Technology and Investment, 2025, 16(2), 68-78
https://www.scirp.org/journal/ti
ISSN Online: 2150-4067
ISSN Print: 2150-4059

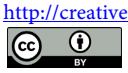
Development and Evaluation of Predictive Machine Learning Models for Crude Oil Supply Chain Logistics in the USA

Maame Korkor Prah[1], Amina Yakubu[2], Lawrence Simon Attah[3], Adeyemi Oluwatoba[4]

[1]Mathematics and Statistics, Austin Peay State University Clarksville, Tennessee, USA
[2]Technology, University of Central Missouri, Warrensburg, MO, USA
[3]Business Administration, Austin Cornell University Ithaca, NY, USA
[4]Dara Analytics, Northwest Missouri State University, Missouri, USA
Email: maamekprah@gmail.com, aminayakubu28@yahoo.com, Sal286@cornell.edu, adeyemioluwatoba25@gmail.com

How to cite this paper: Prah, M. K., Yakubu, A., Attah, L. S., & Oluwatoba, A. (2025). Development and Evaluation of Predictive Machine Learning Models for Crude Oil Supply Chain Logistics in the USA. *Technology and Investment, 16,* 68-78.
https://doi.org/10.4236/ti.2025.162005

Received: March 1, 2025
Accepted: April 20, 2025
Published: April 23, 2025



Abstract

Background and Theoretical Dilemma: The United States of America (USA) is the world's largest consumer of crude oil in the world. Ensuring the sustainability of the role of crude oil in the USA makes the need for effective crude oil supply chain logistics to be important. Therefore, this study tested the predictive ability of two machine learning models such as random forest and support vector machine (SVM) in relation to a classical statistical method such as ARIMA (Autoregressive Integrated Moving Average) for predicting the volume of crude oil import into the USA from 2024 to 2034. **Method**: Crude oil import data used for the prediction were sourced from U.S. Energy Information Administration. The data contained importation data from 1973 to 2023. The performance of the predictive models was tested with mean absolute error (MAE) and Root Mean Square Error (RMSE). **Key Findings and Conclusion**: Among the three predictive approaches used, SVM had the least MAE (265.65) and RMSE (362.91). This was followed by random forest (MAE =479.37; RMSE = 620.75) while ARIMA had the poorest performance (MAE =1670.10; RMSE = 2195.91). This implies that SVM outperformed the other predictive model for determining the import of crude oil from 2023 to 2034. In addition, among the sources from which crude oil is being imported to USA, Iraq, Canada and Russia have the highest feature importance for the random forest model. This implies that machine learning approach not only help predicts the future supply need for crude oil, but also areas where logistic management should be targeted to.

Keywords

Crude Oil, Import, Predictive Modelling

1. Background

Crude oil is an important resource that plays critical role in the United States of America (USA) energy security and economic stability (Oladosu et al., 2022). This critical role makes the management of crude oil supply chain to the USA an important priority for industry leaders and policy makers. Globally, USA is one of the largest consumers of crude oil as it is used to produce essential petroleum products such as gasoline, diesel fuel, heating oil, and jet fuel. The products of crude oil are used to power various sectors of the US economy. In 2023, about 12.933 million barrels of crude oil were produced in the USA per day (b/d) while about 6.478 million b/d were also imported (U.S. Energy Information Administration [EIA], 2023). While the country production has been increasing in recent years, the US still rely on foreign supply to augment what is refined in the country (U.S. EIA, 2024). The reliance on foreign crude oil makes it necessary for the USA to have a sustainable and efficient crude oil supply chain that can withstand disruptions and ensure a steady flow of imports to meet the country's energy needs.

The logistics of crude oil supply chain to the US is subjected to different complexities and challenges. For instance, globalization and interconnectivity has made crude oil supply chain to be at risk of disruptions from natural disasters, geopolitical instability, and economic fluctuations (Golgeci, Yildiz, & Andersson, 2020). For example, the ongoing conflict as a result of Russia invasion of Ukraine disrupt the global crude oil supply chain (Bagchi & Paul, 2023). In order to overcome the vulnerabilities of crude oil supply chain to disruption, there has been continuous advocacy for enhancing supply chain resilience and agility (Dai et al., 2022). Organizations and the government are working towards having an accurate predictive method through which disruption can be predicted as to enhance decision making with respect to mitigation strategy (Tissaoui et al., 2022). Such mitigation can help safeguard the continuous flow of crude oil and its derivatives (Foroutan & Lahmiri, 2024). Methods that are traditionally used for management supply chain are reactive in nature. Reactive approaches to vulnerability management are no longer sufficient in today's volatile, uncertain, complex, and ambiguous (VUCA) environment (Bird, 2018). This has led to an increased interest in predictive analytics and machine learning (ML) as tools for proactive supply chain management (Foroutan & Lahmiri, 2024).

Machine learning models are algorithms that have the capability to improve its predictive performance based on what it learnt from its experience. Example of these models are random forest and Support Vector Machine (SVM). Each of these models have the capability to handle large amount of data in order to generate predictive insights. SVM is defined as a supervised learning method that has the capability to carry out classification, regression and detection of outliers (Guido et al., 2024). Random forest on the other hand is a machine learning approach that reaches as single result after combining the output of multiple decision trees. Both SVM and random forest have the capability to learn from historical data, identify patterns and arrive at new information which can be used to

predict future trends and events (Gunasekaran, Lai, & Cheng, 2008). (2) However, SVM have some setbacks, such as interpretability limitation, high sensitivity of noisy data as well as outliers and challenge of selecting the right Kernel (Sayeed et al. 2024). This makes the need for comparing the predictive performance of SVM with other models.

While machine learning models such as random forest and SVM are valuable for making decision. The accuracy of their predictive performance still remains an issue that is receiving continuous attention (Jaiswal & Samikannu, 2017; Mohapatra, Shreya, & Chinmay, 2020). The need for accurate predictive method for the logistics of crude oil supply chain cannot be overstated. The resilience and agility of the supply chain industry is dependent on making timely and effective mitigative interventions to disruptions. Traditionally, Autoregressive Integrated Moving Average (ARIMA) has received wider applications when predicting historical datasets such as that of supply chain. The simplicity and effectiveness of ARIMA in dealing with time series data makes it to be widely used for prediction (Sonkavde et al., 2023). ARIMA has however been criticized for not being able to handle non-linear data and complex patterns that are characteristics of crude oil supply chains. In view of this limitations, this study aims to compare the predictive accuracy of ARIMA with that of advanced ML models such as RF and SVM for US crude oil imports.

RF and SVM were chosen over other models because; SVM has the ability to capture non-linear relationships through the use of kernel while RF models non-linearity by incorporating diverse decision trees in addition to being resistant to overfitting (Guido et al., 2024). Other models such as Generalized Autoregressive Conditional Heteroskedasticity (GARCH), Vector Autoregression (VAR) and Long Short-Term Memory (LSTM) can only give moderate accuracy in forecasting. Therefore, the rationale behind the evaluation of the predictive performance of SVM and RF in relation to ARIMA (Wang et al., 2017). In addition, the best performing predictive model was used to forecast U.S. crude oil imports from 2024 to 2034.

2. Materials and Method of Analysis

The data used for this study are secondary data of the Total US crude oil imports from Saudi Arabia, Venezuela, Iraq, Other Organization of the Petroleum Exporting Countries (OPEC), Other Non-OPEC Countries, Canada, Mexico and Russia (**Figure 1**). The data contains crude oil supply from the countries and regions from 1973 to 2023 and it was sourced from the Energy Information Administration's Monthly Energy Review (October 2023). Once the data was obtained, it was made to undergo several stages which includes data cleaning, model training, prediction, and model performance evaluation.

2.1. Data Cleaning and Preparation

The excel file of the study dataset was imported into R studio for the data cleaning.

The aim of the data cleaning was to enhance the integrity of the data prior to the training of the model. The first action that was taken after the data was imported was the removal of duplicates. The essence of duplicate removal was to ensure that redundancy was prevented. Missing values were also handled in order to ensure that there is consistency (Kuhn & Johnson, 2019).

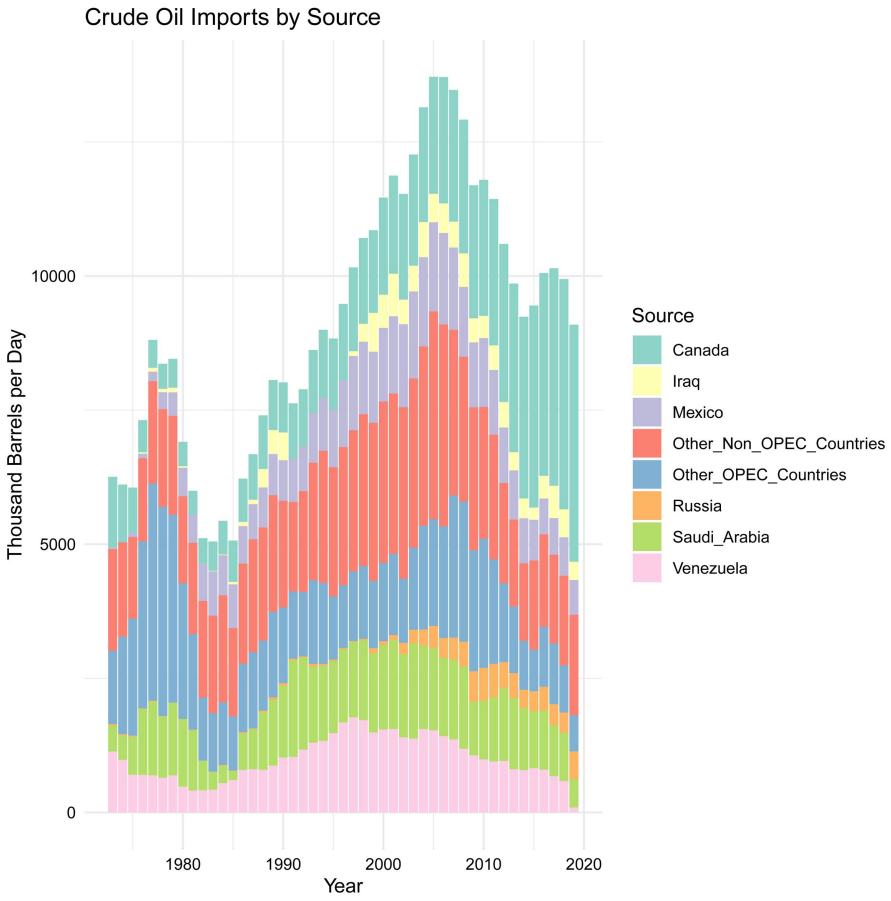


Figure 1. Sources of crude oil import to the US (1973 to 2023).

2.2. Model Training

Each of the models that were examined in the study were trained with method peculiar to their strength. The trend of total crude oil supply into the USA between 1973 to 2023 was examined with the use of R software, auto.arima() function. The function was used because it automatically determines the parameters that are optimal for autoregressive, differencing and moving average components. The ARIMA model was validated by splitting the data into two, with 70% of the data being used for training while the remaining 30% was used to test the model. Random sampling was used to split the data so as to ensure that the model is able to capture underlying trends while also being resilient to any fluctuations in the crude oil importation dataset (Hyndman & Athanasopoulos, 2021).

In the case of the SVM model, non-linear relationship in the study dataset was measured with the use of a radial basis function (RBF) kernel. The classification

accuracy was enhanced by the ability of the kernel function to transform separable data that are non-linear into a higher-dimensional space (Schölkopf & Smola, 2018). The SVM model was also enhanced with training and validation by suing the K-fold cross-validation. The K-fold cross-validation was helpful in optimizing the hyperparameters in addition to the prevention of overfitting (Pandian, 2024). The process used in the validation was to ensure that the SVM model can be generalised beyond the data that was used for the training in order to reduce bias and unreliable predictions.

For the random forest model, feature importance was used to enhance the model performance. The feature importance involves the random shuffling of individual feature values in order to measure the impact of each feature on the performance of the random forest model. Furthermore, uncertainty was reduced in the decision tree splits by evaluating the contribution of each feature with the use of mean decrease in impurity (GeeksforGeeks, 2025). In addition, overfitting was prevented by using cross validation to test the performance of the model. The cross validation also enhances the random forest model reliability by ensuring that its predictive performance was not limited to past pattern, but also future trends (Yates et al., 2023).

2.3. Model Prediction and Performance Evaluation

Once the training was done with the 70% of the dataset, model was used to predict oil imports from the test set. After the test was completed, the performance of the predictions was evaluated with the use of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The RMSE measures the performance of a predictive model by heavily penalizing larger errors with the way their differences are squared before averaging the errors (Kontopoulou et al., 2023). As for the MAE, the focus is on the magnitude of the average error of a predictive model. The implication of this performance evaluator is that, the best performing models are expected to have the lowest RMSE and MAE (Chai & Draxler, 2014; Jierula et al., 2021).

Prediction of the Import to be Supplied to US from 2024 to 2034.

The best performing model was used to forecast the trend of the crude oil supply into the US. The trend is visualized with a line graph. The use of visualization is to aid easy comprehension of the study findings.

3. Results

The actual trend and the predicted trend of the SVM, ARIMA and random forest model are visualized in **Figure 2**. The visualized trend shows that predicted imports of SVM and random forest mimics the actual trend of the supplied crude oil to the USA than that of the ARIMA model. The evaluation of the predictive performance also shows that SVM (MAE = 265.65, RMSE = 362.91) had the best performance while random forest (MAE = 479.37; RMSE = 620.75) had the second-best performance (**Figure 3**). The predictive performance of ARIMA was poor (MAE =1670.10; RMSE = 2195.91) as it reinforces the observed poor mimicry of the actual crude oil import to the USA.

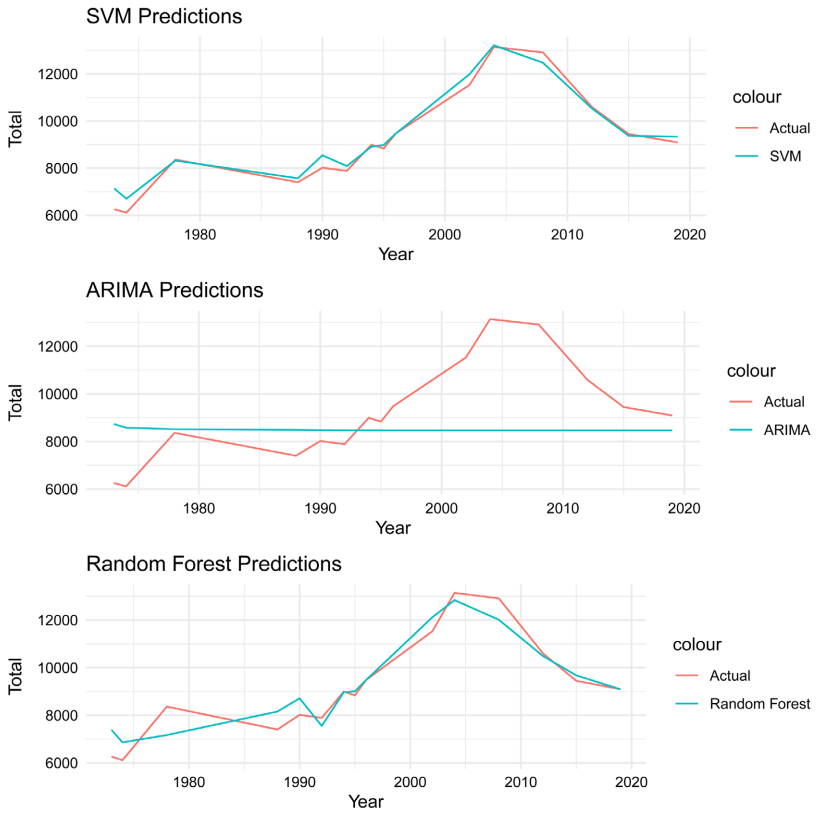


Figure 2. Crude oil import supply predictions of SVM, ARIMA, and random forest versus actual import.

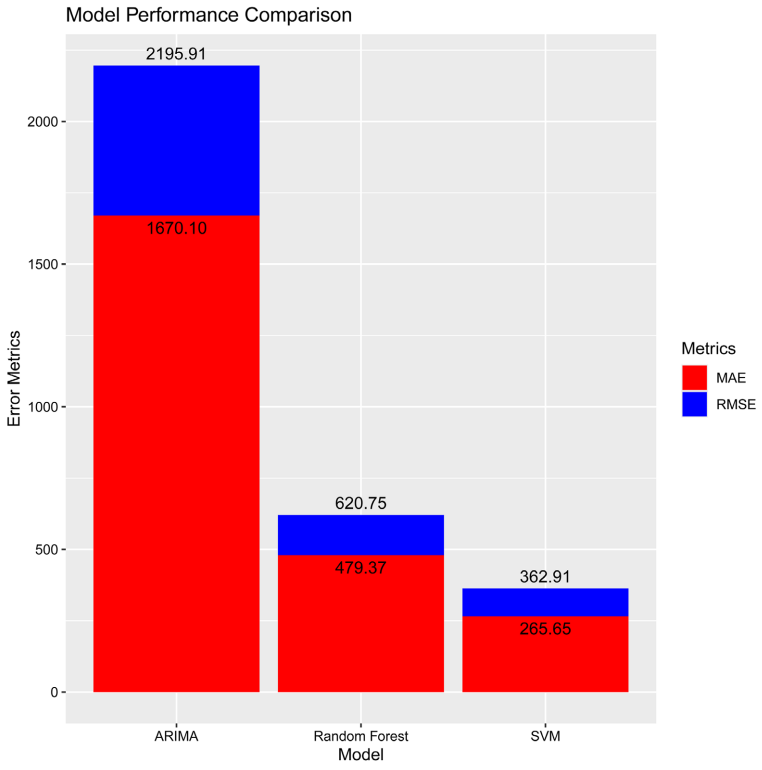


Figure 3. Performance of the models.

Since SVM was the best performing model, it was used to forecast the volume of crude oil import into the USA from 2024 to 2034 and the result is visualized in Figure 4. Based on the result, it is expected that the US would not be importing higher volumes of crude oil like what was imported between 2000 and 2020.

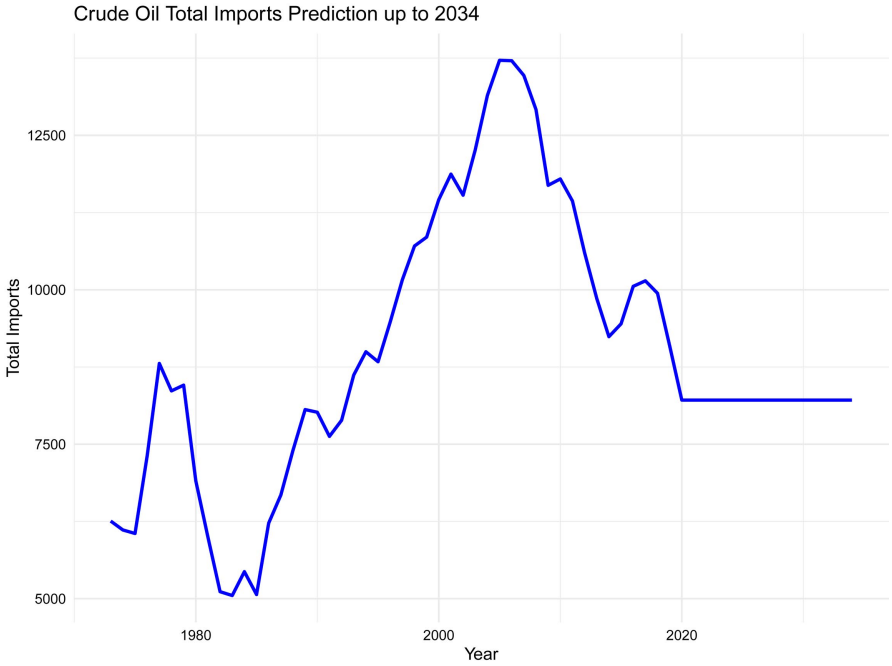


Figure 4. Prediction of crude oil import supply to the US from 2024 to 2034 with SVM.

4. Discussion

The predictive performance of SVM, ARIMA as well as random forest in forecasting crude oil import into the USA were compared in the study. Findings of the study suggest that random forest and SVM had better predictive ability than ARIMA model. Studies of Gajewski et al. (2023) and Zhu (2023) also found that the predictive performance of SVM and random forest for non-linear time series data were better than that of traditional statistics such as ARIMA. ARIMA has been criticized to have difficulty in capturing the dynamics that exist in intricate data sets (Gajewski et al., 2023; Zhu, 2023). This was evident with this study where ARIMA was unable to mimic the trends of the US crude oil import compared to the performance of Random Forest and SVM. When SVM and random forest were compared, SVM had the best performance. This was in line with the findings of Zhu (2023) that observed SVM outperforms other models in handling variables that are complex such as macroeconomic factors and exogenous influences.

Furthermore, the study of Jo et al. (2023) evaluate the predictive performance of ARIMA, vector autoregression model (VAR) vector error correction (VECM), SVM, RF and k-nearest neighbors (KNN) algorithm (KNN) on oil import. Their findings revealed that VECM and SVM were the best performing model. This made the researcher to recommend that the integration of timeseries model with

machine model offers potential for improve predictive performance of crude oil imports. Supporting the recommendation of Jo et al. (2023) were the study of Safari and Davallou (2018), Ning et al. (2022) and He et al. (2022) that all found a hybrid approach that involve time series model and machine learning model to have enhanced predictive performance.

The findings of this study have several implications in the oil and gas sector where accurate forecasting is needed. Economic dependence on crude oil makes the need to ensure that there is accurate forecast that can make planning for future easy. With the accuracy provided by SVM and Random Forest, informed decision can be made on how to adapt to predicted decline in import from any source region. Furthermore, policy makers could have more information on which policy and regulation can be made to sustain future importation of crude oil. Furthermore, with the poor performance of ARIMA, performance of alternative time series model such as VECM can be examined based on Jo et al. (2023) findings. The prospect of VECM combination with SVM for prediction of oil import as a hybrid approach requires assessment in order to validate the findings of other researchers who have recommended hybrid approach (Safari & Davallou, 2018; Ning et al., 2022; He et al., 2022).

Limitation of the Study

A key limitation of the study was that only a single dataset was used for the study. This limits the opportunity to compare other factors such as geopolitical events, economic fluctuations, or environmental policies that could have influence the volume of crude oil imported from the different sources. In addition, this study only compares traditional tool and machine learning model independently. No effort was made to integrate both traditional and machine learning model as to determine whether the hybrid approach could yield better performance. In view of this limitations, future studies should compare multiple datasets that could shed light on other factors that may influence oil importation in the US. Furthermore, there is the need for future study to explore the use of hybrid models (time series combination with machine learning model) to improve forecasting reliability. In addition, the applicability of the model in different settings should be tested to enhance robustness and generalizability of predictive performance of machine learning models.

5. Conclusion

In conclusion, the study demonstrates that machine learning models, particularly SVM and Random Forest, outperform traditional ARIMA models in predicting U.S. crude oil imports. SVM was the best-performing model, closely mimicking the actual import trends and providing reliable forecasts for future crude oil imports, while ARIMA struggled with accuracy. These findings reinforce the growing shift towards machine learning techniques in time-series forecasting, especially in industries with complex data patterns. The superior performance of SVM

highlights its potential for enhancing decision-making, resource optimization, and risk management in the oil and gas sector and beyond. Although, it is recommended that time series models should be integrated with machine learning model such as SVM to determine how enhanced the predictive accuracy can be. The performance of the models can have generalizable implication for countries from USA is importing crude oil from. A more robust data will be needed to test how the prediction can be generalized to other countries that their data were not included in the prediction of this study.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Bagchi, B., & Paul, B. (2023). Effects of Crude oil Price Shocks on Stock Markets and Currency Exchange Rates in the Context of the Russia-Ukraine conflict: Evidence from G7 Countries. *Journal of Risk and Financial Management, 16,* Article 64. https://doi.org/10.3390/jrfm16020064

Bird, R. C. (2018). *VUCA and the Legal Environment of Business*. University of Connecticut.

Chai, T., & Draxler, R. R. (2014). Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)—Arguments against Avoiding RMSE in the Literature. *Geoscientific Model Development, 7,* 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

Dai, L., Xu, L., Lim, M. K., Tseng, M. L., & Tan, K. H. (2022). Enhancing Resilience and Agility of Oil Supply Chain in the Context of COVID-19: A Multi-Methodological Study. *International Journal of Production Economics, 247,* Article ID: 108495. https://doi.org/10.1016/j.ijpe.2022.108495

Foroutan, P., & Lahmiri, S. (2024). Deep Learning Systems for Forecasting the Prices of Crude Oil and Precious Metals. *Financial Innovation, 10,* 1-40. https://doi.org/10.1186/s40854-024-00637-z

Gajewski, P., Čule, B., & Rankovic, N. (2023). Unveiling the Power of ARIMA, Support Vector and Random Forest Regressors for the Future of the Dutch Employment Market. *Journal of Theoretical and Applied Electronic Commerce Research, 18,* 1365-1403. https://doi.org/10.3390/jtaer18030069

GeeksforGeeks (2025). *Feature Importance with Random Forests*. https://www.geeksforgeeks.org/feature-importance-with-random-forests/

Golgeci, I., Yildiz, H. E., & Andersson, U. (2020). The Rising Tensions between Efficiency and Resilience in Global Value Chains in the Post-Covid-19 World. *Transnational Corporations, 27,* 127-141. https://doi.org/10.18356/99b1410f-en

Guido, R., Ferrisi, S., Lofaro, D., & Conforti, D. (2024). An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review. *Information, 15,* Article 235. https://doi.org/10.3390/info15040235

Gunasekaran, A., Lai, K., & Edwincheng, T. (2008). Responsive Supply Chain: A Competitive Strategy in a Networked Economy. *Omega, 36,* 549-564. https://doi.org/10.1016/j.omega.2006.12.002

He, H., Sun, M., Li, X., & Mensah, I. A. (2022). A Novel Crude Oil Price Trend Prediction Method: Machine Learning Classification Algorithm Based on Multi-Modal Data Fea-

tures. *Energy, 244,* Article 122706. https://doi.org/10.1016/j.energy.2021.122706

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

Jaiswal, J. K., & Samikannu, R. (2017). Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression. In *2017 World Congress on Computing and Communication Technologies (WCCCT)* (pp. 65-68). IEEE. https://doi.org/10.1109/wccct.2016.25

Jierula, A., Wang, S., Oh, T., & Wang, P. (2021). Study on Accuracy Metrics for Evaluating the Predictions of Damage Locations in Deep Piles Using Artificial Neural Networks with Acoustic Emission Data. *Applied Sciences, 11,* Article 2314. https://doi.org/10.3390/app11052314

Jo, J., Kim, U., Lee, E., Lee, J., & Kim, S. (2023). A Supply Chain-Oriented Model to Predict Crude Oil Import Prices in South Korea Based on the Hybrid Approach. *Sustainability, 15,* Article 16725. https://doi.org/10.3390/su152416725

Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., & Matsopoulos, G. K. (2023). A Review of ARIMA vs. Machine Learning Approaches for Time Series Forecasting in Data Driven Networks. *Future Internet, 15,* Article 255. https://doi.org/10.3390/fi15080255

Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.

Mohapatra, N., Shreya, K., & Chinmay, A. (2020). Optimization of the Random Forest Algorithm. In *Advances in Data Science and Management* (pp. 201-208). Springer. https://doi.org/10.1007/978-981-15-0077-0_21

Ning, Y., Kazemi, H., & Tahmasebi, P. (2022). A Comparative Machine Learning Study for Time Series Oil Production Forecasting: ARIMA, LSTM, and Prophet. *Computers & Geosciences, 164,* Article 105126. https://doi.org/10.1016/j.cageo.2022.105126

Oladosu, G., Leiby, P., Uria-Martinez, R., & Bowman, D. (2022). Sensitivity of the U.S. Economy to Oil Prices Controlling for Domestic Production and Imports. *Energy Economics, 115,* Article 106355. https://doi.org/10.1016/j.eneco.2022.106355

Pandian, S. (2024). *K-Fold Cross Validation Technique and its Essentials. Analytics Vidhya.* https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/

Safari, A., & Davallou, M. (2018). Oil Price Forecasting Using a Hybrid Model. *Energy, 148,* 49-58. https://doi.org/10.1016/j.energy.2018.01.007

Sayeed, M. A., Rahman, A., Rahman, A., & Rois, R. (2024). On the Interpretability of the SVM Model for Predicting Infant Mortality in Bangladesh. *Journal of Health, Population and Nutrition, 43,* Article No. 170. https://doi.org/10.1186/s41043-024-00646-9

Schölkopf, B., & Smola, A. J. (2018). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond*. MIT Press.

Sonkavde, G., Dharrao, D. S., Bongale, A. M., Deokate, S. T., Doreswamy, D., & Bhat, S. K. (2023). Forecasting Stock Market Prices Using Machine Learning and Deep Learning Models: A Systematic Review, Performance Analysis and Discussion of Implications. *International Journal of Financial Studies, 11,* Article 94. https://doi.org/10.3390/ijfs11030094

Tissaoui, L., El Mhamedi, A., & Benabdelhafid, A. (2022). A Prediction-Based Supply Chain Recovery Strategy under Disruption Risks. *International Journal of Production Research, 61*, 7670-7684. https://doi.org/10.1080/00207543.2022.2161022

U.S. Energy Information Administration (EIA) (2023). *Oil and Petroleum Products Ex-*

plained.
https://www.eia.gov/energyexplained/oil-and-petroleum-products/imports-and-exports.php

U.S. Energy Information Administration (EIA) (2024). *Short-Term Energy Outlook*. https://www.eia.gov/outlooks/steo/

Wang, P., Zhang, H., Qin, Z., & Zhang, G. (2017). A Novel Hybrid-Garch Model Based on ARIMA and SVM for PM 2.5 Concentrations Forecasting. *Atmospheric Pollution Research, 8,* 850-860. https://doi.org/10.1016/j.apr.2017.01.003

Yates, L. A., Aandahl, Z., Richards, S. A., & Brook, B. W. (2023). Cross Validation for Model Selection: A Review with Examples from Ecology. *Ecological Monographs, 93,* e1557. https://doi.org/10.1002/ecm.1557

Zhu, H. (2023). Oil Demand Forecasting in Importing and Exporting Countries: AI-Based Analysis of Endogenous and Exogenous Factors. *Sustainability, 15,* Article 13592. https://doi.org/10.3390/su151813592