Machine Learning Framework for Oil Production Forecasting

Doaa Reyadh Mudheher and Firas Sabar Miften
Department of Computer Science and Artificial Intelligence, College of Education for Pure Sciences, University of Thi-Qar, 64001 Nasiriyah, Iraq
doaa_reyadh@utq.edu.iq, firas@utq.edu.iq

Keywords: Volve Oil Field, SVR, RF, Preprocessing Data, Machine Learning.

Abstract: The prediction of oil well performance and decline rates is important for field management, because these factors can be directly used to model the planning of production as well as operational decision. Classical prediction tools suffer from several difficulties, particularly in terms of the complexity and variable nature of the data. This work proposes an Artificial Intelligence (AI) approach based on machine learning (ML) for predicting oil well production. A case study was based on data from the Volve oil field in Norway. Pre-processing was performed to ensure effective and robust modelling predictions. This included searching for outliers using the IQR method and replacing them with NaNs (missing values). A machine learning based method, Random Forest that learned the entire data to fill in these missing entries was adopted. each feature having missing values were temporarily considered as a target and predicted from the rest of features. Feature relevance was evaluated using the correlation coefficient of each dependent variable with the target (BORE_OIL_VOL). Scaling of all the numerical attributes was carried out and data set has been divided into train set (80%) and test sets (20%). The prediction performance of the forecasting method was improved using an ensemble model of Support Vector Regression (SVR) and Random Forest (RF). The experimental results indicated the new model improved the prediction ability of oil well production, since we achieved to RMSE = 37.7969, R2 = 0.9954 and MAE =17.5485. These metrics demonstrate the reliability of our method. The study shows that the ensemble learning models provide good prediction results of oil well productivity and can act as a promising decision-making approach for oil field management.

1 INTRODUCTION

Forecasting of oil field production rates is extremely critical for petroleum corporations and economists. The production volume is a size and velocity with which the well produces over some period. An accurate production forecast leads to efficient management of the resources, cost control and meets national as well as international energy requirements. Furthermore, accurate prediction of producing rate would be helpful in infill wells position selection and maintenance scheduling. Thus, the process has a strong impact on oil companies and macroeconomics [1].

To overcome these limitations, recent research has increasingly adopted Artificial intelligence and Machine Learning-based approaches, which are well-suited for capturing nonlinear patterns in production data and handling significant variability in operational parameters. Numerous studies have employed algorithms such as Support Vector Machines (SVM), Random Forest (RF), XGBoost, and LightGBM to enhance oil-production forecasting performance [2]. Other studies have developed hybrid ensemble-learning frameworks enhanced by meta-heuristic optimization methods to refine SVM performance and significantly reduce prediction errors [3].

Moreover, recent studies have emphasized the role of integrating reservoir engineering knowledge and operational expertise within machine-learning models to improve interpretability and predictive reliability [4]. Likewise, other studies showed that feature-selection and importance-ranking methods could help alleviate model uncertainty and improve predictive power by filtering out irrelevant attributes from the training process [5].

Although enormous progress has been achieved in the field of production forecasting, most previous studies have mainly concentrated on model building while having limited concern over the systematic assessment of data-preprocessing procedures (e.g. outlier removal, missing-value imputation, feature-correlation analysis and data standardization) that may significantly affect prediction performance. While recent research work has recognized the pivotal role of data quality in machine-learning pipelines, the practical focus on preprocessing

remains lower compared to its proven contribution in model stability and robustness [6], [7].

Production of oil well is one of the major key factors that can affect the efficiency and effectiveness of decision making in reservoir management. The conventional models fail at accuracy because of noise in data, missing values and complex nonlinear relationship. In this sense, the goal of this study is to find prediction performance improvements by combining updated preprocessing methods with ensemble learning models.

In this work, we present a new integrated machine learning framework for predicting oil well production with enhanced data preprocessing and ensemble learning models. Our preprocessing methodology consisting of outlier detection, missing value imputation and feature selection using balanced datasets before model training not only demonstrates better quality values but also predicts more robust results than traditional single models run on oilfield data with real values.

2 RELATED WORKS

In recent times, there has been significant advancement in machine-learning and deep-learning-based methods for prediction of oil-well performance - due to their capability of modeling nonlinear dependencies and complex correlations among the reservoir parameters and operational conditions.

Fang et al. (2024) developed a Random Forest model to obtain the inflow performance of tight-gas wells based-solution using real operation field data, having higher accuracy than conventional decline-curve methods [6]. Likewise, Ma et al. (2024) in a hybrid approach of classical decline analysis, artificial neural networks and choke-related data which provided offshore production forecasting accuracy [3].

To make the model more robust, AlRassas et al. (2023) templated a framework to integrate reservoir-engineering understanding and operational heuristics into machine learning models, making them more interpretable while decreasing prediction uncertainty [4]. Random Forest has been demonstrated as a competent forecasting algorithm for the prediction of shale-gas production, especially for managing sophisticated field datasets by Bhattacharyya and Vyas (2022) [8].

Also, Liu et al. (2024), which utilized XGBoost, Random Forest, and SVR interpretable machine learning models to predict the production based on parameters measured in horizontal plunger gas lift wells. SHAP was used to analyze feature importance and determine the most operational drivers affecting well production [9]

Additionally, Zhu et al. (2023) AutoGluon was used to automatically train and evaluate numerous models, automatically ranking their performance ensemble-based prediction [10].

Direct applications on the Volve field that have validated the usability of real operational data for advanced data-driven modeling too. The Volve Field Research Team (2025) further demonstrated excellent predictive accuracy ($R^2 \approx 0.99$) of the gradient-boosting algorithms (e.g., XGBoost, LightGBM) on Volve production data [5]. Bai et al. (2024) presented a data-mining method with 135 operational and reservoir measurements for analogue reservoirs identification, using such analogues as baseline for training predictive machine learning models to estimate the early production behaving well in data-poor environments [2].

In a recent paper, Wang et al. (2025) applied other machine learning techniques like RF, CNN, SVM and MLP-ANN to predict oil production rate under choke effect that is based on surface data observed in an Iraqi field. The method employed min-max normalization, leverage-based outlier detection and K-fold cross validation (K=5), in which the authors reported that Random Forest gave the best performance ($R^2 = 0.96127$) [1].

Likewise, Tang et al. (2024) looked at the decline curves of over thirty thousand Bakken shale oil wells, Eagle Ford and Permian basins. Model comparison shows that prediction performance is related to production history length: the single decline models are more accurate for short term prediction, and the combined method is superior in long term data. The work highlights the significance of data volume and quality that weighing late-time production history data would enhance prediction certainty [11]. Ghodsi et al. (2025) explored the oil production prediction for Soroush oil field (Iran) using ensemble learning and improved deep learning models. Comparing several ML and DL models (XGBoost, LSSVM and optimised DNNs), they showed a stacking ensemble model which exceeded performance of all individual models revealing the benefit from hybrid ensemble strategies for difficult oil production modelling [12].

Also, Hu et al. (2025) presented hybrid deep learning model (TCN-KAN) to predict oil well production using Volve dataset. High prediction accuracies were also maintained to the efficiency of combined temporal convolutional network and Kolmogorov Arnold Networks, which was in agreement with previous findings demonstrating that

hybrid deep learning models complemented one another to better resolve complex spatio-temporal generation patterns [13]. Zhao et al. (2024) employed a 2D-CNN model to describe the gas wells with geological and engineering cosmetic properties, stating that deep learning approaches can be good at capturing complex relationships between features [14].

Another related study performed by Ma et al. (2025) used artificial intelligence models to predict oil production from gas lifted wells in Iraq. The Random Forest model achieved the highest accuracy, with SHAP-based feature relevance pointing to water-sediment composition, choke diameter and upstream pressure as key parameters. [15]. Zhou et al. (2024) suggest an ensemble approach to DCA of shale gas reservoirs exploiting multiple DCA models with Random Forest meta-learner for better accuracy and stability as compared to conventional DCA approach [16]. Ding et al. 2025) presented the use of a Moth-Flame Optimization method combined with XGBoost model based on 442 completion details from real fields to forecast production decline rates of offshore high water-cut reservoirs. MFO-XGBoost provided better predictive accuracy (R^2 = 0.9128) than several baseline models [17].

Together, the reviewed works demonstrate a significant advancement in data-driven production forecasting particularly through ensemble and hybrid learning models. But earlier studies have typically preferred more complex models rather than careful examination of the contribution data quality elements like outliers and missing values handling, correlation-based attributes selection and normalization. This discrepancy highlights that preprocessing methods are not systematically evaluated. Therefore, in the current study we explore the effect of preprocessing on prediction accuracy by applying our methodology to real Volve field data which we hope will allow an effective and reliable workflow for high-accuracy oil production forecast.

3 MACHINE LEARNING METHODS BACKGROUND

Support Vector Regression (SVR) and Random Forest (RF) are the most commonly used machine-learning algorithms in engineering and for data-driven predictive tasks, due to their capability of modelling complicated nonlinear relationships. RF enhances robustness in predictions (by averaging over several decision trees used to grow it) that helps reducing the variance and achieving good generalization. On the other hand, SVR finds decision boundaries by maximizing the margins between points based on kernel functions, which can efficiently incorporate nonlinear patterns. Both methods have strong predictive properties, and they are particularly suitable for noisy, high dimensional and nonlinear data structures.

3.1 Support Vector Regression (SVR)

Support Vector Regression (SVR) is a learning methodology proposed by Vapnik et al. in the 1990s. This algorithm has been modified to overcome the deficiencies brought by complex linear indivisibility and "curse of dimensionality". The basic principle of the SVR can be overviewed by a kernel function, which can transform the sample data into an acceptable high-dimensional feature space with no-linear transformation from original data input and determine optimal hyperplane via mathematical deduction. The kernel functions act as vital parameters in SVR of different sensitivities for a wide range of data types. Kernel types commonly used are linear, polynomial, Sigmoid and radial ones [6], Figure 1 showing the schematic diagram of SVR.

3.2 Random Forest (RF)

Random Forest (RF) is a widely used and efficient algorithms for classification and regression. It was proposed by Leo Berryman and uses compositing various models to obtain the most precise result.

The fundamental idea behind a Random Forest is to grow multiple Decision Trees. Trees are built at random on the base of training and feature. Training each tree with different data increases diversity, which makes the model better at generalizing. Random forest After we train the random forest every single tree will form decisions independently and make it declare its decision when you give an unseen example to the models. The final results of final classifications are made using major votes between the trees and for regressions we compute their average. This mechanism enables that, even if some trees do wrong choices, their errors when accumulated to the final result have a very small impact on the overall outcome as the random error of each tree is mitigated in such an assembly.

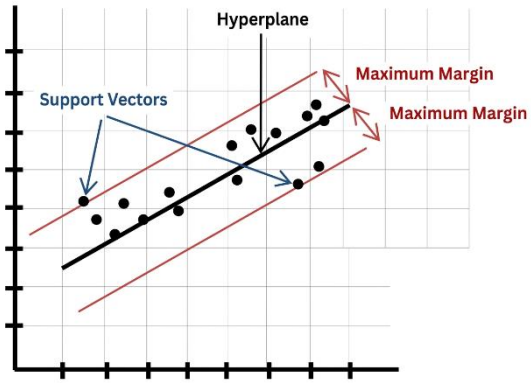

Figure 1: Schematic diagram of SVR.

The obvious superiority of Random Forest, that it can prevent overfitting. It is an inherent way random features and different samplings help model not to over-attend unnecessary details of the input data, so that their generalization ability performs better when coming across new instance Moreover, the ability of random forests to deal with problems that have many features also makes it a very good adversarial tool for many hard complex problems. This technique can also predict the feature contribution to the final decision, giving us several important and non-important data information [1]. Figure 2 showing the Schematic representation of RF.

4 METHODOLOGY

In this section, we first introduce the dataset and then briefly summarize the method. The process consists of data cleaning and preprocessing, feature selection and normalization, and building a hybrid machine-learning model, The detailed steps are presented below:

4.1 Data Description

The Volve oil field dataset, a real production database that has been publicly released by Equinor in 2018 for research and development, is applied in this work. The available dataset consists of well log data from several wells, which can be a reliable platform for analysis and modelling.

These features include (on-stream hour, average downhole pressure & temperature, average differential pressure tubing & annulus pressure, average Choke-size percentage, average wellhead pressure and temperature, choke-size, produced oil, gas & water volume).

The output variable was the rate of oil production (BORE_OIL_VOL), and the other numeric features were input predictors.

The dataset was divided into 80% of samples for training and the next 20% of samples for testing to evaluate model generalization performance.

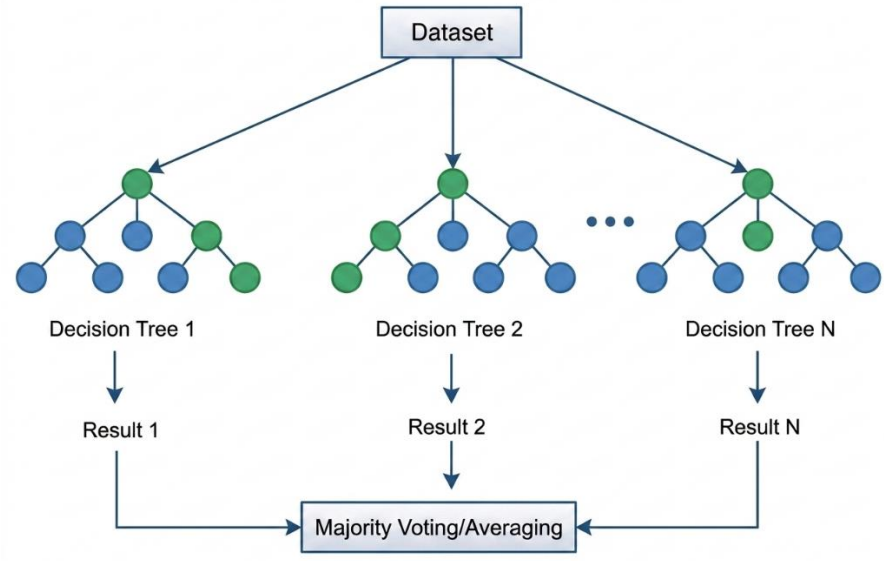

Figure 2: Schematic diagram of RF.

4.2 Data Preprocessing

To ensure that high quality input data used for our model, pre-processing consists of a number of key steps ranging from outlier detection, imputation for missing values, selection of relevant features and numerical attributes standardization. Each step is explained in the following subsections:

4.2.1 Detection and Treatment of outlier

Outliers were detected using the IQR (Interquartile Range) method by identifying an out of range (Q1−1.5×IQR) and (Q3+1.5×IQR). These outliers were replaced with NaN so that it does not affect the quality of the model training.

4.2.2 Missing-Value Imputation (Random Forest Imputation)

A feature-wise Random Forest regression imputation method was used for imputing missing values created due to outlier removal. For each feature column, we treated it as a target and then used the other columns to predict its missing values. This retains complex nonlinearities between features.

4.2.3 Feature Selection

Feature selection performed using the Pearson correlation coefficient measure which is utilized to quantify and isolate the strength of each feature with respect to the target (BORE_OIL_VOL). A correlation cut-off of |r|≥0.10 was set as criteria to keep only the features showing a significant linear contribution. The linearity and weights of the features in the linear combination were assessed by calculating the Pearson correlation coefficients. The values of correlation range from -1 to 1; a value of 1 implies there is perfect positive linear relationship and −1 implies there is negative positive linear relationship. This can be explained using the following equation:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}. \quad (1)$$

Consequently, eight features were selected: (ON_STREAM_HRS,AVG_DOWNHOLE_PRESSURE, AVG_WHP_P, AVG_DOWNHOLE_TEMPERATURE, AVG_WHT_P,DP_CHOKE_SIZE, BORE_GAS_VOL, BORE_WAT_VOL).

On the other hand, three of the features had correlation values that were less than our selected threshold (|r| < 0.10) which led us to remove them: (AVG_DP_TUBING,AVG_ANNULUS_PRESS, AVG_CHOKE_SIZE_P).

Moreover, the target variable has not been included in the list of predictor variables since it is not used as an input attribute. This effectively reduced noise in the dataset and improved the quality of input features for model training, and thus significantly improved predictive performance.

Table 1 presents all features, their correlation coefficients with target feature and whether to select or reject these features in internal expression that we proposed in this study.

Table 1: Correlation coefficients.

Feature	With target	Status	Reason
AVG_DOWNHOLE_PRESSURE	0.214	Selected	Numeric, |r| >= threshold
AVG_DOWNHOLETEMPERATURE	0.235	Selected	Numeric, |r| >= threshold
AVG_WHP_P	0.531	Selected	Numeric, |r| >= threshold
AVG_WHT_P	0.469	Selected	Numeric, |r| >= threshold
BORE_GAS_VOL	0.995	Selected	Numeric, |r| >= threshold
BORE_WAT_VOL	0.161	Selected	Numeric, |r| >= threshold
DP_CHOKESIZE	0.464	Selected	Numeric, |r| >= threshold
ON_STREAMHRS	-0.239	Selected	Numeric, |r| >= threshold
AVG_ANNULUS	-0.084	Excluded	Numeric, |r| < threshold
AVG_CHOKE	0.076	Excluded	Numeric, |r| < threshold
AVG_DPTUBING	0.066	Excluded	Numeric, |r| < threshold
BORE_OIL_VOL	1	Excluded	Target variable (not used as predictor)

4.2.4 Feature Standardization

All numerical features (excluding the target variable) standardized by Z score scaling:

$$z = \frac{x_i - \mu}{\sigma}, \quad (10)$$

μ and σ denote the feature mean and standard deviation. This normalization ensures stable training for kernels and tree-based models.

4.2.5 Train/Test Split

The dataset was split using a Hold-Out strategy with 80 % of samples for training set and 20 % for testing set to check generalization performance of the model.

4.3 Hybrid Machine Learning Framework

In the current work, this stage consists of the following processes: training individual models and then creating a hybrid model on the basis of their outputs blended with proper weights. The purpose of this approach is to take advantage of the merits of both models, and hope that it could result in more robust and accurate predictive performances by training on these multiple views.

4.3.1 Training the Base Models (SVR and RF)

The Support Vector Regression (SVR) with RBF kernel was the first model trained as it is known to predict nonlinear relationships prevalent in oil-well production data well. The hyper parameters of an SVR model, including kernel scale, regularization parameter (C) and ε-insensitive margin, were optimized to find a good trade-off between the flexibility of the model and how its errors are penalized.

A Random Forest regression model with bootstrap aggregation (Bagging) was also fit. Appropriate number of trees and optimal splitting rules (maximum tree depth, minimum of samples in a leaf and some feature samples available to be considered for next split) were identified.

4.3.2 Creating Out-of-Fold Predictions to Get the Best Weight

As a divisor of the training set, we used 5-fold cross-validation setting, the contribution of each model in the final blended output was found empirically. Both models were retrained and their predictions were made on the validation set held out from training for each fold. These predictions are called Out-of-Fold (OOF) predictions for both SVR and RF.

The OOF predictions give an unbiased estimate of how each model will perform in unseen data so they are a good place to start to determine the actual weight that we should give for each model.

As a result of this analysis, an optimum weight RF was learned for the Random Forest model (striding the relative superiority) and a balancing weight (1−wRF) to assign it to the SVR model. This weight was clipped to the range [0, 1] in order to keep numerical stability and avoid from heavily favouring one of the models.

$$w^* = arg\ min_w \sum_{i=1}^{N} \left(y_i - \left(w\, \hat{y}_{RF,i} + (1-w)\hat{y}_{SVR,i} \right) \right)^2. \quad (3)$$

4.3.3 Weighted Blending and Construction of the Hybrid Model

After computing the optimal weight, the prediction of a hybrid model would be evaluated by their linear combination on the outputs of two base learners as follows:

$$Hybrid\ Prediction = W_{\{RF\}} * RF\ Prediction + \left(1 - W_{\{RF\}}\right) * SVR\ Prediction. \quad (4)$$

This weighted average applies a relatively higher weight to the model that performs better under cross-validation. Accordingly:

1) When the Random Forest model gives better performance, WRF increases;
2) In the case where the SVR model is more accurate, then weight reduces resulting in stronger presence of SVR term;
3) When both models actually work well, blending naturally weighs the two of them.

This approach reduce the single-model bias and successfully combines the nonlinear learning of SVR with the robustness properties of Random Forest, compared to using a single model only.

4.4 Performance Evaluation

Model performance on both training and testing sets was using:

1) Root Mean Squared Error (RMSE). Root mean squared error (RMSE), that is the average magnitude of prediction errors, is a widely used error metric. The smaller of RMSE is the more

perfect it predicts.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}. \quad (5)$$

2) Mean Absolute Error (MAE). Is a performance measure that calculates the average of the absolute difference between actual values and forecasts. It indicates the typical size of prediction errors. Smaller MAE means more accurate prediction.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|. \quad (6)$$

3) Coefficient of Determination (R^2). Is a regression metric that compares how good the fitted line describes the actual values. A higher R^2 means that the model explains more variance and better predicts.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}. \quad (7)$$

The general structure of the method proposed is shown in Figure 3. The Volve dataset is preprocessed by performing the following steps: outlier removal, missing value imputation using Random Forest, correlation analysis, feature selection and z-score standardization. Next step is to split the data into 80% training set and 20% testing set.

Then the SVR and Random Forest models are trained and they will be combined through a weighted blending approach. R^2, RMSE and MAE metrics are used to evaluate the final model performance.

5 RESULT ANALYSIS AND DISCUSSION

Data preprocessing and the model's impact is assessed by analyzing the dataset before and after applying our proposed framework. There were outliers and missing values in the raw data which could hamper the stability of models. Outlier removal based on IQR, Random Forest imputation and feature scaling improved data quality and reduced noise. All the metric performance for RMSE, MAE and R^2 was also better than in raw data, therefore it became evident that this preprocessing strategy is able to improve model performance as well.

It has also been observed that performing variable transformations (e.g., scaling and feature selection) can potentially impact the integration of methods by altering data distributions or reducing correlations between features. Thus, the proposed framework adopted the easiest settings that offered the trade-off of better data quality and model performance.

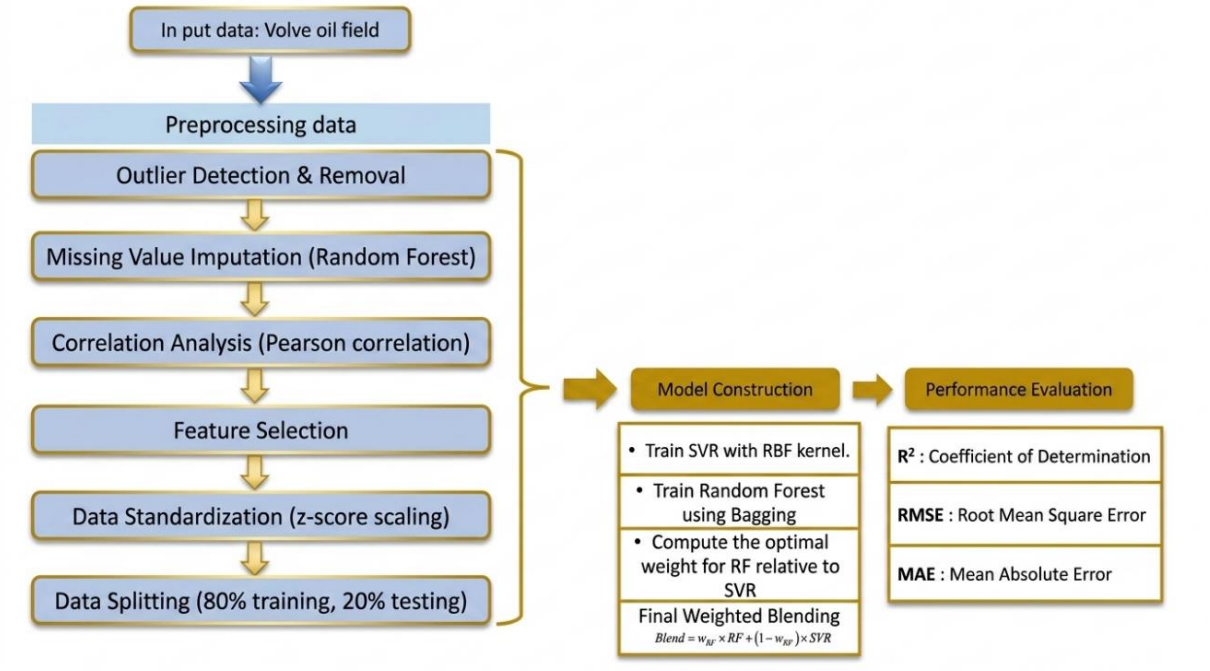

Figure 3: Methodological for hybrid machine learning oil production forecasting (pre-processing, modelling).

The performance of the proposed RF-SVR hybrid approach was well examined in different point of views by means of different statistical parameters including R^2, RMSE, MAE. Results indicated that the joint effects of RF and SVR were effective for enhancing prediction performance than those of the single models.

As showing in Table 2, The hybrid model was trained to form relation between input variables and oil production rates with high accuracy ($R^2 > 0.99$) therefore the model was able to capture complex nonlinear correlation among them. Upon validation with an independent data set, the model performed well and obtained R^2 of close to 0.995, RMSE below 40 and relative error around 5%. Such a stability between training and validating phases suggests good generalization potential and resistance.

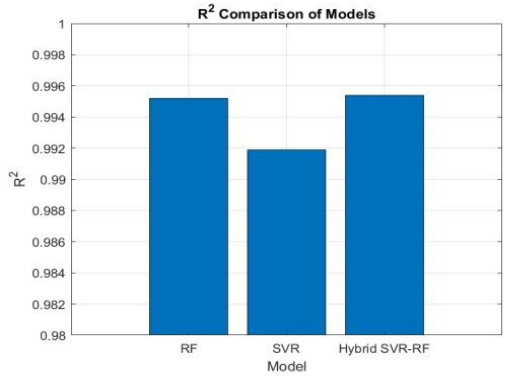

Figure 4: R^2 Comparison of evaluated models.

Table 2: Performance comparison of machine learning models.

Model	Performance metrics		
	R2	RMSE	MAE
RF	0.9952	38.8773	18.3445
SVR	0.9919	50.4240	21.1514
Hybrid SVR-RF Model	0.9954	37.7969	17.5485

The R^2 comparisons for each model are displayed in Figure 4.

To visually analyze the prediction accuracy of the proposed Hybrid SVR-RF model, Figure 5 displays the Actual vs Predicted plot.

Furthermore, due to their complementary nature with one another between the discussed ensemble average (for RF) and kernel-based learning algorithms (for SVR), when both of them were employed together as superior ensembling blending mechanism by utilizing various features which in turn reduced prediction variance and improved consistency. The results show that the hybrid model has great potential for being a reliable decision-making support in prediction of oil well production under various operational conditions.

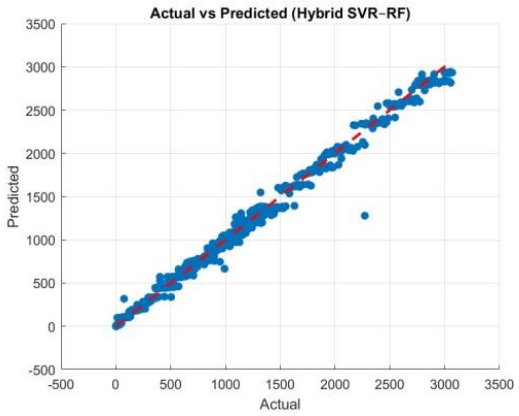

Figure 5: Actual vs. Predicted values for the hybrid model.

The performance of Model developed in this paper demonstrates the ability of the RF-SVR hybrid model to effectively capture complex nonlinear relationships between operational factors and reservoir characteristics in oil production processes. Random Forest (RF) and Support Vector Regression (SVR) performed well in the hybrid approach, they reach an optimal balance between stability and accuracy.

RF making stable against noise and overfitting by ensemble averaging while SVR is to improve the model for capturing small nonlinear pattern that be missed in standard linear models even it is fairly complex.

The values of R^2 and RMSE, MAE are obtained for to demonstrate the prediction performance through combination process blending strategy which increased the predicting capability as compared to individual base models. This is a clear demonstration of how well hybrid learning methodologies work in predicting the wells, given that data complexity and operational variability is one of the bottlenecks. Quality input data enabled better generalization and less error propagation during modelling.

Related studies have also used Random Forest (RF) model for production prediction, but as seen in Table 3, the proposed RF reached the highest R^2 value (0.9952), higher than Fang et al.[6] (0.91), Ma et al.[15] (0.867), and Wang et al. [1] (0.9613). The gain is due to an improved data processing framework that improves both accuracy and predictive stability of the model.

Table 3: Performance comparison of Random Forest (RF) model with previous studies.

Study	Model	R^2
Fang et al. [6]	RF	0.91
Ma et al.[15]	RF	0.867
Wang et al.[1]	RF	0.9613
Proposed RF	RF	0.9952

As shown as in Table 4, by using the same data set (Volve oil field), The proposed Hybrid SVR-RF model achieved the best coefficient of determination as observed in Table 4 with an $R^2 = 0.9954$ which was better than the individual models along with the deep hybrid model mentioned in comparative study [13]. This increase is a result of the potential benefits of applying advanced preprocessing strategies such as outlier removal, imputation of missing values, correlation analysis and feature selection in conjunction with hybrid ensemble learning techniques that could potentially improve.

Table 4: Performance comparison with same dataset.

Study	Model	R^2
Hu et al., 2025 [13]	WOA-TCN-KAN	0.985
Proposed RF	RF	0.9952
Proposed SVR	SVR	0.9919
Proposed Hybrid	SVR-RF	0.9954

In general, the proposed RF-SVR hybrid model shows high potential as a dependable decision-support tool for oil well production forecasting to improve operational planning and production management in petroleum engineering domain.

6 CONCLUSIONS

This research has proposed a new hybrid model that integrates SVR and RF to predict the production of oil well with the real-field data from Volve dataset. The resultant framework successfully combines the nonlinear modelling capacity of SVR with the ensemble robustness of RF, yielding a predictive system that is not only accurate but also stable.

This work is most valuable in highlighting the need to perform careful data pre-processing to improve model performance. Before modelling a pre-processing was performed on all fields such as outlier detection and removal, missing value imputation, correlation analysis and standardization. In this way, these pre-processing steps helped to reduce noise and bias in the data significantly, improve the quality of features used to train models as well as make sure that learning algorithms would learn off from tidy consistent input. Evidence from the work show that careful pre-processing should not be seen as a preparatory measure but an important step that governs model reliability, interpretability and generalization performance.

The Hybrid SVR-RF model showed better predictive power through the higher R^2 values and lower RMSE and MAE errors. Furthermore, it confirms the good performance of weighted hybridization between kernel and ensemble based learning techniques to model nonlinear character and multi feature on oil production dataset. Moreover, the stability of the model to both training and testing datasets further proves its robust characteristics making it applicable for field scale prediction.

The present study thus provides a strong correlation between AI models and domain knowledge from an engineering prospective, which can be harnessed to develop a well informed decision supporting system for improving production optimization and field management.

In the practical significance of oil field operation management, enhance production prediction precision and stability provides a guide for operational plans made by an organization, resource allocation and maintenance scheduling. In practice, improving oil field production forecast accuracy reduces uncertainty and potentially improves decision-making ability.

7 LIMITATIONS AND FUTURE WORKS

Although the proposed model showed a good performance on oil well production prediction, there are some limitations. The training dataset limited to this particular dataset means that a very narrow range of operational features are used, and unless the data used capture all external factors influencing actual production, then they are poorly representative. While these variables offered acceptable measures of model performance, new or finer time data would enhance the models.

In a future study, data from multiple different oil datatypes and some other different operational environment will be tested on this proposed framework to validate it additionally supported with additional tests of generalizability robustness. Moreover, analysis and evaluation of more advanced modelling techniques like intricate deep learning

architectures and ensemble methods can be performed to better distribute complex nonlinear production patterns.

Furthermore, future studies should relate the predictive performance to operational and economic key figures in addition to mere statistical error measures. Integrating operational data with the suggested framework can enhance its practical relevance and lend itself to better decisions for oil field development and production planning processes.

REFERENCES

[1] Q. Wang et al., "Precise prediction of choke oil rate in critical flow condition via surface data," J. Pet. Explor. Prod. Technol., vol. 15, no. 7, pp. 1-18, Jul. 2025, [Online]. Available: https://doi.org/10.1007/s13202-025-02013-8.

[2] W. P. Bai, S. Q. Cheng, X. Y. Guo, Y. Wang, Q. Guo and C. D. Tan, "Oilfield analogy and productivity prediction based on machine learning: Field cases in PL oilfield, China," Pet. Sci., vol. 21, no. 4, pp. 2554-2570, Aug. 2024, [Online]. Available: https://doi.org/10.1016/j.petsci.2024.02.018.

[3] K. Ma, C. Wu, Y. Huang, P. Mu and P. Shi, "Oil well productivity capacity prediction based on support vector machine optimized by improved whale algorithm," J. Pet. Explor. Prod. Technol., vol. 14, no. 12, pp. 3251-3260, Dec. 2024, [Online]. Available: https://doi.org/10.1007/s13202-024-01873-w.

[4] M. AlRassas et al., "Knowledge-based machine learning approaches to predict oil production rate in the oil reservoir," in Springer Series in Geomechanics, Springer, 2024, pp. 282-304, [Online]. Available: https://doi.org/10.1007/978-981-97-0268-8_24.

[5] M. Khan, A. Sakib, A. Samad, M. Shakil Rahaman and M. A. Islam, "Data-driven approach to predict future oil production of an oil field using machine learning techniques," SciEn Publ. Gr., pp. 68-73, Jan. 2025, [Online]. Available: https://doi.org/10.38032/scse.2025.3.16.

[6] M. Fang, H. Shi, H. Li and T. Liu, "Application of machine learning for productivity prediction in tight gas reservoirs," Energies, vol. 17, no. 8, pp. 1-27, Apr. 2024, [Online]. Available: https://doi.org/10.3390/en17081916.

[7] M. Pang, Z. Zhang, Z. Zhou, W. Zhou and Q. Li, "Estimated ultimate recovery prediction of shale gas wells based on stacked integrated learning algorithm," Sci. Rep., vol. 15, no. 1, pp. 1-19, Dec. 2025, [Online]. Available: https://doi.org/10.1038/s41598-025-03801-2.

[8] S. Bhattacharyya and A. Vyas, "Application of machine learning in predicting oil rate decline for Bakken shale oil wells," Sci. Rep., vol. 12, no. 1, Dec. 2022, [Online]. Available: https://doi.org/10.1038/s41598-022-20401-6.

[9] J. Liu et al., "Production prediction and influencing factors analysis of horizontal well plunger gas lift based on interpretable machine learning," Processes, vol. 12, pp. 1-25, 2024, [Online]. Available: https://doi.org/10.3390/pr12091888.

[10] R. Zhu et al., "Well-production forecasting using machine learning with feature selection and automatic hyperparameter optimization," Energies, vol. 18, no. 1, Jan. 2025, [Online]. Available: https://doi.org/10.3390/en18010099.

[11] H. Y. Tang et al., "Production decline curve analysis of shale oil wells: a case study of Bakken, Eagle Ford and Permian," Pet. Sci., vol. 21, no. 6, pp. 4262-4277, Dec. 2024, [Online]. Available: https://doi.org/10.1016/j.petsci.2024.07.029.

[12] M. Ghodsi, P. Vaziri, M. Kanaani and B. Sedaee, "Optimizing oil production forecasts in Iranian oil fields: a comprehensive analysis using ensemble learning techniques," J. Pet. Explor. Prod. Technol., 2025, [Online]. Available: https://doi.org/10.1007/s13202-025-01976-y.

[13] Y. Hu, X. Xin, G. Yu and W. Deng, "Deep insight: an efficient hybrid model for oil well production forecasting using spatio-temporal convolutional networks and Kolmogorov-Arnold networks," Sci. Rep., 2025, [Online]. Available: https://doi.org/10.21203/rs.3.rs-4988083/v1.

[14] N. M. Ibrahim, A. A. Alharbi, T. A. Alzahrani, A. M. Abdulkarim, I. A. Alessa and D. A. Alqahtani, "Well performance classification and prediction: deep learning and machine learning long term regression experiments on oil, gas, and water production," Sensors, vol. 22, no. 14, p. 5326, 2022, https://doi.org/10.3390/s22145326.

[15] F. Ma et al., "Predictive modeling of oil rate for wells under gas lift using machine learning," Sci. Rep., vol. 15, no. 1, pp. 1-26, Dec. 2025, [Online]. Available: https://doi.org/10.1038/s41598-025-12129-w.

[16] Y. Zhou, Z. Gu, C. He, J. Yang and J. Xiong, "An improved decline curve analysis method via ensemble," Energies, vol. 17, 2024, [Online]. Available: https://doi.org/10.3390/en17235910.

[17] Z. Ding et al., "Production decline rate prediction for offshore high water-cut reservoirs by integrating moth-flame optimization with extreme gradient boosting tree," Processes, vol. 13, pp. 1-17, 2025, [Online]. Available: https://doi.org/10.3390/pr13072266.