Time-Series Forecasting of Crude Oil Production Using Hybrid Modeling

Umara Akhtar
Strategic Planning &Information
University of Bolton
Bolton, United Kingdom
U.Akhtar@bolton.ac.uk

Anchal Garg
School of Creative Technologies
University of Bolton
Bolton, United Kingdom
a.garg@bolton.ac.uk

Rossmary Villegas
Kalibrate
Manchester, United Kingdom
rossmaryvill@yahoo.com

***Abstract*—Crude oil is the main energy source, and its demand has been usually growing over years. It has always been an issue in the petroleum industry to forecast the production of crude oil to avoid disruption of supplies and keeping the prices of oil and commodities in control and thereby manage inflation. Hence, it becomes crucial to predict the production of crude oil. This study uses time series data to forecast crude oil production. Traditional statistical Autoregressive Integrated Moving Average (ARIMA). model and deep learning models like Long Short-Term Memory (LSTM), Artificial Neural Network (ANN), and Gated Recurrent Unit (GRU) are used for prediction and comparison. A hybrid technique is used to develop an ARIMA-ANN model to forecast crude oil production more accurately.**

Keywords—Autoregressive Integrated Moving Average (ARIMA). Long Short-Term Memory (LSTM), Artificial Neural Network (ANN), and Gated Recurrent Unit (GRU), Recurrent Neural Network (RNN), Hybrid Model, stationarity

I. INTRODUCTION

Crude oil is the main source of energy, and its demand has been ever increasing. The petroleum industry has struggled with predicting crude oil production that is essential to maintain supplies and adjust prices of petroleum products. This impacts the prices of other related commodities thus influencing the rate of inflation. Therefore, it is essential to forecast crude oil production to keep the prices and inflation under control.

There are several studies on prediction of crude oil prices but comparatively lesser on forecasting crude oil production [1]. Extant studies majorly use a singular statistical, machine learning or deep learning models for forecasting. According to the researchers there is no forecasting model that can be declared the best in every situation [2]. Also, ut is important for the prediction models to have high accuracy and low error rates. There has been evidence in literature that hybrid models give more accurate results compared to traditional single models. Thus, this paper intents to determine which hybrid models can give better accuracy and forecast crude oil production rate more precisely. Additionally, the study also attempts to identify the variables that affect performance of the model and then adjust them for better result.

This paper attempts to answer the following research questions (a) Which forecasting model yields the most precise forecast? (b) What factors influence the performance of the models? (c) Which models should be used together to create a hybrid model to forecast?

The production data is generally time-series data that is recorded at regular time intervals [3]. Due to its temporal sequence characteristic, time series forecasting has proven difficult in several sectors. Thus, finding an efficient model to predict crude oil production using time series data has been a difficult task in the oil and gas industry. Therefore, it is crucial to create a hybrid model that can forecast production levels more precisely.

Time series forecasting makes frequent use of both linear and non-linear models. The purpose of creating a hybrid model is to maximize on the advantages of both forecasting models. The hybrid model can help the oil and gas engineers to accurately predict the oil production. This could help the organisations related to oil and gas industry to take appropriate and timely actions to avoid crisis. It can also help the Government to develop strategies and control costs and inflation.

The study is divided into the following sections: section II discusses relevant work; section III discusses methodology; section IV covers results; section V provides discussion; and section VI presents conclusion.

II. RELATED WORK

The time series forecasting techniques fall into two basic categories: statistical and neural network - based methods. The conventional statistical models such as exponential smoothing, auto regressive, random walk and Auto-Regressive Integrated Moving Average (ARIMA) models have been used for prediction of crude oil production and pricing. However, these models fail to detect nonlinear patterns and perform poorly on large data sets [4]. To reduce the production rates, the petroleum industry has employed a mathematical Decline Curve Analysis (DCA) approach [5]. But the complexity of real-world data sets makes it impossible for mathematical expressions like DCA to be used to solve to time series forecasting problem [6].

The studies have also utilised soft computing method for forecasting. In one of the studies, the crude oil price was forecasted using ARIMA, Support Vector Machines (SVM), and Generalized Autoregressive Conditional Heteroscedastic (GARCH), and SVM outperformed both the models [7]. Studies have used an Artificial Neural Network (ANN) algorithm to

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

predict the oil flow rate from wells [8] and forecast oilfield production [9][10]. Aizenberg et al. used a multi-layer neural network model to forecast oil production [11].

To overcome the limitations of conventional approaches, Deep Long Short-Term Memory model (DLSTM) used for precise forecasting was found to outperform the conventional approaches [12][13]. Another study found that hybrid model yields better accuracy compared to using linear regression and deep belief networks independently [14]. Further, Long Short-Term Memory (LSTM) model was employed and was to outperform the ARIMA model [15].

It is uncommon to find purely linear or non-linear time-series data, an integration of models can produce significant improvements. The efficiency of these models and the disadvantages of statistical and soft computing models to actual problems have been discussed many times. Combining statistical and soft computing methods can solve these issues. In terms of accurate prediction with minimal error, the ARIMA model outperformed the LSTM and GRU models [16]. Since ARIMA performed best on its own, it can beat other forecasting models when paired with soft computing models like ANN. This technique is employed by few experts in forecasting. For the production forecasting of crude oil, [17] combined LSTM with Convolutional Neural Network (CNN). [18] projected Indonesian exports using a hybrid model that combined ARIMA and LSTM.

There are a lot of academic works on time series forecasting of crude oil pricing, but lesser on forecasting of crude oil production. This study adds to the existing body of knowledge by providing a hybrid model for better forecasting of crude oil production. To determine which forecasting model provides greater accuracy, the following models are used: ARIMA, ANN, LSTM, GRU, and Hybrid ARIMA-ANN. The hybrid model is utilized in order to benefit from the efficacy of both statistical and deep learning models in a single framework.

III. Methodology

For time-series forecasting of the crude oil production, various models such as ARIMA, ANN, LSTM and GRU were used. The models were combined to build a hybrid model. All these models were evaluated and one with higher accuracy was chosen as the predictive model.

A. Data Description

This study uses Florida's crude oil production dataset available at the U.S. Energy Information Administration website. It includes time series data from 1981 to 2016. The dataset has 431 rows with two columns, date, and rate of production.

B. Data Preparation

Data cleaning and preprocessing was performed before analysis. Fig. 1 depicts the time series visualized in Python by using Matplotlib package.

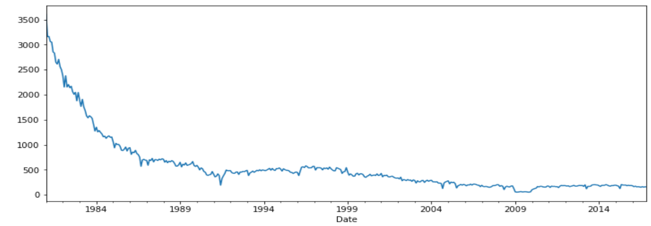


Fig. 1. Crude Oil Field Production in Florida (Thousand Barrels per Day)

C. Check for Stationarity

The Augmented Dickey Fuller (ADF) and Kwiatkowski-Philips-Schmidt-Shin (KPSS) tests were used to check for stationarity of the time-series data. The results of both the tests were combined to confirm the stationarity of the time series. The ADF test indicated that the series is stationary while KPSS showed it to be non-stationary, hence time series data of Florida oil fields was difference stationary.

D. Selecting Forecasting Models

Various time-series forecasting models were evaluated for selection in this study. ARIMA is the most common time-series forecasting model for the univariate data with linear trends [19]. However, linear models such as ARIMA captures only the linear component of the time-series. But in real-life situations, the time-series generally tends to be non-linear. To overcome the limitations of linear models, ANNs have been widely used for time-series forecasting. The neural networks can identify non-linear patterns [20] and hence use of these models is usually encouraged [21]. Time series data is generally complex in nature with longer sequences and significant non-linearity, making forecasting difficult for traditional ANNs [22]. Recurrent Neural Networks (RNNs) are used in time series forecasting to get around these ANN limitations. RNNs can learn short-term sequences, whereas LSTM, an RNN derivative, was developed to learn relatively long sequences [23]. Following the introduction of LSTMs, the Gated Recurrent Unit (GRU), a faster and less memory-intensive variant of RNNs, was developed. The deep learning models also, indeed, have some drawbacks. While RNNs are better at learning shorter patterns, ANNs struggle with learning longer patterns. Although LSTMs can overcome RNNs' shortcomings, their structure is rather complex, and training takes longer [23]. Thus, no single model, whether linear or non-linear, statistical or deep learning model is sufficient for forecasting trends in time-series data. Hence, this study used statistical ARIMA model and other the deep learning methods viz LSTM, GRU, and ANN to identify models with better projection.

IV. Result Analysis

This section presents the results of data analysis. The analysis was performed using various Python libraries, including pandas, keras, NumPy, statsmodel, sciKit-learn, tensorflow, and seaborn.

A. ARIMA Model Development

The optimal model returned by Auto-ARIMA was ARIMA (0,2,2) (0,0,0) [0]. For forecasting, the time series data was split into train set (97%) and test set (3%). The Root Mean Square Error (RMSE) for ARIMA (0,2,2) was calculated. Auto-

ARIMA might not always give the model with low error and hence, parameter tuning was done and RMSE for ARIMA. The RMSE for each of the models is shown in Table 1 along with some other parameter configurations that are used to determine which model performs better.

Table 1: RMSE values for ARIMA model

ARIMA	(0,2,2)	(0,1,2)	(0,1,3)	(0,1,4)
RMSE	8.759	7.850	7.923	7.450

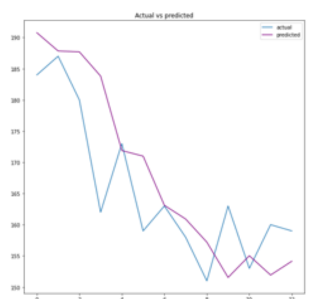
It is evident from Table 1, that ARIMA (0,1,4) is better than the other models. Hence, trends of ARIMA (0,1,4) and ARIMA (0,2,2) were compared (Fig. 2). ARIMA (0,1,4) clearly show more trends and thus outperform (0,2,2).

B. LSTM Model Development

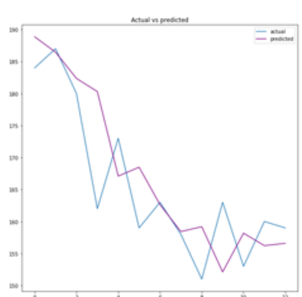
Keras was used to create the LSTM model. The timestep, data samples, and features were used as the matrix input for LSTM. Variable number of epochs were used. RMSE values for LSTM model is given in Table 2.

Table 2: RMSE values for LSTM model

LSTM	(1,2000,10)	(1,2500,10)	(1,3000,10)
RMSE	8.151	5.983	8.166



(a)ARIMA (0,2,2)



(b) ARIMA (0,1,4)

Fig. 2: Comparison of ARIMA models

Table 2 demonstrate that LSTM model with 2500 epochs provide more accurate result with a lower RMSE value. Fig. 3 show the actual vs prediction trend for LSTM LSTM model.

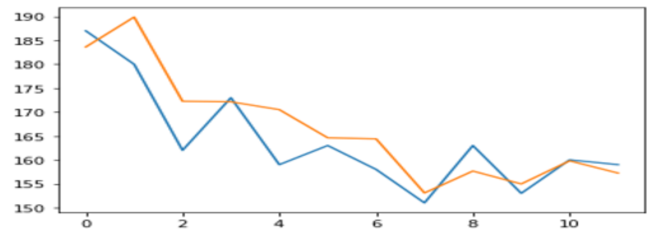


Fig. 3. Predicted vs Expected values for LSTM (1,2500,10)

C. GRU Model Development

The GRU layer was used in place of the LSTM layer, and was tested over a set of epochs, with a batch-size set to 1 and neurons set to 10. Table 3 provides the RMSE for each of the models.

Table 3: RMSE values for GRU model

GRU	(1,1500,10)	(1,2000,10)	(1,2500,10)	(1,3000,10)
RMSE	8.69	8.64	7.95	8.66

Table 3 show that GRU with 2500 epochs perform better than the other three models. Fig. 4 shows a plot of actual and projected oil production for GRU(1,2500,10).

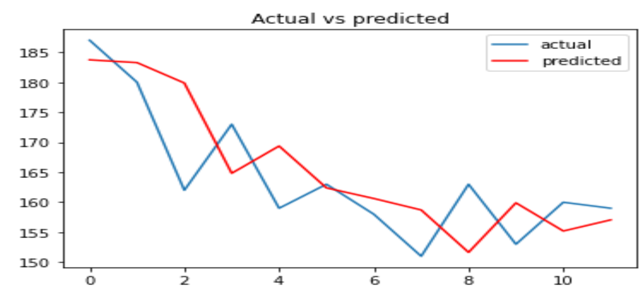


Fig. 4: Predicted vs Expected values for GRU (1,2500,10) model

D. ANN Model Development

For employing ANN, data was normalised, and as usual split into test sets, and converted to numpy arrays. A neural network with two hidden layers with 100 and 50 dimensions, respectively was developed. Window size of 8 was used. Table 4 shows RMSE of the ANN models for 1000, 1500, and 2000 epoch counts.

Table 4: RMSE values for ANN model

Epochs	RMSE
1000	8.11
1500	8.05
2000	8.56

Table 4 demonstrates lower RMSE value for ANN model with 1500 epochs. To test the impact on efficiency of the model, the batch-size was also altered, and the forecasting model with batch-size 20 produced better results. The batch-size was adjusted to 8, and the RMSE value dropped significantly. RMSE of is recorded with batch-size 8. The graph of the actual and projected outcome is shown in Fig. 5.

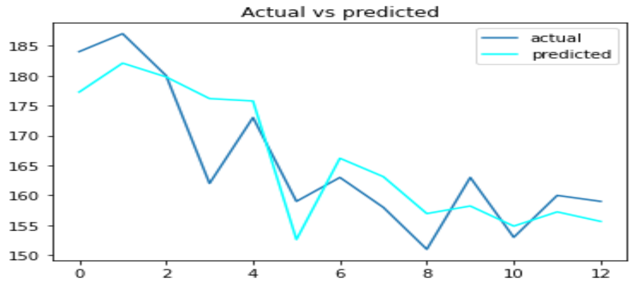


Fig. 5. Predictions made by ANN model with batch-size of 8.

E. Hybrid ARIMA-ANN Model Development

Further for forecasting oil production, different combinations of hybrid model were tested. However, a hybrid forecasting model combining ARIMA and ANN model gave reasonable results. The ARIMA model (0,1,4) was chosen due to higher accuracy. The residuals from ARIMA model were then fed as input to the ANN model. For ANN, epochs of 800 yielded the RMSE value of 1.29. The best RMSE values are determined by testing different numbers of epochs, RMSE values as well as the number of epochs are provided in table 5.

Table 5: RMSE values for Hybrid model

Epochs	1000	1200	1500
RMSE	0.98	0.40	0.30

The hybrid model performed better when epochs are set to 1500 (Fig. 6).

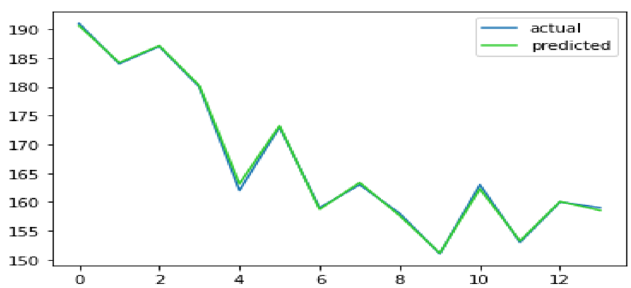


Fig. 6. Predictions made by Hybrid model.

The preceding plot demonstrates how well the hybrid ARIMA-ANN model learns the pattern in oil production and how closely it reflects the actual trend.

V. Discussion

Five forecasting models are evaluated for projecting crude oil production to assess model performance, the hybrid model outperformed the others by achieving the minimum RMSE. In table 6, RMSE for each model is listed.

Table 6: Forecasting accuracy comparison of all Models

Algorithm	RMSE
ARIMA	8.75
LSTM	5.98
GRU	7.95
ANN	5.78
Hybrid (ARIMA-ANN)	0.30

Initially, the ARIMA model's parameter selection was automatically done by utilising the auto-arima method, which provided ARIMA (0,2,2) model with the lowest AIC and an RMSE value of 8.75. Furthermore, tuning was done with various combinations of parameters to see whether better results could be reached utilising manual selection of parameters instead of using auto-arima. The manually chosen ARIMA (0,1,4) model produced lower RMSE of 7.45 as compared to ARIMA(0,2,2). This result demonstrates that utilising the Auto-ARIMA method is not always the best way to choose a model, but it can be a useful starting point. To attain the lowest RMSE with improved forecasts, further configurations can be applied. The RMSE was lower for hybrid ARIMA-ANN with two layers, but it increased when more layers were stacked. As a result, it can be concluded that adding more layers to a model may decrease accuracy rather than improving it. The model becomes more complex as more layers are added, and as a result, more parameters should also be trained. For the current time series, one or two layers are suitable because adding more layers might lead to over-fitting. The size of the dataset affects the hidden layer selection. Therefore, the accuracy of the deep neural network model is not necessarily enhanced by incorporating more layers. All deep learning models took different epoch counts to determine within how many epochs the forecasting model perform better. The prediction accuracy is not always improved by increasing the number of epochs. The efficiency of a model is also influenced by batch-size used in the network. The results from a smaller batch-size may be more effective. In order to address the ANN's difficulty in learning long-term sequences, a window size is specified. ANN model can be affected by a noisy data; this problem can be resolved by incorporating prior time-steps in addition to the present input. Whenever an input window is introduced, the model learns more effectively and produces better outputs. In such forecasting models, choosing optimal parameter combination is quite important.

VI. Conclusion

Time series forecasting has been conducted in this research utilising timeseries data of the actual oil fields in Florida from 1981 to 2016. Five forecasting models are evaluated to determine which one produces better crude oil production forecast. A hybrid model is used to combine the benefits of ARIMA and ANN models. Hybrid ARIMA-ANN forecasting model generated the most accurate forecast, surpassing all the four models. Due to the complex nature of the time series data, we reached the conclusion that an individual forecasting model is not likely to fulfil the criteria.
The integration of deep learning and statistical model massively improved the forecasting accuracy. In future, these results would be improved by carefully altering the parameters. Since LSTM also performed well as an independent model, it can be used in a hybrid architecture with ARIMA model to improve the forecasting results.
In this study there are some limitations. This study uses only total oil production for time-series forecasting, but other factors could also be useful in improving the predictive capability. Future research can make use of multivariate time series analysis to fully analyse the factors affecting crude oil production. The crude oil production data used in this study is the total production of the oil fields of Florida so, it does not provide forecasting for the individual oil fields in Florida. Another limitation of this study is that this study cannot be generalised to other studies and more likely for other time series problems, these models would give different results. Hence these limitations remain, and this study cannot overcome this drawback of time series problem. Finally, the challenge of finding one single robust model for all cases has not been met in this study and this remains a drawback of forecasting techniques.

References

[1] Busari, G. A., & Lim, D. H. (2021). Crude oil price prediction: A comparison between AdaBoost-LSTM and AdaBoost-GRU for improving forecasting performance. *Computers & Chemical Engineering*, *155*, 107513.

[2] Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

[3] Chatfield, C. (2000) *Time-series forecasting*, Chapman and Hall/CRC.

[4] Kamari, A., Mohammadi, A. H., Lee, M., & Bahadori, A. (2017). Decline curve based models for predicting natural gas well performance. *Petroleum*, *3*(2), 242-248.

[5] Längkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern recognition letters*, *42*, 11-24.

[6] Ahmed, R. A., & Shabri, A. B. (2014). Daily crude oil price forecasting model using arima, generalized autoregressive conditional heteroscedastic and support vector machines. *American Journal of Applied Sciences*, *11*(3), 425.

[7] Berneti, S. M., & Shahbazian, M. (2011). An imperialist competitive algorithm artificial neural network method to predict oil flow rate of the wells. *International journal of computer applications*, *26*(10), 47-50.

[8] Liu, Z., Wang, Z., & Wang, C. (2012). Predicting reservoir production based on wavelet analysis-neural network. In *Advances in Computer Science and Information Engineering: Volume 1* (pp. 535-539). Springer Berlin Heidelberg.

[9] Chakra, N. C., Song, K. Y., Gupta, M. M., & Saraf, D. N. (2013). An innovative neural forecast of cumulative oil production from a petroleum reservoir employing higher-order neural networks (HONNs). *Journal of Petroleum Science and Engineering*, *106*, 18-33.

[10] Aizenberg, I., Sheremetov, L., Villa-Vargas, L., & Martinez-Muñoz, J. (2016). Multilayer neural network with multi-valued neurons in time series forecasting of oil production. *Neurocomputing*, *175*, 980-989.

[11] Ma, X., & Liu, Z. (2018). Predicting the oil production using the novel multivariate nonlinear model based on Arps decline model and kernel method. *Neural Computing and Applications*, *29*, 579-591.

[12] Taieb, S. B., Bontempi, G., Atiya, A. F., & Sorjamaa, A. (2012). A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert systems with applications*, *39*(8), 7067-7083.

[13] Xu, W., Peng, H., Zeng, X., Zhou, F., Tian, X., & Peng, X. (2019). A hybrid modelling method for time series forecasting based on a linear regression model and deep learning. *Applied Intelligence*, *49*, 3002-3015.

[14] Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*(1), 75-85.

[15] Yamak, P. T., Yujian, L., & Gadosey, P. K. (2019, December). A comparison between arima, lstm, and gru for time series forecasting. In *Proceedings of the 2019 2nd international conference on algorithms, computing and artificial intelligence* (pp. 49-55).

[16] Abdullayeva, F., & Imamverdiyev, Y. (2019). Development of oil production forecasting method based on deep learning. *Statistics, Optimization & Information Computing*, *7*(4), 826-839.

[17] Dave, E., Leonardo, A., Jeanice, M., & Hanafiah, N. (2021). Forecasting Indonesia exports using a hybrid model ARIMA-LSTM. *Procedia Computer Science*, *179*, 480-487.

[18] Box G.E.P., Jekins G.M., Reinsel G.C. and Ljung G.M. (2015). *Time series analysis: forecasting and control*, John Wiley & Sons.

[19] Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks:: The state of the art. *International journal of forecasting*, *14*(1), 35-62.

[20] Pajo, J. F., Kousiouris, G., Kyriazis, D., Bruschi, R., & Davoli, F. (2021). ANNs going beyond time series forecasting: An urban network perspective. *IEEE Communications Magazine*, *59*(5), 88-94.

[21] Ma, X., & Liu, Z. (2018). Predicting the oil production using the novel multivariate nonlinear model based on Arps decline model and kernel method. *Neural Computing and Applications*, *29*, 579-591.

[22] Shewalkar, A. N. (2018). Comparison of rnn, lstm and gru on speech recognition data.

[23] Balaji, E., Brindha, D., Elumalai, V. K., & Vikrama, R. (2021). Automatic and non-invasive Parkinson's disease diagnosis and severity rating using LSTM network. *Applied Soft Computing*, *108*, 107463.