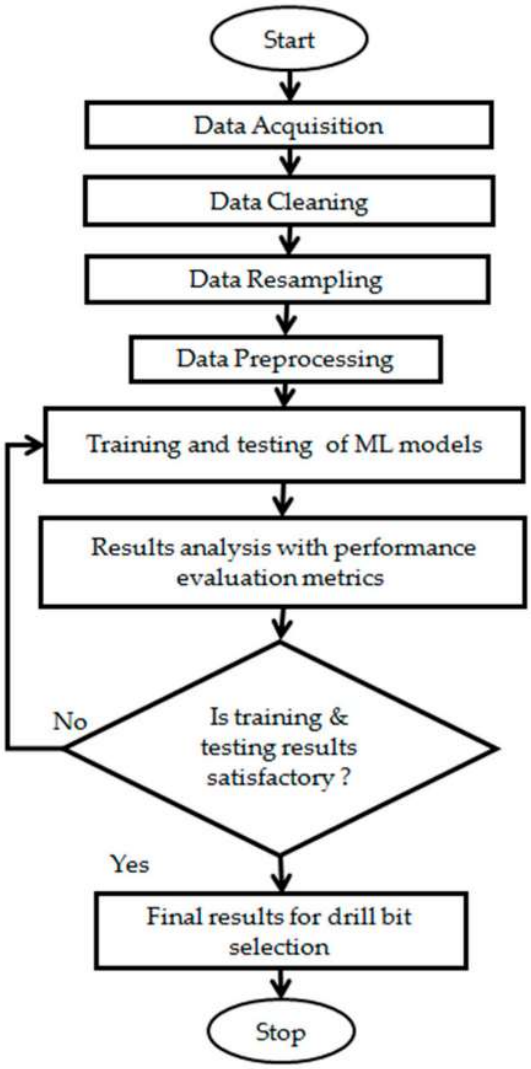


Figure 5.12 A generalized workflow of the drill bit selection process based on the machine learning algorithm.

5.3.8 Performance evaluation metrics

The evaluation metrics play an important role in the assessment of the supervised classifier's performance. Conventionally, accuracy was considered a reliable performance evaluation parameter. However, it becomes unreliable in case of imbalanced data conditions where it does not account for smaller classes. Thus, additional statistical indicators namely, precision, recall, G-means, Matthew coefficient

of correlation (MCC), and F1score are also calculated to determine the performance of classifiers as given below.

$$\text{Accuracy}=\frac{\text{Correctly identifed data samples}}{\text{Total number of data samples}} \tag{5.9}$$

where Accuracy is a widely applied parameter for the performance evaluation of intelligent classifiers.

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{5.10}$$

where FP is correctly classified data samples other than a particular class, and TP is correctly classified data to a particular class.

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{5.11}$$

where FP is correctly classified data samples other than a particular class, FN is the number of samples wrongly classified to a particular class, TP is correctly classified data to a particular class. Higher values of precision and recall have been expected from every classifier.

$$\text{F1}_{\text{scores}}=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{5.12}$$

where $F1_{scores}$ values have been estimated to ensure the authentication of precision and recall results. This parameter is widely applied in the area of information retrieval. All the above-mentioned parameters are influenced by data imbalance issues and may mislead classification results [142]. Therefore, MCC and G-mean have been calculated to ensure the reliability of the accuracy parameter.

$$MCC=\frac{SC\times TS-\sum_{k}^{K}PC_k\times TC_k}{\sqrt{\left(TS^2-\sum_{k}^{K}PC_k^2\right)\times\left(TS^2-\sum_{k}^{K}TC_k^2\right)}} \tag{5.13}$$

where PC_K is the number of iterations in which K class has been predicted, TC_K is the number of iterations in which K class is correctly predicted, SC is the number of data samples correctly classified and TS is the number of all the data samples considered in the classification task. MCC parameter has been utilized for ensuring that the classification results are reliable and unaffected by data imbalance issues [142]. G-mean is also a performance indicator parameter that is not affected by data imbalance. Kubat et al. (1997) proposed G-mean as given below [145].

$$G - mean = \sqrt{TP_{rate}.TN_{rate}} \tag{5.14}$$

where TP_{rate} is the true positive rate and TN_{rate} is the true negative rate. Both of these parameters are expected to be high concurrently for good classification results. Figure 5.10 shows a generalized workflow of the drill bit selection process based on the RF algorithm.

5.4 Results and discussion

This section discusses the results obtained while selecting different types of the drill bit through machine learning models. Two ensemble methods namely, AdaBoost and RF, have been investigated for handling the complex multiclass imbalanced data problem associated with the intelligent drill bit selection process. Validation curves have been generated to identify the stable regions existing in ranges of various models' parameters as shown in Figures 5.5, 5.6, 5.7, and 5.8. A detailed description of drill bit types has been provided in Table A-1 of the appendix. Two data-driven experimental scenarios have been simulated to test the intelligent bit selection approach. In the first experimental scenario, machine learning models were trained and tested on the combined dataset obtained from eight wells using 10-FCV. The input data utilized for training and testing of various machine learning models contain uneven training

samples belonging to various classes as shown in Figure 5.2. In the case of imbalanced data, classification accuracy becomes unreliable and unfit for the performance evaluation of machine learning models. Thus, average values of recall, precision, F1score, G-mean, and MCC, have been determined to examine the overall performance of various machine learning models. Table 5.5 shows the classification performance of standard classifiers for bit selection. It can be observed from Table 5.5 that the performance of NBC and KNN are the lowest among all the other classifiers. NBC has failed to learn about hidden dependencies or patterns among diverse variables present inside the training data samples related to smaller classes. Smaller classes have existed sparsely in the training data space due to data scarcity which also harms the performance of KNC with testing data.

MLP model has been trained on 70% of input data along with 15% for validation and 15% as testing data. The optimum number of neurons in the hidden layer of MLP was determined based on minimum training error after several iterations as shown in Table A-2 in the appendix. MLP is three layers of a popular neural network with a backpropagation (BP) paradigm in its internal architecture for the training phase. BP trains the MLP network iteratively by adjusting the weights associated with each variable present inside training data. The weights adaptation is dependent on the length of the gradient vector calculated for error minimization in the training phase. The expected length of the gradient vector is dependent on the number of samples present for each class. During imbalanced data conditions, majority classes dominate the whole error minimization process during the training phase and produce larger errors for minority classes. Thus, the performance of MLP is adversely affected by the imbalance condition. MLP, NBC, and KNN classifiers are prone to become biased for majority classes in imbalanced data conditions. However, MLP has emerged as the second-best

performing single supervise classifier for drill bit selection followed by SVC in the first place as shown in Table 5.5. SVC is known to have some level of immunity for imbalanced data condition but become biased to majority classes in critically high imbalance condition. All of the above said supervised models fail to provide proper generalization and become unreliable for the selection of drill bit type. Therefore, ensemble methods have been investigated for drill bit selection to achieve better model generalization.

Table 5.5 The performance of machine learning classifiers for drill bit selection in the first experimental scenario.

Classifiers	**Training Accuracy**	**Testing Accuracy**	**Precision**	**Recall**	**$F1_{score}$**	**MCC**	**G-mean**
NBC	56.15	55.001	0.639	0.550	0.517	0.516	0.61
KNC	63.00	60.00	0.623	0.61	0.60	0.566	0.54
MLP	72.12	71.06	0.711	0.711	0.708	0.678	0.70
SVC	0.83	0.82	0.79	0.78	0.81	0.83	0.81
AdaBoost	0.96	0.90	0.90	0.90	0.91	0.90	0.90
Random forest	0.97	0.91	0.92	0.92	0.92	0.91	0.91

AdaBoost and RF are the two ensemble methods that have been utilized for handling the imbalanced offset wells data for drill bit selection. RF has achieved higher testing accuracy than AdaBoost for bit selection as shown in Table 5.5. Although, ensemble methods have given much better results as compared to single supervised classifiers still they are affected by the imbalance conditions. Thus, both of these

techniques were modified to enhance their capability of imbalanced data classification. AdaBoost and RF have been combined with an undersampling technique that reduced the data samples from majority classes to make the whole dataset balance. However, this approach has degraded the performance of both ensemble methods as their classification accuracies are heavily dependent upon the majority class samples as shown in Table 5.6. In imbalanced data condition, classifiers normally ignore smaller classes as fewer data samples are available during the training phase that makes difficult for intelligent paradigms to learn and identify any hidden pattern between variables. This technique may produce satisfactory results when a reasonable amount of data samples is present in the smaller classes.

Table 5.6 Modified ensemble classifiers for the classification of imbalanced drilling data.

Modified Ensembles	**Training Accuracy**	**Testing Accuracy**	**Precision**	**Recall**	**$F1_{score}$**	**MCC**	**G-mean**
Under Sampling AdaBoost (USA)	0.95	0.74	0.80	0.70	0.75	0.88	0.89
Under sampling RF (USRF)	0.80	0.77	0.70	0.60	0.64	0.70	0.72
Weighted Class RF (WCRF)	0.93	0.92	0.93	0.92	0.92	0.92	0.96
Weighted Bootstrap Class RF (WBCRF)	0.93	0.93	0.93	0.92	0.92	0.92	0.97

In the second approach, classes are assigned weights to focus the classification operation on the samples of minority classes. The weights will be adjusted according to the inverse relationship with class frequencies in the training data. This will result in the formation of a weighted class RF classifier (WCRF) for the classification of imbalanced

data. It can be observed from Table 5.6 that the WCRF has given a better performance than standard RF for drill bit selection in terms of precision, recall, MCC, and G-mean. This indicates that WCRF has a greater generalization ability than standard RF. In the third approach, separate weight adjustment of classes has been performed based on its distribution in every bootstrap sample in place of the whole training dataset. Such a configuration of RF is known as RF with a bootstrap class weighting (WBCRF) classifier. This classifier contains the benefits of both data resampling and weighting techniques that are quite useful for compensating the impact of imbalance conditions. Training and testing results of WBCRF have shown a slight performance improvement when compared with WCRF. Both these approaches have provided higher MCC and G-mean values than the standard RF paradigm as shown in Table 5.6 and Figure 5.12. WBCRF technique has given the best classification results as compared to other classifiers considered in this study.

In the second experimental scenario, three subsets have been created from eight wells data containing data samples belonging to 17.5”, 12.25”, and 8.5” individual wellbore sections. All the earlier applied classifiers have been trained and tested on these subsets. The performance of conventional classifiers has been evaluated in terms of precision, recall, G-mean, and F1 score for every BT to understand the effect of imbalanced data. Tables 5.7, 5.8, and 5.9 contain drill bit selection test results for the 17.5” section subset. It can be observed from the abovementioned tables that bit type 16 (minority class) is hard to predict due to a lesser number of data samples available during the training phase. KNC, SVC, and MLP have also failed to identify BT 16 due to scarcity of data samples as shown in Tables 5.7, 5.8, and 5.9. Thus, G-mean values are recorded to be zero for SVC, KNC, and MLP as it is clear identification of the development of unreliable biased majority class classifiers.

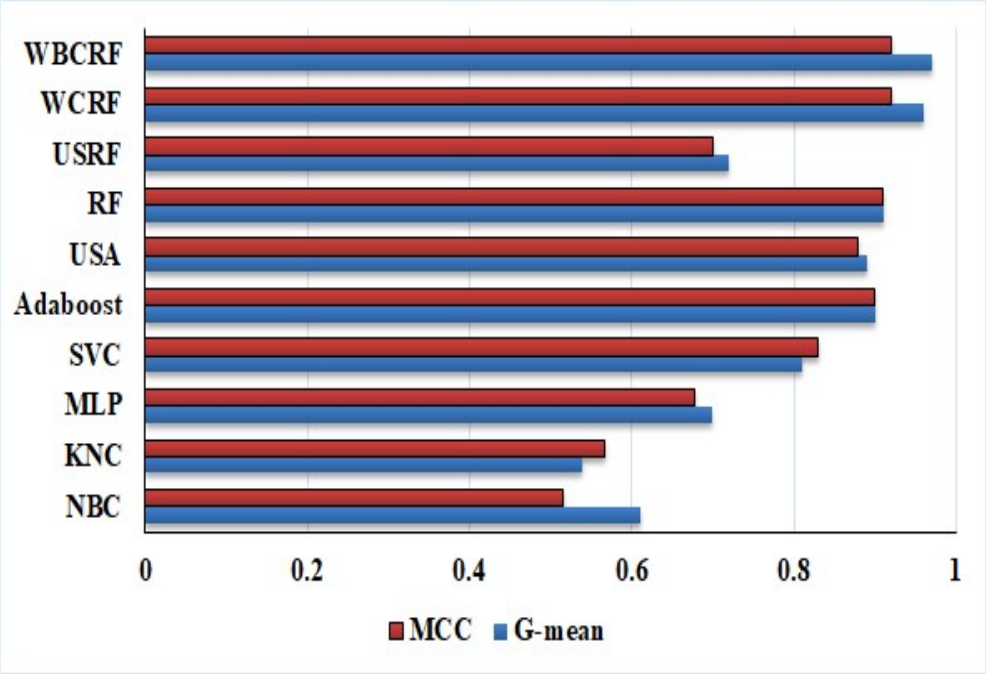


Figure 5.13 MCC and G-mean scores of machine learning models considered in this study for the first experimental scenario.

Drill bit 16 is intentionally discussed for understanding the effects of data imbalance arise while drilling through the thin lithofacies layer. The subsurface formations have varied thickness patterns in their natural state which results in random unequal data samples for the training phase. Therefore, the uneven distribution of training data has been particularly considered to evaluate the worst to the best performance of each classifier. Uneven data samples for various bit types (class labels) in training data make classification difficult for machine learning models [139]. Drill bit selection has been formulated as a multiclass classification problem with 19 diverse bit types as class labels as shown in Table A-1. However, large fluctuation in the values of precision, recall, F1score can be observed from Tables 5.8 and 5.9. Standard RF has shown good classification performance even for bit type 16 due to the presence of random bootstrap resampling technique in its internal architecture. RF has given the best prediction performance for 17.5” section subset with good immunity to data

imbalance conditions. Further, WBCRF has also been evaluated for 17.5” datasets that have given more accurate results with stable values for precision, recall, and F1 score.

Table 5.7 The screening of 17.5” bits through RF and WBCRF models.

RF				**WBCRF**			
Bit type	**Precision**	**Recall**	**F1score**	**Bit type**	**Precision**	**Recall**	**F1 score**
1	1.00	1.00	1.00	1	1.00	1.00	1.00
9	0.99	1.00	0.99	9	1.00	1.00	1.00
15	0.96	1.00	0.98	15	0.96	1.00	0.98
16	1.00	0.40	0.57	16	1.00	0.70	0.71
19	1.00	0.99	0.97	19	1.00	1.00	0.99
Average	0.99	0.88	0.91	Average	0.99	0.94	0.93
Accuracy	0.98	G-mean	0.83	Accuracy	0.99	G-mean	0.85

Table 5.8 The screening of 17.5” bits through KNC and NBC models.

KNC				**NBC**			
Bit type	**Precision**	**Recall**	**F1score**	**Bit type**	**Precision**	**Recall**	**F1 score**
1	0.98	0.93	0.96	1	0.99	0.77	0.87
9	0.68	0.63	0.65	9	0.60	0.60	0.60
15	0.67	0.92	0.92	15	0.47	0.86	0.61
16	0.00	0.00	0.00	16	0.50	1.00	0.67
19	0.57	0.56	0.56	19	0.77	0.72	0.74
Average	0.58	0.61	0.59	Average	0.67	0.79	0.70
Accuracy	0.81	G-mean	00	Accuracy	0.75	G-mean	0.779

Table 5.9 The screening of 17.5” bits through MLP and SVC models.

SVC				**MLP**			
Bit type	**Precision**	**Recall**	**F1score**	**Bit type**	**Precision**	**Recall**	**F1 score**
1	1.00	0.98	0.99	1	0.99	0.95	0.97
9	0.91	0.74	0.82	9	0.79	0.81	0.80
15	0.86	0.98	0.92	15	0.80	0.98	0.88
16	0.0	0.0	0.0	16	0.00	0.00	0.00
19	0.79	0.94	0.86	19	0.83	0.80	0.81
Average	0.71	0.73	0.72	Average	0.68	0.71	0.69
Accuracy	0.92	G-mean	0.00	Accuracy	89	G-mean	0.00

In the data subset of the 12.25” section, performance for every classifier has been recorded as shown in Tables 5.10, 5.11, and 5.12. Higher fluctuations in the values of precision, recall, and F1 score has been recorded in classification results. This indicates that these sections are physically challenging drilling zones. RF and WBCRF have given impressive results for the classification of these critical geological zones as shown in Tables 5.10, 5.11, and 5.12. In this section, BTs 6 and 11 are minority classes that are hard to classify. However, only MLP becomes a bias classifier as it fails to classify any samples for BT 11 as shown in Figure 5.13.

Table 5.10 The screening of 12.25” bits through RF and WBCRF models.

RF				**WBCRF**			
Bit type	**Precision**	**Recall**	**F1score**	**Bit type**	**Precision**	**Recall**	**F1 score**
2	0.96	0.98	0.97	2	0.96	0.99	0.97
4	0.99	0.96	0.97	4	0.98	0.96	0.97
6	0.86	0.86	0.86	6	1.00	0.71	0.83
10	0.79	0.81	0.80	10	0.80	0.86	0.83
11	0.80	0.57	0.67	11	1.00	0.73	0.72
17	0.75	0.80	0.77	17	0.78	0.83	0.81
Average	0.86	0.83	0.84	Average	0.92	0.85	0.86
Accuracy	0.91	G-mean	0.82	Accuracy	0.92	G-mean	0.84

Table 5.11.The screening of 12.25" bit through KNC and NBC models.

KNC				**NBC**			
Bit type	**Precision**	**Recall**	**F1score**	**Bit type**	**Precision**	**Recall**	**F1 score**
2	0.80	0.84	0.82	2	0.73	0.44	0.55
4	0.78	0.86	0.81	4	0.69	0.56	0.62
6	0.62	0.71	0.67	6	0.67	0.86	0.75
10	0.85	0.79	0.81	10	0.38	0.60	0.46
11	1.00	0.43	0.60	11	0.22	0.86	0.35
17	0.78	0.60	0.68	17	0.38	0.47	0.42
Average	0.81	0.70	0.73	Average	0.51	0.63	0.52
Accuracy	0.79	G-mean	0.685	Accuracy	0.53	G-mean	0.60

Table 5.12 The screening of 12.25" bits through SVC and MLP classifier.

SVC				**MLP**			
Bit type	**Precision**	**Recall**	**F1 score**	**Bit type**	**Precision**	**Recall**	**F1 score**
2	0.98	0.98	0.98	2	0.81	0.85	0.83
4	1.00	0.99	0.99	4	0.77	0.94	0.85
6	0.86	0.86	0.86	6	0.75	0.43	0.55
10	0.94	0.74	0.83	10	0.79	0.81	0.80
11	0.75	0.86	0.80	11	0.00	0.00	0.00
17	0.74	0.97	0.84	17	0.93	0.43	0.59
Average	0.88	0.90	0.88	Average	0.67	0.58	0.79
Accuracy	0.94	G-mean	0.80	Accuracy	0.79	G-mean	0.00

The geological lithofacies existing in section 8.5" are found to be the most challenging formations for drilling operations due to several faults and unstable zones existing along its depths. The 8.5" section formations have been reported to be unstable because they are made up of softer rocks such as claystone, sandstone, siltstone, tuff, marl, limestone, and argillaceous clay contents. Certain incidents of gas leaks and drill string stuck ups were also recorded while drilling 8.5" section of the wells with high stick slips conditions in its upper formations. Polycrystalline diamond compact bits

(PDC) were primarily utilized for drilling softer 8.5" section because of their higher ROP values and stable drilling operation (Table A-1). However, it becomes difficult for the driller to choose the right PDC bit type as varieties of bit models are available while planning for drilling operations. The pattern recognition has become difficult in the 8.5" section as the performance of all the classifiers has shown more fluctuations in their precision and recall values due to heterogeneity of lithofacies as shown in Tables 5.13, 5.14, and 5.15. Here, BT 12 and 13 are the minority classes for which KNC and MLP failed to identify any samples while SVC and NBC have shown poor prediction performance as shown in Figure 5.13. Finally, worst-to-best accuracy of various classifiers in second data-driven scenario can be given as: WBCRF (0.92-0.99), RF (0.91-0.98), SVC (0.88-0.94), MLP (0.74-0.89), KNC (0.61-0.81), and NBC (0.53-0.75). RF and WBCRF have shown great immunity for data imbalance conditions and successfully maintained their performance even in the critical 8.5" section. Recently, hybrid drill bits (e.g. Kymera) have been developed that combined the properties of conventional PDC bit and roller cone bit types [142]. These hybrid bits seem to be a good solution for drilling problematic 8.5" section while maintaining the stability of drilling operations.

Table 5.13 The selection of 8.5" bits through RF and WBCRF models.

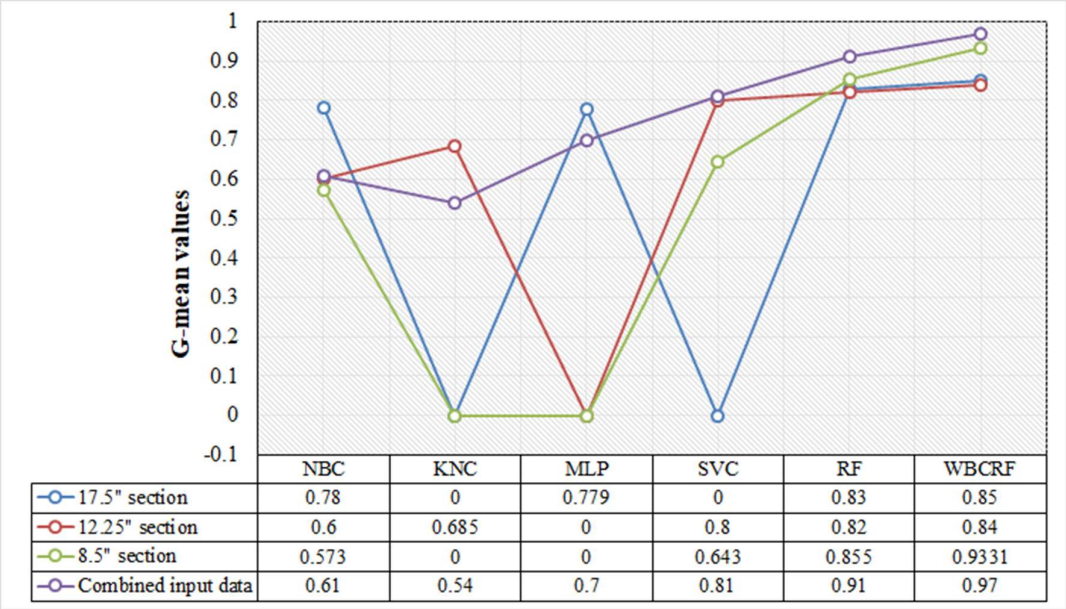
RF				**WBCRF**			
Bit type	**Precision**	**Recall**	**F1score**	**Bit type**	**Precision**	**Recall**	**F1 score**
3	0.91	0.99	0.95	3	0.92	0.99	0.95
5	0.98	0.87	0.92	5	0.98	0.89	0.93
12	0.93	0.81	0.87	12	0.93	0.81	0.87
13	1.00	0.67	0.80	13	1.00	1.00	1.00
18	0.89	0.98	0.93	18	0.94	0.98	0.96
Average	0.94	0.86	0.93	Average	0.95	0.93	0.94
Accuracy	0.93	G-mean	0.855	Accuracy	0.94	G-mean	0.931

Table 5.14 The selection of 8.5” drill bits through KNC and NBC models.

KNC				**NBC**			
Bit type	**Precision**	**Recall**	**F1score**	**Bit type**	**Precision**	**Recall**	**F1 score**
3	0.74	0.92	0.82	3	0.71	0.93	0.80
5	0.52	0.42	0.46	5	0.80	0.51	0.62
12	0.24	0.25	0.24	12	0.39	0.69	0.50
13	0.00	0.00	0.00	13	0.43	0.33	0.38
18	0.60	0.61	0.60	18	0.74	0.57	0.64
Average	0.42	0.44	0.43	Average	0.61	0.61	0.59
Accuracy	0.61	G-mean	0.00	Accuracy	0.68	G-mean	0.573

Table 5.15 The selection of 8.5” bits through SVC and MLP classifiers.

SVC				**MLP**			
Bit type	**Precision**	**Recall**	**F1score**	**Bit Type**	**Precision**	**Recall**	**F1 score**
3	1.00	0.99	0.99	3	0.79	1.00	0.88
5	0.96	0.84	0.89	5	0.90	0.49	0.64
12	0.77	0.62	0.62	12	0.57	0.50	0.53
13	0.67	0.22	0.33	13	0.00	0.00	0.00
18	0.72	0.96	0.82	18	0.65	0.86	0.74
Average	0.82	0.73	0.75	Average	0.58	0.57	0.56
Accuracy	0.88	G-mean	0.643	Accuracy	0.74	G-mean	0.00



	NBC	KNC	MLP	SVC	RF	WBCRF
17.5" section	0.78	0	0.779	0	0.83	0.85
12.25" section	0.6	0.685	0	0.8	0.82	0.84
8.5" section	0.573	0	0	0.643	0.855	0.9331
Combined input data	0.61	0.54	0.7	0.81	0.91	0.97

Figure 5.14 Summary of G-mean scores achieved by intelligent models for both experimental scenarios.

5.5 Summary

A novel data-driven approach has been proposed using the fusion of data resampling technique and ensemble method for handling the imbalance issues of complex drilling data. The problem of imbalanced training data results in the development of unreliable biased classifiers that are unfit for practical field applications. Two experimental data-driven scenarios have been specially designed and tested to confirm the generalization of the proposed approach for drill bit selection. An extensive comparative study has been performed to evaluate the performance of popular classifiers for the screening of drill bits. After a meticulous comparison of results, the following important conclusions can be drawn as given below:

- WBCRF technique has given the most impressive performance during automatic bit type selection with testing accuracy ranges from 92% to 99%, and G-mean (0.84–0.97) for various experimental scenarios.

- The large fluctuations in the performance of classifiers have been recorded in the 8.5” section in terms of precision, recall, and F1 scores. It is observed that drill bit selection becomes difficult in the lower formations due to uncertainty in subsurface conditions.
- Data imbalance condition exists due to the drilling of thin lithofacies that harm the performance of classifiers.
- WBCRF has given good prediction results for screening drill bit even for critical drilling zones.
- The performance of conventional classifiers is largely affected by data imbalance issues. Conventional classifiers can’t be trusted for the drill bit selection, especially for critical drilling zones.
- RF has shown great immunity for data imbalance conditions and successfully maintained its performance even in the critical 8.5” section.
- The proposed approach can also be applied over any other oil and gas fields to automate the drill bit selection, which will minimize human error, time, and drilling cost.
- The combination of ensemble methods with the data resampling technique results in modified ensemble classifiers that are found to be efficient even in highly imbalanced conditions.

Chapter 6

Assessment of Machine Learning Models for Forecasting of Hydrocarbon Production

6.1 Introduction

Hydrocarbon production forecasting is an important task for reservoir engineers to measure the performance of installed production systems. It also plays a vital role in the estimation of the remaining hydrocarbon inside the producing reservoir formations, optimization of production operations, reservoir management, and business planning [147]. Continuous recording and monitoring of daily hydrocarbon production data are usually done to forecast future well production. However, it is a challenging task due to the reservoir's heterogeneity and complex interactions of the reservoir with hydrocarbon production systems [148]. The production of multiphase fluid through surface choke is also influenced by the behavior of the producing reservoir formation in static and dynamic conditions [148]. Accurate assessment of reservoir properties, itself, is a problematic issue and its heterogeneity adds uncertainties in all types of reservoir measurements [149-150]. In the petroleum domain, production forecasting has always been considered a thought-provoking and popular problematic task due to the complexity of acquired production data [151].

Wellhead chokes are widely installed to control the oil and gas flow rates on the surface, to maintain downhole pressure, and also to produce back pressure that protects the reservoir from formation damage [152-155]. In order to regulate the flow rate for meeting various regulations, chokes are installed to minimize various problems owing to varying production rates which may be slugging of surface equipment, avoid excess

sand production caused due to high drawdown, and water/gas coning [156-157]. A significant part of the production optimization relies heavily on the study of the flow behavior through the chokes i.e. whether the flow is subsonic or sonic [158]. The critical pressure ratio is calculated to distinguish between sub-sonic and sonic flow conditions. It is approximately 0.55 for natural oil and gas above which subsonic condition prevails. When the fluid velocity in a choke matches the traveling velocity of sound in the fluid under in-situ conditions then such type of flow is termed sonic flow [159]. Under sonic flow conditions, the pressure wave downstream of the choke cannot go upstream through the choke because the medium is traveling in the reverse direction at a similar velocity [8].

Several correlations have been developed theoretically or empirically using experimental or field production data to study the simultaneous oil and gas flow behavior in sonic and sub-sonic conditions through chokes. Tangren et al. [48] contributed the first study on wellhead chokes and their effects on the production rate of hydrocarbons for the continuous liquid phase. Gilbert [49] correlated oil production rate with wellhead surface choke size, gas oil ration, and wellhead pressure. Ros [50] reported that a correlation existed between upstream pressure, restriction size of choke, and flow rate of hydrocarbon. Several other researchers also proposed similar correlations for the oil production rate using diverse field data [51-55]. Al-Attar and Abdul Majeed [56] tested several proposed correlations to provide the best fitting for East Baghdad Oil field production data. They found that the revised correlation was similar to the Gilbert equation with different constants values. Mirzaei-Paiaman and Salavati [42] proposed a newer correlation for the flow of oil through wellhead chokes using Persian oil field data. All the correlations have been developed either theoretically or empirically for wellhead chokes and multiphase hydrocarbon production based on

experimental data or field data [42][49-54]. Theoretical correlations developed using field data require a large number of parameters collection from fields which is a time-consuming and costly affair. On the other side, experimentally developed empirical correlations lack generalizability due to the limited range of experimental data. Therefore, advanced machine learning techniques have been employed to model the production rate of hydrocarbon with the surface installed chokes.

Researchers have widely utilized machine learning techniques to correlate variables that have complexities in their relationships. Machine learning techniques have achieved more reliable and generalized prediction models for several engineering domains such as reservoir characterization, drilling automation, etc. Morzaei-paiaman and Salavati, [42] applied the Artificial neural networks (ANN) model for the estimation of the oil production flow rate. He also compared the prediction results of ANN with the correlations proposed by Mirzaei-Paiaman [43], Gilbert [49], Ros [50], Achong [51], and Baxendell [52] to prove that ANN results are more accurate than empirical and theoretical correlations. Elhaj et al. [53] studied ANN along with Fuzzy logic, Functional networks, Decision tree, and Support vector machines for single gas flow rate forecasting. Choubineh et al. [35] applied hybrid ANN training-based optimization for modeling the hydrocarbon flow rate. Table 6.1 contains popular correlations for the determination of oil and gas flow rate.

Production forecasting has always grabbed the attention of reservoir engineers, however, only limited applications of machine learning models can be found in the literature. In this chapter, a comparative investigation of performances has been done among five popular machine learning models (viz. ANN, SVR, LSSVR, extremely randomized tree (ExtraTree), and RF) in quest of higher prediction accuracy for production forecasting. All of these techniques are widely accepted and reported for

estimation purposes in reservoir characterization and drilling automation [160-162]. The performance of the above-mentioned machine learning models has been evaluated using production data obtained from the Norwegian Volve oil and gas field. ExtraTree and Random forest paradigms have been applied the first time for production forecasting as per the knowledge of the authors. This study also examines the importance and contribution of each input (predictor) variable existing in production data for the pattern estimation of hydrocarbon production rate.

6.2 Random forest and ExtraTree

Machine learning models are pattern recognition techniques that are primarily utilized to solve problems involving detection, identification, and estimation tasks. These intelligent techniques are capable of finding the relationship between field variables having complex datasets [163]. In this study, five popular machine learning algorithms have been investigated for the estimation of the daily production of the Norwegian oil and gas field. This section provides a brief introduction of Random forest and Extremely Randomized Trees models utilized for the prediction of daily oil and gas production.

6.2.1 Random forest and Extremely randomized trees

Random forest is one of the most popular meta-learner developed by Leo Breiman [164]. It can be applied for categorical data as a classification model and also for continuous response target data as a regression model. Random forest is computationally appealing in nature due to fast computational speed, lesser tuning parameters, easily estimable generalization error, handling of high dimensional data, can measure predictor variable importance, etc. [165]. It is an ensemble of Decision trees in which each tree is dependent on random input variables. It uses the bootstrap technique for the generation of random data subsets with the replacement with the training phase. The estimation function is defined in terms of the Loss function that is

needed to be minimized. In the case of regression, minimizing expected error and considering square error loss gives conditional expectations. RF is an ensemble-based meta-learner that is composed of several base learners viz. decision trees. These base learners are combined together in a single architecture known as ensemble learners. The outcomes of base learners are averaged out to provide the final estimation results in regression. Decision Trees are utilized as base learners to split the predictor data space on individual variables. The root node of the tree contains all the predictor data space. The non-partitioning nodes are called terminal nodes that have the final partition of input data space. In the case of regression, the splitting is done based on the mean squared residual at the node. The best possible split in regression is determined by sorting the values of the predictor and splitting every distinct consecutive value of pairs. The algorithm of Random forest is given below as suggested by Breiman[164], Cutler, *et al.*[175] and Ho [166].

ExtraTree is also an ensemble learner containing a forest of unpruned Decision trees similar to a Random forest ensemble with certain constructional dissimilarities. It has two major differences as compared to other decision/regression trees based paradigms. First, each node is split depending upon a fully random cut-point. Secondly, it utilizes whole learning data samples for growing forest trees randomly with random samples instead of a bootstrap replica. ExtraTree utilizes strong randomization for the training of trees to reduce error due to bias-variance [167-168]. Machine learning researchers reported that ExtraTree may outperform RF in certain cases of pattern estimation [168]. Thus, ExtraTree has also been considered in this study to challenge the Random forest algorithm. A detailed explanation of ExtraTree can be found in the research work of Geurst *et al.* [168].

6.3 Case study of Volve oil and gas field production

This study utilizes the technical production data downloaded from the Equinor website for the development and testing of machine learning models to estimate daily production rates. The downloaded production data in a format of excel contains 6488 production data samples of seven production wellbores of the Volve field. The data samples having null and missing values were removed from downloaded production data as they were not suitable for modeling purposes. Downloaded production data were reduced to 4167 data samples that were utilized as input data for the training and testing of machine learning models in this study. The production data were collected from seven wellbores of Volve fields viz. 15/9-F-1C, 15/9-F-11H, 15/9-F-12H, 15/9-F-14H, 15/9-F-15D, 15/9-F-4AH, and 15/9-F-5AH. Table 6.2 contains a summary of various research works related to production forecasting through surface installed chokes. The statistical description of Volve Production data is shown in Table 6.3. Figure 6.1 shows the location of the Volve oil and gas field at the North Sea adapted from Ravasi et al. [169].

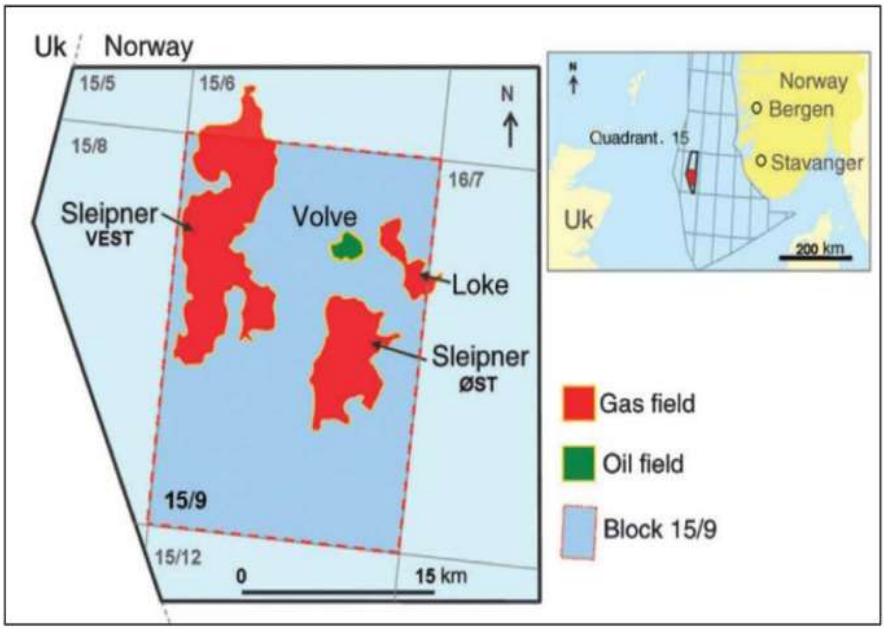


Figure 6.1 Volve oil and gas field located at North Sea [169].

Table 6.1 Popular correlations for the determination of oil and gas flow rate.

S. No.	Empirical Correlation	Publication
1.	$Q_{Rate} = 0.1 * \frac{WHP * CS^{1.89}}{GOR^{0.546}}$	Gilbert (1945)
2.	$Q_{Rate} = 9.56 * \frac{\mathrm{THP} * CS^{1.93}}{GOR^{0.546}}$	Baxendell 1957
3.	$Q_{Rate} = 0.574 * \frac{WHP * CS^{2}}{GOR^{0.5}}$	Ros (1960)
4.	$Q_{Rate} = \frac{\mathrm{THP} * CS^{1}}{0.262 * GOR^{0.65}}$	Achong, (1961)
5.	$CS = \frac{20.696 * Q_{RATE}^{o.483} * \gamma^{0.707}}{P_{wh}^{0.474}}$	Al-Towailib (1992)
6.	$Q_{rate} = \frac{0.087607 * P_{WH.} d^{1.9215}}{GOR^{0.5334}}$	Mirzaei-Paiaman and Salavati (2013)

Table 6.2 Summary of various research works related to production forecasting through surface installed chokes.

S. No.	Author(s)	Year	Methods	Input Variables	Production Data
1.	Tangren	1949	Theoretical	Pressure ratio, fluid Velocity, volume ratio, density ratio	Laboratory
2.	Gilbert	1954	Empirical	WHP, CS and GOR	California field
3.	Baxendell	1957	Empirical	THP, CS, and GOR	Laboratory
4.	Ros	1960	Theoretical	WHP, CS, and GOR	Laboratory
5.	Poettmann	1963	Empirical	GOR, WHP, and CS	Field data (108)
6.	Omana	1969	Empirical	CS, Upstream Pressure, and WHT	Laboratory
7.	Fourtunati	1972	Theoretical	GOR, downstream pressure, and WHP	Field data
8.	Ashford	1974	Theoretical	CS, GOR, pressure, temperature, discharge coefficient, specific gravity	Oil field test

9.	Al-Attar	1988	Empirical	GOR, WOR, oil gravity, gas gravity, CS, upstream pressure, and average temperature.	Baghdad field (155)
10.	Surbey	1988	Empirical	Choke setting, upstream pressure, and temperature	Laboratory
11.	Osman	1990	Empirical	CS, WHP, and WHT	Laboratory
12.	Perkins	1993	Empirical	Pressure, temperature, gas specific gravity, oil specific gravity, etc.	Middle East field data (1432)
13.	Al-Towailib	1994	Empirical	GOR, upstream pressure, CS, and downstream pressure	Middle east filed data (3554)
14.	Mirzaei-Paiaman and Salavati	2013	Empirical	Upstream pressure, CS, GOR, oil specific gravity, and gas specific gravity	Persian oil and gas fie;
15.	Al-Khalifa et al.	2013	ANNs	GOR, CS, and upstream pressure	Various field data (4031)
16.	Nejatian et al.	2014	LS-SVM	Reynolds number, the ratio of choke diameter to pipe diameter, and choke flow coefficient	Southwest Louisiana (512)
17.	Gorjaei et al.	2015	PSO-LSSVM	GOR, CS, and WHP	Iranian fields (276)
18.	Elhaj et al.	2015	ANN, Fuzzy logic, SVM, Functional network, and Decision tree	CS, upstream and downstream pressures, tubing temperature, and the specific gravity of the gas	Sudan (276)
19.	Mirazaei-Paiaman and Salavati	2013	ANNs	WHP, CS, and GOR	Southwest Iran and Northern Persian Gulf
20.	Ghorbani et al.	2017	Firefly Optimization algorithm	GOR, CS, and WHP	Ghawar field
21.	Choubineh et al.	2017	Teaching learning based optimization and ANNs	CS, specific gravities of oil and gas, GOR, WHP, and WHT	South Iran (113)

21.	Rashid et al.	2019	ANNs	WHP, GOR, and CS	Unknown Field (276)
22.	Wang et al.	2019	Deep neural networks	Big database: well information, reservoir thickness, hydraulic fracture parameters, fracturing system, and proppant information	Bakken shale reservoir data

Table 6.3 The statistical description of Volve production data utilized in this study.

S. No.	Predictor variables	Units	Maximum	Minimum
1.	Downhole pressure (DP)	psi	308.1	0
2.	Downhole temperature (DT)	celsius	107.51	0
3.	Tubing size (TS)	inch	259.09	26.12
4.	Annulus pressure (AP)	bar	30.0198	0.06
5.	Wellhead pressure (WHP)	bar	120.889	0
6.	Wellhead temperature (WHT)	celsius	92.0711	7.04
7.	Choke size (CS)	(1/64) inch	106	0.5
8.	Gas oil ratio (GOR)	scf/STB	6177.5	103
9.	Oil production rate (OPR)	STB/day	5889	4.3
10.	Gas production rate (GPR)	scf/day	83598	856

6.4 Methodology

A rigorous comparative study has been performed to evaluate the performance of five machine learning techniques viz. ANNs, SVR, LSSVR, ExtraTree, and Random forest. The data-driven experimental workflow for production forecasting is broadly partitioned into two stages i.e. (a) data preprocessing and (b) model preparation. Both of the experimental stages are explained in detail as given below.

6.4.1 Data preprocessing

Data processing of input data helps the machine learning algorithms to understand and learn the hidden patterns in a better way to ensure that all the input variables will obtain

equal importance from machine learning models. A dataset of 4167 samples has been utilized after eliminating null and garbage values for modeling hydrocarbon production rate (PR) through popular machine learning techniques viz. (ANN, SVR, LSSVR, ExtraTree, and RF). Machine learning models have been developed for production forecasting from surface measured predictor variables viz. DP, DT, TS, AP, WHP, WHT, CS, and GOR with production rate (PR) as a target/ response variable as shown in Table 6.3. These variables are recorded on the surface of producing wells that are used for modeling oil and gas flow rates. Surface measured predictor variables are modeled independently for OPR and GPR forecasting to investigate the behavior of machine learning models. Initially, input production data were normalized to reduce the effects of larger values on smaller ones. Normalization scales down all the values of predictor variables between zero and one. The formula for normalizing production data is given below.

$$X_{Norm} = \frac{X - X_{Min}}{X_{Max} - X_{Min}} \tag{6.1}$$

where X_{Max} is the maximum value of predictor variable X, X_{Min} is the minimum value of input variable X and X_{Norm} is the normalized value of X.

6.4.2 Attribute selection

After the normalization of input data, different attributes were statistically tested to find any redundant variables that could be eliminated to reduce dimensionality and computation time. Table 6.4 presents the statistical analysis of input predictor variables to understand their contribution to the estimation of OPR. The input variables were provided one by one in each trial to measure their impact on the estimation of OPR. The effects of each input variable were studied using conventional statistical analysis using correlation of coefficient (CC) calculation between predictor variables and OPR. Table

6.4 shows the effects of adding predictor variables one by one in the training data on the performance of the RF technique.

Table 6.4 Trail test to identify input variables that are contributing to oil production estimation.

S. No.	Input parameters	Training CC	Training RMSE	Testing CC	Testing RMSE
1.	DP, DT	0.9873	0.0405	0.9002	0.1091
2.	DP,DT, TS	0.9947	0.0263	0.9567	0.0266
3.	DP, DT, TS, AP	0.9958	0.0203	0.9672	0.0632
4.	DP, DT, TS, AP, WHP	0.996	0.0226	0.9687	0.0619
5.	DP, DT, TS, AP, WHP, WHT	0.9977	0.0169	0.9829	0.0459
6.	DP, DT, TS, AP, WHP, WHT, CS	0.9979	0.0164	0.9839	0.0446
7.	DP,DT, TS, AP, WHP, WHT, CS, GOR	0.9979	0.0164	0.9843	0.044

Similar results can also be found using attribute selection algorithms such as the Relief algorithm [170], a forest of trees [168], etc. for identifying and eliminating irrelevant attributes from production data. Statistical analysis for the selection of important attributes is a time-consuming affair. Therefore, algorithm-based attribute selection techniques are more recommended over conventional statistical analysis as they consume less time. In this study, the Relief algorithm has been utilized for understanding the importance of different predictor variables for the pattern estimation of oil and gas flow rates. It assigned weights and ranks to all the input predictor variables according to their contribution to the pattern estimation of production rate. It also helps to identify conditional dependencies and the existing correlation between predictor variables and target outputs (OPR/GPR)[168]. The rank and weights assigned

by the Relief algorithm are case-specific in nature and may vary with production data of different reservoirs. Figure 6.2 shows the contribution of each predictor variable for the pattern estimation of PR utilizing the relief algorithm. All the predictor variables were assigned positive weights by the relief algorithm as shown in Figure 6.2, therefore, each of them was utilized for training and testing of machine learning models. Enthusiastic readers are advised to refer to cited research papers for a detailed explanation of the relief algorithm [168-170].

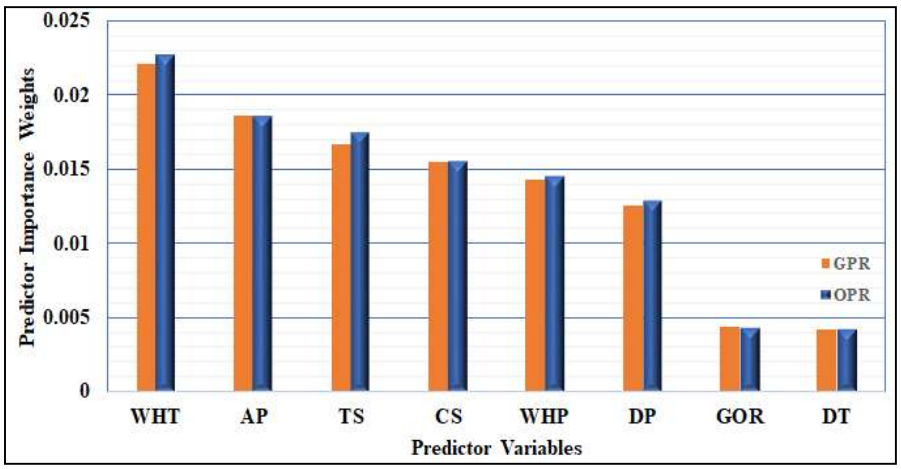


Figure 6.2 Predictor variables arranged according to their contribution for production forecasting.

6.4.3 Model preparation

Researchers have always shown concerns about overfitting and underfitting conditions that reduce the generalizability of computationally intelligent machine learning models during classification and regression tasks. It is suggested that the chances of overfitting or underfitting can be reduced using cross-validation techniques [171]. Several cross-validation strategies such as K-fold cross-validation, hold out validation, leave one out validation, etc. are available for splitting pre-processed input data into training and testing subsets. After attribute selection, the input data were organized into random

training and testing subsets using a 10-fold cross-validation technique to reduce the chances of overfitting and underfitting conditions. In 10-fold cross-validation, the pre-processed production data were split randomly into ten equal-sized smaller subsets of data (K=10). Out of ten subsets, nine subsets were utilized for the training of a machine learning model and the tenth subset for validation of the trained machine learning model. This whole procedure was repeated ten times until all the data subsets were acted at least once for the validation of the trained machine learning model. All the machine learning models were trained and tested on these ten pairs of subsets iteratively and final results were decided by averaging their performances with these subsets (K=10). K-fold cross-validation has been reported to be maximum helpful in variance error reduction as compared to other cross-validation techniques. During the training phase, parameters of each intelligent model were optimized using a Grid search optimizer to minimize the prediction error, to achieve the best possible performance, and to accomplish optimally tuned models. Proper optimization of models' parameters is an essential step to maintain their performance for the prediction of unseen test datasets. The performance of machine learning paradigms was evaluated by three statistical performance indicators, viz. the coefficient of correlation (CC), root mean square error (RMSE), and mean absolute error (MAE), as given below.

A. Coefficient of correlation (CC)

$$CC = 1 - \frac{\sum_{i=1}^{n}\left(PR_m - PR_p\right)^2}{\sum_{i=1}^{n}\left[PR_m - \frac{1}{n}\sum_{i=1}^{n}\left(PR_m\right)\right]_i^2} \tag{6.2}$$

where PR_m and PR_p are measured and predicted production rate of oil or gas. CC measures the correlation between PRm and PR_p variables which ranges between one

and zero. Values nearer to one represent good correlation and vice versa. CC indicates towards generalization performance of machine learning models.

B. Root mean square error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(PR_m - PR_p\right)^2} \tag{6.3}$$

where, RMSE is an error function, primarily utilized for measuring the prediction error of the machine learning model and has an inverse relation with the prediction performance of the model.

C. Mean absolute error

$$MAE = \frac{1}{n}\sum_{i=0}^{n}\left|\frac{\left|PR_m - PR_p\right|}{PR_m}\right| \tag{6.4}$$

MAE is also an error function that helps to compare the predictions with measured target variables. Readers should keep in mind that all the values are calculated after the normalization of predictor variables and response variables data. Therefore, RMSE and MAE have also been reported on a similar scale accordingly. Figure 6.3 shows a generalized data driven architecture for production forecasting.

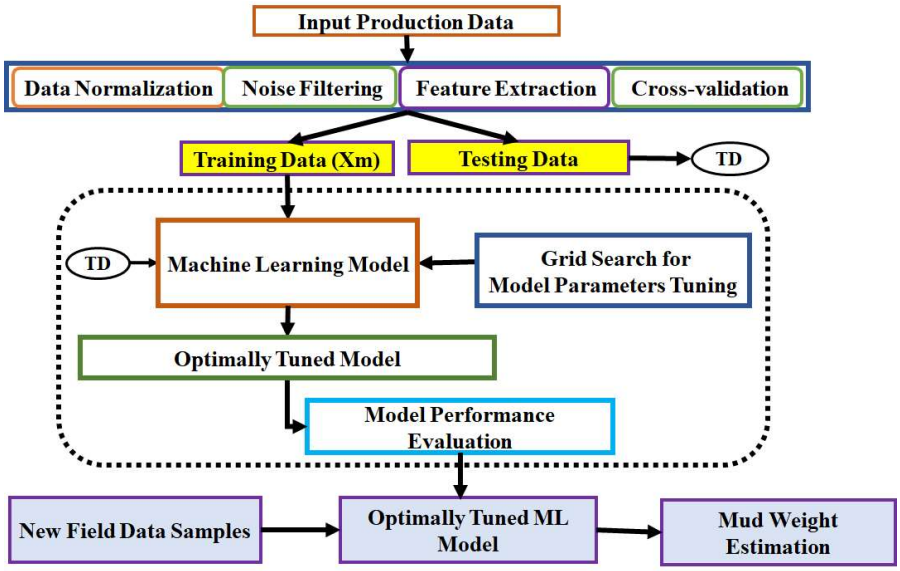


Figure 6.3 A Generalized data driven architecture for production forecasting.

6.5 Results and discussion

Initially, multiple linear regression (MLR) models have been developed to correlate predictor variables with the oil and gas production rate. These preliminary MLR models are also considered in this study as reported in several earlier research works [172-174] [175]. The equation (6.5) and (6.6) models OPR and GPR using surface measured predictor variables obtained from Volve data.

$$OPR=-1.3806*DP+1.6776*DT-0.156*AP+0.8177*WHT+0.9163*CS-1.4682*GOR-0.8194 \quad (6.5)$$

$$GPR= -1.4021*DP1.6672-0.1545*AP+0.8295*WHT+0.933*CS-1.2516*GOR-0.82 \quad (6.6)$$

However, the estimation accuracy of the MLR model, for forecasting OPR, was found to be the lowest (training CC=0.7816 and testing CC=0.7724) and the highest prediction error. (Training/Testing RMSE = 0.1557/0.1585, and Training/Testing MAE = 0.1069/0.1074). ANNs are one of the most widely applied machine learning models in

the oil and gas industry. Several configurations of ANNs were trained and tested for the estimation of PR. The pre-processed input data were partitioned into training data (70%), validation data (15%), and testing data (15%). Partitioned datasets were provided to feedforward backpropagation ANN model for its training, validation, and testing phases. During the training phase, several trials were performed for the selection of appropriate transfer function, network function, number of hidden layers, and neurons. Table 6.5 contains optimized values of different parameters of the ANN model for production forecasting. The configurations 8-10-1 were found the most suitable architecture of ANN for the estimation of daily hydrocarbon production as shown in Figures 6.4 and 6.5.

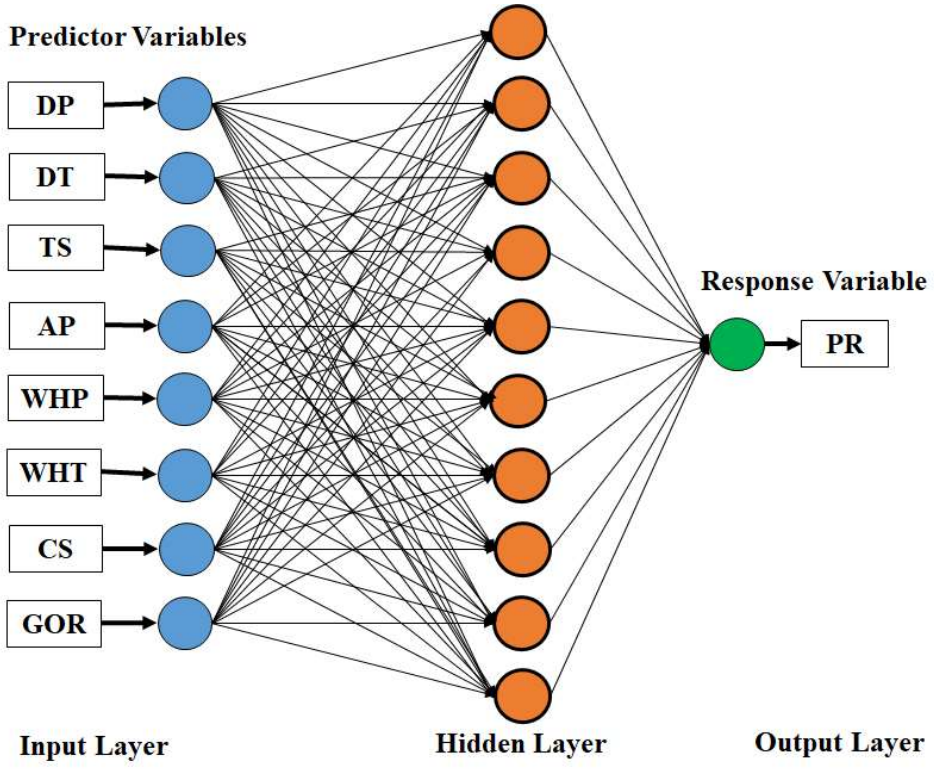
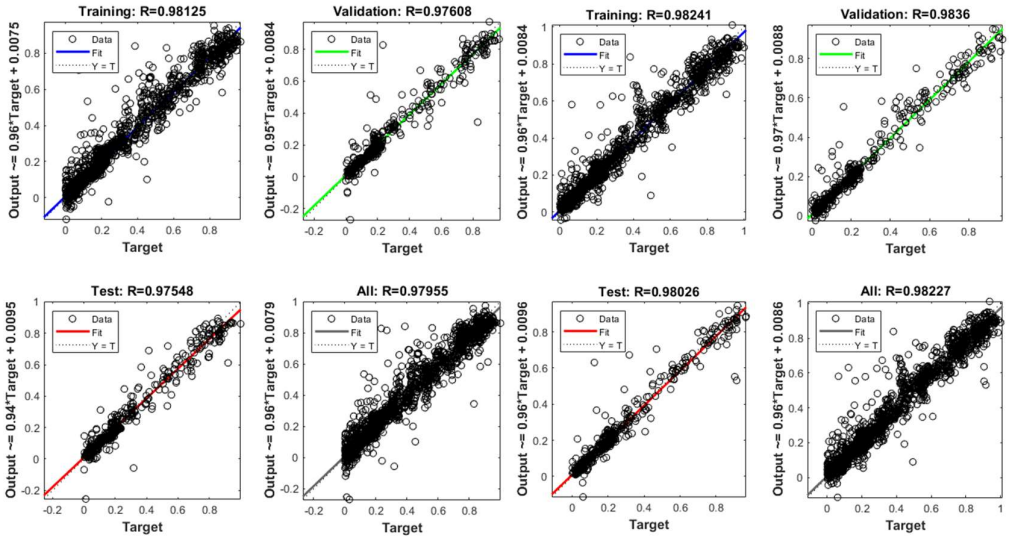
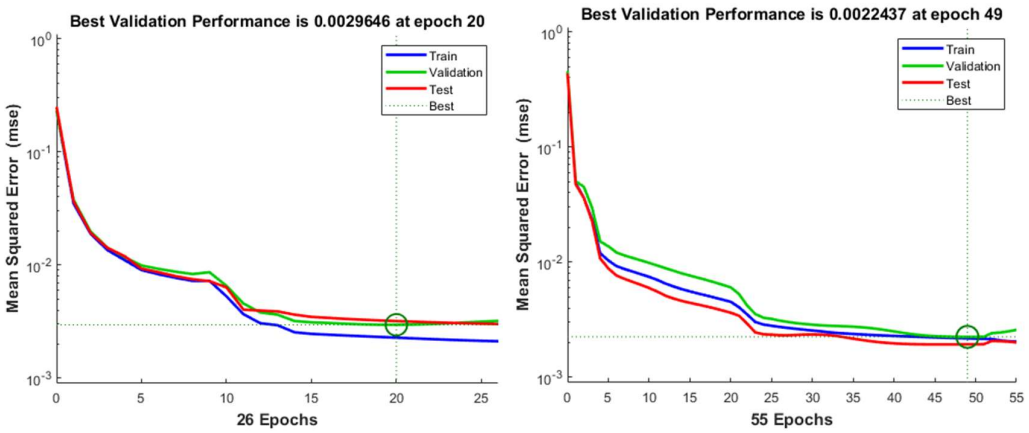


Figure 6.4 ANNs architecture (8-10-1) found suitable for production forecasting of oil and gas flow through the surface installed choke.



(a) Regression plot for OPR (b) Regression plot for GPR



(c) Performance plot for OPR. (d) Performance plot for GPR.

Figure 6.5 Regression and Performance plots for training (70%), validation (15%), and testing (15%) of ANN (8-10-1) for production forecasting.

Figure 6.5(a) and 6.5(b) show regression plots of ANN with training, validation, and testing datasets during oil and gas forecasting. Training function and activation-transfer function were also tested along with the number of neurons. It was found that Trainlm and Tansig functions combination gave the best prediction results with a high convergence rate. Figure 6.6 shows the effects of increasing the number of neurons in the hidden layer architecture of ANN on its performance during the training and testing

phase. Figure 6.5(c) and 6.5(d) shows the effects of epoch or iterations on the performance of ANN models for production forecasting. ANN is one of the most popular machine learning algorithms in the petroleum domain, however, it has its own limitations such as overfitting, stuck up in local minima/ maxima, etc. [163]. Therefore, SVR was considered to develop a more generalized and reliable model for the estimation of PR. The prediction results of ANN for production forecasting are summarized in Table 6.6 and 6.7.

Table 6.5 The optimized values of different parameters of machine learning models implemented for production forecasting.

S. No.	Regression Models	Model Parameter	Search Range	Optimized value
1.	ANN	Transfer Function	N/A	Tansig
		Training function	N/A	Trainlm
		Number of hidden layers	1-10	1
		Adaption learning function	N/A	Learngdm
		Number of neurons	1-100	10
		Configuration	N/A	8-10-1
2.	SVR	Kernel function	Linear, Polynomial, and RBF kernel	RBF function
		Regularization parameter C	10-1000	100
		Epsilon	0.00014-13.96	0.01
		Gamma	0.001-10	0.5
		Box Constraints	0.001-1000	0.001
3.	LSSVR	Kernel function	Linear, polynomial, RBF kernel	RBF kernel
		Regularization parameter	1-1000	20
		Gamma	0.001-10	0.8

4.	ExtraTree	Number of randomly selected attributes at each node	0-9	8
		Minimum samples required for the split at each node	0-300	2
		Maximum depth of the tree	0-100	None*
		Number of base learners or estimators	1-100	50
		Minimum samples leaf size	1-400	1
5.	RF	Maximum tree depth	1-None	None*
		Number of base learners or estimators	1-100	40
		Minimum samples leaf size	1-300	1
		Minimum sample required for the split at each node	0-300	2

The whole pre-processed input data were partitioned into 10 random training and testing datasets using a 10-fold cross-validation technique to avoid overfitting and underfitting conditions of machine learning models. The limitations of ANNs were overcome by applying SVR and LSSVR models for production forecasting of oil and gas. Primarily, the selection of kernel function was done before the application of the SVR and LSSVR model for forecasting. Appropriate kernel function helps SVR and LSSVR to handle the high dimensionality and nonlinearity of production data efficiently. RBF kernel was found suitable after several trials with different kernel functions for production forecasting. LSSVR and SVR were also constituted of various hyperplane model parameters such as gamma, regularization parameter (C), epsilon, etc. that were required to be optimized to enhance model performance. These parameters were optimized during the training phase using the Grid search optimizer available in the Matlab 2019b Platform. The hyperplane model parameters were optimized using a grid search algorithm which is a widely reported tuning algorithm in various domains

[176-177]. Grid search utilizes the hill-climbing approach to determine optimum parametric values of machine learning models. The iteration stops after reaching the best possible values of model parameters and can also extend its search range if values exist at the search boundary. The optimized values of SVR and LSSVR parameters are shown in Table 6.5. Figure 6.7 shows a cross-plot between actual and predicted values of daily oil and gas production for training and testing of the LSSVR model. The performance of optimized LSSVR was found better than conventional SVR and ANNs for OPR forecasting but it failed to outperform the ANN in the case of GPR estimation. Therefore, ensemble models were investigated to provide a more generalized model for the production estimation of oil and gas rates.

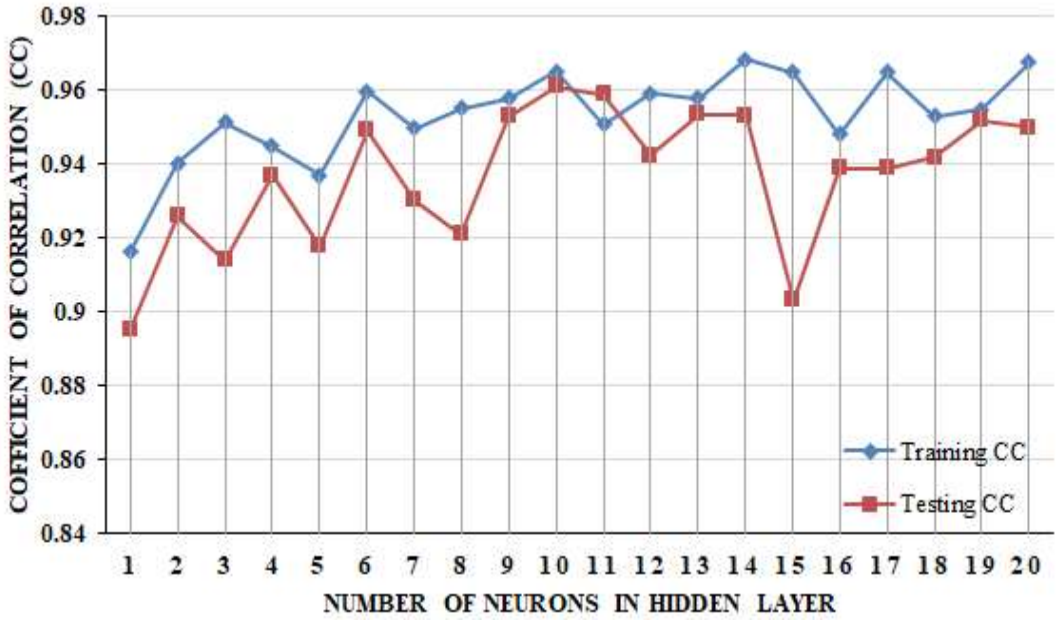
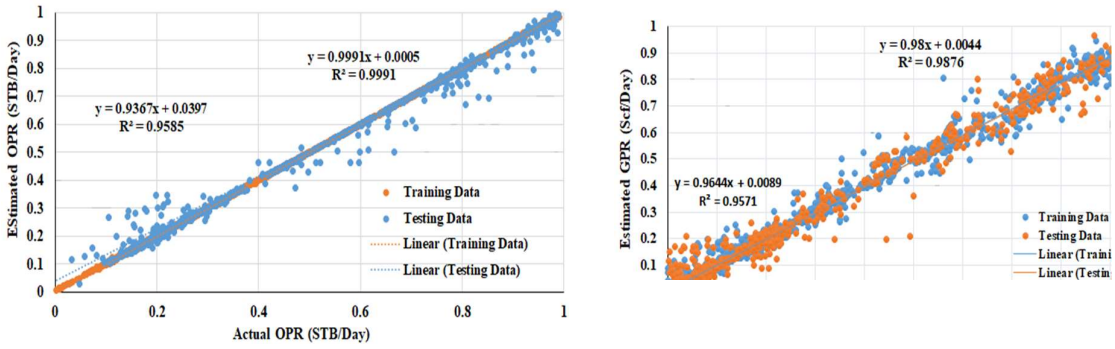
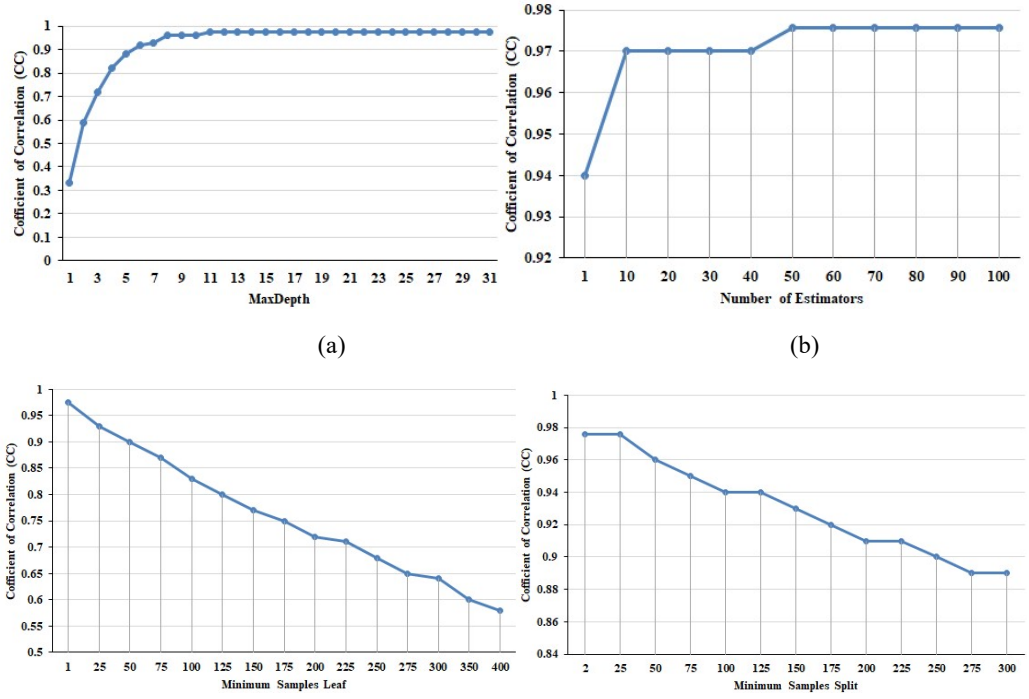


Figure 6.6 Effects of the increasing number of neurons in the hidden layer of ANN architecture for production forecasting.



(a) (b)

Figure 6.7 Coefficient of correlations for (a) OPR (b) GPR using LSSVR with training and testing datasets for production forecasting.



(a) (b)

(c) (d)

Figure 6.8 Effects of parameters' variation on the performance of ExtraTree (a) Variation of maximum depth. (b) Variation of the number of estimators (c) Variation of minimum samples leaf and (d) Variation of minimum samples split.

In the quest for higher estimation accuracy, ExtraTree and RF ensembles were tested for production forecasting of oil/gas and also to outperform ANN, and LSSVR. ExtraTree algorithm employed for production forecasting was based on Geurts *et al.*

(2006) research work [168]. The main parameters on which the performance of ExtraTree model primarily depends are the minimum number of data samples required for splitting any node, the number of decision trees or estimators (base learners), the maximum depth of the tree, etc. Figure 6.8 shows the effects of the variation of four important parameters on the performance of ExtraTree. The optimum values of these parameters are shown in Table 6.5. The training and testing results clearly reveal that ExtraTree outperformed the ANN, SVR, and LSSVR for the estimation of oil and gas production. Further, an RF ensemble was implemented to challenge the performance of ExtraTree for production forecasting. The performance of RF was also dependent upon its model parameters such as the number of estimators, minimum split at each tree node, minimum samples required for splitting, learning rate, etc.

In this study, 40 decision trees have been utilized for the development of an RF ensemble estimator to predict the oil and gas production rate. Certain researchers suggested that large numbers of base learners viz. decision tree increases stability for petrophysical applications, [178]. However, the performances of RF and ExtraTree were saturated after 40 and 50 estimators for production forecasting as shown in Figures 6.7 and 6.8. The complexity, stability, and computational cost of both ensembles were also controlled by the maximum depth of decision trees or estimators. The maximum depth parameter of RF and ExtraTree was set at the default value “None” which allowed nodes of the decision tree to expand until every leaf contained the minimum number of samples that could not be split. The optimum values of each parameter for Random forest are given in Table 6.5. Figure 6.9 depicts the performance of RF with varying values of four important parameters. Training and testing estimation accuracies of RF are found to be the highest among all other paradigms applied in this study as shown in

Table 6.6 and 6.7. It can be observed from Figure 6.9 that Random forest has minimum training and testing RMSE for forecasting oil and gas production.

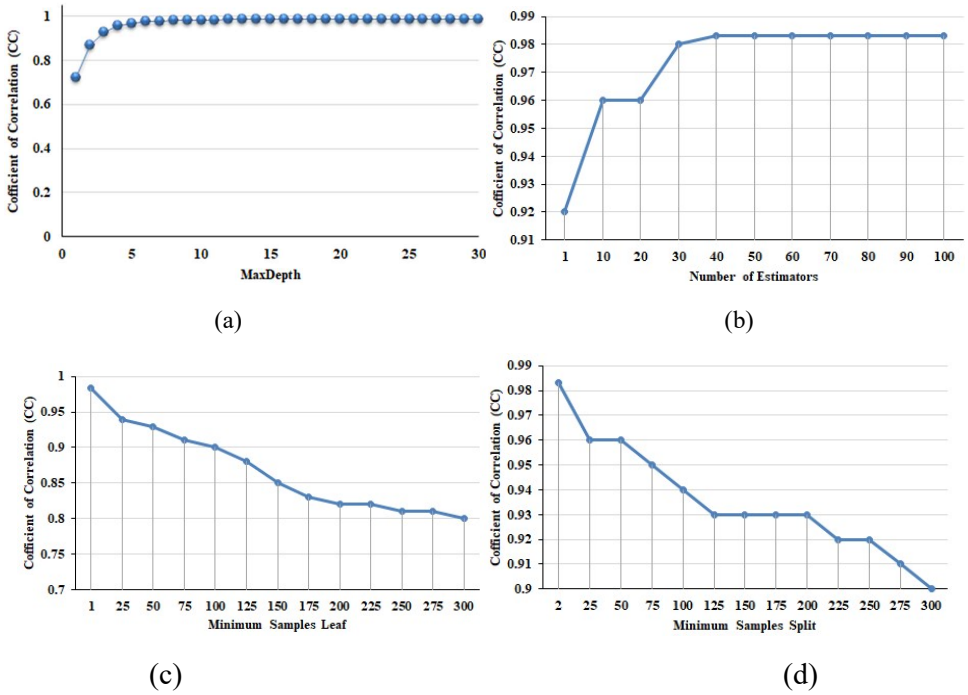


Figure 6.9 Effects of parameters' variation on the performance of Random forest (a) Variation of maximum depth (b) Variation of the number of estimators (c) Variation of minimum samples leaf and (d) Variation of minimum samples split.

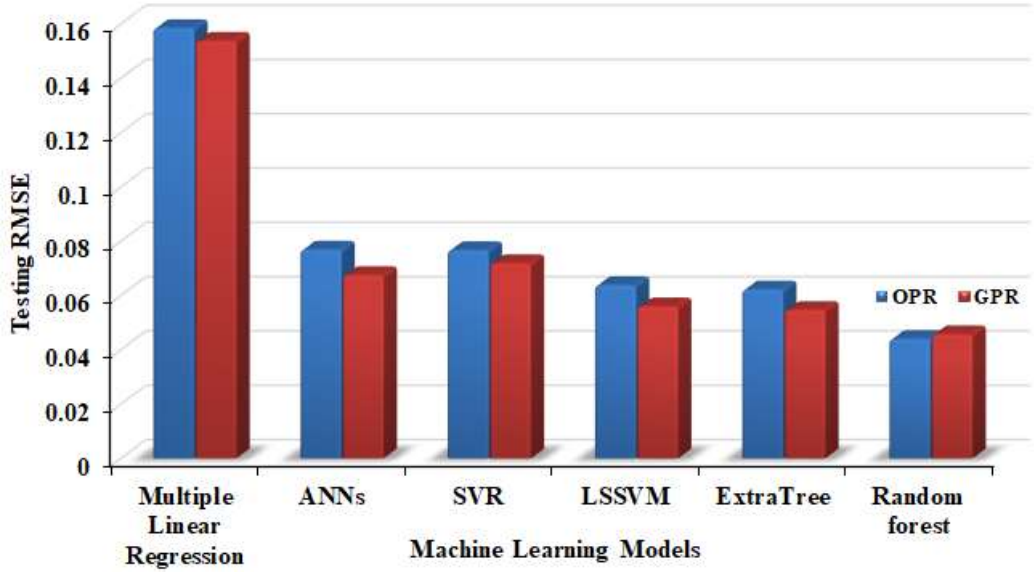


Figure 6.10 Comparison of RMSE occurred during estimation of OPR and GPR utilizing machine learning models under study.

Table 6.6. The estimation accuracy and errors recorded for different techniques utilized for oil production forecasting.

S. No.	Estimation Techniques	Training CC	Training RMSE	Training MAE	Testing CC	Testing RMSE	Testing MAE
1.	Multiple Linear Regression	0.7816	0.1557	0.1069	0.7724	0.1585	0.1074
2.	ANN	0.9656	0.0663	0.044	0.9514	0.0769	0.0595
3.	SVR	0.9678	0.0628	0.0457	0.9533	0.0767	0.0563
4	LSSVM	0.9991	0.0020	0.0015	0.9585	0.0638	0.0192
5.	ExtraTree	0.9958	0.0230	0.0057	0.9687	0.0622	0.0191
6.	Random forest	0.9979	0.0164	0.0073	0.9843	0.044	0.0154

Table 6.7. The estimation accuracy and error recorded for different techniques utilized for gas production forecasting.

S. No.	Estimation Techniques	Training CC	Training RMSE	Training MAE	Testing CC	Testing RMSE	Testing MAE
1.	Multiple linear regression	0.7881	0.1534	0.1055	0.7869	0.1538	0.105
2.	ANN	0.9651	0.0670	0.045	0.9608	0.0675	0.0311
3.	SVR	0.9728	0.0572	0.0298	0.9507	0.0719	0.0313
4.	LSSVM	0.9876	0.0275	0.0130	0.9571	0.0560	0.0202
5.	ExtraTree	0.9957	0.023	0.0061	0.9757	0.0547	0.0182
6.	Random forest	0.9979	0.0164	0.0073	0.9831	0.0457	0.0162

6.6 Summary

A rigorous investigation of five machine learning paradigms has been done in the quest for higher estimation accuracy in production forecasting. Volve oil and gas production field data were utilized for training and testing of machine learning models. The production data contained eight input variables that were pre-processed before the development and testing of various models. RF and ExtraTree are the first time implemented for production forecasting and have shown improvement in estimation accuracy as compared to LSSVR, SVR, and ANN. The relevance of each predictor variable is also studied using statistical analysis and the Relief algorithm. The contribution of this study is to identify the most suitable and robust intelligent forecasting model for daily hydrocarbon production through surface chokes. Machine learning paradigms are found useful in correlating surface measured predictor variables with daily oil and gas production rates with high estimation CC (approximately 0.98) and minimum prediction errors. Random forest and ExtraTree ensembles have outperformed the popular estimation models, viz. ANNs, SVR and LSSVR, for production forecasting. The performance of LSSVR has been found slightly better than conventional SVR for production forecasting. However, proper tuning of the model's parameters is essential for its impressive performance. WHT has been identified as a maximum contributing variable for the estimation of daily hydrocarbon production using the Relief algorithm. The analysis of estimation results indicates that RF and ExtraTree are powerful techniques for production forecasting.

Chapter 7

Conclusions and Future Scope

7.1 General

The increased trends towards, measurements-while-drilling and smart-well technology had led to a digitalization boom and more automation, in the oil and gas industry. Monitoring and control of various petroleum field operations are performed using information acquired from the installed sensors. These sensor-based measurements produce a large amount of field data that are needed to be analyzed properly. Conventionally, field data are interpreted by experienced experts to extract useful information. Extraction of useful information from the acquired data poses various challenges as discussed in the earlier chapters. Further, with the advent of measurements-while-drilling and smart-well technologies, there is a large increase in the volume of data generated and to be analyzed. These data demand advanced computational tools to be employed for their processing and analysis. Therefore, data-driven machine learning models are found to be a more suitable candidate for processing complex oil and gas field data. These models can provide real-world solutions to several complex petroleum problems and will expedite automation of various field operations. The application of these data-driven models requires comprehensive investigations before their deployment at the real field level.

Quantitative lithofacies modeling is one of the most challenging parts of reservoir characterization that involves the identification of subsurface lithofacies [1]. Understanding of petrophysical properties of rocks and their spatial distribution in association with lithofacies are essential for the development of a reservoir model to produce hydrocarbon. Several intelligent machine learning models were applied for the

automatic identification of lithofacies through computational processing of well logs data. However, these methods require much improvement in their classification accuracy and generalization performance before they can be used in the real field scenario.

Drill bits and ROP are the important drilling parameters that are needed to be optimized for the success of drilling operations due to their large impact on operational efficacy and cost. Selecting the right bit types for drilling operations is still one of the most challenging tasks due to its dependency on various factors. Recently, data-driven intelligent models have been utilized to find suitable types of drill bits. However, none of them have considered the problem of imbalanced data that will naturally occur due to the varying thickness of subsurface lithofacies. The actual field data contain the uneven distribution of data samples that result in a complex imbalanced multiclass classification problem during drill bit selection.

Hydrocarbon production forecasting is an important task for reservoir management and production optimization. Several theoretical and empirical correlations have been proposed to estimate the hydrocarbon production rate. However, these methods are found to be less accurate and unreliable for hydrocarbon flow rate predictions. Data-driven models can be used for the estimation of the hydrocarbon flow rate, however, only limited applications of machine learning models are found in the literature on this important issue. Further, various input variables of production data are required to be examined for their importance and contribution in the estimation of hydrocarbon production rate.

To improve the performance of prevailing machine learning models, more reliable and accurate data-driven workflows, which recognize targeted operational parameters, have been incorporated in this thesis work. Recently, hybrid computational models,

such as ensemble methods, etc., are introduced for processing complex petroleum data [5]. These techniques are of significant importance, especially, when a high classification or estimation accuracy is targeted, as they can increase the generalization capabilities of the existing machine learning model by enhancing its modeling strategy [5]. The research work carried out in this thesis focuses on the applications of ensemble methods for lithofacies identification, suitable drill bit selection, optimization of drilling rate of penetration, and estimation of hydrocarbon production rate using diverse petroleum data. The main contributions of the research work carried out in this thesis are as follows.

- A new homogeneous ensemble-based workflow has been proposed for automatic identification and recognition of geological lithofacies for unconventional mudstone reservoirs.
- A novel application of heterogeneous ensemble methods has been performed to achieve a better generalization performance for the complex geological mudstone lithofacies.
- A novel method based on Response surface analysis and Artificial bee colony has been proposed for drill bit selection utilizing optimum values of drilling penetration rate (ROP).
- An innovative adaptation in ensemble methods along with resampling techniques has been proposed to resolve the imbalanced data issue encountered during the drill-bit selection process.
- A comprehensive study has been performed using ensemble methods to obtain the best performing forecasting models for oil and gas production through surface installed chokes.

- Systematic workflows and guidelines have been provided for the pre-processing of data and implementation of all proposed methods.

This chapter aims to highlight the main findings of the work carried out in this thesis and to make suggestions for future research work. This section concludes with an informative summary of all the work carried out and presented in the thesis.

7.2 Lithofacies modeling using homogeneous ensemble methods

Big data-driven ensemble methods, which is a novel approach for lithofacies modeling, have been critically examined in this study for their efficacy in enhancing the classification performance of the existing supervised classifiers, used in quantitative lithofacies modeling. Since the performance of ensemble methods is greatly influenced by the selection of base classifiers, therefore five most popular ensemble methods have been combined with seven base classifiers to examine their suitability in the modeling of geological lithofacies. Standard statistical metrics have been used to authenticate the classification performances of ensemble and base classifier combinations for quantitative lithofacies modeling. The research work carried out in the study directed to the following conclusions:

- Selection of the base classifiers for ensemble learners is found to be very crucial for the lithofacies modeling using well-logs data.
- Most of the ensemble methods have outperformed the corresponding single classifier based techniques found in the literature for quantitative lithofacies modeling.
- The suitable selection of base classifiers for ensembles have resulted in the following pairs: Bagging-CART/C4.5, AdaBoost-C4.5, Rotation forest-SVM, Random subspace-SVM, and DECORATE-C4.5

- The classification performance of three base classifiers are found to be in close competition with each other (CART, C4.5, and SVM) for quantitative lithofacies modeling.
- RBF is found to be the worse base classifier when used with any ensemble method. Therefore, it should be avoided as a base classifier for quantitative lithofacies modeling.
- SVM has recorded the highest overall classification accuracy with Random subspace ensemble for quantitative lithofacies modeling. Hence, it is found to be the most suitable ensemble base combination for quantitative lithofacies modeling among various HoEMs.

7.3 Lithofacies modeling using heterogeneous ensemble methods

In this study, two HEMs, namely Voting and Stacking, ensembles have been applied for the quantitative modeling of mudstone lithofacies using Kansas oil-field data. RF, gradient boosting (GB), SVM, and MLP have been incorporated as base classifiers in the applied HEMs architecture. A comprehensive comparison has also been performed among these classifiers for lithofacies identification. Multiple wells data have been considered to achieve better-generalized results for lithofacies modeling. A rigorous facies-wise comparison has been made between Stacking and Voting ensembles for the detection and identification of lithofacies. Stacking has shown nearly 4% and 2% improvement in test accuracy as compared to SVM and RF. Four popular machine learning algorithms have been combined in HEMs as base classifiers to provide more accurate and generalized results. The individual performance of the abovementioned classifiers has been evaluated with proper parameter optimization in their stable search ranges. The validation curve and grid search algorithm have been properly utilized for

the model parameters tuning to achieve maximum classification accuracy. The research work carried out in this study has led to the following conclusions.

- The performance of HEMs depends upon the selection of efficient base classifiers for quantitative lithofacies modeling.
- Validation curve has been found as an efficient measure for identifying stable search range for machine learning parameters.
- The Stacking ensemble has shown great potential to extract lithofacies information from well logs data.
- The training and testing classification accuracies of HEMs have been found highest among the other classifiers used in this study.
- DP layer is found to be the most challenging facies among all the nine target lithofacies. The Stacking ensemble has given the highest individual identification accuracy for all the layers of lithofacies.
- Prediction accuracy of individual facies ranges from 67.9 to 95.8% (worst to best possible testing accuracy), and maximum overall accuracy is (training=92.78% and testing=88.32%) obtained for Stacking ensemble.
- HEMs have shown their potential for quantitative lithofacies modeling and have outperformed the other classifiers.
- A combination of diverse base classifiers will lead to higher accuracy and better model generalization.

The analysis of results reveals that HEMs are practical and more accurate models, with a significant improvement in classification accuracy for lithofacies identification, as compared to the individual base classifiers.

7.4 Intelligent drill bit selection using Response surface methodology and Artificial bee colony

In this work, Response surface methodology (RSM) and Artificial bee colony (ABC) have been combined to choose different drill bit types based on the optimum values of ROP. RSM has been applied for the generation of the ROP objective function and further optimized it with ABC to search the optimum values of ROP and drill bit types. The performance of the proposed approach was also compared with the prevailing ANN, ANN, and GA models for drill bit selection and drilling optimization. This work provides an alternate intelligent approach for bit selection as compared to the popular ANN and GA combination model. The research work carried out in this study has led to the following conclusions.

- The proposed drill bit selection method was found to be more accurate than existing models.
- The correlation coefficient of the RSM objective function is found to be 81.23% while 85.5 % has been found for ANN during the estimation of ROP.
- The ROP objective function developed through RSM was found to be less complex than the ANN-based objective function due to the absence of an internal exponential function.
- It is observed that ANN needs more computational time for the generation and optimization of the ROP objective function.
- These models were found to be case-specific data-dependent models and involve calibration with other field data.

- This approach has presented positive results for the sustainable development of a more efficient, robust, reliable, and economical approach to achieve drilling optimization and cost reduction.

The proposed approach can effectively reduce the overall time and expenses involve in drilling operations that a company invests in a field by smartly selecting the optimum parameters of any newly planned oil and well.

7.5 Drill bit selection using ensemble methods and data resampling techniques

A novel data-driven approach has been proposed using the fusion of data resampling technique and ensemble method for handling the imbalance issues of complex drilling data. The problem of imbalanced training data results in the development of unreliable biased classifiers that are unfit for practical field applications. Two experimental data-driven scenarios have been specially designed and tested to confirm the generalization of the proposed approach for drill bit selection. An extensive comparative study has been performed to evaluate the performance of popular classifiers for the screening of drill bits. After a thorough comparison of results, the following important conclusions are drawn.

- Data imbalance condition exists due to the drilling of thin lithofacies which deteriorate the performance of the applied classifier.
- The performance of conventional classifiers is largely affected by data imbalance issues. Conventional classifiers can't be trusted for the drill bit selection, especially for critical drilling zones.
- The combination of ensemble methods with the data resampling technique results in modified ensemble classifiers that are found to be efficient even in highly imbalanced conditions.

- WBCRF technique has given the most impressive performance during automatic bit type selection with testing accuracy ranges from 92% to 99%, and *G-mean* (0.84–0.97) for various experimental scenarios.
- The large fluctuations in the performance of classifiers have been recorded in the 8.5” section in terms of precision, recall, and F1 scores. It is observed that drill bit selection becomes difficult in the lower formations due to uncertainty in subsurface conditions.
- WBCRF has given good prediction results for screening drill bit even for critical drilling zones.
- RF has shown great immunity for data imbalance conditions and successfully maintained its performance even in the critical 8.5” section.

The present study shows that the ensemble methods have great potential for automatic drill bit selection. The proposed approach can also be applied over any other oil and gas fields to automate the drill bit selection, which will minimize human error, time, and drilling cost.

7.6 Assessment of ML models for forecasting of hydrocarbon production

A rigorous investigation of five machine learning paradigms has been done in the quest for higher estimation accuracy in production forecasting. Volve oil and gas production field data were utilized for training and testing of machine learning models. The production data contained eight input variables that were pre-processed before the development and testing of various models. Random forest and ExtraTree are the first time implemented for production forecasting and have shown improvement in estimation accuracy as compared to LSSVR, SVR, and ANNs. The relevance of each predictor variable is also studied using statistical analysis and Relief algorithm. The contribution of this study is to identify the most suitable and robust intelligent

forecasting model for daily hydrocarbon production through surface chokes. The research work carried out in this study has led to the following conclusions.

- Ensemble paradigms are found to be an efficient means in correlating surface measured predictor variables with daily oil and gas production rates.
- Random forest and ExtraTree ensembles have outperformed the popular estimation models, viz. ANNs, SVR and LSSVR, for production forecasting.
- Random forest ensemble has given very high estimation accuracy with coefficient of correlation close to 0.98 and minimum prediction errors (RMSE value =0.045).
- The performance of LSSVR has been found slightly better than conventional SVR for production forecasting. However, proper tuning of the model's parameters is essential to obtain such performance.
- WHT has been identified as a maximum contributing variable for the estimation of daily hydrocarbon production using the Relief algorithm.

The analysis of estimation results indicates that Random forest and ExtraTree are powerful techniques for production forecasting. The estimated daily production of oil and gas, through surface chokes, has been found quite consistent with the field measured production data using Random forest ensemble.

7.7 Future scope

As a consequence of the work carried out in this thesis on the assessment of machine learning models for reservoir characterization and drilling automation, the following future areas of research are identified.

1. Ensemble methods are successfully investigated for depth-wise recognition of geological lithofacies. To achieve higher accuracy, these results are needed to be

integrated with seismic data of the geological formation. A deep learning neural network can be examined for developing a 3-D reservoir model that utilizes geochemical data, well logs, and pulsed neutron spectroscopy log data together.

2. More research is required to investigate other ensemble approaches such as bucket of models, cascading, random committee, clustering ensemble, etc. for their feasible implementation for quantitative lithofacies modeling.
3. Heterogeneous ensemble methods have been utilized for extracting useful geological facies information from well logs related to complex mud rock lithofacies. This study has used five popular base classifiers, other base classifiers and their combinations can be included in future research work for different mudrock reservoirs.
4. An ideal intelligent data-driven reservoir model should automatically estimate its properties along with depth-wise facies layers integrating well logs, geochemical data, seismic data, and pulsed neutron spectroscopy log. Multiagent-based systems can be explored for the integration of the different data sources for developing quantitative lithofacies modeling.
5. Drill bits selection has been performed based on two data-driven approaches. Initially, drill bits were selected based on the optimum values of the drilling penetration rate. Theoretically, this approach has given impressive results along with cost minimization of drill bits utilized for drilling the wells. This approach is required to be tested in real field drilling operations to test its efficacy in real-time scenarios.
6. Drill bit selection has also been formulated as a multiclass classification problem having an imbalanced data issue. Data resampling and boosting techniques have been combined with ensemble methods to tackle the formulated classification

problem. However, other approaches such as ‘adaptive algorithm’ and ‘cost-sensitive learning’ can also be explored to solve data imbalanced issues. These models are also needed to be tested on real field conditions.

7. The reinforcement learning approach can be explored for the automation of various drilling processes. Further, the reinforcement learning approach can be investigated with streaming drilling data for automatic decision-making in real-time.
8. Chapter 6 investigates several machine learning models for the production forecasting of hydrocarbon. The estimated daily production of oil and gas, through surface chokes, has been found quite consistent with the field measured production data using RF ensemble. However, there are several other types of intelligent techniques that can also be explored for modeling and optimization of multiphase flow through the surface chokes.

Few technical limitations exist in this thesis work are described as follows: (a) The proposed data-driven frameworks are required to be tested on real-field streaming data to evaluate the adaptability of applied machine learning models. (b) Multi-variety data sources must be considered for the pattern recognition tasks to increase their operational reliability in decision making. (c) The complexities of intelligent algorithms, data-driven frameworks, and associated processes make error diagnosis difficult. (d) Time constraints with big training data. (e) These data-driven applications require a large number of statistical tests for their verifications at a theoretical or conceptual level.

References

[1] S. Chaki, A. Routray, W.K. Mohanty, M., and Jenamani, "A novel multiclass SVM based framework to classify lithology from well logs: a real-world application", In IEEE INDICON.2015.

[2] P. Avseth, and T. Mukerji, T., "Seismic lithofacies classification from well logs using Statistical rock physics" *Petrophysics*, 2002, v. 43, pp. 70-81.

[3] B. Bhattacharya, R.C. Timothy, and M. Pal, "Comparison of Supervised and unsupervised approaches for mudstone lithofacies classification: case studies from the Bakken and Mahatango-Marcellus Shale, USA". *Natural Gas Science and Engineering*, 2016, v. 33, pp. 1119-1133.

[4] L. Wang, and C. A. Alexander, "Big data in the design and manufacturing engineering," *American Journal of Economics and Business Administration*, 2015, v. 7(2), pp. 60-67.

[5] F. Anifowose, J. Labadin, and A. Abdulraheem, "Improving the prediction of petroleum reservoir characterization with a stacked generalization ensemble model of support vector machines," *Applied Soft Computing*, 2014, v. 26, 483-496.

[6] H. Wolff, J. Pelissier-Combescure, "FACILOG-automatic eletrofacies determination", In Society of Professional Well Log Analysts Annual Logging Symposium, Paper FF. 1982.

[7] Y.Z. Ma, "Lithofacies clustering using principal component analysis and neural network", *Geosciences*, 2011, v. 43. pp. 401- 419.

[8] M. Raeesi, A. Moradzadeh, F.D. Ardejani, and M. Rahimi, M., "Classification and identification of hydrocarbon reservoir lithofacies and their heterogeneity using seismic attributes, logs data, and artificial neural networks," *Journal of Petroleum Science and Engineering*, 2012, v. 82-83, pp. 151-165.

[9] P. Bestagini, V. Lipari, and S. Tubaro, "A machine learning approach to facies classification using well logs." SEG Technical Program Expanded Abstracts, 2017.

[10] T. G. Dietterich, "An Experimental comparison of three methods for constructing ensembles of decision trees: Bagging, Boosting, and Randomization," Machine learning, 2002, v. 40(2), pp. 139-157.

[11] M. Skurichina, and R. P. W. Duin, “Bagging, boosting and the random subspace method for linear classifiers,” *Pattern analysis and applications*, 2001, v. 5, pp. 121-135.

[12] A. C. Aplin, and J.H.S. Macquaker, “Mudstone diversity: origin and implications for source, seal, and reservoir properties in petroleum systems,” *AAPG Bulletin*, 2011, v. 95(12), pp. 2031–2059.

[13] D. R. Spain, G. D. Merletti, and W. Dawson, “Beyond volumetric: unconventional petrophysics for efficient resource appraisal (example from the Khazzan field, Sultanate of Oman),” presented in Proceedings of 5th SPE Middle East unconventional resources conference and exhibition, 2015, Muscat, Oman.

[14] M. P. Sesmero, A. I. Ledezma, and A. Sanchis, “Generating ensembles of heterogeneous classifiers using stacked generalization,” Wiley Interdisciplinary Review Data Min Knowledge Discovery, 2015, v. 5(1), pp. 21–34.

[15] Bahari, A.; Seyed, A.B. “Drilling cost optimization in a hydrocarbon field by a combination of comparative and mathematical methods”, *Petroleum Science*, 2009, v. 6, pp. 451–463.

[16] Perrin, V.P.; Mensa-Wilmot, G.; Alexander, W.L. Drilling index-a new approach to bit performance evaluation. In Proceedings of the SPE/IADC Drilling Conference, Amsterdam, The Netherlands, 4–6 March 1997.

[17] R. Teale, “The concept of specific energy in rock drilling”, Int. *J. Rock Mech. Min. Sci. Geomech.* Abstr. 1965, v. 2, pp. 57–73 .

[18] A.K. Abbas, Assi, A.H.; Abbas, H.; Almubarak, H.; Al Saba, M. “Drill Bit “Selection Optimization Based on Rate of Penetration: Application of Artificial Neural Networks and Genetic Algorithms”, In Proceedings of the Abu Dhabi International Petroleum Exhibition & Conference, Abu Dhabi, UAE, 11 November 2019.

[19] W.J. Hightower, “Proper selection of drill bits and their use”, In Proceedings of the SPE Mechanical Engineering Aspects of Drilling and Production Symposium, Fort Worth, TX, USA, 23 March 1964.

[20] A.T. Bourgoyne, Jr.; Young, F.S., Jr. “A multiple regression approach to optimal drilling and abnormal pressure detection”, *Soc. Pet. Eng. J.* 1974, v. 14, pp. 371–384.

[21] H. Rabia, Farrelly, M.; Barr, M.V. “A new approach to drill bit selection”, In Proceedings of the SPE European Petroleum Conference, London, UK, 20 October 1985; pp. 20–22.

[22] M.J. Fear, Meany, N.C.; Evans, J.M. “An expert system for drill bit selection”, In Proceedings of the SPE/IADC Drilling Conference, Dallas, TX, USA, 15–18 February 1994.

[23] V. Uboldi, Civolani, L.; Zausa. F. “Rock strength measurements on cuttings as input data for optimizing drill bit selection”, In Proceedings of the SPE Annual Technical Conference and Exhibition, Houston, TX, USA, 3–6 October 1999.

[24] K. Bybee, “A New Approach for Drill-Bit Selection”’ *J. Pet. Technol.* 2000, v. 52, pp. 27–28.

[25] H.I Bilgesu, A.l-Rashidi, A.F; Aminian, K.; Ameri, S. “A new approach for drill bit selection”, In Proceedings of SPE Eastern Regional Meeting, London, UK, 17–19 October 2000.

[26] S. Yılmaz, Demircioglu, C.; Akin, S. “Application of artificial neural networks to optimum bit selection”, *Comput. Geosci*. 2002, v. 28, pp. 261–269.

[27] S. Edalatkhah, Rasoul, R.; Hashemi, A. “Bit selection optimization using artificial intelligence systems”, *Pet. Sci. Technol*. 2010, v. 28, pp. 1946–1956.

[28] M.S. Momeni, Ridha, S.; Hosseini, S.J.; Meyghani, B.; Emamian, S.S. Bit selection using field drilling data and mathematical investigation. In IOP Conference Series: Materials Science and Engineering; Penang, Malaysia, 2018; 328, 012008.

[29] M. Momeni, Hosseini, S.J.; Ridha, S.; Laruccia, M.B.; Liu, X. An optimum drill bit selection technique using artificial neural networks and genetic algorithms to increase the rate of penetration. *J. Eng. Sci. Technol*. 2018, v. 13, pp. 361–372.

[30] Kůrková, V. Kolmogorov’s theorem and multilayer neural networks. Neural Netw. 1992, v. 5, pp. 501–506.

[31] G. Villarrubia, De Paz, J.F.; Chamoso, P.; De la Prieta, F. Artificial neural networks used in optimization problems. *Neurocomputing* 2018, v. 272, pp. 10–16.

[32] R. Polikar,Pattern Recognition. Wiley Encyclopedia of Biomedical Engineering; *John Wiley & Sons, Inc*.: New York, NY, USA, 2006.

[33] A. Choubineh, Ghorbani, H., Wood, D. A., Robab Moosavi, S., Khalafi, E., & Sadatshojaei, E. “Improved predictions of wellhead choke liquid critical-flow

rates: Modelling based on hybrid neural network training learning based optimization". *Fuel*, 2017, v. 207, pp. 547–560.

[34] H. Parvizi, Rezaei-Gomari S., and Nabhani F., "Robust and flexible hydrocarbon production forecasting considering the heterogeneity impact for hydraulically fractured wells", Energy *& Fuels*, 2017, v. 31(8).

[35] A.Choubineh, Ghorbani, H., Wood, D. A., Robab Moosavi, S., Khalafi, E., & Sadatshojaei, E. "Improved predictions of wellhead choke liquid critical-flow rates: Modelling based on hybrid neural network training learning based optimization". *Fuel*, 2017, v. 207, pp. 547–560.

[36] H. Parvizi, Rezaei-Gomari S., and Nabhani F., "Robust and flexible hydrocarbon production forecasting considering the heterogeneity impact for hydraulically fractured wells". *Energy & Fuels*, 2017, v. 31(8).

[37] S. Chaki, Routray, A., Mohanty, W. K., and Jenamani, M., 2015. "A novel multiclass SVM based framework to classify lithology from well logs: a real-world application". Presented in Annual IEEE India Conference (INDICON).

[38] S.Wang, Chen, Z., Chen, S., "Applicability of deep neural networks on production forecasting in Bakken shale reservoirs", *J. of Petro. Sci. and Eng.*,2019, v. 179, pp.112-125,

[39] P. Panja, Velasco, R., Pathak, M., and Deo, M., "Application of artificial intelligence to forecast hydrocarbon production from shales". *Petroleum*, 2018, 4(1), v. 75–89.

[40] B. Guo, A.S. Al-Bemani, and A. Ghalambor, "Improvement in Sachdeva's multiphase choke flow model using field data". *J. Can. Petro. Tech*., 2007, 46, 5.

[41] H.R. Nasriani, and A. Kalantari, "Two-Phase Flow Choke Performance in High Rate Gas Condensate Wells. In Proceedings of SPE Asia Pacific Oil and Gas Conference and Exhibition,Jakarta, Malaysia, 2011

[42] A. Mirzaei-Paiaman and S. Salavati, "The application of artificial neural networks for the prediction of oil production flow rate, energy sources, Part A: Recovery, *Utilization, and Environmental Effects*, 2013, v. 34(19), pp. 1834-1843.

[43] A. Mirzaei-Paiaman, "An empirical correlation governing gas-condensate flow through chokes". *Petrol. Sci. Tech.*,2013, v. 31, pp. 368–379.

[44] G. Boyun, Al-Bemani, A., and Ali, G. “Applicability of Sachdeva’s Choke Flow Model in Southwest Louisiana Gas Condensate Wells”. In Proceedings of SPE Gas Technology Symposium, 2002.

[45] B. Guo, 2011 Petroleum Production Engineering, a Computer-Assisted Approach, Gulf Professional Publishing.

[46] J. D. Jansen, and Currie P.K., “Modeling and optimization of oil and gas production systems”. UDelft, Lecture notes for course production optimization, Version 5c 2004.

[47] J. P. Brill, and H. D. Beggs. “Two-Phase Flow through Pipes”, Sixth Edition. Tulsa, OK: Tulsa University Press, 1991.

[48] R. F. Tangren, Dodge C.H., Seifert H.S., “Compressibility effects in two phase flow”. *J. Appl. Phys*.,1949, v. 20, pp. 637–645.

[49] W. E. Gilbert, 1954. “Flowing and gas-lift well performance”. *API Drill. Prod. Prac*., v. 20, pp. 126–157.

[50] N.C.J. Ros, “An analysis of critical simultaneous gas/liquid flow through a restriction and its application to flow metering”, *Appl. Sci. Res*., 1960, v. 9, 374.

[51] I. Achong, “Revised bean performance formula for Lake Maracaibo wells, Internal Company Report, Shell Oil Co., Houston, TX. 1961.

[52] P. B. Baxendell, “Bean performance—lake wells”. Shell Internal Report.1957.

[53] M.A. Elhaj, F. Anifowose, and A. Abdulraheem A. “Single gas flow prediction through chokes using artificial intelligence techniques”, SPE Saudi Arab. Sect. Annu. Tech. Symp. Exhib. *Society of Petroleum Engineers*, 2015.

[54] O. E. Agwu, Akpabio, J. U., S. B. Alabi, A. Dosunmu, A., “Artificial intelligence techniques and their applications in drilling fluid engineering: A review”. *J of Petro. Sci*. and Eng., 2018, v. 160, pp. 300-315.

[55] I. Nejatia, M. Kanani, M. Arabloo, A. Bahadori, S. Zendehboudi, “Prediction of natural gas flow through chokes using support vector machine algorithm”, *J. Nat. Gas Sci. and Eng*. 2014, v. 18, pp. 155–163.

[56] R.G. Gorjaei, Songolzadeh R., Torkaman M., Safari M., Zargar G., “A novel PSO-LSSVM model for predicting liquid rate of two phase flow through wellhead chokes”, *J. Nat. Ga. Sci. Eng*., 2015, v. 24, pp. 228–237.

[57] S. Tewari, and U. D. Dwivedi, “Assessment of Big Data Analytics Based Ensemble Estimator Module for the Real-Time Prediction of Reservoir

Recovery Factor", In proceedings of SPE Middle East Oil and Gas Show and Conference, 18-21 March, Manama, Bahrain, 2019.

[58] D. F. Merriam, "The geologic history of Kansas. Kansas Geological Survey, Bulletin, 2019, v. 162, pp. 317 http://www.kgs.ku.edu/Publications/Bulletins/162/index.html

[59] K.D. Newell, "Overview of petroleum geology and production in Kansas. Kansas Geology Survey Bulliten, 1987a, v. 237. http://www.kgs.ku.edu/Publications/Bulletins/237/Newell2/overview.pdf

[60] K.D. Newell, "Sub-Chattanooga subcrop map of Salina basin, Kansas", Kansas Geological Survey, Open-file Report, 1987b, pp. 87-4. http://www.kgs.ku.edu/Publications/OFR/1987/OFR87_4/index.html

[61] F.J. Adler, W. M. Caplan, M.P. Carlson, E.D.Goebel, H.T. Henslee, I.C. Hicks, T.G. Larson, M.H. McCracken, M.C. Parker, B. Rascoe, M.W. Schramm, J.S.Wells, "Future petroleum provinces of the midcontinent. In: Cram IH (ed) Future petroleum provinces of the United States—their geology and potential", *American Association of Petroleum Geologists*, Memoir, Tulsa, 1971, pp 985–1120

[62] J.M. Jewett, D.F. Merriam, "Geologic framework of Kansas—a review for geophysicists. In: Hambleton WW (ed) Proceedings of a symposium on geophysics in Kansas, Kansas Geological Survey Bulletin, 1959, v. 137, pp 9–52.

[63] Allaud, Louis, M. Martin, "Schlumberger: the History of a Technique," New York: *Wiley*, 1977.

[64] J. Kittler, "Combining classifiers: A theoretical framework*," Pattern Analysis & Application,* 1998, v. 1, pp. 18–27.

[65] L.I. Kuncheva, "Classifier Ensembles for Changing Environments". In: Roli F., Kittler J., Windeatt T. (eds) Multiple Classifier Systems. MCS 2004. Lecture Notes in Computer Science, v. 3077. Springer, Berlin, Heidelberg.

[66] L. Kuncheva, C. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," Machine Learning, 2003, v. 51(2), pp. 181-207.

[67] R. Polikar "Ensemble based systems in decision making," IEEE Circuits and Systems Magazine, 2006, v. 6(3), pp. 21–45.

[68] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, 1989, v. 84, pp. 165-175.

[69] G. An, "The effects of adding noise during backpropagation training on a generalization performance," *Neural Computation*, 2006, v. 8, pp. 643-674.

[70] L. Breiman, "Bagging predictors" Machine Learning, 1996, v. 24(2), pp. 123-140.

[71] Y. Freund, and R. E. Schapire, "Experiments with a new boosting algorithm," presented in Thirteenth International Conference in Machine learning, 1996.

[72] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 20(8), pp. 832-844.

[73] P. Melville, and R. J. Moonet, "Creating diversity in an ensemble using artificial data," *Information Fusion*, 2005, v. 6(1), pp. 99-111.

[74] D. H. Wolpert, "Stacked generalization," *Neural Networks.* 1992, v. 5(2), pp. 241–259.

[75] F. Anifowose, J. Labadin, A. Abdulraheem, "Improving the prediction of petroleum reservoir characterization with a stacked generalization ensemble model of support vector machines," *Applied Soft Computing*. 2015, v. 26, pp. 483–496.

[76] Kansas Geological Survey (KGS) (2009). http://www.kgs.ku.edu/Magellan/Logs/

[77] S. Tewari , U.D. Dwivedi, "A real world investigation of Twin SVM classifier for classification of petroleum drilling data". In: Proceeding of IEEE TENSYMP 2019 symposium, Kolkata, India, 2019.

[78] M. B. Diaz, K. Y. Kim, T. H. Kang, H.S. Shin, "Drilling data from an enhanced geothermal project and its pre-processing ROP forecasting improvement," *Geothermics*. 2018, v. 72, pp. 348–357.

[79] R. Kohavi, "A study of cross validation and bootstrap for accuracy estimation and model selection," In proceedings of 14th International Joint Conference on Artificial Intelligence. 1995, v. 2, pp. 1137–1143.

[80] J. Friedman, "Greedy function approximation: a gradient boosting machine," *The Annals of Statistics*, v. 29(5), pp. 1189–1232.

[81] T.K. Ho, “Random decision forest,” proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, 1995, pp 278–282.

[82] C. Cortes, and V. Vapnik, Support-vector networks. Machine Learning, 1995, v. 20(3), pp. 273–297.

[83] A. Bahari, and A.B. Seyed, “Drilling cost optimization in a hydrocarbon field by a combination of comparative and mathematical methods”, *Petroleum Science*. 2009, v. 6, pp. 451–463.

[84] R. Teale, “The concept of specific energy in rock drilling. *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts*. 1965. v. 2(1), pp. 57-73. 965.

[85] V.P. Perrin, G. Mensa-Wilmot, and W.L. Alexander, “Drilling index-a new approach to bit performance evaluation”, In Proceedings of SPE/IADC drilling conference, Amsterdam, Netherlands 4-6 March,1997.

[86] V. Kůrková, “Kolmogorov’s theorem and multilayer neural networks,” *Neural Networks*. 1992, v. 5(3), pp. 501–506.

[87] G. Villarrubia, J.F. De Paz, P. Chamoso, and F. De la Prieta, “Artificial neural networks used in optimization problems,” *Neurocomputing*, 2018, v. 272, pp. 10-16.

[88] R. Polikar, “Pattern recognition”, Wiley Encyclopedia of Biomedical Engineering. *John Wiley & Sons*, Inc. 2006.

[89] A. Muthiah, and R. Rajkumar, “A comparison of artificial bee colony algorithm and genetic algorithm to minimize the makespan for job shop scheduling,” *Procedia Engineering*. 2014, v. 97, pp. 1745-1754.

[90] R. H. Myers, A. I. Khuri, and H. C. Jr. Walter, “Response surface methodology”, 1989, v. 31(2), pp. 137-157.

[91] D. C. Montgomery, “Design and analysis of experiments”. John wiley & sons, Inc. 2014.

[92] Z. N. M. Alqattan, and R.A. Abdullah, “Comparison between Artificial Bee Colony and Particle Swarm Optimization Algorithms for Protein Structure Prediction Problem”, *Lecture Notes in Computer Science*, 2013, pp. 331–340.

[93] D. Karaboga, B. Basturk, “A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm,” *Journal of Global Optimization*. 2007, v. 39(3), pp. 459–471.

[94] R. Hecht-Nielsen, “Kolmogorov’s mapping neural network existence theorem,” In Proceedings of the IEEE First International Conference on Neural Networks. San Diego, CA, IEEE, 1987, v. 989, pp. 11–14.

[95] G. E. P. Box; K.B.Wilson, “On the Experimental Attainment of Optimum Conditions,” In: Kotz S., Johnson N.L. (eds) Breakthroughs in Statistics. Springer Series in Statistics (Perspectives in Statistics). *Springer*, New York, Morgantown, West Virginia, 1993.

[96] R. H. Myers, D.C. Montgomery, C.M. Anderson-Cook, “Response Surface Methodology: Process and Product Optimization Using Designed Experiments”, *John Wiley& Sons. Inc*., New York, NY, 1995, pp.134-174.

[97] D. Karaboga, “An idea based on honey bee swarm for numerical optimization,” Technical report-tr06, Erciyes University, Engineering Faculty, Computer Engineering Department. 2005. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=E32274EB05825D919807935CF722606F?doi=10.1.1.714.4934&rep=rep1&type=pdf. Accessed 29 October 2020.

[98] Y. Zhao,Noorbakhsh, A., Koopialipoor, M. et al. A new methodology for optimization and prediction of rate of penetration during drilling operations. *Engineering with Computers,* 2020, v. 36, pp. 587–595.

[99] B. Nozohour-leilabady, B. Fazelabdolabadi, “On the application of artificial bee colony (ABC) algorithm for optimization of well placements in fractured reservoirs; efficiency comparison with the particle swarm optimization (PSO) methodology”, *Petroleum,* 2016, v. 2(1), pp. 79–89.

[100] M. Koopialipoor, E. N. Ghaleini, M. Haghighi et al. “Overbreak prediction and optimization in tunnel using neural network and bee colony techniques”, *Engineering with Computers*, 2019, v. 35, pp. 1191–1202.

[101] E.N. Ghaleini, M. Koopialipoor, M. Momenzadeh, M.E. Sarafraz, E.T. Mohamad, B. Gordan, “A combination of artificial bee colony and neural network for approximating the safety factor of retaining walls”, *Engineering with Computer*s. 2018. 1–12.

[102] A. Ahmad, S. F. M. Razali, Z. S. Mohamed, and A. El-shafie, “The application of artificial bee colony and gravitational search algorithm in reservoir optimization”, *Water Resource Management*, 2018, v. 30(7), pp. 2497–2516.

[103] Equinor website database. https://www.equinor.com/en/how-and-why/digitalisation-in-our-dna/ volve-field-data-village-download.html. (Accessed on 17 August 2020.)

[104] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators", *Neural Networks*, 1989, v. 2(5), pp.359–366.

[105] K. Zorlu; C. Gokceoglu, F. Ocakoglu, H. A. Nefeslioglu, and S. Acikalin, "Prediction of uniaxial compressive strength of sandstones using petrography-based models", *Engineering Geology*, 2008, v. 96(3), pp.141–158.

[106] H. Demuth, M. Beale, and M. Hagan, "Artificial Neural Networks & Genetic Algorithms User's Guide", Revised for Matlab Programming, 2007, v. 5.

[107] R. Hecht-Nielsen Kolmogorov's mapping neural network existence theorem. In: Proceedings of the international joint conference in neural networks, 1989, v. 3, pp. 11–14.

[108] B. D. Ripley, "Statistical aspects of neural networks Networks and Chaos: Statistical and Probabilistic Aspects", 1993, v. 50, pp. 40–123.

[109] J. D. Paola, "Neural network classification of multispectral imagery", Master Tezi University, Arizona. 1994.

[110] C. Wang, "A theory of generalization in learning machines with neural applications". Ph.D. thesis, The University of Pennsylvania, U.S.A., 1994.

[111] T. Masters, Practical neural network recipes in C++. Morgan Kaufmann, Burlington, 1993.

[112] I. Kanellopoulos, G. G. Wilkinson, "Strategies and best practice for neural network image classification", *International Journal of Remote Sensing*. 1997,v. 18(4), pp. 711–725.

[113] I. Kaastra, and M. Boyd, "Designing a neural network for forecasting financial and economic time series" *Neurocomputing*, 1996, v. 10(3), pp. 215–236.

[114] D. Lindenmayer, M.A. Burgman, "Monitoring, assessment, and indicators", *Practical Conservation Biology*, 2005. pp. 401-424.

[115] S. Schlotzhauer, Elementary statistics using JMP. SAS Publishing. 2007.

[116] D. Karaboga, and B. Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm", *Journal of Global Optimization*. 2007, v. 39(3), pp. 459–471.

[117] R. Pessier, and M. Damschen, “Hybrid bits offer distinct advantages in selected roller-cone and pdc-bit applications”, SPE Drilling and Completion, 2011, v. 26(01), pp. 96–103.

[118] Longadge, R.; Dongre, S. Class imbalance problem in data mining review. arXiv 2013, arXiv:1305.1707. Available online:http://arxiv.org/abs/1305.1707 (accessed on 16 August 2019).

[119] Ghorbani, R.; Ghousi, R. Comparing Different Resampling Methods in Predicting Students’ Performance Using Machine Learning Techniques. IEEE Access 2020, 8, 67899–67911.

[120] Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 2002, 16, 321–357.

[121] Lhassan, T.; Aljurf, M.; Al-Mohanna, F.; Shoukri, M. Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method. J. Inform. Data Min. 2016, 1, 1–11.

[122] C. Zang, and Y. Ma, “Ensemble machine learning: Methods and application”, *Springer publication*, 2012, v. 8, pp. 332

[123] Y. Sun, A. K. C.Wong, M.S. Kamel, “Classification of imbalanced data: a review”, *International Journal of Pattern Recognition and Artificial Intelligence*, 2009, v. 23(04), pp. 687–719.

[124] L. Breiman, “Random forests”, *Machine Learning*, 2001, v. 45, pp. 5–32.

[125] A. Cutler, D. R. Cutler, J. R. Stevens, “Random forests”, In Ensemble Machine Learning, 2nd ed.; Zhang, C., Ma, Y.Q., Eds., *Springer*, New York, 2012, pp.157-175.

[126] T.K. Ho, “Random decision forests”, In Proceedings of 3rd International conference on document analysis and recognition. 14-16 August, 1995.

[127] C. Chen , A. Liaw, L. Bremain, “Using Random forest to learn Imbalanced data”, Report number 666, 2004.https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf (Accessed online 01-11-2020).

[128] T. Cover, and P. Hart, “Nearest neighbor pattern classification”, *IEEE Transactions on Information Theory*, 1967, v. 13(1), pp. 21–27.

[129] A. McCallum, and K. Nigam, “A comparison of event models for naive bayes text classification”, In Proceedings of AAAI’98 Workshop on Learning for Text Categorization, 1948, pp. 41–48.

[130] V. R. Balaji, S. T. Suganthi, R. Rajadevi, V.K. Kumar, B.S. Balaji, S. Pandiyan, “Skin disease detection and segmentation using dynamic graph cut algorithm and classification through Naive Bayes Classifier”, *Measurement,* 1920. pp. 107922.

[131] S. Haykin, Neural networks: A comprehensive foundation. Prentice Hall PTR, Upper Saddle River, 1998.

[132] V. Vapnik, “The nature of statistical learning”, *Springer*-Verlag New York, 2000.

[133] C. Cortes, and V. Vapnik, “Support-vector networks”, *Machine learning*, 1995, v. 20(3), pp. 273-297.

[134] Equinor website database. https://www.equinor.com/en/how-and-why/digitalisation-in-our-dna/volve-field-data-village-download.html. (Accessed on 17August, 2020.)

[135] E.A Løken, J. Løkkevik, and D. Sui, “Data-driven approaches tests on a laboratory drilling system”, *Journal of Petroleum Exploration Production Technology*, 2020, v. 10, pp. 3043–3055.

[136] G. N. Karadzhova, “Drilling efficiency and stability comparison between tricone, PDC and Kymera Drill Bits”, Master’s thesis, University of Stavanger, 2014.

[137] O. A. Akisanmi, “Automatic management of rate of penetration in heterogeneous formation rocks”, Master’s thesis, University of Stavanger, 2016.

[138] E. Kenneth, and S. C. Russel, “Innovative ability to change drilling responses of a PDC bit at the rigsite using interchangeable depth of-cut control features”, In Proceedings of IADC/SPE Drilling conference and exhibition, 2016, SPE-178808-MS.

[139] Jayadeva, K. Kumar, and G. R. Naik, “TwinSVM gesture classification using the surface Electromyogram”, *IEEE Transactions on Information Technology in Biomedicine*, 2010, v. 14(2), pp. 301–308.

[140] M. B. Diaz, K. Y. Kim, T. H. Kang, and H. S. Shin, “Drilling data from an enhanced geothermal project and its pre-processing ROP forecasting improvement”, *Geothermics*. 2018, v. 72, pp. 348–357.

[141] Z. Mustaffa, and Y. Yusof, “A comparison of normalization techniques in predicting dengue outbreak”, In Proceedings of International Conference on Business and Economics Research, IACSIT Press, Kuala Lumpur, Malaysia, 2010.

[142] S. Tewari, and U. D. Dwivedi, “A comparative study of heterogeneous ensemble methods for the identification of geological lithofacies”, *Journal of Petroleum Exploration and Production* Technology, 2020, pp. 1-20.

[143] C. Gan, W. H. Cao, M. Wu, X. Chen, Y.L. Hu, K. Z. Liu, et al. “Prediction of drilling rate of penetration (ROP) using hybrid support vector regression: a case study on the Shennongjia area, Central China”’ *Journal of Petroleum Science and Engineering*, 2019, v. 181, pp. 1060200.

[144] R. Kohavi, “A study of cross validation and bootstrap for accuracy estimation and model selection”, Proceedings of the 14th International Joint Conference Artificial Intelligence. 1995, v. 2, pp. 1137–1143.

[145] M. Kubat, R. Holte, and S. Matwin, “Learning when negative examples abound”’. In Proceedings of the 9th European Conference on Machine Learning. LNCS, Springer, London, 1995, pp. 146–153.

[146] R. Pessier, and M. Damschen, “Hybrid bits offer distinct advantages in selected roller-cone and pdc-bit applications”, *SPE Drilling and Completion*, 2011, v. 26(01), pp. 96–103.

[147] A. Choubineh, H. Ghorbani, H., D. A. Wood, S. Robab Moosavi, E. Khalafi, and E. Sadatshojaei, “Improved predictions of wellhead choke liquid critical-flow rates: Modelling based on hybrid neural network training learning based optimization”. *Fuel*, 2017, v. 207, pp. 547–560.

[148] H. Parvizi, S. Rezaei-Gomari, and F. Nabhani, “Robust and flexible hydrocarbon production forecasting considering the heterogeneity impact for hydraulically fractured wells”. *Energy Fuels*, 2019, 31(8), pp. 8481–8488.

[149] S.Chaki, Routray, A., Mohanty, W. K., and Jenamani, M., 2015. “A novel multiclass SVM based framework to classify lithology from well logs: a real-world application”. Presented in Annual IEEE India Conference (INDICON).

[150] S. Wang, Z. Chen, and S. Chen, “Applicability of deep neural networks on production forecasting in Bakken shale reservoirs”, *J. of Petro. Sci. and Eng*., 2019, v. 179, pp.112-125.

[151] P. Panja, R. Velasco, M. Pathak, and M. Deo, “Application of artificial intelligence to forecast hydrocarbon production from shales”. *Petroleum*, 2018, 4(1), pp. 75–89.

[152] B. Guo, A. S. Al-Bemani, and A. Ghalambor, “Improvement in Sachdeva’s multiphase choke flow model using field data”. *J. Can. Petro. Tech*., 2007, v. 46, pp. 5.

[153] H. R. Nasriani, and A.S.L. Kalantari, “Two-Phase Flow Choke Performance in High Rate Gas Condensate Wells”, In Proceedings of SPE Asia Pacific Oil and Gas Conference and Exhibition, 2011.

[154] A. Mirzaei-Paiaman and S. Salavati S., “The application of artificial neural networks for the prediction of oil production flow rate, energy sources, Part A: Recovery, *Utilization, and Environmental Effects*, 2013, v. 34(19), pp. 1834-1843.

[155] A. Mirzaei-Paiaman, “An empirical correlation governing gas-condensate flow through chokes”. *Petrol. Sci. Tech*., 2013, v. 31, pp. 368–379.

[156] G. Boyun, A. Al-Bemani, and G. Ali, “Applicability of Sachdeva’s Choke Flow Model in Southwest Louisiana Gas Condensate Wells”. Proceedings of SPE Gas Technology Symposium, Alberta, Canada, 2002.

[157] B. Guo., Petroleum Production Engineering, a Computer-Assisted Approach, Gulf Professional Publishing, 2011.

[158] J. D. Jansen, and P.K. Currie, “Modeling and optimization of oil and gas production systems”. UDelft, Lecture notes for course production optimization, Version 5c, 2004.

[159] J. P. Brill, and H.D. Beggs, “Two-Phase Flow through Pipes”, Sixth Edition. Tulsa, OK: Tulsa University Press. 1991.

[160] Y. Xie, C. Zhu, W. Zhou, Z. Li, X. Liu, and M. Tu, “Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances” *Journal of Petroleum Science and Engineering,* 2017,v. 160, pp. 182-193.

[161] S. Tewari, U.D. Dwivedi, “Ensemble-based Big Data Analytics of Lithofacies for Automatic Development of Petroleum Reservoirs”, *Comp. & Indus.* Eng. 2018, v.128, pp. 937-947.

[162] C. Hedge, H. Daigle, H. Millwater, and H. Gray, "Analysis of rate of penetration in drilling using physic-based and data-driven models". *J. Petro. Sci. and Eng.*, 2017, v. 159, pp. 295-306.

[163] O. E. Agwu, J. U. Akpabio, S. B. Alabi, and A. Dosunmu, "Artificial intelligence techniques and their applications in drilling fluid engineering: A review". *J of Petro. Sci. and Eng*., 2018, v.160, pp. 300-315.

[164] L. Breiman, "Random Forests". *Mach. Learn*. 2001, v. 45, pp.5-32.

[165] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random Forests", *Ensemble Machine Learning*, 2012, pp. 157–175.

[166] T. K. Ho, "Random decision forests. In Document analysis and recognition", Proceedings of the third international conference, Montreal, Quebec, Canada, 1995, v. 1, pp. 278–282.

[167] L. Breiman, "Randomizing outputs to increase prediction accuracy". *Mach. Learn*., 2000, v. 40(3), pp. 229–242.

[168] P. Geurts, D. Ernst and L. Wehenkel, "Extremely Randomized Trees". *Mach. Learn*. 2006, v. 63, 3-42.

[169] M. Ravasi, I. Vasconcelos, A. Curtis, and A. Kristi, A., "Vector-acoustic reverse time migration of Volve ocean-bottomcable data set without up/down decomposed wavefields", *Geophysics*, 2015, v. 80(4), pp.S137-S150.

[170] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics.*2018, v. 85, pp. 189-203

[171] R. Kohavi, "A study of cross validation and bootstrap for accuracy estimation and model selection". Presented at International Joint Conference on Artificial Intelligence (IJCAI) Conference. 1995. v. 2, pp. 1137-1143.

[172] H.H Al-Attar., and G.H. Abdul-Majeed, "Revised bean performance equation for East Baghdad oil wells". *SPE Prod. Eng*. 1988, v. 3, pp.127–131.

[173] A. Mirzaei-Paiaman and S. Salavati, "The application of artificial neural networks for the prediction of oil production flow rate, energy sources, Part A: Recovery, *Utilization, and Environmental Effects*,2013, v. 34(19), pp. 1834-1843.

[174] A. Mirzaei-Paiaman, "An empirical correlation governing gas-condensate flow through chokes". *Petrol. Sci. Tech*., 2013, v. 31, pp. 368–379.

[175] H. H., Al-Attar, and Abdul-Majeed G.H., "Revised bean performance equation for East Baghdad oil wells". *SPE Prod. Eng*.1988, v. 3, pp.127–131.

[176] X. Kong, Y. Sun, R. Su, and X. Shi, "Real-time eutrophication status evaluation of coastal waters using support vector machine with grid search algorithm". *Marine Pollution Bulletin*, 2017, v. 119, pp. 307–319.

[177] H. A. Fayed, and A. F. Atiya, "Speed up grid-search for parameter selection of support vector machines". *Applied Soft Computing,* 2019, v. 80, pp. 202-210.

[178] F. A. Anifowose, J. Labadin, A. Abdulraheem, "Ensemble machine learning: An untapped modeling paradigm for petroleum reservoir characterization". *J of Petro. Sci. and Eng*., 2017, v. 151, pp. 480-487.

Appendix A

Table A1. Different drill bit utilized for drilling of Volve wells at different depths.

Wells	Depth In-Out (m)	Bit Type	IADC Code	Bit Size (Inch)
F-4	260–1360	1	M115 PDC	17.5
	1360–2770	2	M422 PDC	12.25
	2770–3510	3	M222 PDC	8.5
F-5	230–1415	4	M115 PDC	17.5
	1415–2930	5	M223 PDC	12.25
	2930–3785	6	M323 PDC	8.5
F-7	217–307	1	M115 PDC	17.5
	915–1080	6	M115A PDC	12.25
F-9	216–918	10	Smith XR+VEC MTMZ2069	17.5
F-10	146–207	7	XR+VEC MZ25069	36
	207–1400	8	XR+MG04B	26
	1400–1463	9	Reed RSRT16M-C9	17.5
	1463–2616	9	Reed RSR716M-C9	17.5
	2616–2825	10	Smith MDI716	12.25
	2825–3319	10	SmithMDI716	12.25
	3319–3442	11	Reed Hycalog RSX8195219S960	12.25
	3442–3695	9	Reed Hycalog RSR716D	8.5
	3695–4911	13	Reed Hycalog RSR816H-C1	8.5
	4911–5311	13	Reed Hycalog RST816H-C1	8.5
F-12	365–1365	14	M415 PDC	26
	1365–2510	15	M322 PDC	17.5
	2510–2570	16	135 Milled Tooth	17.5

	2570–3110	2	M422 PDC	12.25
	3110–3515	3	M222 PDC	8.5
F-14	251–1369	14	M415M PDC	26”
	1369–2513	5	M322 PDC	17.5”
	2513–2573	16	MT 135	17.5”
	2573–3114	2	PDC M422	12.25”
	3114–3520	3	PDC M222	8.5”
F-15	144–226	1	M115 PDC	17.5”
	226–1378	6	M115A PDC	26”
	1378–1381	12	Reed Hycalog RST816H-C2	17.5”
	1381–2480	19	M333 PDC	17.5”
	2480–2536	18	M332 PDC	12.25”
	2536–3670	5	M323 PDC	8.5
	3670–4090	5	M323 PDC	8.5”

Table A2. Selection of the optimum number of neurons in the hidden layer of MLP based on minimum training error.

Hidden Layer Neurons.	**Neural Network**	**Average Percentage Error**
1	15-1-19	12.95
2	15-2-19	10.724
3	15-3-19	12.1
4	15-4-19	7.42
5	15-5-19	3.97
6	15-6-19	3.26
7	15-7-19	3.6
8	15-8-19	3.44
9	15-9-19	3.09
10	15-10-19	3.88
11	15-11-19	4.59
12	15-12-19	2.91
13	15-13-19	3.67

14	15-14-19	1.55
15	15-15-19	2.91
16	15-16-19	1.85
17	15-17-19	2.38
18	15-18-19	3.53
19	15-19-19	3.03
20	15-20-19	3.83

Table A3. A comparative study of significant methods applied for drill bit selection.

S. No.	Publication	Techniques	Field Details	Data Types	Advantages	Limitations
1.	Rabia (1985)[6]	Cost per foot	Unavailable	Operational drilling parameters	Simple to apply and empirical in nature	Unfit for horizontal and multilateral drilling operation and low accuracy
2.	Rabia et al. (1986)[5]	Specific Energy	Southern North Sea	Operational drilling parameters	Simple to apply	Based on only three operational parameter and low accuracy
4.	Hightower, (1964)[8]	Offset Well logs	County of East Texas	Well logs	Easy application	Indirect measurement of rock properties with high chances of error risk
5.	Perrin et al. (1997)[10]	drilling index	Unavailable	Operational drilling parameters	empirical correlation and can be easily	Low accuracy and high chances of

					applied	error
6.	Xu et al. (1997)[12]	Empirical modeling	Unavailable	Mud logging data, operational drilling data	Improved cost per foot model	Mathematically complex and requires more data
7.	Mensa-Wilmot et al. (1999)[11]	Formation drillability parameter	Unavailable	Rock mechanical and geologic properties	Indirect measurements, and more accurate than empirical correlations	High level of uncertainty in well logs data due to hydrocarbon reservoir heterogeneity
10.	Uboldi et al. (1999)[13]	Rock strength measurements and indentation technique	Southern Italy, near Apennines chain	Rock mechanical and geologic properties	Accurate measurement of core properties	Require testing lab, costly and time-consuming
11.	Bilgesu et al. (2000)[2]	ANN	Middle East field	BS, WOB, RPM, pump rate, DT, and BT	High prediction accuracy	Not immune to imbalanced data, noise, overfitting and underfitting problems
12.	Yılmaz, et al. (2002)[17]	ANN and fractal geostatistics	southeast Turkey.	rock bit data	High prediction accuracy	Not immune to imbalanced data, noise, overfitting and underfitting problems
13.	Bataee et al. (2010)[7]	bit dullness	Shadegan oil field	Human experience	Simple, empirical, and organized.	Requires human visual expertise with high chances of error

14.	Edalatkhah, Rasoul, and Hashemi, (2010)[18]	ANN and Genetic algorithm	South Pars Field	Drilling operational data	More accurate than individual ANN model	Not immune to imbalanced data, noise, overfitting and underfitting problems
15.	Hou, Chien, and Yuan, (2014)[19]	ANN	Tarim Oilfield,	offset wells data, drillability, and lithofacies information	More accurate than empirical models	Not immune to imbalanced data, noise, overfitting and underfitting problems
16.	Sherbeny et al. (2016)[14]	Image logs and mineralogy logs	Unavailable	Image data and mineralogy data	Accurate method	Application in limited lithofacies, computationally challenging, and costly technology.
17.	Nabilou (2016)[20]	Resistance against Drilling	Southwest of Iran	Geo-Mechanical data	More accurate than the empirical correlation	Application in limited lithofacies, computationally challenging and costly technology
18.	Mardiana and Noviasta (2017)[15]	Rock strength analysis and Dynamic Finite-Element Analysis (FEA) Modeling	Unavailable	Offset well logs data	More accurate than the empirical correlation	Application in limited lithofacies, computationally challenging and costly technology
19.	Momeni et al. (2018)[22]	ANN	Unavailable	drilling bit records from offset	More accurate than the empirical	Not immune to imbalanced data, noise,

				wells	correlation	overfitting and underfitting problems
20.	Cornel and Vazquez (2020)[16]	Rock Strength Analysis and bit dull grading approach	South West of Wandoan, Queensland	dull grading and bit records	More accurate than the empirical correlations	Application in limited lithofacies, computationally challenging and costly technology
21.	Abbas et al. (2019)[23]	ANN, Genetic algorithm and Mechanical earth model	Unavailable	Operational drilling parameters	More accurate than individual ANN model and empirical models	Not immune to imbalanced data, noise, overfitting and underfitting problems
22.	Proposed Approach	Ensemble methods and Resampling techniques	North Sea	Operational drilling parameters and mud logging data	More reliable, stable, and accurate than previous models	Need to be tested on field with streaming data conditions.

LIST OF PUBLICATIONS

Journal Publications

1. **Tewari, S.,** Dwivedi, U.D., and Biswas, S. (2021) A Novel Application of Ensemble Methods with Data Resampling Techniques for Drill Bit Selection in the Oil and Gas Industry. Energies 2021, 14, 432.

2. **Tewari, S.,** and Dwivedi, U.D. (2021). Assessment of Machine Learning Models for Forecasting Hydrocarbon Production through Surface Chokes. Petroleum Science, Under Review.

3. **Tewari, S.,** Dwivedi, U.D., and Biswas, S. (2021) Intelligent Drilling of Oil and Gas Wells using Response Surface Methodology and Artificial Bee Colony, Sustainability, Accepted.

4. **Tewari, S.,** and Dwivedi, U. D. (2020). A comparative study of heterogeneous ensemble methods for the identification of geological lithofacies. Journal of Petroleum Exploration and Production Technology, 1-20.

5. **Tewari, S.,** and Dwivedi, U. D. (2019). Ensemble-based big data analytics of lithofacies for automatic development of petroleum reservoirs. Computers & Industrial Engineering, 128, 937-947.

Conference Publications

1. **Tewari, S.,** and Dwivedi U. D., 2019. A Real-World Investigation of TwinSVM for the Classification of Petroleum Drilling Data. IEEE R 10, Symposium. IEEE Xplore Digital library.
2. **Tewari, S**. 2019. Ensemble Methods: An Intelligent Modelling Approach for Oil & Gas Industry, EAGE SCOPE, Magzine.
3. **Tewari, S.,** and Dwivedi, U. D., 2019. Assessment of Big Data Analytics Based Ensemble Estimator Module for the Real-Time Prediction of Reservoir Recovery Factor. SPE-194996-MS. SPE Middle East Oil and Gas Show and Conference, 18-21 March, Manama, Bahrain.

4. **Tewari, S**. (2019). Assessment of Data-Driven Ensemble Methods for Conserving Wellbore Stability in Deviated Wells. SPE SPE-123456-MS, Annual Technical Conference and Exhibition, Calgary, Canada.
5. **Tewari, S**., and Dwivedi, U. D., 2018. A Novel Automatic Detection and Diagnosis Module for Quantitative Lithofacies Modelling. SPE-192747-MS. Abu Dhabi International Petroleum Exhibition & Conference, 12-15 November, Abu Dhabi, UAE.
6. **Tewari, S.** (2019). Assessment of Data-Driven Ensemble Methods for Conserving Wellbore Stability in Deviated Wells. SPE SPE-123456-MS, Annual Technical Conference and Exhibition, Calgary, Canada.
7. **Tewari, S.** and Dwivedi, U.D. 2019. Assessment of Big Data Analytics for the Optimization of Drilling Efficiency in Real field Applications, In Proceeding of PetroTech 2019, Delhi, India.
8. **Tewari, S.** and Dwivedi, U.D. 2017. Development and testing of Nu SVR model for drilling mud density estimation of HPHT wells, In Proceedings of Challenges and Prospects of Petroleum Production and Processing Industries, IIT Dhanbad.

LIST OF WORKSHOPS/WEBINARS/ SHORT TERM TRAINING COURSES ATTENDED

1. Attended a workshop on “Advanced Pattern Recognition Techniques” at the department of computer science and engineering, M.N.I.T., India. 2018.

2. Attended a workshop on “Flow Assurance in the Petroleum Industry” at the department of ocean engineering, GIAN, Indian Institute of Technology, Chennai, 2017.

3. Attended a workshop on “Multi-objective Optimization Using Metaheuristics” at the department of industrial engineering, Indian Institute of Technology, Kanpur, India.

4. Attended a workshop on “Petroleum reservoir and characterization” at the department of ocean engineering, GIAN, Indian Institute of Technology, Chennai, 2017.

5. Attend a workshop on “Hydraulic Fracking” at the department of petroleum engineering, IIT Dhanbad, 2015.

6. Attended a short term training program in structural health monitoring 2021.

LIST OF AWARDS AND ACHIEVEMENTS

1. Winner, (Ist Place) South Asia Pacific Regional Paper Contest 2019 organized by Society of Petroleum Engineers.
2. Selected for representing South Asia Region in International Paper Contest, ATCE, 2019.
3. Nominated for Young Professional of The Year 2018-2019 by SPE ADIPEC 2019.
4. Featured as Emerging Young Researcher in EAGE Magazine published from Suez University.
5. Campus placement in multinational Algo8 Company for the role A.I. Solution Architect Fellow for the design and deployment of intelligent technologies to upstream and downstream of oil and gas industry.