**ABSTRACT**

In oil fields where direct monitoring of oil output is impractical, it is never simple to predict the rate of production correctly. Empirical correlations and analytical techniques based on decline curves are utilized in these situations to estimate the rate of oil production. There could occasionally be serious mistakes made when employing these processes, producing findings that are incorrect. This research's objective was to forecast oil output using an artificial neural network (ANN).

The study included 7880 data sets, 70% of which were used for training and 30% for testing. The input parameters used are ON STREAM HRS, AVG DP TUBING, AVG WHP P, AVG WHT P, and flowing bottom-hole pressure. The output is BORE OIL VOLUME. The developed ANN model predicts oil production based on ON STREAM HRS, AVG DOWNHOLE PRESSURE, AVG DP TUBING, AVG WHP P, AVG WHT P, and flowing bottom-hole pressure. The created ANN model's accuracy was compared to that of the other 12 models in order to determine which method is more accurate for estimating the rate of oil production. The corresponding correlation coefficients for the models were 0.8425, 0.8364, 0.8342, 0.8223, 0.8195, 0.7917, 0.7763, 0.7655, 0.7566, 0.6675, 0.9589.

Compared to the other models, the ANN model has a greater accuracy in predicting oil production.

**CHAPTER ONE**

**INTRODUCTION**

**1.1 Background of Study**

The prediction of oil production from the reservoir is a crucial step in the field of petroleum reservoir engineering. Under a variety of operating and maintenance scenarios, including well operations and completion, artificial lift, workover, production, and injection operations, this calculation of reserves entails a significant investment of money, time, and technology. The amount of oil in the reservoir needs to be estimated pretty precisely, however the geological and fluid parameters of the reservoirs are highly heterogeneous and nonlinear. As a result, it is challenging to predict the future oil production accurately. Numerous static and dynamic factors, including rock porosity and permeability (static parameters), reservoir pressure, and fluid saturation, affect how much oil is produced from a reservoir.

The forecasting of oil output from a reservoir would be more precise with these static and dynamic factors accessible. All of the parameter data aren't usually available, though. The accuracy of forecasting is decreased by the restricted data access from the oil fields.

Various forecasting approaches, including soft computing technologies and decline curve analysis, have been developed in the past (Tamhane et al., 2000). Because of their ability to deal with nonlinearities and time-varying conditions, artificial intelligence techniques including genetic algorithms, fuzzy inference systems, and neural computing have been widely used in the petroleum industry (Mohaghegh, 2001).

The capacity of neural networks (NN) to learn and adapt to new dynamic contexts makes it one of the most appealing artificial intelligence techniques for dealing with nonlinearities in parameter estimation as well as in production forecasting (Weiss et al., 2002). Many studies have demonstrated the effective use of NN in the field of oil exploration and development, including pattern recognition in well test analysis (Al-Kaabi and Lee, 1993), reservoir history matching (Maschio et al., 2010), prediction of phase behavior (Habiballah et al., 1996), prediction of natural gas production in the United States (Al-Fattah and Startzman, 2001), and reservoir characterization by mapping the complex nonlinear in reservoirs (Mohaghegh et al., 2001).

**1.2 Statement of the Problem**

The oil and gas sector places a lot of importance on quantifying uncertainty. Production projections and reserve estimations will always contain some degree of uncertainty, which can be rather significant early in an oil and gas well's production life. Overconfident and pessimistic estimates of reserves and profitability might result from incorrect measurement of uncertainty.

McVay and Dossary (2014) claim that even little overconfidence and optimism can result in a portfolio's performance falling short by more than 30%. Early reserve and resource evaluation is crucial for the best possible development. According to McKinney et al. (2002), a suboptimal development strategy might cause 50% of potential asset value losses.

The most used technique for estimating reserves and forecasting production performance is decline curve analysis (DCA). But certain wells exhibit complicated characteristics, and as a result, the decline curve does not accurately reflect the production history.

Supertight shale formations and water-flooded oil and gas fields are two examples. Alternative approaches are needed to evaluate the uncertainty in production forecasts and reserves estimations in these situations when deterministic projections of future output and reserves might differ greatly from the actual values.

Most oil corporations employ simulation to do this. This approach is accurate, but it takes a long time since it takes a lot of computing resources to do the work. In this study, the researcher propose an alternate strategy that will considerably simplify this phase and drastically cut the required processing resources.

While simulation procedures take a lot of time and provide uneven accuracy, our approach makes reliable predictions quickly without sacrificing precision. The given dataset will enable the AI system to generate these predictions using machine learning (ML) and Deep Learning (DL), and the trained system will enable the AI to anticipate the output of oil wells based on a few geological parameters.

**1.3 Aim and Objective**

**Aim:**

The aim of the study is to forecast oil output from reservoirs using artificial neural networks.

**Objectives:**

The following goals will help to accomplish the aim of this research project.

1. Creation of reliable artificial neural network models that can accurately predict production over the field's useful life. It must be possible for the resilient neural network models to forecast oil output while taking changes in reservoir characteristics and well configurations into account.
2. Utilizing field data to validate ANN models and ascertain their predicting accuracy and constraints.

**1.4 Significance of the Study**

An estimate of the gross and net quantities (oil, gas, and water) that are anticipated to be generated from a hydrocarbon accumulation during the course of the accumulation's remaining life, as well as the fluids injected to create these volumes, is known as a production forecast (water, gas, and steam). (Petrowiki, 2016). Production forecasts are a critical component of petroleum engineering and geology. Production forecasts are used to determine the long-term viability of a field, as well as its potential to become an economic assets.

Production engineers may identify anomalous values and any defects in large oil well systems by using accurate oil production flow rate predictions to spot unusual values. (George, 2021) One of the most challenging tasks in production engineering is predicting oil production flow rates. This requires accurate data about the well geometry, reservoir properties, and fluid flow behavior. The Volve Field case study provides a real-world example where we can use artificial neural networks (ANNs) to make predictions on oil production flow rate from measurements taken at various stages of development.

**1.5 Scope of Study**

The goal of this project is to estimate the output of an oil well given a set of well characteristics that are utilized as independent variables, such as the rate of gas production, the pressure in the tubing, and the pressure in the flowing bottom hole, while the rate of oil production is kept constant. Following that, it utilizes 30% of the datasets for testing and validation and 70% of the datasets from a field in the Niger Delta for training. In the Python Google Collab environment, the code was tested. Two separate approaches were used to complete the project:

1. The regression approach, which forecasts data using linear regression, and
2. Artificial Neural Networks (ANN)

**CHAPTER TWO**

**LITERATURE REVIEW**

* 1. **Conventional Petroleum Accumulation**

Petroleum is made up of hydrocarbons that are the byproducts of fossilized organic matter, mostly plankton, which has been buried under silt to create source rocks. Liquids and gases have formed as a result of the high temperature at great depths. Petroleum was discharged from the source beds because it was less dense than the surrounding water. It then moved upward through porous rock, including some sandstone and limestone, until it was eventually blocked by nonporous rock, such shale or solid limestone. The bending, faulting, and erosion of the Earth's crust led to the formation of geologic structures that eventually contained petroleum reserves.

Petroleum can exist either separately or in combination in gaseous, liquid, or near-solid forms. The more solid phase may be referred to as bitumen, tar, pitch, or asphalt, whereas the liquid phase is frequently referred to as crude oil. When both phases coexist, the liquid often covers the more-solid phase while the gas typically rests on top of the liquid.

New sources of rock oil were actively sought once it was learned in the 19th century that rock oil would provide a distilled product (kerosene) suited for lanterns. It is now widely accepted that Edwin Laurentine Drake's 1859 well at Titusville, Pennsylvania, United States, was the first one to be dug particularly for oil. The success of this well, which was dug close to an oil seep, inspired more drilling nearby and eventually brought about similar exploration elsewhere. By the turn of the century, other governments and nations had begun digging their own oil wells due to the rising global demand for petroleum goods.

Around the world, 150 million barrels of crude oil were produced in 1900. The majority (80%) of the remaining was manufactured in the United States, with half of it being generated in Russia.

The second decade of the 20th century saw the introduction and expansion of vehicle use, which increased demand for petroleum-based products. In 1925 and 1940, the annual production was more than one billion barrels. By the final decade of the 20th century, more than 20 billion barrels were being produced annually by almost a million wells spread across more than 100 nations.

* 1. **Petroleum Prospecting and Exploration**

The original Drake well was constructed not far from a known crude oil surface seepage. Such seepages were the sole verifiable signs of subsurface oil and gas production for many years. New techniques for assessing the potential of subsurface rock formations were developed, nevertheless, as demand increased. To effectively explore for oil today, data from seismic surveys, geologic framing, geochemistry, petrophysics, Geographic Information Systems (GIS) data collection, geostatistics, drilling, reservoir engineering, and other surface and subsurface investigative techniques must be integrated. The main technique for searching for petroleum is geophysical exploration, which includes seismic analysis. In more complicated and difficult exploration conditions, such sub-salt structures and deep water, gravity and magnetic field approaches are also historically trustworthy evaluation techniques.

Beginning with GIS, gravity, magnetic, and seismic surveys enable geoscientists to efficiently narrow their search for target assets to investigate, decreasing the risks involved with exploration drilling. There are three primary types of exploratory methods: (1) surface approaches, such as geologic feature mapping, made possible by GIS, (2) area surveys of gravity and magnetic fields, and (3) seismographic methods. These approaches reveal the existence or absence of subsurface characteristics that are favorable for petroleum accumulations. There is still no technique to anticipate the presence of profitable subterranean oil reserves with 100 percent accuracy.

Another strategy relies on surface signs of potential subsurface rock formations. Subsurface folds and faults in rock formations are sometimes replicated in surface structures. A modest protrusion in an otherwise level ground surface, for example, may suggest the existence of subsurface salt domes. Oil and gas accumulations are frequently caused by uplift and faulting in the geological formations surrounding these domes.

* + 1. **Surface Techniques**

Crude oil seeps can appear as a tar-like deposit in a low terrain, such as the oil springs mentioned by Marco Polo near Baku, Azerbaijan, on the Caspian Sea. They usually appear as a thin skim of oil on little streams that run through a region. This latter phenomena is what gave rise to the name Oil Creek in Pennsylvania, where Drake's well was dug. Natural gas seeps are seldom visible, but sensors may detect natural gas concentrations in air as low as 1 part in 100,000. Similar devices have been used to detect traces of gas in saltwater. These geochemical surface prospecting approaches are inapplicable to the vast majority of petroleum reservoirs with no surface leakage.

* + 1. **Gravity and Magnetic Survey**

Although gravity is essentially constant at the Earth's surface, it is somewhat higher where solid rock formations are close to the surface. As a result, gravitational force rises over the tops of anticlinal (arch-shaped) folds while decreasing over the tops of salt domes. A sensitive sensor called as a gravimeter may detect very slight variations in gravitational force. Measurements are taken on a precise grid across a vast region, and the findings are plotted and evaluated to determine the existence of prospective oil or gas reserves.

Magnetic surveys make advantage of the magnetic characteristics of particular types of rock, which impact the Earth's normal magnetic field when they are close to the surface. Sensitive devices are now again utilized to map abnormalities across broad regions. Surveys are frequently conducted from aircraft over land and from ocean-going boats across continental shelves. Magnetotellurics (MT), a related technology, measures the natural electromagnetic field at the Earth's surface. Anomalies are caused by the various electrical resistivities of rock formations, which are then mapped and interpreted to represent subterranean geologic characteristics. Before more expensive and time-consuming seismic surveys, MT is becoming a more cost-effective filter for identifying a petroleum play (a group of oil fields or petroleum deposits with comparable geology features). MT is sensitive to what is present within the strata of the Earth.

Subsalts (salts whose bases have not been totally neutralized by acid) are particularly resistant to electromagnetic waves, whereas porous rocks are frequently conductive due to the saltwater and brines trapped inside them. Anomalies such as salt domes are used by petroleum geologists to identify possible stratigraphic traps for petroleum.

* + 1. **Seismographic techniques**

Even if there are no surface indications of the presence of large geologic anomalies such as anticlines (arch-shaped folds in subterranean layers of rock), fault blocks (sections of rock layers separated by a fracture or break), and salt domes, the survey methods described above can reveal their presence. They cannot, however, be depended on to detect smaller, less visible traps and unconformities (gaps) in the stratigraphic arrangement of rock strata that may contain petroleum reserves. Seismic surveying, which takes use of the sound-transmitting and sound-reflecting qualities of subsurface rock formations, may identify and locate these. Seismic waves move at varying speeds through different types of rock formations and are reflected by interfaces between them. Typically, the sound-wave source is a tiny explosion in a shallow drilled hole. To pick up and record the transmitted and reflected sound-wave arrivals, microphones are set at varying distances and orientations from the explosive site.

The operation is done at regular intervals throughout a large region. An skilled seismologist can then evaluate the collected data to determine the outlines of the subsurface formations.

The primary difference between offshore and land-based seismic data collecting is the manner of setup. One of the most crucial components of offshore seismic surveys is knowing where the ship and receivers are at all times, which is made possible by transferring real-time global positioning system (GPS) information from satellites to GPS reference and monitoring stations, and subsequently to the ship. Real-time readings have become an integral aspect of the seismic sound-wave capture, data processing, and analysis process.

**2.3. Oil Production Prediction**

Hydrocarbon exploration and production is a global sector that is both technically difficult and financially risky. Because well performance prediction is crucial for profit making, it begins in the early phases of production to anticipate future recovery. If a well's expected output does not meet a certain level, it might be remedied or even shut down to prevent additional loss.

For future production prediction, it is usual practice to compare early stage production data from a new well with historical production data from surrounding wells. Type curve matching is a well-known method for estimating cumulative production. Analogous wells are initially categorized based on the similarity of their cumulative production curves, and the average production profiles for each group of wells are determined as the type curves.  (Mohaghegh & Gaskari, 2009; Sproule, 2015)

The early cumulative output of a new well is then matched to the various type curves, and the closest curve is chosen by engineers to estimate the new well's future well performance. Matching type curves is efficient but subjective. It is difficult to replicate the research because various engineers may produce different estimates for a certain well based on their visual interpretations of the type curves. Another difficulty is that cumulative production curves tend to appear quite similar to one another over time, making it difficult to differentiate performance variances in individual wells.

According to the research literature, data mining techniques are gaining popularity in the oil industry for production data analysis because they are objective, resistant to poor data quality, and more accurate than statistical approaches (Mohaghegh et al., 2008; Ma et al., 2015; Zhong et al. 2015). Different data mining methods, such as Support Vector Machines, Random Forests, and Boosted Regression Trees, have shown to be efficient and useful tools for understanding oil and gas operation and production (LaFollette et al. 2012; Esmaili et al. 2013; Zhong et al. 2015). However, the time-series production data in these research were either averaged or represented by their maximum or cumulative values.

These statistical measurements oversimplify production curves and do not reflect each well's whole production pattern. As a result, an effective and understandable data-driven prediction technique based on time-series production data will be investigated.

**2.4 Artificial Neural Network**

A complicated non-linear relationship between input and target variables can be identified or approximated using an artificial neural network (ANN), a technique in artificial intelligence. Artificial neural networks are computational models inspired by biological systems that are capable of machine learning and pattern recognition. They are based on human central nervous systems, particularly the brain, and are used in computer science and related disciplines. Over the past few decades, artificial neural networks have attracted a lot of attention and are effectively being used in a variety of scientific fields, including physics, engineering, economics, and geology. These problems are more about classification, control and complex behavior prediction (Nikravesh, et al. 2003; Graupe 2007).

The neurons (nodes) in ANN's hidden layers use nonlinear activation functions to process a weighted sum of input data. An artificial neural network (ANN) is created by teaching a network to reflect the inherent connections contained in data. The network is trained using a data set that contains both input (or predicting) and target variables. The unknown network parameters (weights and biases) are estimated using an inverse problem procedure that minimizes the discrepancy between the network output and the known values of the target variables.

The concept of a neural network dates back to 1943, when mathematician Walter Pitts and neurophysiologist Warren McCulloch published a paper on the potential functions of neurons (McCulloch and Pitts. 1943). To describe how neurons works in the brain worked, they created a simple neural network model using electrical circuits. It was only in the 1950s that it became feasible to simulate a fictitious neural network, thanks to developments in computer science and technology. It has been noted in past decades that neural networks made up of numerous neurons can be more effective than those made up of only one neuron. The gradient descent-based learning principles for large neural networks were created and popularized by McClelland and James in 1986. Recently, scientists have made enormous advances in neural networks and created cutting-edge methods like self-organizing maps and associative memories (Mehrotra, et al. 1997). Numerous sources, including Mehrotra et al. (1997), Poulton (2001), and Suh (2012), give thorough background material on the development of artificial neural networks.

Any neural network is taught using a learning algorithm and a training data set. Unsupervised learning and supervised learning are the two main categories of neural network learning methods. The supervised learning is used to uncover a hidden structure in unlabeled data.

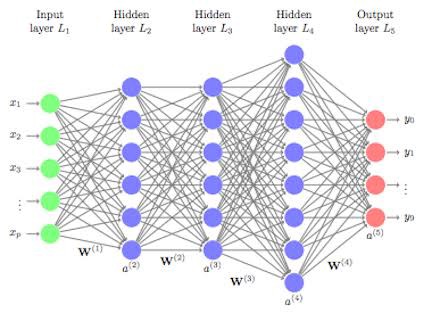
The goal is to classify, find features or regularities in the training data, or both. A cluster analysis is the most common use of unsupervised learning. In contrast, the supervised learning method requires that target values be provided. The input vector requires a training dataset, and will generate the rules according to the desired output by adjusting the weights.

The weights are then used to process the test data set's inputs. The weights will be adjusted after providing the desired output to the net to match the model to the desired goal. The learning process iteration will be continued till the intended goal is realized.

**2.5 Artificial Neural Networks Architecture**

An artificial neural network is formed of several artificial neurons that are coupled together according to a certain network architecture. The two most prevalent forms of network architectures are Single Layer Perceptron (SLP) and Multilayer Perceptron (MLP) which comprises of one input and output layer with any number of hidden layers, as seen in Fig. 2.1.

The number of layers and neurons (nodes) in each layer must be chosen when designing a neural network. There is a trade-off between accuracy and overfitting of the data: with insufficient neurons, the mismatch between network predictions and actual values of the target variables could not be minimized, while with excessive neurons, the overfitting of the network parameters could occur.



**Fig.2.1: Multilayer Perceptron (MLP) neural network configuration**

**2.6 Application of Artificial Neural Networks in Petroleum Engineering**

Petroleum engineering has used neural networks for a variety of purposes, including classification (Stundner and Al-Thuwaini. 2001), reservoir characterization or property prediction (Tang, et al. 2011; Raeesi, et al. 2012; Aminian, et al. 2003; Chaki, et al. 2014), proxy for recovery performance prediction (Aw), and more (Mohaghegh. 2000; Maschio and Schiozer. 2014), as well as the design or optimization of industrial processes and well trajectory. (Stoisits, et al. 1999; Ayala H and Ertekin. 2007; Artun, et al. 2012; Yeten, et al. 2002; Oberwinkler, et al. 2004; Malallah and Nashawi. 2005; Zangl, et al. 2006).

Specifically, neural networks have been used in recent years as a proxy model to perform EOR (enhanced oil recovery) screening, predict heavy oil recoveries, characterize reservoir properties in unconventional reservoirs, and more (Queipo, et al. 2002; Popa, et al. 2011; Popa and Cassidy. 2012; Ahmadloo, et al. 2010; Fedutenko, et al. 2012; Fedutenko, et al 2014; Mohammadpoor, et al. 2012).

The overall size of the training data set and the input vector dimension continue to influence how many free parameters (i.e., weights and bias connections) are present in a hidden layer.

There are several correlations between training data set size and user-defined error parameters in the literature (Haykin, 2001), and extrapolation detection methods are used in some of these relationships. Several design challenges pertaining to the creation of ANNs have recently been reviewed, and can be found in  Al-Bulushi et al. (2012).

Most studies on network architecture design adhere to a few broad guidelines and arrive at the ideal network architecture by trial and error, however others have suggested generalizing training to include learning an acceptable architecture through Bayesian inference, where Poulton (2001) discusses probabilistic networks as an approach.

The usage of functional neural networks, where the activation functions and weight connections connected with the neurons might fluctuate and be approximated through training, was also recently advocated in the literature (El-Sebakhy, et al. 2012). The majority of projects make use of back-propagation or other gradient-based optimization techniques as the learning algorithm, only a few authors have considered the application of more generic global optimization algorithms like a genetic algorithm optimization approaches like the learning algorithm (Saemi, et al. 2007).

Numerous studies also stress the significance of data pre-processing, which includes normalization and outlier detection for input and target data utilized in training (Tang, et al. 2011). Additionally, experimental design is frequently carried out in applications where answers from thorough flow simulations are utilized to train a network that would act as a stand-in for reservoir performance prediction. This reduces redundant training data and cuts down on computation time (Queipo, et al. 2002).

In order to account for heterogeneity and uncertainty in reservoir parameters, several authors have also suggested include indices in the set of predictive variables, such as a Dykstra Parsons coefficient and the permeability ratio of neighboring strata (Ahmadloo, et al. 2010).

With the use of prior information (supervised learning), several researchers in the oil sector are using machine learning and deep learning to anticipate production of water, gas, and oil as well as to identify patterns in well test data. This section will discuss work that uses machine learning to forecast oil, water, and gas output as well as the well pressure and well oil/gas/water ratio. With the currently employed techniques, such as simulation, the well production forecast process is time-consuming and expensive in terms of computer resources. Due to the fact that our dataset is a continuous (time-series) problem type, we had to thoroughly investigate and investigate the likelihood of getting good results with our dataset (Abdullayeva, F. 2019)

The methods now in use have a bottleneck effect when creating new wells. Numerous ML and DL strategies have been applied to improve the performance of production prediction as technology has advanced. Additionally, the preliminary search and comprehension of the pertinent machine learning research addressing well-performed classification and prediction served as the foundation for the organization of this part. Additionally, the research were carefully examined, and numerous concepts connected to the core theme were gathered from several papers. The literature research was then divided into various sub-areas of oil production that were pertinent to our topic.

**2.6.1. Oil Flowrate Prediction**

According to operating parameters and choke size, the volume of oil flowing was determined in a recent artificial intelligence research. For critical and subcritical flow, the study employed two distinct datasets: choke size, upstream pressure, temperature, gas/oil ratio, and water cut for critical flow. Along with downstream pressure, the subcritical dataset included the same data as the critical flow dataset. Then, a range of AI technologies, including artificial neural networks, fuzzy logic, and functional networks, as well as widely used techniques like Gilbert correlation, were put to the test. The ANN system demonstrated notable accuracy with a correlation coefficient of 0.89 when compared to current correlation techniques, which have a maximum correlation value of 0.30 (al Ajmi, et. Al 2015).

An orifice flow meter's oil flow rate was recently predicted using machine learning algorithms, according to a study. From a dataset of 1037 data records, inputs for pressure, temperature, viscosity, square root of differential pressure, and oil-specific gravity were taken. The employed machine learning (ML) techniques, including the adaptive neuro-fuzzy inference system (ANFIS), least squares support vector machine (LSSVM), radial basis function (RBF), multilayer perceptron (MLP), and gene-expression programming (GEP), all produced results with high levels of accuracy, with correlation coefficients varying from 0.90 to 0.99 (Ghorbani, H. 2020).

According to the authors, tight gas fields constitute a significant source of hydrocarbons for the energy sector. Large amounts of information are produced by the process. As a consequence, data may be assessed, forecasted, and utilized to spot patterns between dependent and independent variables using machine learning techniques. In this study, the generalized linear model (GLM) and the artificial neural network (ANN) were both used. To determine how well-planned new wells will recover is the study's main objective. A dataset of 224 wells was evaluated for production data. These models' predictions were contrasted with the actual initial gas production rate of the wells. A GLM model had a mean square error of 1.57 while an ANN model had a mean square error of 1.24, according to research.

Additionally, the reservoir thickness was shown to be responsible for 36.5 percent of the initial gas output, followed by flowback rate (29%), according to the ANN model's performance index. As a consequence, the ANN model beat the GLM model when it came to projecting gas production and looking back (Ghorbani, H 2020).

The authors discussed how to anticipate the flow rate of oil production in circumstances when direct measurement is challenging, which is a challenge encountered by petroleum engineers across the world. In order to provide a novel methodology for calculating oil flow rate in two-phase oil and gas flow through wellhead chokes, this study uses an artificial neural network approach.

As a function of the choke upstream pressure, choke size, and generating gas-to-oil ratio, the oil flow rate is estimated using the artificial neural networks (ANNs) model. The accuracy of the proposed model was tested to many well-known empirical correlations, and the findings showed that the projected oil flow rates of the new model are quite near to actual observed data. Additionally, this model estimates the oil flow rate more quickly and accurately than empirical correlations with more than three input parameters, with an accuracy rate of about 87% (Mirzaei-Paiaman, A.; Salavati, S 2012)

**2.6.2. Production of Wells**

A research has been proposed that describes a hybrid model for time series prediction of oil output based on a combination of CNN and LSTM networks. The model's CNN layer is applied to the current time window first, extracting the features, and then the LSTM is used to forecast the relationship between the time windows. The goal of this study is to use deep neural networks to build a system that can estimate oil output accurately based on well debt data. They conclude that the CNN + LSTM model's root mean squared logarithmic error (RMSLE) value is 0.186891 (Abdullayeva, F. 2019).

The authors presented their study as not only constructing but also evaluating a proxy model utilizing field data from 1239 horizontal wells. As a consequence of this investigation, the following conclusions were reached. Initially, the datasets were investigated using the exploratory data analysis (EDA) approach, which comprised outlier analysis, categorical and numerical variable analysis, and correlation analysis. From a total of 1239 wells, 1150 were used as training data using EDA. Second, principal component analysis (PCA) was used to minimize the size of the input variable. Finally, VIA employed the RF, GBM, and XGBoost algorithms to find independent factors with a significant influence on cumulative gas production during a 12-month projection (CGP12). Using this strategy, the average relevance ranking of the independent variables was calculated.

Finally, hyperparameter sensitivity analysis is utilized to build more predictive DNN models (Han, D.; Kwon, S. 2021)

The author reviewed a previous paper that described the use of RNN-LSTM to predict oil, gas, and water outputs of wells based on injection patterns in a time-series way. The RNN-LSTM model estimated oil, water, and gas output with a first-year accuracy of more than 90% and production values with a 73.63 percent accuracy for up to 5 years (Pal, M. 2021).

The authors described their work by proposing an automated method for assisting technical teams in increasing data quality in production data analysis through the use of machine learning techniques, improving reliable production forecasting, lowering operating costs, and optimizing drilling schedules.

Input characteristics for this study included reservoir pressure, water cut, wellhead pressure, choke size, and rate. The following machine learning techniques were effectively employed to remove noisy data and outliers. For instance, the K-means clustering technique is employed to identify outliers or anomalies. Second, at the analysis stage, support vector regression (SVR) is a potent method for reducing noise from data. The aforementioned techniques can increase the precision of production forecasts (Doan, T.T.; van Vo, M. 2020)

The study examined how supporters predicted oil, water, and gas output in reservoirs with water injection using artificial neural networks. The researchers used the Bayesian regularization method to train the ANN model.

The ANN model attained a coefficient of determination of higher than 0.9. The delay in the time-step term is a crucial component that might enhance the model's prediction, which the researchers also found since other models disregard it (Negash, B.M.; Yaw, A.D. 2020)

The idea that various factors, including geology and completion, play a role in gas output has been the subject of a study proposal. Using machine learning techniques, we can develop a production prediction model and identify the key factor impacting production. There were 159 horizontal wells that were aimed towards the Duvernay formation. Then, the crucial variables were discovered using Pearson correlation and grey-connection analysis. Finally, three statistical models were produced using multiple linear regression (MLR), support vector regression (SVR), and Gaussian process regression (GPRFor the first six months of production, total oil and gas output was predicted (Guo, Z.; Wang et. Al 2021)

A study that uses the attention mechanism and combines it with the convolutional neural network (CNN) and the long short-term memory neural network to create a hybrid model has been proposed (attention-CNN-LSTM). Two wells in a southern Chinese oilfield provided daily average data for this investigation. Each day of the 23 years that the T1 well data were gathered represents a unique record, while the T2 well data were only collected for 17.5 years. Support vector regression (SVR), back-propagation neural network (BP), standard CNN LSTM, and other techniques have also been investigated. With an RMSE (root mean square error) score of 0.315 and 0.402, an MAE score of 0.218 and 0.303, and a MAPE (mean absolute percentage error) score of 0.008 and 0.005 for T1 and T2 wells, respectively, attention-CNN-LSTM appears to have the lowest score uncontested among all other models, though, according to three different comparison measures (Jabbari, M.; et. Al 2021).

The suggested approach for calculating oil production from a single well is discussed in a recent paper. They applied the LSTM technique. To get a better outcome, the model has been altered and adjusted. Several optimization techniques, such the batch normalization layer, which increases prediction accuracy, can speed up training and improve the model's generalization skills. From January 1973 to May 1995, 1275 wells are included in the dataset. The utilized dataset also included 43 characteristics. Additionally, the author devised a data-labeling technique based on the various water cut phases. Each well must be a part of at least one of the four stages. The program also provides other labeling options, such well type. They chose 65 wells from the dataset, which had been split into 90/10, 1409 samples for the training, after the tool had preprocessed the data. For the test set, the remaining 157 samples were taken into account. The author employed RF and SVM in addition to the LSTM to compare them. RMSE, MAS, and R2 are the employed metrics. With an RMSE score of 383.33, the LSTM has the lowest value, indicating higher performance. In addition, LSTM's value in MAE is 285.34, which is the lowest. The LSTM approach has the greatest R2 of all of the techniques, at 0.786 (Xia, L.; Shun et. al 2020).

The study states that the hydraulic fracture design parameters, fracture network features, such as fracture spacing and fracture conductivity, and simulation data from 370 reservoir simulations were used to build the three ANN architectures. The output of shale gas is greatly impacted by these variables. Based on an additional set of 92 simulations, the testing outcomes confirmed a strong connection between input and goal functions, with an R2 better than 0.86. Additionally, throughout the course of 1, 5, 10, and 15 years, there was good agreement between observed and forecast total gas output, with an R2 > 0.94 and error rates < 15%. For wells that have been producing for a brief period of time, using the peak production rate can increase forecast accuracy (Nguyen-Le, V.; Shin, H. 2020).

**2.6.3. Prediction of Well Production Enhancement**

A research that examines a special deep recurrent neural network built for use by petroleum engineers to examine the characteristics of active oil wells has been offered. Additionally, this study examines the most widely used time-series models using the deep structure of a GRU recurrent neural network, which serves as the model's foundation. The purpose of this study is to evaluate and contrast the layered DGRU model's robustness and accuracy to the most widely used time-series models. They end up with a collection of 21,000 observations of US oil production data spanning 90 years after pre-processing the information. Training 40%, Validation 20%, and Testing 20% make up the three subgroups that comprise the whole dataset collection. Additionally, this method can be used to assess the effectiveness of both the trained model and the stacked deep GRU model using the testing dataset (unknown new data). The DGRU model also assessed the performance and accuracy of other time-series models. Multi-RNN, single GRU, Multi-GRU/ANN, and Multi-LSTM models are among those used in this work. Using R2 measurement and a 10-step prediction process, the total accuracy performance results were as follows:

• GRU: 55.078%

• GRU/ANN Multi: 72%

51% for Multi LSTM

• DGR Stacked: 70%

They discovered that the model's design is a great deal more straightforward than that of LSTM and RNN. Our DGRU model performs better than other traditional models due to its low level of complexity and capacity to handle long interval time-series information. As a consequence, our DGRU model may be utilized to predict long-term relationships in a complex time series dataset (Al-Shabandar et. Al 2021).

**2.6.4. Prediction of Pressure Gradients**

The usage of a multilayer feed-forward ANN was described in order to estimate the horizontal oil-water flow pressure gradient utilizing five inputs: oil superficial velocity, water superficial velocity, pipe diameter, pipe roughness, and oil viscosity. The data from 765 experiments were separated into three sections: training (60%), validation (20%), and testing (20%).

When compared to the other techniques investigated (two-fluid model, homogeneous model, and correlation), the ANN model had an APE of 0.30 percent, an AAPE of 2.9 percent, and an SD of 7.6 percent, which is extremely low (Al-Wahaibi, T.; Mjalli, F.S. 2014).

The authors presented their work by developing and implementing machine learning techniques. The pressure gradient in a liquid-liquid flow was calculated using six different methods, including support vector machine (SVM), Gaussian process (GP), random forest (RF), artificial neural network (ANN), k-nearest neighbor (kNN), and a fusion model. Using SVM, seven predictor values were picked as the best collection of predictors with the fewest errors. The seven predictors include oil and water velocities, FP, input diameter, oil and water density, and oil viscosity. With the exception of the ML-fusion model, which had a (p 0.05), the GP model had the best prediction accuracy in compared to other models (Wahid, M.F. et. al 2021)

**2.6.5 Fault Prediction**

The early development of a machine learning (ML) model for early fault prediction of a centrifugal pump in the oil and gas sector that is straightforward and deployable was studied. Process and equipment sensors on the chosen machinery were utilized to review past data in real-time. Raw sensor data, mostly from temperature, pressure, and vibration probes, was denoised, pre-processed, and sequentially coded in order to train the model.

In this study, support vector machine (SVM) and multilayer perceptron are two techniques that are compared (MLP). High accuracy was recorded by the SVM algorithm at 98.1%. The MLP method outperformed the SVM algorithm in terms of total accuracy, recording 98.2%. The main goal of this work is not to build very accurate ML models, but rather to show that it is feasible to make accurate predictions using a straightforward and understandable ML approach. Overall, the findings imply that the suggested algorithms deliver good overall classification performance and are effective in determining the health condition of the monitored machine (Orrù, P.F. et. al 2020).

A research has been suggested that addresses an unique hybrid LSTM-SAE learning approach that will be utilized to improve fault detection accuracy while addressing the drawbacks of RNN training and the usage of a single methodology alone. All of the vibration indications are provided by the DCS of the electrical generator. Approximately 2000 samples were divided into a training dataset (80%) and a testing dataset (20%) before being input into the suggested DL architecture. These samples are separated into two groups: training and testing datasets with errors and datasets without errors. The sample data were selected and utilized to train and test RNN-LSTM in preparation for use in the fault detection stage of the suggested DL architecture. The following key contributions to the present work were intended for this study:

• A unique DL fault detection framework is developed based on the RNN-LSTM, SAE, and particle swarm optimization (PSO) techniques to balance the three processes of parameter optimization, fault feature extraction, and fault detection.

• The suggested DL technique can not only adapt the essential features, but also materialize patterns without preserving the past sequence inputs, thanks to an unique hybrid mathematical approach that addresses RNN training restrictions such as fading error, deficit, gradient vanishing, and backflow.

roaches, with a 99.67% detection accuracy. Furthermore, the findings show that the proposed deep learning system can detect errors in industrial data without using labels. This can help data engineers properly extract characteristics and avoid relying on human expertise based on the unsurprised fault detection technique. Furthermore, according to this study, the suggested framework has the greatest accuracy when compared to other techniques, which are RNN and ANN, which have accuracy of (69.0%) and (49.5%), respectively. When compared to previous techniques, the proposed DL framework is more time efficient and has a quicker detection time (0.17 s) (Alrifaey, M. et. al 2021).

**2.6.6. Prediction of Bottom-Hole Pressure**

It is necessary to compute the bottom hole pressure in order to determine the pressure drop in wells (BHP). To estimate the BHP, the author used ANN with a hybrid genetic algorithm and particle swarm optimization (HGAPSO). Oil flow rate, gas flow rate, water flow rate, oil API, depth, tubing diameter, surface temperature, wellhead pressure, and bottom-hole temperature are the nine inputs that are employed (Ahmadi, M.A. 2019).

Results from testing this model on a population of 150 had a maximum inaccuracy of 10%. Although other approaches, including GA-ANN and PSO-ANN, have been tried, the hybrid approach HGAPSO-ANN has demonstrated the highest success rate. The article (Khamehchi, E.; Bemani 2021) demonstrates a different way to compute BHP. The author employed extreme learning machine (ELM) and gradient tree boosting (GTB) (ELM). Eight variables made up the dataset that was used: actual vertical depth, average deviation, average angle, wellhead pressure, oil flow rate, water flow rate, and gas flow rate. The combination of these two techniques produced a mean relative error that was less than 4%. A recent research also examines the use of ANN, KNN, and random forests to forecast the flowing bottom hole pressure (FBHP). An error rate of 2.5%, 3.6%, and 4% for ANN, random forest, and KNN, respectively, was obtained using these methods with the following nine inputs: flowing oil rate, flowing gas rate, flowing water rate, production tubing internal diameter, well perforation depth, oil gravity, surface temperature, well bottom hole temperature, and wellhead pressure (Sami, N.A.; Ibrahim, D.S. 2021)

**2.7. Summary of Related Work**

In conclusion, complicated non-linear situations cannot be handled by machine learning. However, deep learning is best used for complicated non-linear problems, thus as a first step, our system will utilize deep learning to estimate the rates of production for oil, water, and gas. After completing the initial stage, we intend to enhance the system to provide more deep-learning-supported solutions for forecasting critical factors in oil extraction, such as the oil/water ratio, well pressure, and flow rate. We intend to investigate the use of ANN because it is ideal for our issue type. Other approaches, like as RNN and CNN, are less feasible; CNN excels in image recognition and may be used with RNN to enhance RNN module results. However, additional experimentation with other ways will be required to improve the system's efficiency.

**CHAPTER THREE**

**MATERIALS AND METHODS**

**3.1 Choosing the Best Environment**

Knowing the data analytics tools and the intricacy of the Volve dataset, the Python programming language was selected to develop models and conduct computations. Python is often regarded as the greatest scripting language for data mining and analysis. Furthermore, it has a large community, so any difficulties encountered may be readily addressed owing to the knowledge given on specialized forums. This ecosystem includes a plethora of strong libraries spanning from fundamental statistics to complicated machine learning techniques. All libraries shine in terms of performance, productivity, and collaboration, making the entire process of data handling and visualization relatively simple as compared to other languages.

**3.2. Data Acquisition**

Data is a major factor in artificial neural network's forecasting ability. In other words, ANN processes a large volume of data that spans a wide range of operational circumstances. If poor, noisy, and insignificant amounts of data are expected, ANN is clearly not the best option. For pattern recognition, prediction, system identification, and control, ANNs have emerged as the technology of choice because of their capacity to handle complicated, non-linear connections between parameters.

The initial hurdle was familiarizing oneself with the available data due to the enormous quantity of the collection.

The oil well production data includes the date of production, codes and names as determined by the Norwegian Petroleum Directorate, including the well field code, name, and facility. Following these are the following: Pressure Data (Downhole Pressure & Temperature, Average Tubing & Annular Pressure Details), Flow Data (Volumetric Data for Produced Gas, Oil, Water, and Injected Water), Flow Kind (e.g. Spherical), and Other Data (Onstream Hours, Type of Well, Choke Details (e.g. Choke Size, DP Ratio [Pressure drop/downstream pressure]),were gathered from an abandoned field since the validity of the data is crucial to the accuracy of the machine learning process.

**3.3 Selection of Best Input and Target Parameters**

It is generally agreed that in neural networks, having enough data available is one thing, but knowing the best parameters to use for the best outcome is another, as having more input parameters than necessary will result in a large network size and a corresponding decrease in learning speed and efficiency (Goda et al., 2005).

As a result of the critical analysis and independence examination, ON STREAM HRS, AVG DOWNHOLE PRESSURE, AVG DP TUBING, AVG WHP P, AVG WHT P were used as primary input parameters, while BORE OIL VOL was the target output, as knowing the full influence of the parameters is the primary point of concern in ANN model development.

**3.3 Data Pre-Processing**

Normalization method before presenting the input data to the network is mostly a decent drill. Mixing variables with large amounts and small amounts blur the  
learning algorithm on the prominence of an individual variable and may drive it to finally discard the variable with the lesser amount (Tymvios et al., 2008). To ensure that the selected data for modeling distinguish distinctive operational varieties and off data set will be separated from the list. The data set will be grouped in two i.e. training, and testing data sets.

Consequently, the considered parameters were normalized, using the min-max method, to speed up learning and achieve convergence and in order for the neural network to be more efficient and, in overall, yield better predictions.

Hence, the available data were normalized into the range of 0 to 1 by using Equation 3.1 shown here:

Equation 3.1

Where Xn = normalized value, Xmin = minimum of original values, Xmax = maximum of original values, and X = original value.

The data used in this research are the raw production data of an actual oilfields, so it is highly possible to include noise as in influencing factor. As such, it is not appropriate to use the raw production data in the learning of NN because NN requires extremely low learning rates. Thus, a preprocessing scenario consists of four steps has been incorporated before the use of the raw production data in the experiments of this research work.

**Step 1: Reduce noise from raw data**

A previous research which utilized linear and multiple regression to solve the same problem and on whose recommendation is the base of this study had provided with us a cleaned data and had carefully gone through each step of data cleaning in the study. A comparison was made between our cleaned and their cleaned data, a noticeable amount of outliers were removed from their data which made us come to a conclusion of using that for our research work.

**Step 2: Transform raw data to stationary data**

As explained in earlier, time series data are non- stationary data, which may, in fact, exhibit a specific trend (J.D. Cryer, K.-S. Chan 2008). Of course, stationary data is easier to model and will very likely result in more skillful forecasts. In the current preprocessing step, we removed the trend property in the data; whether it is increasing or decreasing trend. Later on, we added the trend back to forecasts in order to return the forecasting problem into the original scale and calculate a comparable error score. A standard way to remove the trend is by the differencing of the data. That is the observation from the previous time step (t-1) is subtracted from the current observation (t) (J.D. Cryer, K.-S. Chan 2008).

**Step 3: Transform data into supervised learning**

We use one-step ahead forecast, where the next time step (t+1) is predicted. We divide the time series into input (x) and output (y) using lag time method, and specifically, in the study we have used different sizes of lag from lag1 to lag6.

**Step 4: Transform data into the problem scale**

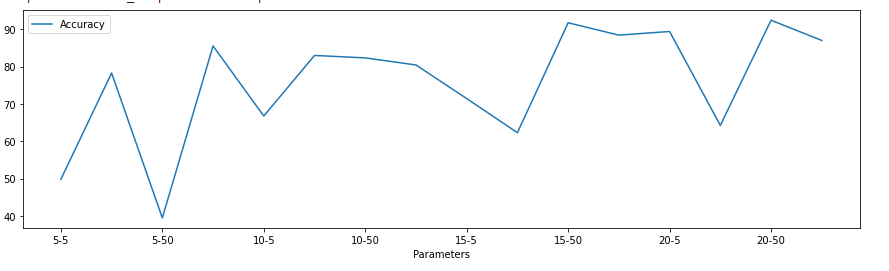
Like other neural networks, ANN expects data to be within the scale of the activation function used by the network. The default activation function for ANN is the reLu ReLU function is its derivative both are monotonic. The function returns 0 if it receives any negative input, but for any positive value x, it returns that value back. Thus it gives an output that has a range from 0 to infinity.

This is the preferred range for the dataset. Later on, we transformed the scaled data back in order to return the forecasting problem into the original scale.

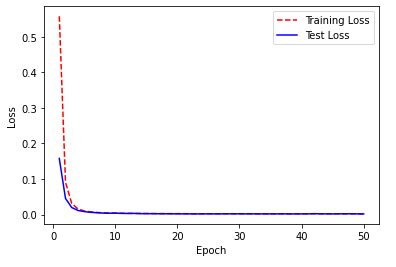
**3.4. Training of an Artificial Neural Network**

It is an optimization procedure where an error function is decreased by regulating  
ANN weights. In the training period, a set of recognized input-output configurations are continually offered to the network in order to demonstrate the network. The weights are constantly adjusted in order to decrease the error. The system learns how to handle different situations so that the network model can appropriately handle the similar scenarios in the future. Thus error is deduced from comparing output of an ANN with the actual output provided by the user. The training is to be conducted until error is reduced and better performance of ANN is achieved. A both stable and convergent ANN is desired from a training phase.

To train the network, Python was used along with the TensorFlow and Keras libraries. The normalized inputs and target were selected and fed to the system. The network was trained over 100 epochs with a learning rate of 0.001 with the Adam optimizer. The Adam optimizer combines the best traits of other optimizers that use root mean square methods to improve their ability to work with noisy data. The default parameters of the Adam optimizer work well enough to train networks effectively. Figures 3.1 and 3.2 highlight the improvements in model accuracy and model loss of the network throughout training.



**Fig 3.1: Visualizing the results of parameter trials for ANN**



**Fig 3.2: Model Loss vs training epoch of the training data**

**3.5 Design of Network Architecture**

The number of input and output parameters is simply determined by the number of neurons in the input and output layers. ANN models generally include three layers: input, hidden, and output. Each of these levels is made up of neurons (sometimes called nodes). The input layer is the initial layer, and it contains neurons that represent the inputs or parameters of a particular issue.

The computing process is started in the second layer, which is hidden. The output layer, which in our case is one neuron named BORE OIL, represents the labels or outputs of the problem. The number of neurons in the output layer depends on the nature of the problem. Two hidden layers with five neurons each and one output layer with one neuron were thoughtfully chosen for the work's hidden layer configuration.

Rectified Linear Unit (ReLu) activation functions were employed in both hidden layers of the regression neural network. ReLu returns zero for negative input, hence this stopped any negative target predictions from happening. Mean Squared Error (MSE) with a mean reduction was the loss function utilized for regression. Adam was selected to serve as the optimizer.

**3.6 Artificial Neural Network (ANN) Validation and Testing**

A validation dataset is a subset of data from model training that is used to evaluate model skill while tweaking the model's hyper-parameters (Wikipedia, 2009). Validation datasets are used for regularization by early stopping: when the error on the validation dataset grows, this is a symptom of over-fitting to the training dataset and should be stopped. When comparing or selecting amongst final models, a test dataset is a sample of data that has been held back from the model's training and is instead utilized to provide an impartial evaluation of the quality of the final tuned model. This data set is solely used to verify the final solution and check the network's true predictive power.

For regression, no validation set is employed because early stopping did not increase the regression model's performance. Unless otherwise stated, all accuracy and error distributions are calculated on the test set. Because the train/test split influences the findings, each experiment has a total of 10 tests, and the means of the errors/accuracy are determined.

**CHAPTER FOUR**

**RESULTS AND DISCUSSIONs**

**4.1 Artificial Neural Network (ANN) Model**

Through the optimization of the weight matrices linked to the predictors, ANNs learn to more accurately approximation the target variable. Comparatively to the feedforward component, optimizing the model weights is a significantly less straightforward task.

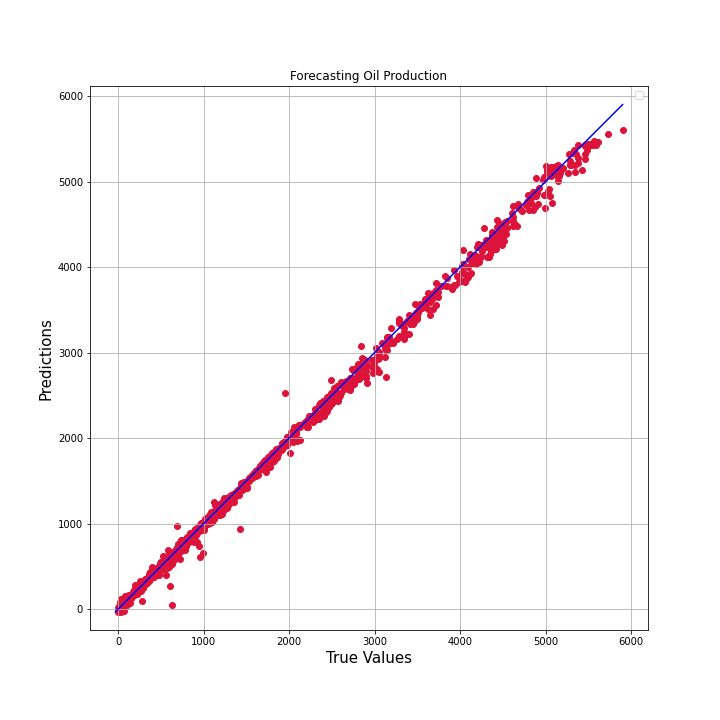
Backpropagation is a sophisticated mathematical procedure that calculates the gradient of the weight matrix by combining the Chain Rule with partial derivatives of the associated variables.

The backpropagation algorithm is used by Stochastic Gradient Descent (SGD) to modify the model weights by minimizing the cost function at a given rate of convergence (learning rate) over a predetermined number of training repetitions (epochs). The Neural Network model generates an accuracy of 91.83% on our dataset.

The network employed had two hidden layers, five neurons, and one output layer. This result was reached after various ANN model configurations (8-2-1, 8-3-1, 8-4-1, 8-5-1, 8-6-1, and 8-7-1) were examined to accurately and efficiently estimate oil production.

It is always vital to test several configurations to make sure that the most trustworthy ANN structure is used to identify trends and relationships in the set of data being used. Predictions that are suitable to the cases and conditions of interest can subsequently be made using the well-trained ANN model, which has the best performance criteria.

The Regression Analysis Plots demonstrate how well the linear regression fit to the data points, comparing the projected output to the actual target. It can be concluded that the ANN model has functioned flawlessly thanks to an overall correlation coefficient of 0.9183. Additionally, comparisons made in an MS Excel analysis with other models also support the same results.



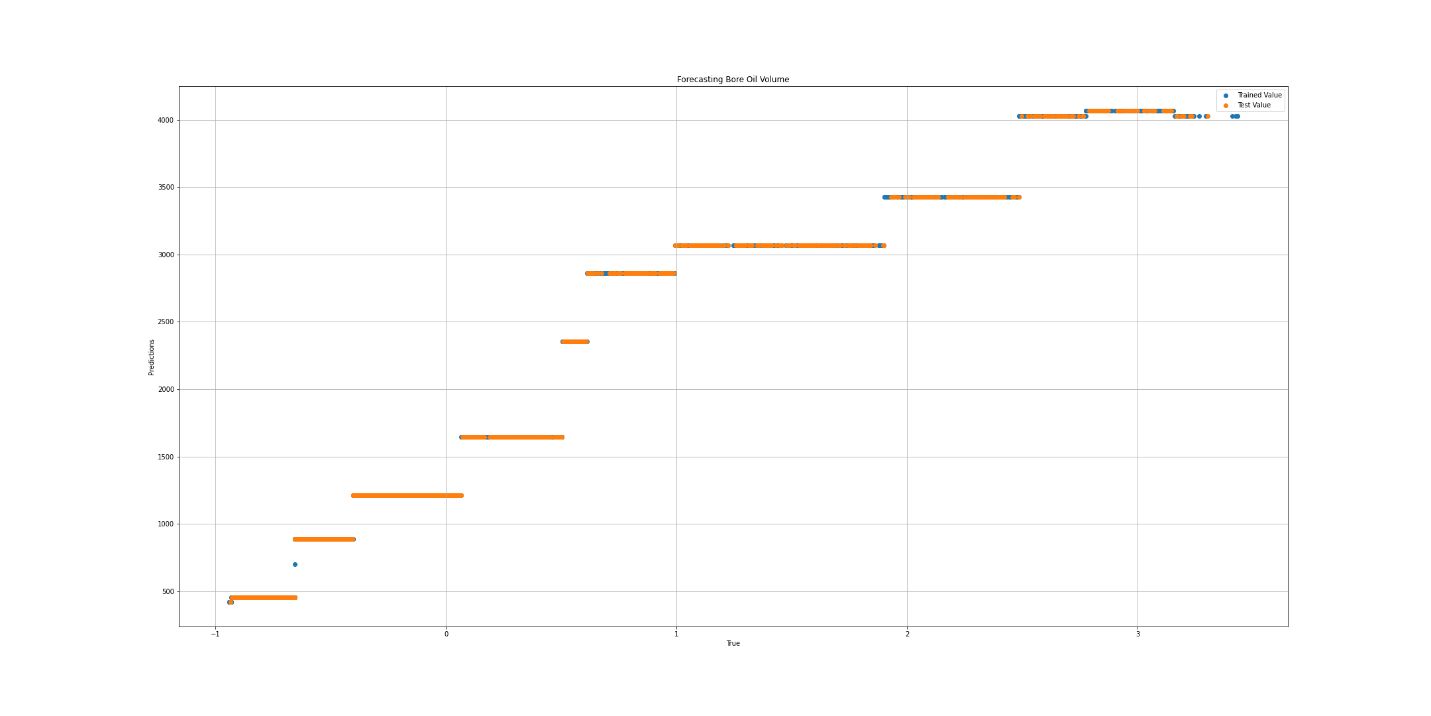
**Fig. 1. Production data vs. prediction using Vanilla Neural Network**

**4.2 Reference models**

Toward a fair evaluation, we will compare the proposed ANN model with different reference models that vary from statistical methods, machine learning methods, and hybrid (statistical and machine learning) methods. The reference models are:

1. **AdaBoosting:** One of the earliest effective boosting methods was AdaBoost, which is short for "Adaptive Boosting," an ensemble machine learning algorithm. AdaBoost combines the forecasts from brief decision trees with only one level, or "decision stumps." In order to employ a large number of weak models and modify their forecasts by adding more weak models, decision stump algorithms are used. Starting with a single decision tree, the training algorithm looks for examples in the training dataset that were incorrectly categorised and gives such examples greater weight. The same data is used to train a different tree that is now weighted by misclassification errors. Until the necessary amount of trees are added, this process is repeated.

We generated the resulting plot using the Adabooting regression model on our dataset, with a model accuracy of 79.16%.

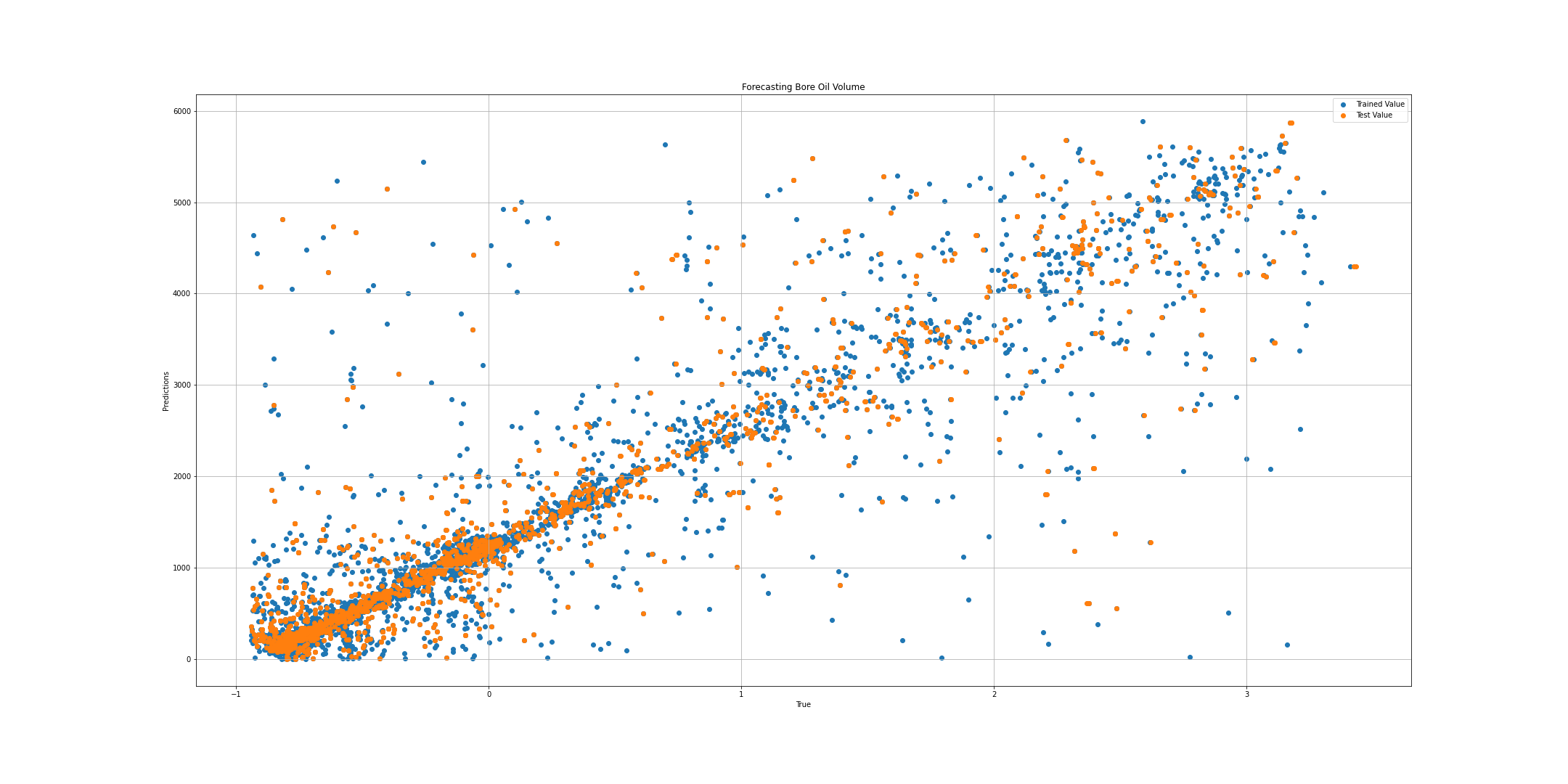


**Fig. 4.2.** Production data vs. prediction using AdaBoosting

1. **Decision Tree:** Decision Tree is a form of white box ML algorithm. It shares intrinsic decision-making logic that black box algorithms such as Neural Network do not have. Its training period is less than that of the neural network approach.

The temporal complexity of decision trees is proportional to the amount of records and characteristics in the input data. The decision tree is a non-parametric or distribution-free strategy that does not rely on probability distribution assumptions. Decision trees are capable of handling high-dimensional data with excellent accuracy.

Using the Decision Tree model on our dataset, we were able to generate the following plot, a model accuracy of 66.47%.

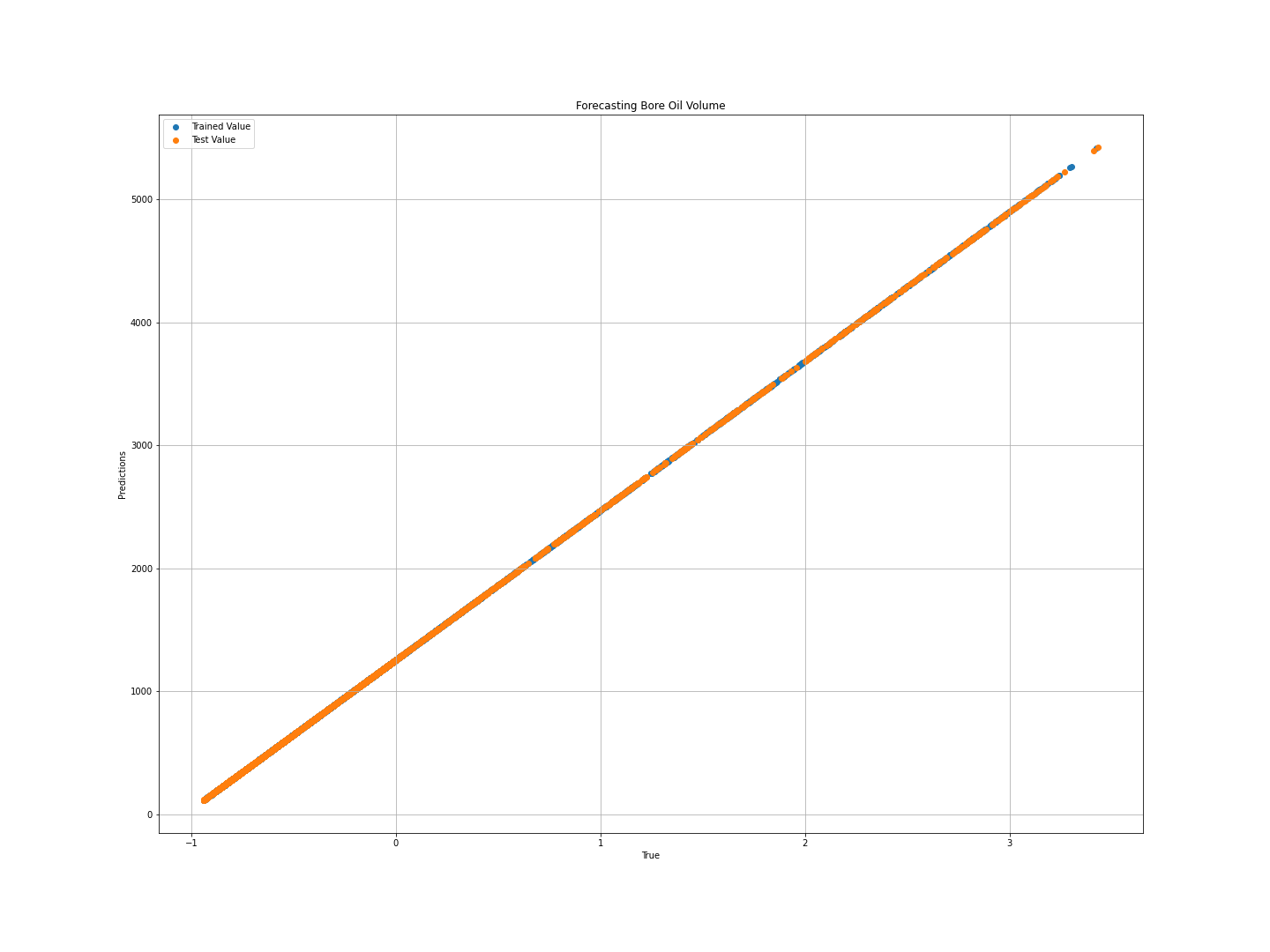


**Fig. 4.3.** Production data vs. prediction using Decision Tree

1. **LASSO:** This is an acronym that stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is a type of regularization.

Lasso Regression is a sort of regularized linear regression with an L1 penalty. This reduces the coefficients for input variables that do not contribute significantly to the prediction task. This penalty allows some coefficient values to be set to zero, thus removing input variables from the model and offering a sort of automated feature selection.

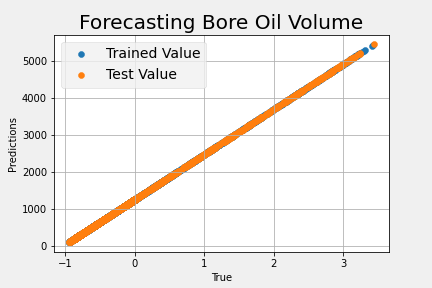
We were able to construct the following figure using the Decision Tree model on our dataset, with a model accuracy of 82.230%.



**Fig. 4.4.** Production data vs. prediction using Lasso

1. **Ridge:** Another regularization method that applies L2 regularization is ridge regression or Tikhonov regularization. By adding the penalty (amount of shrinkage) equal to the square of the coefficients' magnitudes, it also affects the loss function.

The following plot below, with a model accuracy of 83.42%, was produced using the Ridge model on our dataset.

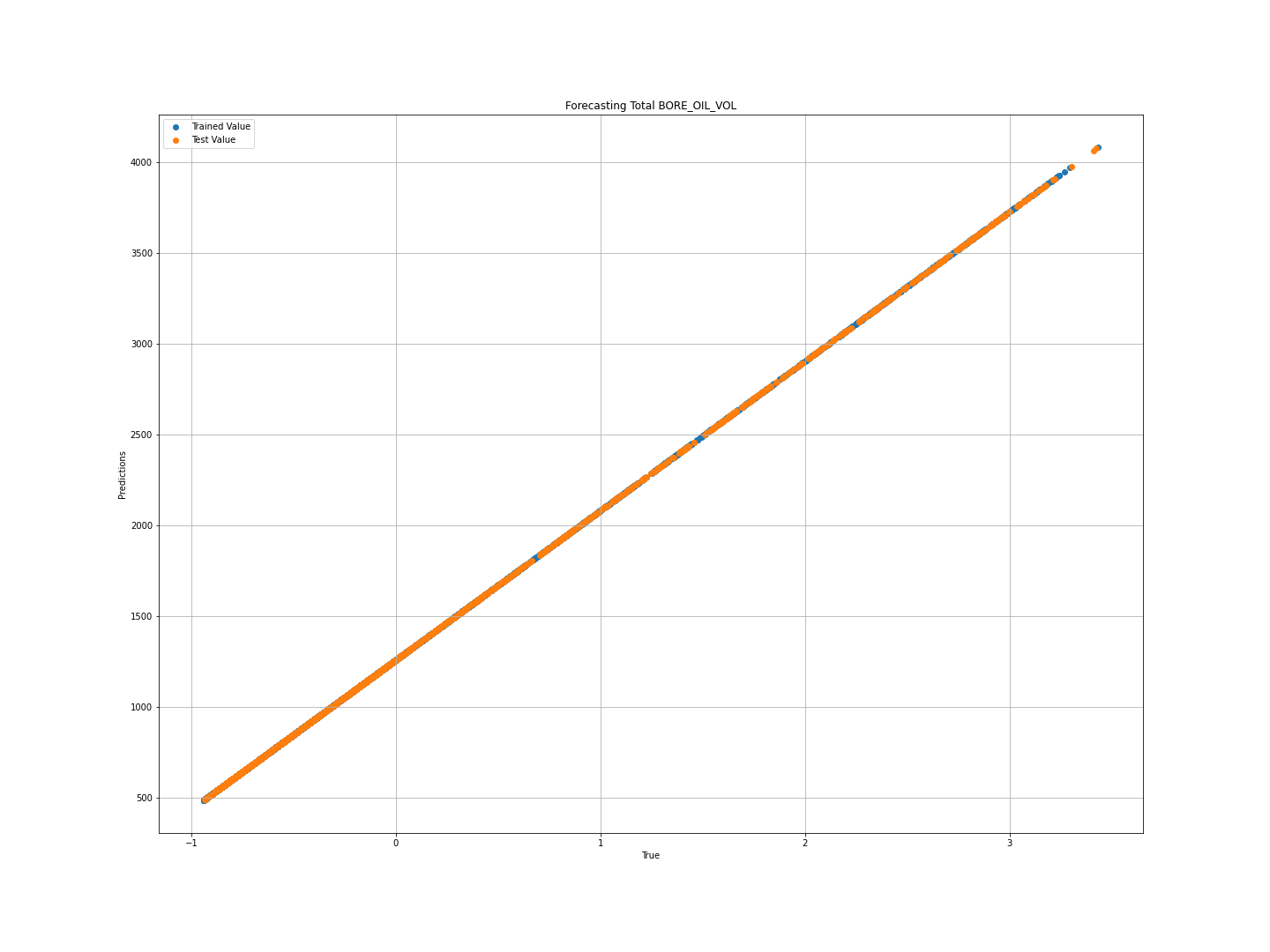


**Fig. 4.5.** Production data vs. prediction using Ridge

1. **ElasticNet Regression:** The Elastic-Net regression method is a regularized regression technique that linearly integrates the L1 and L2 penalties from the Lasso and Ridge regression techniques. When several connected traits exist, it is helpful. The distinction between Lass and Elastic-Net is that although Lasso is more likely to choose one of these traits at random, Elastic-Net is more likely to choose both at the same time.

The lasso technique has several drawbacks, but the elastic net approach addresses these by allowing the inclusion of "n" numbers of variables up until saturation, whereas lasso only takes a few samples for high-dimensional data. When the variables are grouped into highly linked groups, lasso usually picks one variable from these groups and ignores the others completely.

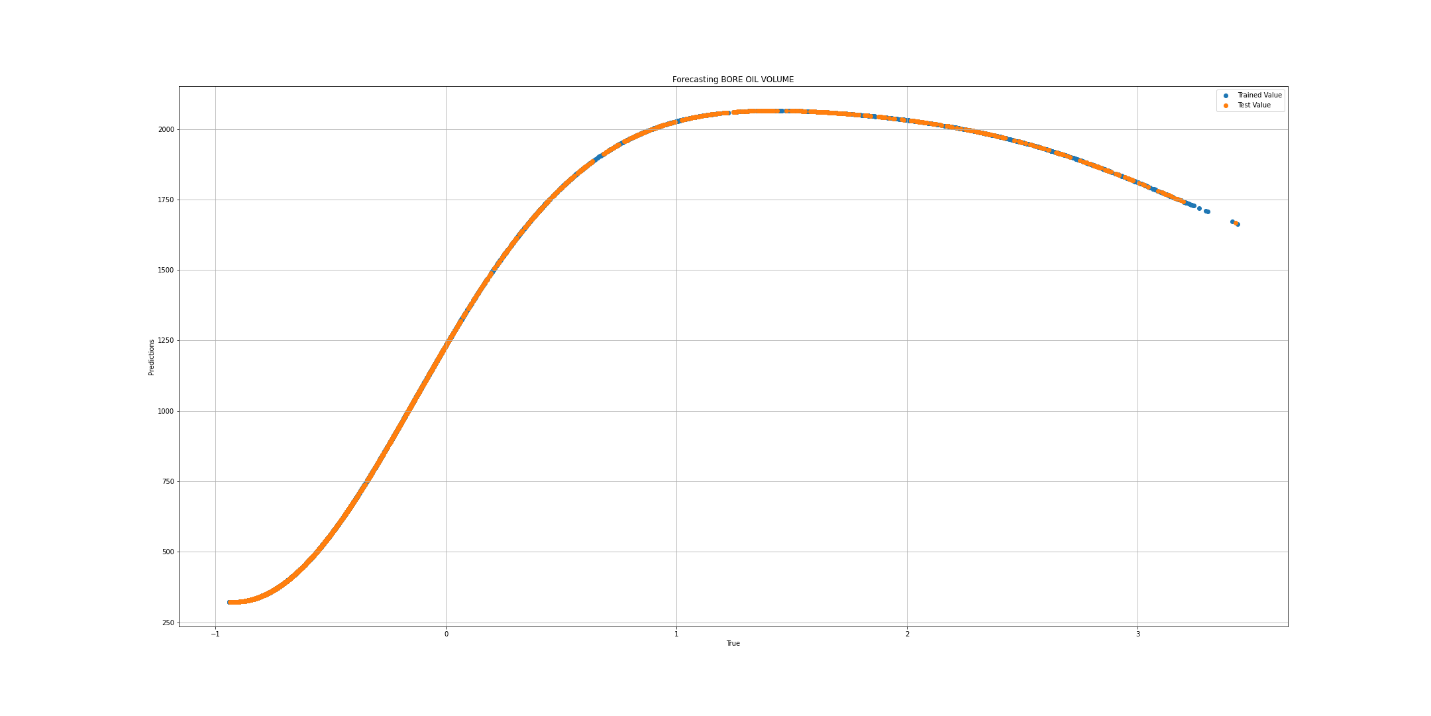
We were able to generate the following plot using the Decision Tree model on our dataset, with a model accuracy of 75.66%



**Fig. 4.6.** Production data vs. prediction using ElasticNet Regression

1. **SVR:** Support Vector Regression (SVR) is a regression function generalized by Support Vector Machines, a machine learning model used for continuous data categorization. Support Vector Regression is a form of Support Vector Machine that allows for both linear and nonlinear regression.

SVR necessitates the following training data: X, Y, which spans the domain of interest and is complemented with solutions in that domain. This model on our dataset produced an accuracy prediction of 52.16%.

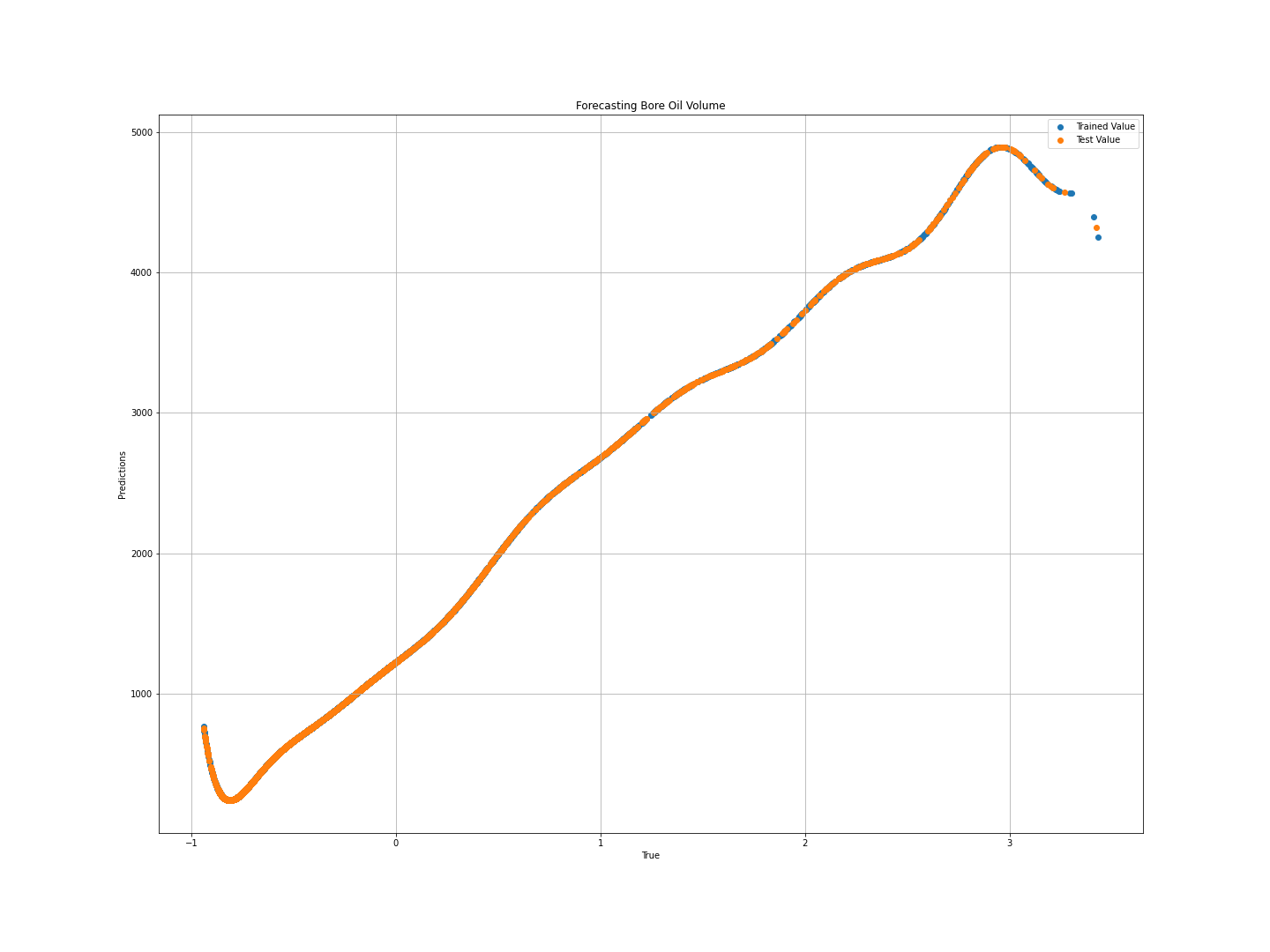
****

**Fig. 4.7.** Production data vs. prediction using SVR

1. **Gaussian Process Regressor:** Gaussian Process Regression (GPR) is a nonparametric, Bayesian regression technique that is making waves in the field of machine learning.

Unlike many popular supervised machine learning methods, the Bayesian methodology infers a probability distribution over all possible values.

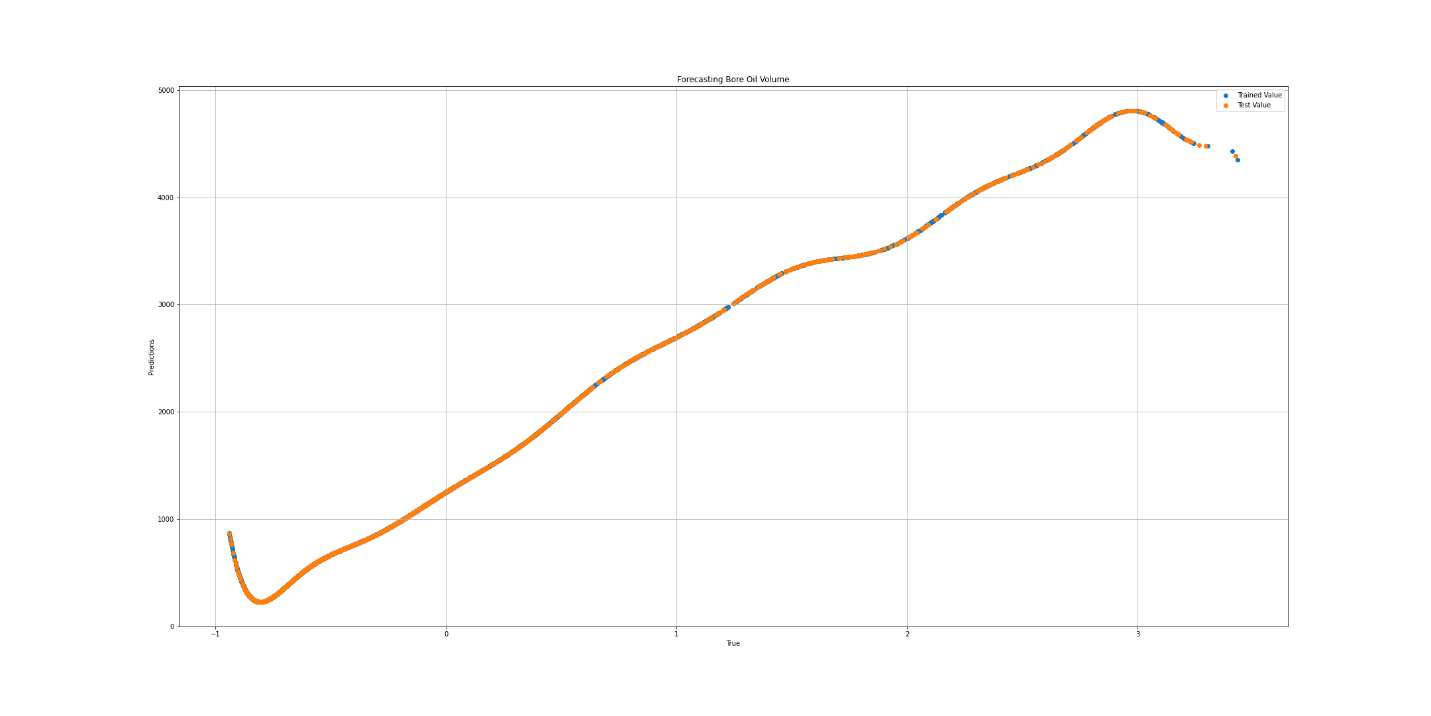
Because Gaussian process regression is nonparametric (i.e. not constrained by a functional form), it calculates the probability distribution across all admissible functions that fit the data rather than the probability distribution of parameters of a single function. The accuracy of this model on our dataset is 81.95%

****

**Fig. 4.8.** Production data vs. prediction using Gaussian Process Regressor

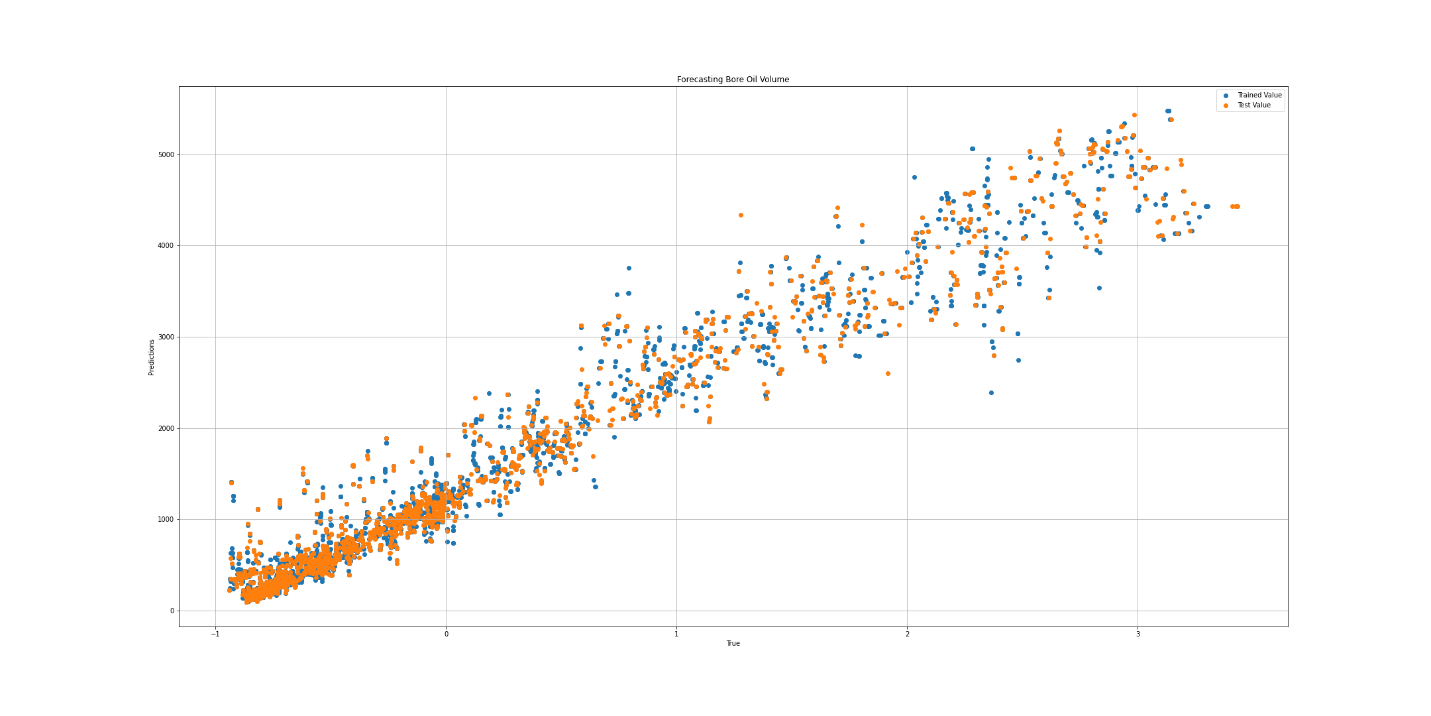
1. **Gradient Boosting Regressor:** By merging weak learners or weak predictive models, the Gradient Boosting technique is utilized to construct an ensemble model. The gradient boosting approach may be used to train regression and classification models.

Gradient boosting creates an additive mode by employing several fixed-size decision trees as weak learners or weak predictive models. The option n estimators determines how many decision trees will be utilized in the boosting phases. Gradient boosting varies from AdaBoost in that AdaBoost uses decision stumps (one node and two leaves), whereas Gradient Boosting uses decision trees of fixed size. On our dataset the accuracy of the model is 83.64%

****

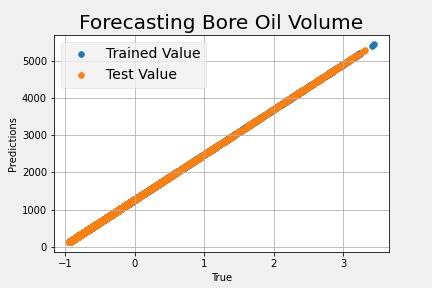
**Fig. 4.9.** Production data vs. prediction using Gradient Boosting Regressor

1. **KNeighbors Regressor:** The K Nearest Neighbors Algorithm is a simple algorithm that collects all available data and predicts the numerical target using a similarity metric (e.g., distance functions). KNN has been utilized as a non-parametric approach in statistical estimates and pattern recognition since the early 1970s. On our dataset the accuracy of the model is 76.55%

****

**Fig. 4.10.** Production data vs. prediction using KNeighbors

1. **Linear\_Regression:** The relationship between a scalar response and one or more explanatory factors (also known as dependent and independent variables) can be modeled linearly using the technique of regression. In linear regression, the connections are modeled using linear predictor functions, and the model's unobserved parameters are inferred from the data. We refer to these models as linear models. Based on this model is a model accuracy of 84.254% is derived.

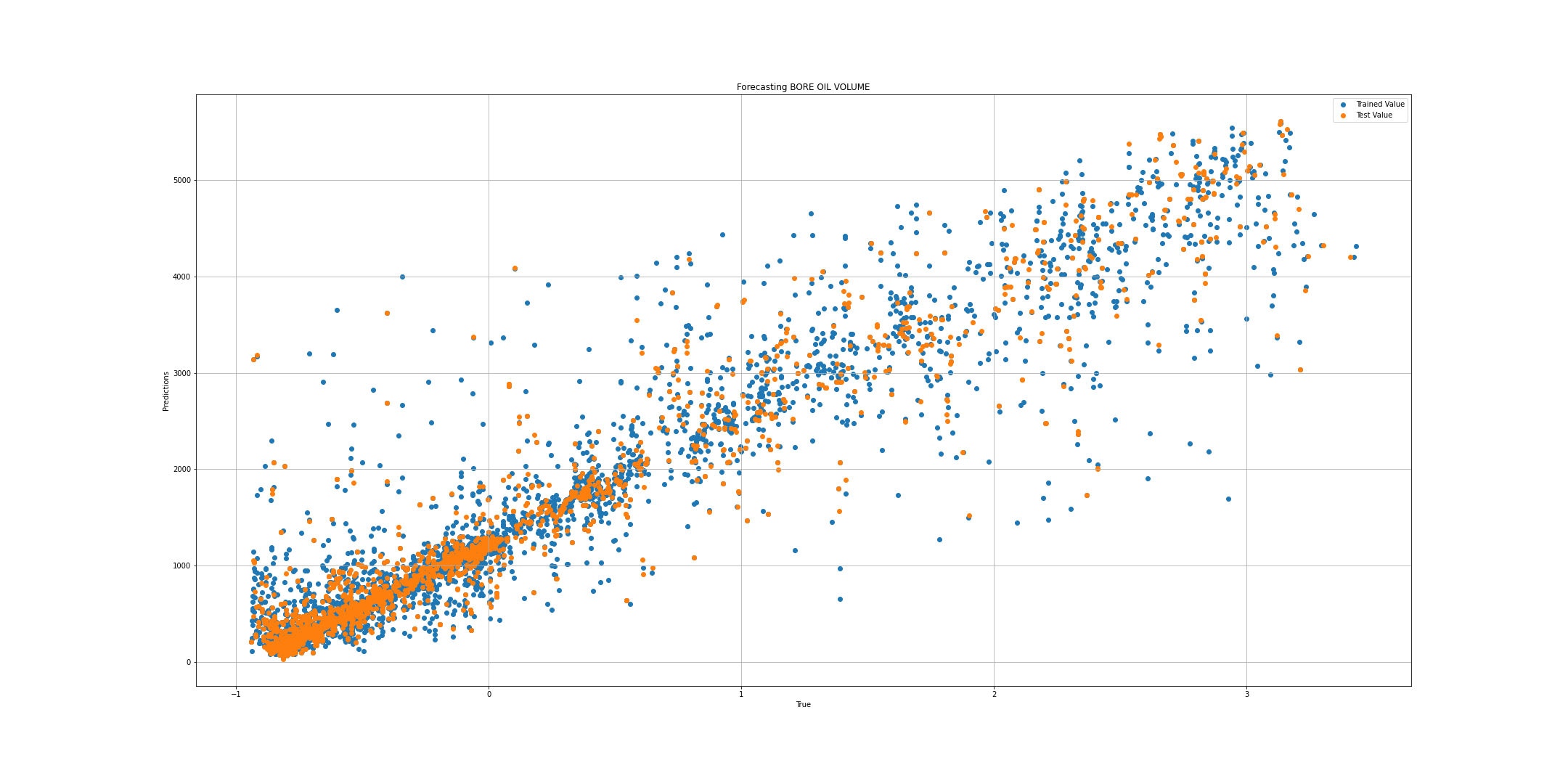
****

**Fig. 4.11.** Production data vs. prediction using Linear Regresion

1. **Random\_Forest:** A large number of decision trees are built during the training phase of the random forests or random decision forests ensemble learning approach, which is used for classification, regression, and other tasks. The class that the majority of the trees choose is the output of the random forest for classification problems. The mean or average forecast of each individual tree is returned for regression tasks. The tendency of decision trees to overfit their training set is corrected by random decision forests.

Although they frequently outperform decision trees, gradient enhanced trees are more accurate than random forests.

However, their effectiveness may be impacted by data peculiarities. 77.63% model accuracy was derived from this model.

****

**Fig. 4.12.** Production data vs. prediction using Random Forest

**4.3. Measures of Prediction Accuracy**

Scale-dependent errors and percentage errors are the two types of errors that are often quantified in the literature to quantify forecasting accuracy and performance evaluation of the predictions.

1. **Scale-dependent Errors:** The extent of these errors matches that of the underlying data. Because of this, it is a restriction that series on various scales cannot be compared using accuracy measurements that are based only on this inaccuracy. Therefore, the root mean square error (RMSE) (R.J. Hyndman, A.B. Koehler, 2006) provides the basis for the recognized scale-dependent measure and may be expressed as follows:

Equation 4.1

Where, yobsi is the current observation and ypredi is its predicted value.

1. **Mean Absolute Error:** The average of the absolute difference between the dataset's actual and anticipated values is represented by the Mean Absolute Error. It calculates the dataset's residuals' average.

Equation 4.2

Where yi is the predicted value of y and y the actual y values

1. **Mean Square Error:** The mean of the squared difference between the data set's original and predicted values is known as the mean squared error. It calculates the residuals' variance.

Equation 4.3

1. **Coefficient of Determination (R2):** The coefficient of determination, often known as R-squared, shows the fraction of the variation in the dependent variable that the linear regression model explains. It is a scale-free score, which means that regardless matter whether the values are little or large, the value of R square will be less than one.

Equation 4.4

R- Squared and RMSE both quantify how well a regression model fits a dataset. The RMSE measures how effectively a regression model predicts the absolute value of a response variable, whereas R- Squared measures how well the predictor variables explain the variance in the response variable. The table below gives a compassion on the measures of prediction accuracy of all the models developed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Models** | **Accuracy** | **RMSE** | **MSE** | **R2** | **MAE** |
| ANN Regression | 91.83% | 57.3131757 | 3284.8 | 0.9183 | 32.36482 |
| Linear Regression | 84.25% | 2.0948 | 293935.1 | 0.8425 | 270.6317 |
| Gradient Boosting | 83.64% | 564.265357 | 294399.6 | 0.8364 | 261.1862 |
| Ridge | 83.42% | 571.350729 | 309000.7 | 0.8342 | 267.4 |
| Lasso | 82.23% | 570.918614 | 309969.4 | 0.8223 | 258.197 |
| Gaussian Process Regression | 81.95% | 556.173235 | 315947.8 | 0.8195 | 261.881 |
| Ada Boosting | 79.16% | 608.389854 | 370138.2 | 0.7917 | 397.9778 |
| Random Forest | 77.63% | 260.011632 | 404780.1 | 0.7763 | 318.4598 |
| KNN | 76.55% | 483.933049 | 418809.2 | 0.7655 | 319.1827 |
| ElasticNet | 75.66% | 666.589853 | 17.566 | 0.7566 | 468.4899 |
| DecisionTree | 66.75% | 767.630656 | 589256.8 | 0.6675 | 381.5057 |
| SVR | 52.16% | 957.243721 | 847966.3 | 0.9589 | 464.0944 |

**Table 4.1:** Comparison of Model Prediction Accuracy

**CHAPTER FIVE**

**CONCLUSION AND RECOMMENDATIONS**

**5.1 Conclusion**

The application of Artificial Neural Network (ANN) for the prediction of the oil production has been elaborately demonstrated. By feeding concerned data to the network, ANN was effectively trained to oil production. To make an objective assessment, the effectiveness of the suggested strategy is contrasted with a number of widely used techniques, either statistical or soft computing. Different models were analyzed and the vanilla neural network was finally settled with. The empirical findings reveal that the suggested model performs better than other conventional techniques using various measurement criteria.

Again, the output of the model provided good matches with original experimental data and the work shows the reliability and applicability of the ANN in oil production prediction. The developed model in the work gives universal access and is quite easy.

**5.2 Recommendations**

Numerous difficulties were faced over the course of this undertaking. The following suggestions have been made in order to improve the oil production prediction model created in this study in the future:

1. The Artificial Neural Network, a machine learning tool, forms the foundation for this research (ANN). The quality and quantity of data used have a significant impact on this tool's prediction process. The skewness of the dataset was one of the biggest problems that were faced throughout the actual work. Therefore, it is advised that more research be conducted on this work using adequate, reliable data.
2. This research is restricted to predicting oil production. It is advised that more research be done on the forecast of gas output and water cut. It's important to remember that water is generated with oil in every oil and/or gas producing well. The production of oil and gas is viable up to a certain water-cut ratio that is also economically viable, but when the water percentage rises and passes that threshold, the well is abandoned. While the specific well may still be producing, it is now producing more water than hydrocarbon fluid, or it may have once been producing, but it has now entered its decline phase and is close to being abandoned. The statistics on these wells' water output might confirm the same conclusion.

**REFERENCES**

Abdullayeva, F.; Imamverdiyev, Y. Development of Oil Production Forecasting Method Based on Deep Learning. Optim. Inf. Comput. 2019, 7, 826–839.

Ahmadi, M.A.; Chen, Z. Machine Learning Models to Predict Bottom Hole Pressure in Multi-Phase Flow in Vertical Oil Production Wells. Can. J. Chem. Eng. 2019, 97, 2928–2940

Ahmadloo, F., Asghari, K., & Renouf, G. (2010). A new diagnostic tool for performance evaluation of heavy oil waterfloods: case study of western Canadian heavy oil reservoirs. Anaheim, CA, USA, SPE Western Regional Meeting, May.

AI-Kaabi A.U., McVay, D.A. and Lee W.J. “Using an Expert System to Identify a Well Test Interpretation Model”. Journal of Petroleum Technology, (May), pp. 654-661, 1990.

al Ajmi, M.D.; Alarifi, S.A.; Mahsoon, A.H. Improving Multiphase Choke Performance Prediction and Well Production Test Validation Using Artificial Intelligence: A New Milestone. In Proceedings of the Society of Petroleum Engineers—SPE Digital Energy Conference and Exhibition 2015, The Woodlands, TX, USA, 3–5 March 2015.

Al-Bulushi, N.I., King, P.R., Blunt, M.J. *et al.* Artificial neural networks workflow and its application in the petroleum industry. *Neural Comput & Applic* **21**, 409–421 (2012). <https://doi.org/10.1007/s00521-010-0501-6>

Al-Fattah, S. M., and R. A. Startzman. "Neural Network Approach Predicts U.S. Natural Gas Production." Paper presented at the SPE Production and Operations Symposium, Oklahoma City, Oklahoma, March 2001. doi: https://doi.org/10.2118/67260-MS

Alrifaey, M.; Lim, W.H.; Ang, C.K. A Novel Deep Learning Framework Based RNN-SAE for Fault Detection of Electrical Gas Generator. IEEE Access 2021, 9, 21433–21442.

Al-Shabandar, R.; Jaddoa, A.; Liatsis, P.; Hussain, A.J. A Deep Gated Recurrent Neural Network for Petroleum Production Forecasting. Mach. Learn. Appl. 2021, 3, 100013.

Al-Wahaibi, T.; Mjalli, F.S. Prediction of Horizontal Oil-Water Flow Pressure Gradient Using Artificial Intelligence Techniques. Chem. Eng. Commun. 2014, 201, 209–224

Amaechi, U.C.; Ikpeka, P.M.; Xianlin, M.; Ugwu, J.O. Application of Machine Learning Models in Predicting Initial Gas Production Rate from Tight Gas Reservoirs. Rud.-Geol.-Naft. Zb. 2019, 34, 29–40.

Aminian, K., Ameri, S., Oyerokun, A., & Thomas, B. (2003). Prediction of flow units and permeability using artificial neural networks. SPE Western Regional/AAPG Pacific Section Joint Meeting.

Artun, E., Ertekin, T., Watson, R., & Miller, B. (2012). Designing cyclic pressure pulsing in naturally fractured reservoirs using an inverse looking recurrent neural network. Computers & Geosciences, 38(1), 68-79.

Ayala H, L. F., & Ertekin, T. (2007). Neuro-simulation analysis of pressure maintenance operations in gas condensate reservoirs. Journal of Petroleum Science & Engineering, 58(1-2), 207-226.

Chaki, S., Verma, A. K., Routray, A., Mohanty, W. K., & Jenamani, M. (2014). Well tops guided prediction of reservoir properties using modular neural network concept: a case study from western onshore, India. Journal of Petroleum Science and Engineering, 123, 155-163.

Doan, T.T.; van Vo, M. Using Machine Learning Techniques for Enhancing Production Forecast in North Malay Basin. In Springer Series in Geomechanics and Geoengineering; Springer: Singapore, 2020; pp. 114–121

El-Sebakhy, E. A., Asparouhov, O., Abdulraheem, A., Al-Majed, A., Wu, D., Latinski, K., Raharja, I. Functional networks as a new data mining predictive paradigm to predict permeability in a carbonate reservoir, Expert Systems with Applications, Volume 39, Issue 12, 2012, Pages 10359-10375, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2012.01.157>.

Fedutenko, E., Yang, C., Card, C., & Nghiem, L. X. (2012). Forecasting SAGD Process Under Geological Uncertainties Using Data-Driven Proxy Model. SPE Heavy Oil Conference Canada.

Fedutenko, E., Yang, C., Card, C., & Nghiem, L. X. (2014). Time-dependent neural network based proxy modeling of SAGD process. SPE Heavy Oil Conference-Canada.

George, Alexander "Predicting Oil Production Flow Rate Using Artificial Neural Networks - The Volve Field Case." Paper presented at the SPE Nigeria Annual International Conference and Exhibition, Lagos, Nigeria, August 2021. doi: <https://doi.org/10.2118/208258-MS>

Ghorbani, H.; Wood, D.A.; Choubineh, A.; Tatar, A.; Abarghoyi, P.G.; Madani, M.; Mohamadian, N. Prediction of Oil Flow Rate through an Orifice Flow Meter: Artificial Intelligence Alternatives Compared. Petroleum 2020, 6, 404–414.

Graupe, D. (2007). Principles of Artificial Neural Networks (Vol. 2). Singapore: World Scientific Publishing Company.

Guo, Z.; Wang, H.; Kong, X.; Shen, L.; Jia, Y. Machine Learning-Based Production Prediction Model and Its Application in Duvernay Formation. Energies 2021, 14, 5509.

Habiballah, W. A., Startzman, R. A., and M. A. Barrufet. "Use of Neural Networks for Prediction of Vapor/Liquid Equilibrium K Values for Light-Hydrocarbon Mixtures." SPE Res Eng 11 (1996): 121–126. doi: https://doi.org/10.2118/28597-PA

Han, D.; Kwon, S. Application of Machine Learning Method of Data-driven Deep Learning Model to Predict Well Production Rate in the Shale Gas Reservoirs. Energies 2021, 14, 3629

Haykin, S. S. (2001). Neural networks: a comprehensive foundation: Tsinghua University Press. J. Forecast. 22 (4) (2006) 679–688.

Jabbari, M.; Khushaba, R.; Nazarpour, K.; Han, S.; Zhong, X.; Shao, H.; Pan, S.; Wang, J.; Zhou, W. Prediction on Production of Oil Well with Attention-CNN-LSTM. J. Phys. Conf. Ser. 2021, 2030, 012038

Khamehchi, E.; Bemani, A. Prediction of Pressure in Different Two-Phase Flow Conditions: Machine Learning Applications. Measurement 2021, 173, 108665

Malallah, A., & Nashawi, I. S. (2005). Estimating the fracture gradient coefficient using neural networks for a field in the Middle East. Journal of Petroleum Science and Engineering, 49(3), 193-211.

Maschio C, D.J. Schiozer, C.P.V. Carvalho. A new methodology to reduce uncertainties in reservoir simulation models using observed data and sampling techniques. J. Pet. Sci. Eng., 72 (1–2) (2010), pp. 110-119

Maschio, C., & Schiozer, D. J. (2014). Bayesian history matching using artificial neural network and Markov Chain Monte Carlo. Journal of Petroleum Science and Engineering, 123, 62- 71.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. MIT Press, Cambridge MA, 3-44

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of mathematical biophysics, 5(4), 115-133.

McKinney, P.D., Rushing, J.A., and Sanders, L.A. 2002. Applied Reservoir Characterization for Maximizing Reserve Growth and Profitability in Tight Gas Sands: A Paradigm Shift in Development Strategies for Low-Permeability Gas Reservoirs. Paper presented at the SPE Gas Technology Symposium, Calgary, Alberta, Canada. Copyright 2002, Society of Petroleum Engineers Inc. 00075708. DOI: 10.2118/75708-ms.

McVay, D.A., Dossary, M.N. 2014. The Value of Assessing Uncertainty. SPE Journal 6 (2): pp. 100-110. DOI: 10.2118/160189-pa.

Mehrotra, K., Mohan, C. K., & Ranka, S. (1997). Elements of artificial neural networks: MIT press.

Mirzaei-Paiaman, A.; Salavati, S. The Application of Artificial Neural Networks for the Prediction of Oil Production Flow Rate. Energy Sources 2012, 34, 1834–1843

Mohaghegh, S. (2000). Virtual-Intelligence Applications in Petroleum Engineering: Part IArtificial Neural Networks. Journal of Petroleum Technology, 52(9), 64-73.

Mohaghegh, S., Gaskari, R., Popa, A., Ameri, S., Wolhart, S., 2001. Identifying best practices in hydraulic fracturing using virtual intelligence techniques. In: SPE 72385 Proceedings of the SPE Eastern Regional Conference and Exhibition, October, pp. 17–19.

Mohammadpoor, M., Firouz, Q., Reza, A., & Torabi, F. (2012). Implementing simulation and artificial intelligence tools to optimize the performance of the CO2 sequestration in coalbed methane reservoirs. Carbon Management Technology Conference.

N.C. Chakra, K.-Y. Song, M.M. Gupta, D.N. Saraf, An innovative neural forecast of cumulative oil production from a petroleum reservoir employing higher-order neural networks (HONNs), J. Pet. Sci. Eng. 106 (2013) 18–33

Negash, B.M.; Yaw, A.D. Artificial Neural Network Based Production Forecasting for a Hydrocarbon Reservoir under Water Injection. Pet. Explor. Dev. 2020, 47, 383–392.

Nguyen-Le, V.; Shin, H. Artificial Neural Network Prediction Models for Montney Shale Gas Production Profile Based on Reservoir and Fracture Network Parameters. Energy 2022, 244, 123150.

Nikravesh, M., Zadeh, L. A., & Aminzadeh, F. (2003). Soft computing and intelligent data analysis in oil exploration (Vol. 51): Elsevier.

Oberwinkler, C., Ruthammer, G., Zangl, G., & Economides, M. J. (2004). New tools for fracture design optimization. SPE International Symposium and Exhibition on Formation Damage Control.

Orrù, P.F.; Zoccheddu, A.; Sassu, L.; Mattia, C.; Cozza, R.; Arena, S. Machine Learning Approach Using MLP and SVM Algorithms for the Fault Prediction of a Centrifugal Pump in the Oil and Gas Industry. Sustainability 2020, 12, 4776.

Pal, M. On Application of Machine Learning Method for History Matching and Forecasting of Times Series Data from Hydrocarbon Recovery Process Using Water Flooding. Pet. Sci. Technol. 2021, 39, 519–549.

Popa, A. S., & Cassidy, S. D. (2012). Artificial intelligence for heavy oil assets: The evolution of solutions and organization capability. SPE Annual Technical Conference and Exhibition.

Popa, A. S., Cassidy, S. D., & Mercer, M. (2011). A data mining approach to unlock potential from an old heavy oil field. SPE Western North American Region Meeting.

Poulton, M. M. (2001). Computational neural networks for geophysical data processing (Vol. 30): Elsevier.

Queipo, N. V., Goicochea, J. V., & Pintos, S. (2002). Surrogate modeling-based optimization of SAGD processes. Journal of Petroleum Science and Engineering, 35(1), 83-93.

R.J. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, Int.

Raeesi, M., Moradzadeh, A., Ardejani, F. D., & Rahimi, M. (2012). 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, 82, 151-165.

Sami, N.A.; Ibrahim, D.S. Forecasting Multiphase Flowing Bottom-Hole Pressure of Vertical Oil Wells Using Three Machine Learning Techniques. Pet. Res. 2021, 6, 417–422.

Stoisits, R. F., Crawford, K. D., MacAllister, D. J., McCormack, M. D., Lawal, A. S., & Ogbe, D. O. (1999). Production optimization at the Kuparuk River field utilizing neural networks and genetic algorithms. SPE mid-continent operations symposium.

Stundner, M., & Al-Thuwaini, J. S. (2001). How data-driven modeling methods like neural networks can help to integrate different types of data into reservoir management. SPE Middle East Oil Show.

Suh, S. C. (2012). Practical applications of data mining: Jones & Bartlett Publishers.

Tamhane, D., Wong, P.M., Aminzadeh, F., and M. Nikravesh. "Soft Computing for Intelligent Reservoir Characterization." Paper presented at the SPE Asia Pacific Conference on Integrated Modelling for Asset Management, Yokohama, Japan, and April 2000. doi: <https://doi.org/10.2118/59397-MS>

Tang, H., Meddaugh, W. S., & Toomey, N. (2011). Using an artificial-neural-network method to predict carbonate well log facies successfully. SPE Reservoir Evaluation & Engineering, 14(01), 35-44.

Wahid, M.F.; Tafreshi, R.; Khan, Z.; Retnanto, A. Prediction of Pressure Gradient for Oil-Water Flow: A Comprehensive Analysis on the Performance of Machine Learning Algorithms. J. Pet. Sci. Eng. 2021, 208, 109265.

Xia, L.; Shun, X.; Jiewen, W.; Lan, M. Predicting Oil Production in Single Well Using Recurrent Neural Network. In Proceedings of the 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering, ICBAIE, Fuzhou, China, 12–14 June 2020; pp. 423–430.

Yeten, B., Durlofsky, L. J., & Aziz, K. (2002). Optimization of nonconventional well type, location and trajectory. SPE Annual Technical Conference and Exhibition.

Zangl, G., Graf, T., & Al-Kinani, A. (2006). Proxy modeling in production optimization. SPE Europec/EAGE Annual Conference and Exhibition.