automation

Article

Integrating Machine Learning with Intelligent Control Systems for Flow Rate Forecasting in Oil Well Operations

Bibars Amangeldy [1,2], Nurdaulet Tasmurzayev [1,2,*], Shona Shinassylov [1,2], Aksultan Mukhanbet [1,2] and Yedil Nurakhov [1,3]

[1] Joldasbekov Institute of Mechanics and Engineering, Almaty 050010, Kazakhstan; a.s.bibars@gmail.com (B.A.); shona.shinassylov.87@gmail.com (S.S.); mukhanbetaksultan0414@gmail.com (A.M.); y.nurakhov@gmail.com (Y.N.)
[2] Faculty of Information Technologies, Al-Farabi Kazakh National University, Almaty 050010, Kazakhstan
[3] Department of Information Technology, Astana IT University, Astana 050040, Kazakhstan
[*] Correspondence: tasmurzayev.n@gmail.com

Abstract: This study addresses the integration of machine learning (ML) with supervisory control and data acquisition (SCADA) systems to enhance predictive maintenance and operational efficiency in oil well monitoring. We investigated the applicability of advanced ML models, including Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and Momentum LSTM (MLSTM), on a dataset of 21,644 operational records. These models were trained to predict a critical operational parameter, FlowRate, which is essential for operational integrity and efficiency. Our results demonstrate substantial improvements in predictive accuracy: the LSTM model achieved an R^2 score of 0.9720, the BiLSTM model reached 0.9725, and the MLSTM model topped at 0.9726, all with exceptionally low Mean Absolute Errors (MAEs) around 0.0090 for LSTM and 0.0089 for BiLSTM and MLSTM. These high R^2 values indicate that our models can explain over 97% of the variance in the dataset, reflecting significant predictive accuracy. Such performance underscores the potential of integrating ML with SCADA systems for real-time applications in the oil and gas industry. This study quantifies ML's integration benefits and sets the stage for further advancements in autonomous well-monitoring systems.

Keywords: machine learning; SCADA systems; oil well monitoring; LSTM; bidirectional LSTM; GRU; transformers; predictive maintenance; flow rate prediction; oil and gas industry; operational efficiency; time series forecasting

check for updates

Citation: Amangeldy, B.; Tasmurzayev, N.; Shinassylov, S.; Mukhanbet, A.; Nurakhov, Y. Integrating Machine Learning with Intelligent Control Systems for Flow Rate Forecasting in Oil Well Operations. *Automation* **2024**, *5*, 343–359. https://doi.org/10.3390/automation5030021

Academic Editors: Luis Gomes, Luís Brito Palma, Bruno J. N. Guerreiro and Anikó Costa

Received: 17 June 2024
Revised: 19 July 2024
Accepted: 25 July 2024
Published: 1 August 2024

1. Introduction

The oil and gas industry continuously pursues enhanced efficiency, cost reduction, and improved safety in its operations. A crucial aspect of achieving these goals lies in optimizing oil well monitoring and production forecasting. These elements are essential for efficient reservoir management, enabling timely intervention and maximizing hydrocarbon recovery [1,2]. Traditionally, the industry relied on decline curve analysis and reservoir simulation models to forecast production [3]. However, these methods often fall short, being unable to accurately represent the complex, nonlinear, and dynamic nature inherent in oil well production [4,5]. This challenge is further amplified in unconventional reservoirs with their intricate flow channels and fluid phase behavior [6]. Additionally, frequent manual operations can introduce further complexities that traditional methods struggle to accommodate [7].

The convergence of advancements in sensor technology, data analytics, and artificial intelligence (AI) has opened up exciting new possibilities for addressing these challenges [8,9]. The widespread adoption of supervisory control and data acquisition (SCADA) systems enables the continuous monitoring of crucial well parameters [10–14]. This provides a constant stream of data, which is ideal for in-depth analysis and data-driven decision-making [15]. Simultaneously, the evolution of machine learning (ML) algorithms has provided the industry with powerful tools capable of deciphering complex patterns and relationships within these large, multifaceted datasets [16,17]. This convergence has sparked a surge in research exploring the application of ML to optimize numerous facets of the oil and gas sector [17,18].

Early research recognized the potential of ML for optimizing various aspects of oil and gas operations. Studies explored the use of ANNs for predicting experimental parameters related to friction and wear, providing valuable insights for forensic preparation [19]. Machine learning techniques such as discriminant analysis, Bayesian modeling, and transfer learning were also employed for predicting experimental parameters and classifying well performance [20].

Accurately predicting oil and gas production is paramount for efficient reservoir management and optimizing recovery. Recognizing the limitations of traditional approaches [3,6,7], researchers began exploring ML techniques for more accurate and reliable forecasts. Support Vector Machines (SVMs), known for their ability to handle high-dimensional data and capture complex relationships, emerged as a promising tool for predicting oil production [21]. Researchers successfully employed SVMs to forecast oil production rates, demonstrating their effectiveness in scenarios with sparse data or complex reservoir characteristics [22].

Artificial Neural Networks (ANNs) also gained significant traction due to their ability to learn complex, nonlinear relationships from data, making them well-suited for modeling the dynamic behavior of oil reservoirs [9,23]. Studies demonstrated the successful application of ANNs for predicting well performance, forecasting production rates, and estimating parameters such as lubrication film thickness and lubrication regimes [19,24]. However, the limitations of conventional ANNs in handling temporal dependencies inherent in time-series data became apparent, prompting the exploration of more advanced architectures such as Recurrent Neural Networks (RNNs), particularly the Long Short-Term Memory (LSTM) architecture [2,25]. LSTM networks, designed to excel in capturing long-term dependencies within sequential data, proved ideal for modeling the temporal evolution of oil well production [25–27].

Numerous studies demonstrated the superior performance of LSTM networks in predicting oil and gas production rates compared to traditional methods and other ML algorithms [24,28]. Researchers explored LSTM networks for predicting oil, gas, and water production rates, demonstrating their effectiveness in handling the complexities of multiphase flow [3]. The application of LSTM networks extended to forecasting production in unconventional reservoirs, where traditional methods struggled to capture the unique production dynamics [6,23].

Moreover, researchers recognized the potential of RNNs, specifically LSTM networks, in equipment management, particularly for predictive maintenance. Studies demonstrated the successful application of LSTM algorithms for fault prognosis models, specifically in predicting the maintenance needs of critical rotating equipment such as air booster compressor motors [29]. These models leverage sensor data to identify patterns indicative of impending failures, enabling proactive maintenance and minimizing costly downtime, a crucial aspect for optimizing operations and reducing costs in the oil and gas industry.

To further enhance the performance of LSTM networks, researchers experimented with incorporating attention mechanisms. These mechanisms allowed the models to focus on specific time steps or features within the input sequence, leading to the development of hybrid models such as LSTM–Attention, which further improved prediction accuracy [26,28].

The availability of larger datasets and increased computational power led to the exploration of deeper and more complex LSTM architectures [23,27]. Researchers found that stacking multiple LSTM layers could improve the model's ability to learn intricate temporal patterns within the data [23,30]. The combination of LSTMs with other neural network architectures, such as Multilayer Perceptrons (MLPs), resulted in hybrid models including GRU-MLP, further enhancing the accuracy of production forecasting in shale gas reservoirs [31].

Beyond production forecasting, ML found significant application in anomaly detection, which is crucial for maintaining operational integrity and ensuring safety [32]. Researchers developed ML-based models to analyze network traffic in oil and gas well-monitoring systems, detecting anomalies that could indicate potential security threats or equipment failures [11,15]. The use of ML for anomaly detection extended to monitoring pipeline integrity [33]. By analyzing sensor data, ML models could identify subtle deviations from normal operating conditions, indicating potential leaks or other anomalies requiring attention.

Furthermore, the increasing focus on AI-driven predictive maintenance signifies a paradigm shift in equipment management within the oil and gas industry. This approach leverages machine learning algorithms to analyze equipment data, predicting maintenance needs before breakdowns occur [34]. By proactively addressing potential issues, oil and gas companies can minimize downtime, reduce maintenance costs, and enhance overall equipment reliability and safety, leading to substantial cost savings and improved operational efficiency.

In recent times, transformer architectures have been explored, representing a burgeoning area of research in oil and gas production forecasting. Transformers, originally designed for natural language processing tasks, have demonstrated impressive capabilities in handling long-range dependencies in sequential data, surpassing even the performance of LSTM models in certain scenarios [35].

While the integration of ML into oil well monitoring and production forecasting has achieved promising results, several challenges persist. Access to large, high-quality datasets is paramount for training robust and reliable ML models [17,21]. The scarcity of publicly available oil well data, often closely guarded by companies as proprietary information, presents a significant hurdle to research and model development [8,21,35]. Additionally, the computational demands of training and deploying sophisticated ML models, particularly deep learning architectures such as LSTM networks, can be substantial [3,7,36]. This can pose challenges for their practical implementation, especially in resource-constrained environments.

This research addresses these challenges by developing a robust system for accurate flow rate forecasting in oil well operations. We propose a framework that harnesses the predictive power of LSTM, BiLSTM, and MLSTM models while integrating the real-time monitoring capabilities of SCADA systems. Leveraging a publicly available oil well dataset [37], we demonstrate the efficacy of these models in capturing complex temporal patterns and achieving high prediction accuracy. The integration with SCADA systems facilitates real-time decision-making, empowering operators to proactively adjust operational parameters and optimize well performance. Furthermore, we evaluate the feasibility and benefits of deploying these models within a real-world oil well monitoring system, paving the way for a more efficient and reliable approach to production optimization and predictive maintenance.

2. Materials and Methods

In this section, we describe the comprehensive ML workflow implemented in our study for forecasting flow rates in oil well operations. As illustrated in Figure 1, the process begins with data collection from GitHub, securing a valuable dataset for analysis. Following this, the data undergo preprocessing and analysis, where Unix timestamps are converted to a readable date–time format, and essential data visualizations are created to understand the underlying patterns.

Subsequently, the core of the workflow involves modeling and forecasting, where various ML models, such as LSTM, BiLSTM, MLSTM, GRU, and Transformers are built and trained. These models are then validated on test datasets to ensure their effectiveness. Finally, the models are rigorously evaluated using metrics such as R^2, MAE, and MSE to assess their performance. This evaluation helps in comparing different models and selecting the best one for accurate flow rate prediction. The detailed steps and methodologies of this workflow are discussed in the following sections, starting with the dataset description and preprocessing techniques.

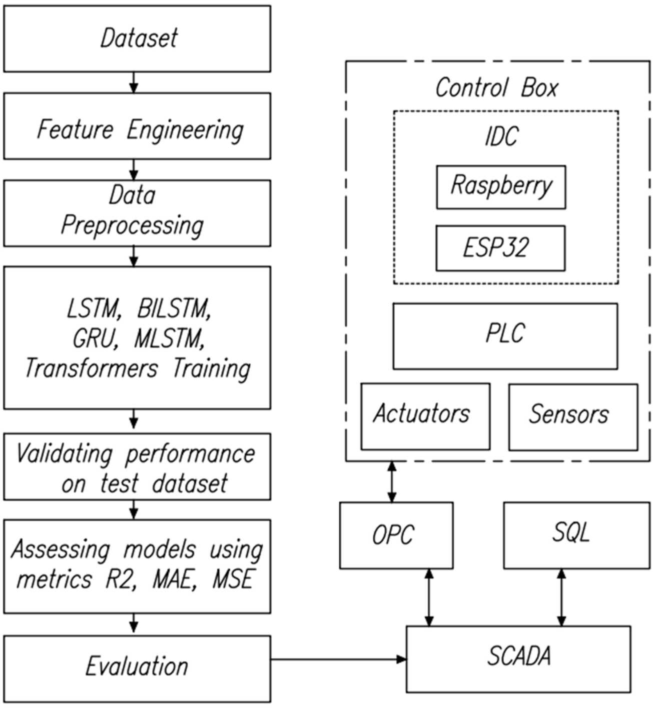

Dataset
Feature Engineering
Data Preprocessing
LSTM, BILSTM, GRU, MLSTM, Transformers Training
Validating performance on test dataset
Assessing models using metrics R2, MAE, MSE
Evaluation
Control Box
IDC
Raspberry
ESP32
PLC
Actuators
Sensors
OPC
SQL
SCADA

Figure 1. Workflow of the system.

In this architecture, the SCADA (supervisory control and data acquisition) system is the central framework for process supervisory management and operational control. It allows for real-time monitoring and the adjustment of operational parameters from a centralized location and interfaces with the control box to implement control commands.

The control box features a PLC (Programmable Logic Controller) that handles real-time control tasks based on programmed operational scenarios, ensuring rapid response to changes. The PCB SMD Device enhances the interface and processing capabilities, supporting the PLC with faster and more reliable data processing.

Integrated into the architecture for operational reporting and data storage, the SQL database stores historical data for trend analysis, performance assessments, and operational auditing, along with configuration data that can be utilized by the SCADA system to dynamically adjust control strategies.

Sensor data are primarily used for condition monitoring and sent to the SCADA system for operational visibility and immediate decision-making. Control commands from the SCADA system are transmitted to the control box, where they are executed. Non-critical operational data for long-term storage are forwarded to the SQL database for analysis and reporting.

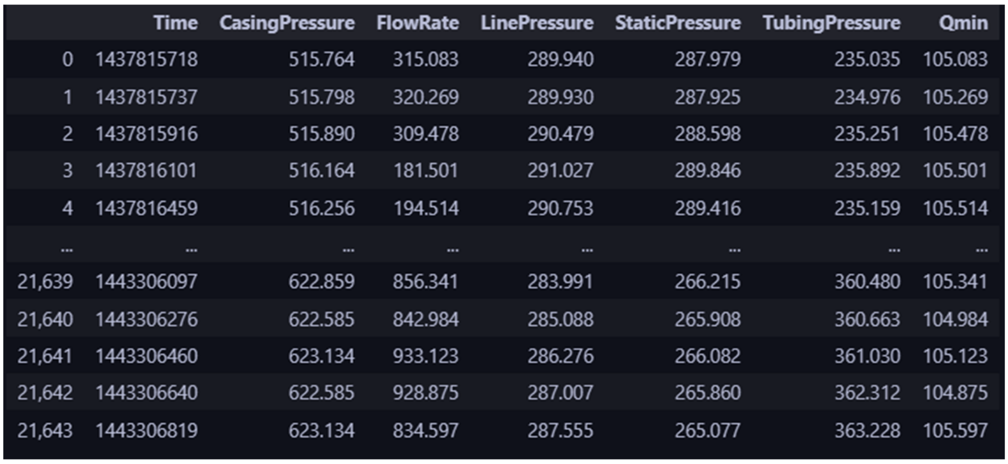
2.1. Dataset Description

The dataset used in this study consists of 21,644 records, each capturing crucial parameters related to oil well operations, such as FlowRate, CasingPressure, and TubingPressure (shown in Figure 2). In our study, we utilized a rare and valuable dataset from oil wells, which is available as an open-source repository on GitHub, curated by the scientist Shivankur Kapoor [37]. Access to such data is typically scarce and obtaining it independently is a challenging endeavor. These data points were collected in real-time through sensors and SCADA systems strategically positioned at oil wells. The variables we used from the dataset are as follows:

1. FlowRate: Represents the flow rate of oil through the well, providing a measure of production levels. This is the target parameter.

2. CasingPressure: Measures the pressure within the well's casing, offering insights into the well's structural integrity and overall performance. This was used as a feature.

3. DateTime: A timestamp indicating when each data point was recorded, allowing for accurate time-series analysis.



	Time	CasingPressure	FlowRate	LinePressure	StaticPressure	TubingPressure	Qmin
0	1437815718	515.764	315.083	289.940	287.979	235.035	105.083
1	1437815737	515.798	320.269	289.930	287.925	234.976	105.269
2	1437815916	515.890	309.478	290.479	288.598	235.251	105.478
3	1437816101	516.164	181.501	291.027	289.846	235.892	105.501
4	1437816459	516.256	194.514	290.753	289.416	235.159	105.514
...	...	...	...	...	...	...	...
21,639	1443306097	622.859	856.341	283.991	266.215	360.480	105.341
21,640	1443306276	622.585	842.984	285.088	265.908	360.663	104.984
21,641	1443306460	623.134	933.123	286.276	266.082	361.030	105.123
21,642	1443306640	622.585	928.875	287.007	265.860	362.312	104.875
21,643	1443306819	623.134	834.597	287.555	265.077	363.228	105.597

Figure 2. General view of dataset.

In addition to these core parameters, the dataset also contains derived features created during preprocessing to enrich the analysis. These include pressure differentials and historical averages. The raw data underwent thorough cleaning to remove duplicates and null values, ensuring consistency and reliability. Furthermore, all features were normalized to the range [0, 1] using MinMaxScaler to enhance the performance of machine learning models.

While machine learning models can be sensitive to noise, training on data that reflect real-world conditions, including inherent noise, often improves generalization. This approach aligns with the findings of Yue et al. [38], who demonstrated that removing noise from oil reservoir data could negatively impact model performance. Our study also supports this principle; we did not clean the noise present in the dataset. Our LSTM-based models, known for their ability to filter irrelevant information due to their forget gate, achieved high predictive accuracy (R^2 scores consistently above 0.97), even with noise present in the dataset. Preliminary sensitivity analyses confirmed that removing noise did not significantly improve performance. While our models demonstrated robustness to the noise levels present in the current dataset, future work may explore additional noise mitigation strategies, such as rolling mean, moving average filters, Savitzky–Golay filters, or wavelet denoising [39], to ensure resilience in potentially noisier scenarios.

The dataset's comprehensive structure enables a holistic view of both the hydraulic and electrical aspects of oil well operations. The data were used to build and validate a digital twin framework aimed at improving operational forecasts and real-time decision-making. Specifically, this dataset allows for accurate prediction of FlowRate and the detection of anomalies in CasingPressure. These predictions are crucial for optimizing well performance, particularly in challenging operating conditions. Additionally, by monitoring CasingPressure and TubingPressure, predictive maintenance strategies can be implemented, identifying potential issues before they escalate into operational downtimes.

In general, this dataset was selected due to its comprehensive representation of oil well operations. For this study, we focused on the casingPressure as a feature, DateTime for time series, and FlowRate for target. Our objective was to predict future FlowRate values based on historical data, using the DateTime feature to ensure accurate time-series analysis.

In the preprocessing stage, the time column was converted from Unix timestamp to DateTime format. This adjustment facilitates better feature extraction, allowing us to derive additional temporal features, such as the year, month, day, hour, and day of the week. These features can enhance the model's ability to capture temporal patterns and improve its predictive performance. Additionally, the DateTime format improves human readability, making it easier to debug and interpret the data.

A correlation analysis was conducted to assess the relationships between different features, revealing a significant positive correlation between FlowRate and other operational parameters such as CasingPressure, as shown in Figure 3.

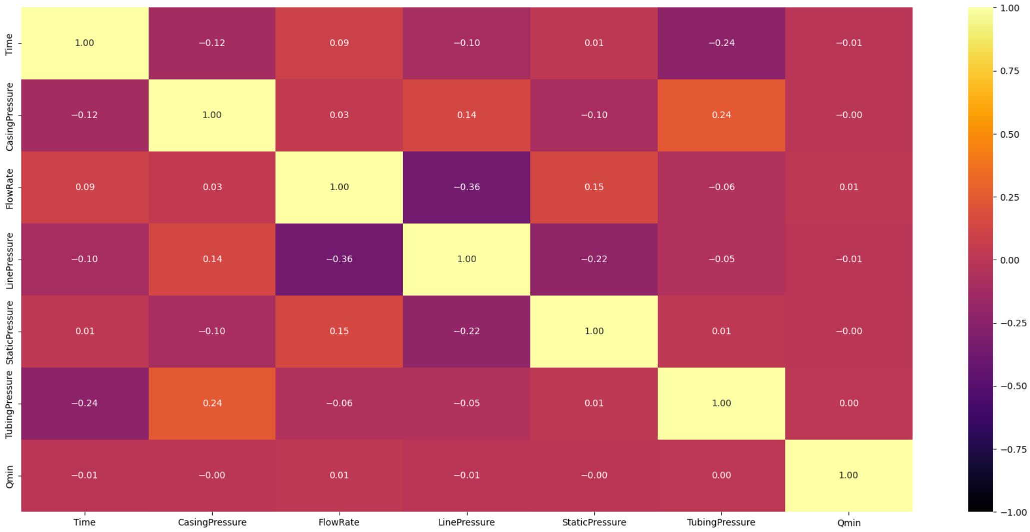


Figure 3. Correlation matrix of variables.

2.2. Machine Learning Algorithms

Five distinct machine learning algorithms, including Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BILSTM), Momentum LSTM, Gated Recurrent Unit (GRU), and Transformers, were employed for data analysis. These models were trained and tested using the dataset focused on predicting future FlowRate values. Training utilized 80% of the dataset, with the remaining 20% designated for testing, and 15% of the training data were used for validation.

LSTM is a type of recurrent neural network (RNN) designed to model sequences and long-term dependencies effectively. Unlike traditional RNNs, LSTM models can maintain information over long periods using memory cells. Each memory cell is controlled by three gates: input, forget, and output gates.

BILSTM networks extend the functionality of LSTM networks by allowing information to flow in both forward and backward directions. This architecture is particularly effective for tasks that require context from both past and future data points.

In a BILSTM network, as you can see in Figure 4, each time step receives two hidden states: one from the forward pass (left to right) and one from the backward pass (right to left). These two hidden states are concatenated and passed to the next layer.

Here, x_t is the input at time t, h_{t-1^f} and h_{t-1^b} are the hidden states from the previous forward and backward passes, W and b are weight matrices and biases, and σ and tanh represent the sigmoid and hyperbolic tangent activation functions, respectively.

MLSTM networks incorporate the concept of momentum into the traditional LSTM architecture to accelerate convergence during training. The momentum term helps the model learn faster by smoothing the gradients and guiding the optimization process in the right direction.

MLSTM retains the same gating mechanisms as traditional LSTM while adding a momentum term to accelerate learning.

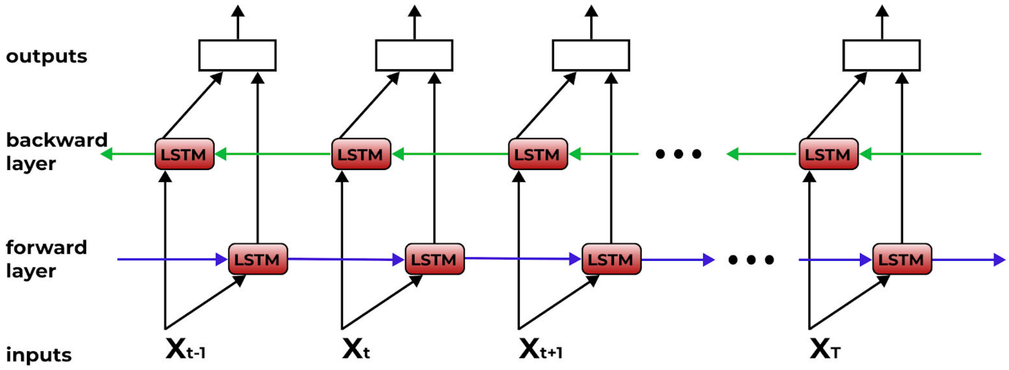


Figure 4. Unrolled BILSTM.

Momentum update:

Let g_t represent the gradients at time step t and ϑ_{t-1} represent the previous velocity (momentum), then the new velocity ϑ_t is updated using the following:

$$\vartheta_t = \beta * \vartheta_{t-1} + (1 - \beta) * g_t$$

where β is the momentum factor.

The updated weights and biases are then calculated using the velocity, as follows:

$$\theta_t = \theta_{t-1} - \Delta * \vartheta_t$$

where θ_t represents the model parameters (weights and biases) at time step t, Δ is the learning rate.

MLSTM retains the benefits of traditional LSTMs while potentially improving training speed and model performance.

GRUs are a type of recurrent neural network similar to Long Short-Term Memory (LSTM) networks but with a simpler architecture. GRUs have fewer gates and do not have a separate cell state, which makes them computationally more efficient, while retaining long-term memory capabilities.

GRUs consist of two main gates: the update gate and the reset gate. These gates control the flow of information and allow the model to maintain or forget previous states.

Update gate (z_t):

$$z_t = \sigma(W_z * [h_{t-1}, x_t] + b_z)$$

This gate determines how much of the previous hidden state should be carried forward to the current time step.

Reset gate (r_t):

$$r_t = \sigma(W_r * [h_{t-1}, x_t] + b_r)$$

The reset gate controls how much of the past information should be forgotten.

Candidate hidden state ($\widetilde{h}_t$):

$$\widetilde{h}_t = tanh(W_h * [r_t * h_{t-1}, x_t] + b_h)$$

The candidate hidden state computes the potential new state, incorporating the reset gate.

Hidden state update (h_t):

$$h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h}_t$$

The new hidden state h_t is a combination of the previous hidden state and the candidate hidden state, weighted by the update gate.

In general, GRUs perform the following operations:

1. Calculate gates: computes the update and reset gates.
2. Update gate (z_t): controls the amount of previous state retained.
3. Reset gate (r_t): determines how much past information is forgotten.
4. Compute candidate hidden state: calculates the candidate hidden state, $\widetilde{h}_t$, using the reset gate.
5. Update hidden state: calculates the new hidden state by combining the previous state and candidate state weighted by the update gate.

Transformers are a type of neural network architecture that excel at handling sequential data and were initially designed for natural language processing tasks. They leverage self-attention mechanisms to capture long-range dependencies, making them suitable for time-series forecasting.

The architecture of a Transformer (as shown in Figure 5) consists of an encoder and a decoder. For time-series forecasting, only the encoder is typically used. Here is how the self-attention mechanism works:

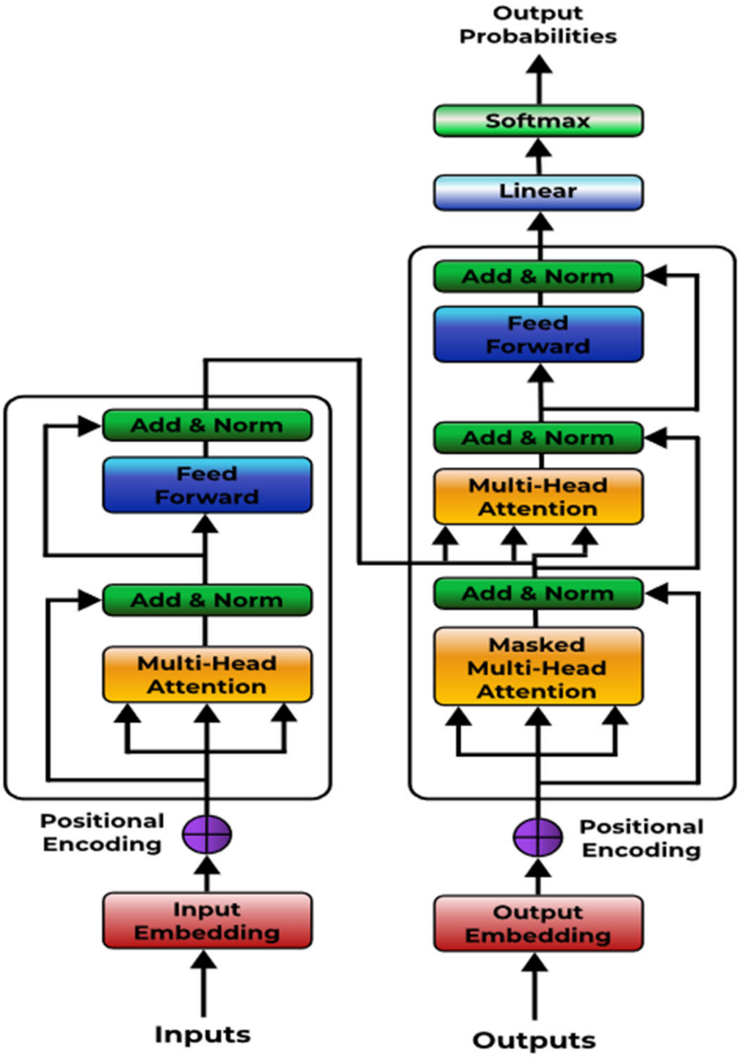


Figure 5. Transformer architecture.

Transformers can handle long-range dependencies effectively using self-attention mechanisms. By capturing relationships between time steps without relying on recurrence, they can provide accurate forecasts for time-series data such as FlowRate in oil well operations.

3. Experiments and Results

In the realm of oil well monitoring, the control box (see Figure 6) stands as a pivotal component, orchestrating a seamless integration of hardware and software to manage and control field operations effectively. At the heart of this sophisticated apparatus is the Programmable Logic Controller (PLC), which processes incoming data from various sensors, including those monitoring flow and pressure, to execute control logic in real-time.

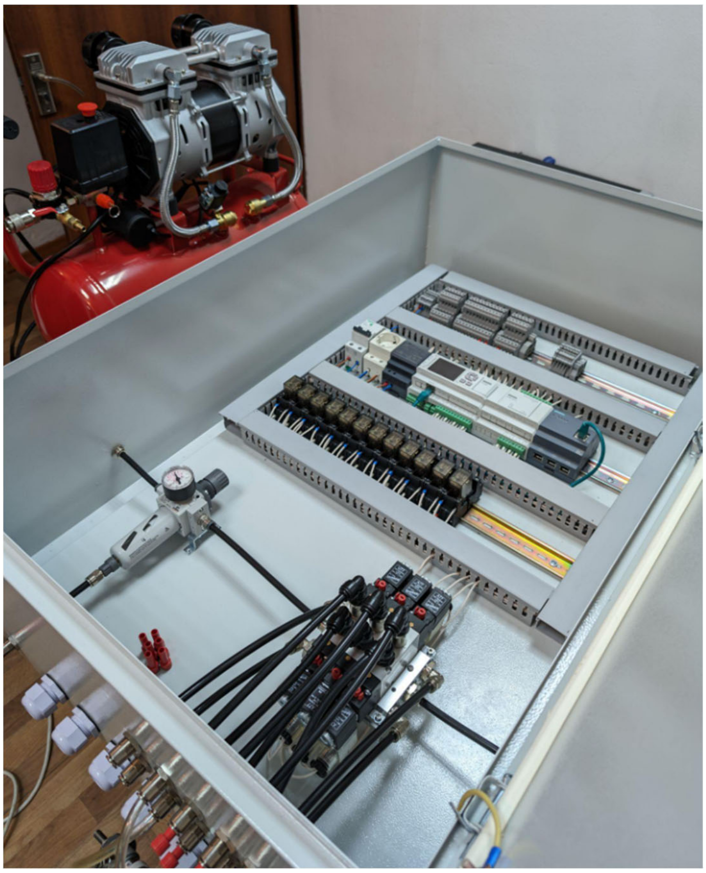


Figure 6. Overview of the control box.

The relays within the control box are crucial for switching operations, enabling the control of higher power loads that exceed the PLC's direct output capabilities. These are complemented by extension modules that augment the PLC's input and output range, thereby enhancing the system's scalability and accommodating a broader array of sensors and control mechanisms. This setup facilitates the handling of both analog and digital signals—where analog inputs capture continuous data from sensors, digital inputs manage binary signals from switches, and status indicators ensure a comprehensive monitoring environment.

High-speed counting inputs in the control box are meticulously designed to capture rapid pulse signals from turbine meters, which are indispensable for accurate flow measurement. Conversely, Pulse Width Modulation (PWM) outputs play a critical role in regulating the operation of devices such as pumps, which manage the dynamic fluid flows within the well, ensuring precise control over such critical operational parameters.

One of the pivotal functionalities of the control box is the management of pneumatic valves, which regulate oil flow and maintain pressure within safe limits. These valves are adjusted based on predictive insights provided by advanced machine learning models, which process historical and real-time data to predict operational parameters. This predictive capability allows for preemptive adjustments, optimizing performance, and averting potential operational hazards.

Moreover, the control box is fortified with Ethernet connectivity and supports the Modbus TCP protocol, a staple in industrial communications, which facilitates robust integration with other industrial systems and enables effective remote monitoring and management. Embedded within the control box is a specialized PCB (as depicted in Figure 7), equipped with an SMD device running a Linux operating system. This configuration not only handles complex computations and data processing but also acts as a crucial link between the PLC and the network, providing a stable and versatile platform for the development of custom applications and the integration of machine learning algorithms.

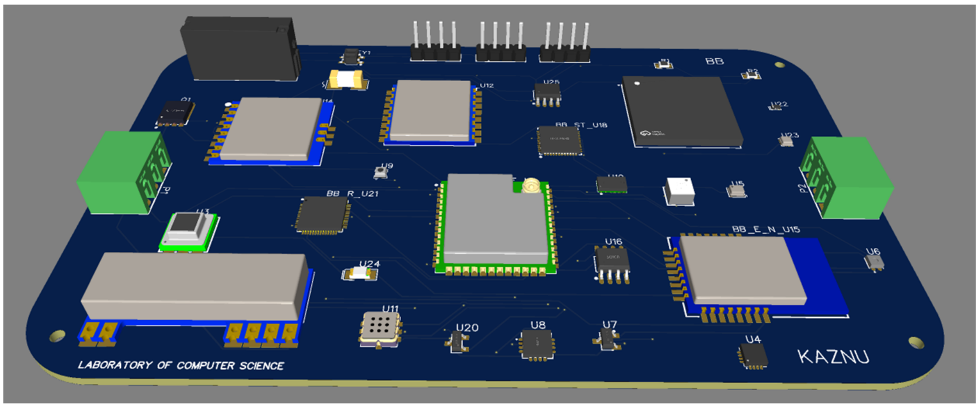


Figure 7. Overview of the intelligent data collection device.

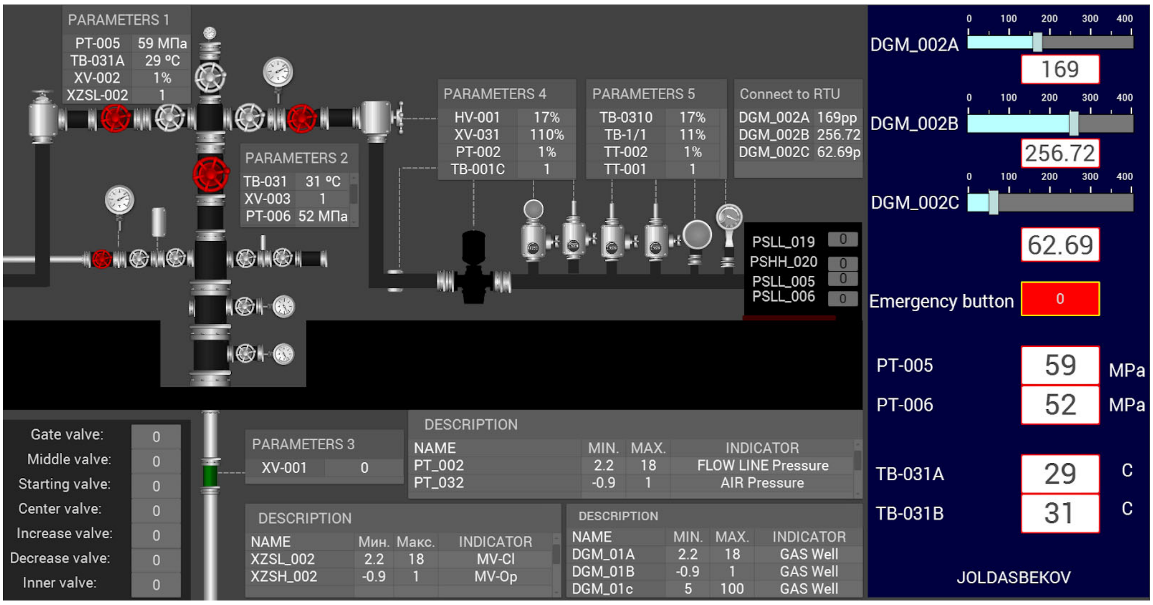
The PCB SMD device, intricately designed for the control box, incorporates cutting-edge components such as the Raspberry Pi Pico 2040 and the Espressif ESP32. These microcontrollers offer robust processing capabilities and flexibility for various monitoring tasks. The Raspberry Pi Pico 2040, built on the ARM Cortex-M0+ architecture, handles complex computations efficiently, while the ESP32 facilitates Wi-Fi connectivity, enabling seamless TCP/IP communications. This setup ensures robust data acquisition from an array of sensors strategically positioned throughout the oil well system. The integration of Wi-Fi connectivity allows for real-time data transmission, enhancing the system's responsiveness and enabling immediate operational adjustments based on data-driven insights. This sophisticated electronic framework is crucial for optimizing operational efficiency and reliability in oil well monitoring, reflecting the advanced technological integration characteristic of modern industrial systems.

The schematic (refer to Figure 8) delineates the SCADA system's critical role in monitoring and controlling oil well operations. Integrated with an SQL database and an OPC server, the system effectively manages essential parameters, such as flow rate and pressures, to ensure operational efficiency and safety. The SCADA system utilizes predicted values from external machine learning models to optimize responses and actions, enhancing its capability to maintain operational integrity. This setup underscores the system's strategic functionality and robust connectivity in managing oil well operations.

To prepare the data for training machine learning models, the data were divided into three main sets. A training dataset that was 80% of the original dataset was used to train the models. To tune the hyperparameters and prevent overfitting, the validation dataset, which was 15% of the training dataset, was used with the validation split parameter. The remaining 20% of the original dataset was allocated to the test dataset and used to evaluate the final performance of the model.

Data preprocessing included handling missing values, which were imputed with means for numerical features, and normalizing feature values using the min-max method so that the feature values ranged from 0 to 1.

Each model was trained using the following configuration: 10 epochs, a batch size of 64, and a verbose parameter of 1 to output the training progress. For validation, a dataset comprising 15% of the training dataset was used.



PARAMETERS 1

PT-005	59 МПа
TB-031A	29 °C
XV-002	1%
XZSL-002	1

PARAMETERS 2

TB-031	31 °C
XV-003	1
PT-006	52 МПа

PARAMETERS 4

HV-001	17%
XV-031	110%
PT-002	1%
TB-001C	1

PARAMETERS 5

TB-0310	17%
TB-1/1	11%
TT-002	1%
TT-001	1

Connect to RTU

DGM_002A	169pp
DGM_002B	256.72
DGM_002C	62.69p

PSLL_019	0
PSHH_020	0
PSLL_005	0
PSLL_006	0

Gate valve:	0
Middle valve:	0
Starting valve:	0
Center valve:	0
Increase valve:	0
Decrease valve:	0
Inner valve:	0

PARAMETERS 3

XV-001	0

DESCRIPTION

NAME	MIN.	MAX.	INDICATOR
PT_002	2.2	18	FLOW LINE Pressure
PT_032	-0.9	1	AIR Pressure

DESCRIPTION

NAME	Мин.	Макс.	INDICATOR
XZSL_002	2.2	18	MV-Cl
XZSH_002	-0.9	1	MV-Op

DESCRIPTION

NAME	MIN.	MAX.	INDICATOR
DGM_01A	2.2	18	GAS Well
DGM_01B	-0.9	1	GAS Well
DGM_01c	5	100	GAS Well

DGM_002A 169

DGM_002B 256.72

DGM_002C 62.69

Emergency button 0

PT-005	59	MPa
PT-006	52	MPa
TB-031A	29	C
TB-031B	31	C

JOLDASBEKOV

Figure 8. Supervisory control and data acquisition system.

Each model was compiled using the following configuration: a mean squared error (mean_squared_error) loss function, an Adam optimizer with a learning rate of 1×10^{-4}, and an R^2 score evaluation metric.

According to Table 1, the models performed well on both test datasets. All models achieved R^2 values above 0.96 on the test datasets, indicating that they were able to explain most of the variance of the target variable. Momentum LSTM and Bidirectional LSTM performed the best, with R^2 values very close to 1. Additionally, the GRU and standard LSTM models showed a performance level no worse than MLSTM and BILSTM.

Table 1. The statistical metrics of the proposed models.

Model	R^2 Score	MAE	MSE
LSTM	0.9720	0.0090	0.0001
BILSTM	0.9725	0.0089	0.0001
MLSTM	0.9726	0.0089	0.0001
GRU	0.9724	0.0089	0.0001
Transformers	0.9685	0.0096	0.0001

Transformers, while achieving a notable R^2 value of 0.9685, performed comparatively worse than the recurrent models. As described in the methodology section, the sequential nature of the data favors recurrent models, so these results are quite natural. Despite this, Transformers still demonstrated relatively strong capabilities, with results that are competitive.

As a result of comparing the performance of various machine learning model algorithms based on the R^2 score metric, the following conclusions can be drawn:

A comparison of different modeling algorithms shows that LSTM (see Figure 9), BILSTM, MLSTM, and GRU demonstrate high performance and similar coefficient of determination (R^2) values (0.9720, 0.9725, 0.9726, and 0.9724, respectively), indicating their ability to accurately predict. The mean absolute error (MAE) for all these models also remains low, at approximately 0.0089, and the mean square error (MSE) remains minimal, at approximately 0.0001. These results indicate that all four algorithms have good predictive ability and are well-trained on the provided data.

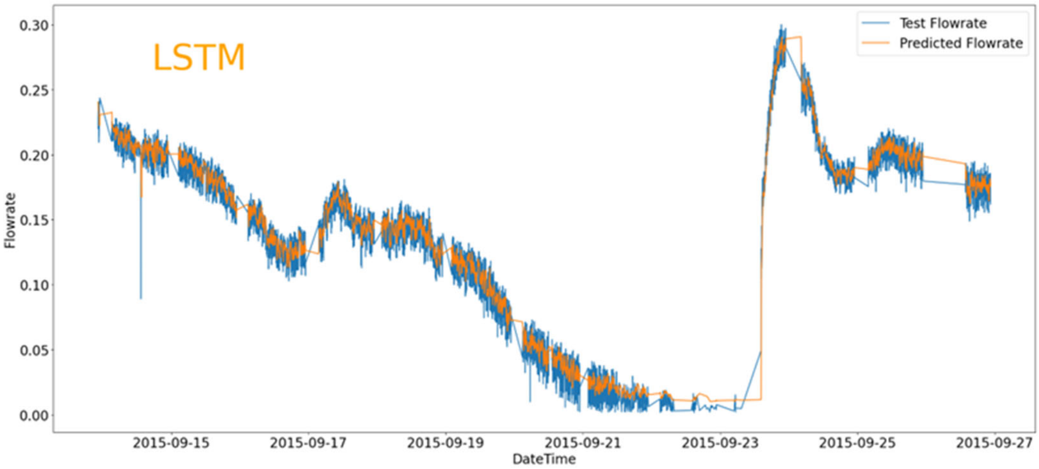


Figure 9. Comparison of test and predicted results of LSTM.

On the other hand, the Transformers-based model showed slightly lower performance compared to LSTM, BILSTM, MLSTM, and GRU as shown in Figure 10, having a coefficient of determination (R^2) equal to 0.9685. This may indicate that, for a given dataset and application context, the Transformers architecture is not as effective as the other algorithms considered. Additionally, the mean absolute error for the Transformers model is 0.0096, which is slightly higher than that of the other models, while the root mean square error remains the same, at 0.0001.

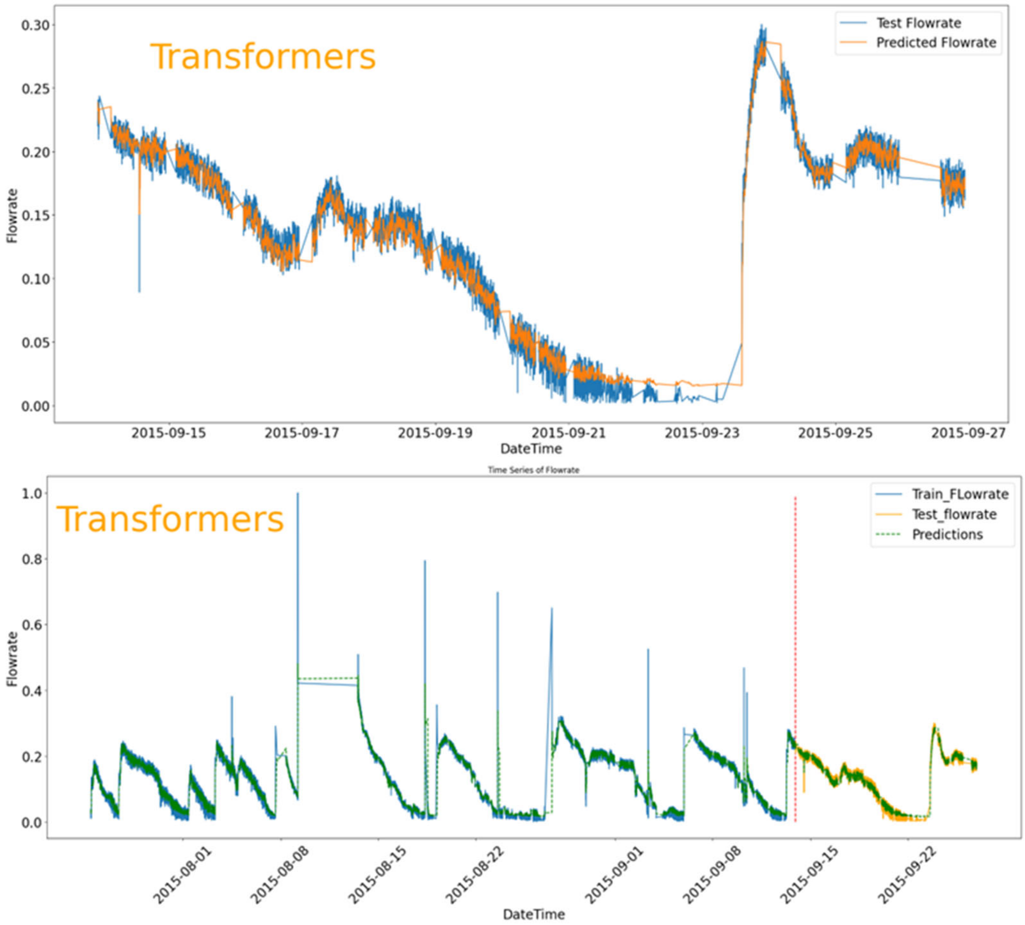


Figure 10. Comparison of test and forecast results, as well as forecasting future results for a month.

Thus, based on the results provided, it can be concluded that LSTM, BILSTM, MLSTM, and GRU are preferable options for this prediction task than the Transformers-based model.

In general, all the listed models have high performance according to the R^2 score metric, which confirms their effectiveness in predicting dependent variables in the data (see Figure 11).

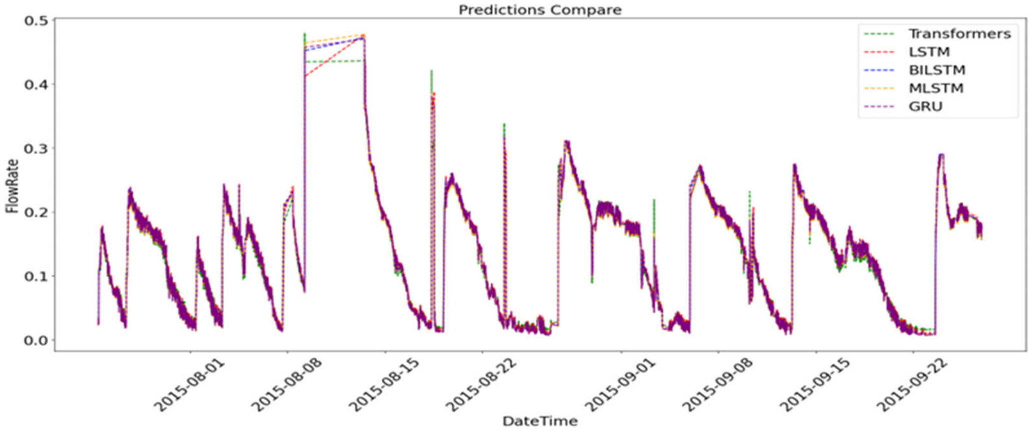


Figure 11. Comparative graph of results of all algorithms.

This study revealed that the LSTM model and its variants, including BiLSTM and GRU, exhibited superior performance in predicting oil well operational parameters, with R^2 values consistently above 0.97. This indicates a high level of accuracy in modeling and predicting crucial operational metrics, which is essential for efficient well management. The Transformer model, although slightly less effective, with an R^2 value of 0.9685, still demonstrated substantial predictive capabilities, suggesting its potential utility in scenarios requiring the processing of large datasets with complex patterns.

Moreover, the implementation of these ML models within SCADA systems facilitated enhanced real-time decision-making. By integrating predictive models directly into operational workflows, the systems could preemptively adjust to changing conditions, thereby preventing potential failures and optimizing production efficiency. This integration not only underscores the robustness of machine learning models in operational settings but also highlights the evolving landscape of industrial monitoring systems, where data-driven insights are becoming central to operational strategies.

4. Discussion and Limitations

The integration of machine learning models into SCADA systems has facilitated real-time decision-making, allowing systems to proactively adapt to changing conditions. For example, the use of predictive models has made it possible to proactively adjust pneumatic, electromagnetic valves, and other control mechanisms, thereby maintaining pressure within safe limits and ensuring optimal flow rates. Moreover, this study highlights the robustness and versatility of recurrent neural network models in processing time series data. The slightly lower performance of the Transformer model indicates that, although it is a powerful tool for certain applications, the sequential nature of oil well production data favors the use of recurrent models.

We conducted an analytical study utilizing the Ishikawa diagram (refer to Figure 12) to systematically visualize the factors impacting the integration of machine learning with intelligent control systems for cost forecasting in the oil industry. The diagram facilitated the identification and organization of potential causes of problems related to this integration process.

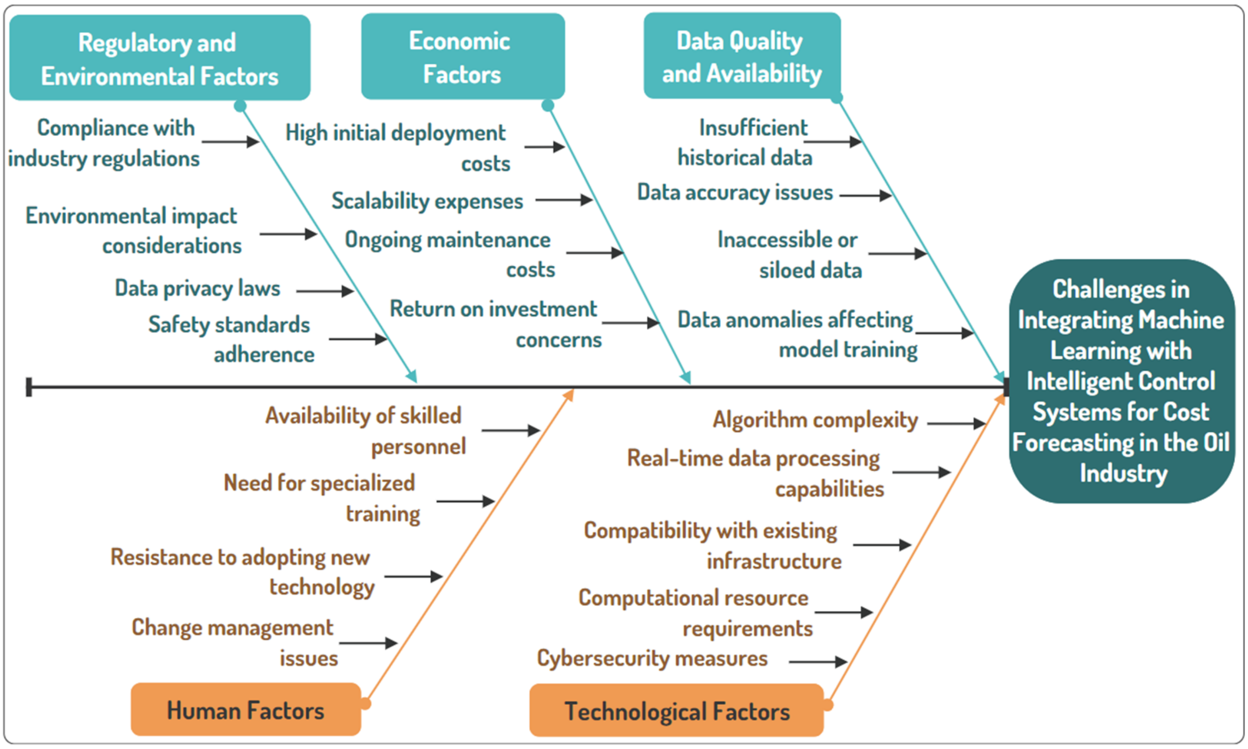

Regulatory and Environmental Factors
Economic Factors
Data Quality and Availability
Compliance with industry regulations
Environmental impact considerations
Data privacy laws
Safety standards adherence
High initial deployment costs
Scalability expenses
Ongoing maintenance costs
Return on investment concerns
Insufficient historical data
Data accuracy issues
Inaccessible or siloed data
Data anomalies affecting model training
Challenges in Integrating Machine Learning with Intelligent Control Systems for Cost Forecasting in the Oil Industry
Availability of skilled personnel
Need for specialized training
Resistance to adopting new technology
Change management issues
Algorithm complexity
Real-time data processing capabilities
Compatibility with existing infrastructure
Computational resource requirements
Cybersecurity measures
Human Factors
Technological Factors

Figure 12. Analysis of the effectiveness of integrating machine learning with intelligent control systems for cost forecasting in the oil industry according to the Ishikawa diagram.

Although the results are promising, this study has several limitations that deserve consideration. The dataset, although comprehensive, is limited to specific operational parameters from a single source, which may not fully reflect the variability and complexity of oil well operations worldwide. This may affect the generalizability of the results. This highlights the need for more extensive testing of unseen data and cross-validation techniques. Implementing these models in real-time in actual field operations presents technical and logistical challenges, as it requires robust infrastructures to deploy and integrate into existing SCADA systems. Moreover, the computational resources required to train and deploy these advanced machine-learning models are significant, which may limit their practicality in resource-constrained environments.

In addition, it is important to acknowledge the limitations of using a dataset encompassing 21,644 records from a specific oil well. This limited scope may not fully represent the vast heterogeneity and dynamic nature of oil well operations in diverse geological formations and under varying conditions.

Future research should address these limitations by including more diverse datasets, testing the performance of models under different operational conditions, developing efficient deployment frameworks, and optimizing models to improve computational efficiency. Additionally, addressing the challenges of data shift and concept drift through techniques such as continuous learning and transfer learning is crucial for ensuring the long-term reliability and generalizability of ML-based predictive models in real-world oil well applications.

5. Conclusions

This study demonstrates the effectiveness of integrating advanced machine learning models, particularly LSTM-based architectures, with SCADA systems for predictive maintenance and operational optimization in oil well operations. The LSTM, BiLSTM, and GRU models consistently achieve R^2 values above 0.97, highlighting their superior performance in accurate flow rate prediction and anomaly detection.

The successful implementation of these machine learning models in SCADA systems contributed to more efficient real-time decision-making, significantly reducing operational downtime and optimizing production efficiency. The results of this study highlight the need to integrate such predictive models in the oil and gas industry to improve overall operating performance.

Oil and gas companies should consider implementing advanced machine learning models to improve their SCADA systems. By leveraging the forecasting capabilities of LSTM, BiLSTM, and GRU models, companies can achieve more accurate flow forecasting and timely anomaly detection, leading to improved maintenance strategies and process optimization. It is recommended to invest in the necessary computing infrastructure and expertise to effectively implement and support these models.

Future research should focus on expanding the data set to include more diverse operational parameters from different well types and geographic locations to increase the generalizability of the results. In addition, further research should focus on improving machine learning models to improve their reliability and efficiency, as well as on reducing computational resource requirements. Studying real-time implementation mechanisms and testing the performance of models under different operating conditions will be critical to the wider adoption of these technologies in the oil and gas industry.

Author Contributions: Conceptualization, B.A., N.T., A.M., Y.N. and S.S.; methodology, B.A., N.T., A.M., Y.N. and S.S.; software, B.A.; validation, B.A., N.T. and A.M.; formal analysis, B.A.; investigation, B.A.; resources, B.A.; data curation, B.A.; writing—original draft preparation, B.A.; writing—review and editing, B.A., N.T., A.M., Y.N. and S.S.; visualization, B.A.; supervision, B.A.; project administration, B.A.; funding acquisition, N.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a grant from the Ministry of Education and Science of the Republic of Kazakhstan BR18574136 "Development of deep learning and intellectual analysis methods for solving complex problems of mechanics and robotics" (2022–2024).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this research are available from the open-source repository referenced in [37].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Huang, R.; Wei, C.; Wang, B.; Yang, J.; Xu, X.; Wu, S.; Huang, S. Well Performance Prediction Based on Long Short-Term Memory (LSTM) Neural Network. *J. Pet. Sci. Eng.* **2022**, *208*, 109686. [CrossRef]
2. 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. [CrossRef]
3. Panja, P.; Jia, W.; McPherson, B. Prediction of Well Performance in SACROC Field Using Stacked Long Short-Term Memory (LSTM) Network. *Expert Syst. Appl.* **2022**, *205*, 117670. [CrossRef]
4. Song, X.; Liu, Y.; Xue, L.; Wang, J.; Zhang, J.; Wang, J.; Jiang, L.; Cheng, Z. Time-Series Well Performance Prediction Based on Long Short-Term Memory (LSTM) Neural Network Model. *J. Pet. Sci. Eng.* **2020**, *186*, 106682. [CrossRef]
5. Xue, L.; Wang, J.; Han, J.; Yang, M.; Mwasmwasa, M.S.; Nanguka, F. Gas Well Performance Prediction Using Deep Learning Jointly Driven by Decline Curve Analysis Model and Production Data. *Adv. Geo-Energy Res.* **2023**, *8*, 159–169. [CrossRef]
6. Ning, Y.; Kazemi, H.; Tahmasebi, P. A Comparative Machine Learning Study for Time Series Oil Production Forecasting: ARIMA, LSTM, and Prophet. *Comput. Geosci.* **2022**, *164*, 105126. [CrossRef]
7. Fan, D.; Sun, H.; Yao, J.; Zhang, K.; Yan, X.; Sun, Z. Well Production Forecasting Based on ARIMA-LSTM Model Considering Manual Operations. *Energy* **2021**, *220*, 119708. [CrossRef]
8. Kandziora, C. *Applying Artificial Intelligence to Optimize Oil and Gas Production*; OnePetro: Richardson, TX, USA, 2019.
9. Temizel, C.; Canbaz, C.H.; Palabiyik, Y.; Aydin, H.; Tran, M.; Ozyurtkan, M.H.; Yurukcu, M.; Johnson, P. *A Thorough Review of Machine Learning Applications in Oil and Gas Industry*; OnePetro: Richardson, TX, USA, 2021.
10. Rashad, O.; Attallah, O.; Morsi, I. A Smart PLC-SCADA Framework for Monitoring Petroleum Products Terminals in Industry 4.0 via Machine Learning. *Meas. Control* **2022**, *55*, 830–848. [CrossRef]

11. Phillips, B.; Gamess, E.; Krishnaprasad, S. An Evaluation of Machine Learning-Based Anomaly Detection in a SCADA System Using the Modbus Protocol. In Proceedings of the 2020 ACM Southeast Conference, Tampa, FL, USA, 2–4 April 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 188–196.
12. Chen, Y.; Miao, B.; Wang, Y.; Huang, Y.; Jiang, Y.; Shi, X. A Deep Regression Method for Gas Well Liquid Loading Prediction. *SPE J.* **2024**, *29*, 1847–1861. [CrossRef]
13. Eso, A.; Omorogiuwa, D. Real-Time Effective Monitoring and Control in Oil and Gas Industry Using SCADA Technology as a Management Tool. *J. Altern. Renew. Energy Sources* **2022**, *8*, 22–38. [CrossRef]
14. Al-Fadhli, M.; Zaher, A. A Smart SCADA System for Oil Refineries. In Proceedings of the 2018 International Conference on Computing Sciences and Engineering (ICCSE), Kuwait, Kuwait, 11–13 March 2018; pp. 1–6.
15. He, X.; Robards, E.; Gamble, R.; Papa, M. Anomaly Detection Sensors for a Modbus-Based Oil and Gas Well-Monitoring System. In Proceedings of the 2019 2nd International Conference on Data Intelligence and Security (ICDIS), South Padre Island, TX, USA, 28–30 June 2019; pp. 1–8.
16. Chen, Y.; Huang, Y.; Miao, B.; Shi, X.; Li, P. Adaptive Anomaly Detection-Based Liquid Loading Prediction in Shale Gas Wells. *J. Pet. Sci. Eng.* **2022**, *214*, 110522. [CrossRef]
17. Tariq, Z.; Aljawad, M.S.; Hasan, A.; Murtaza, M.; Mohammed, E.; El-Husseiny, A.; Alarifi, S.A.; Mahmoud, M.; Abdulraheem, A. A Systematic Review of Data Science and Machine Learning Applications to the Oil and Gas Industry. *J. Pet. Explor. Prod. Technol.* **2021**, *11*, 4339–4374. [CrossRef]
18. R Azmi, P.A.; Yusoff, M.; Mohd Sallehud-din, M.T. A Review of Predictive Analytics Models in the Oil and Gas Industries. *Sensors* **2024**, *24*, 4013. [CrossRef] [PubMed]
19. Rahman, M.H.; Shahriar, S.; Menezes, P.L. Recent Progress of Machine Learning Algorithms for the Oil and Lubricant Industry. *Lubricants* **2023**, *11*, 289. [CrossRef]
20. Ibrahim, N.M.; Alharbi, A.A.; Alzahrani, T.A.; Abdulkarim, A.M.; Alessa, I.A.; Hameed, A.M.; Albabtain, A.S.; Alqahtani, D.A.; Alsawwaf, M.K.; Almuqhim, A.A. Well Performance Classification and Prediction: Deep Learning and Machine Learning Long Term Regression Experiments on Oil, Gas, and Water Production. *Sensors* **2022**, *22*, 5326. [CrossRef]
21. Osah, U.; Howell, J. Predicting Oil Field Performance Using Machine Learning Programming: A Comparative Case Study from the UK Continental Shelf. *Pet. Geosci.* **2023**, *29*, petgeo2022-071. [CrossRef]
22. Rathnayake, S.; Rajora, A.; Firouzi, M. A Machine Learning-Based Predictive Model for Real-Time Monitoring of Flowing Bottom-Hole Pressure of Gas Wells. *Fuel* **2022**, *317*, 123524. [CrossRef]
23. He, D.; Qu, Y.; Sheng, G.; Wang, B.; Yan, X.; Tao, Z.; Lei, M. Oil Production Rate Forecasting by SA-LSTM Model in Tight Reservoirs. *Lithosphere* **2024**, *2024*, lithosphere_2023_197. [CrossRef]
24. Yan, C.; Qiu, Y.; Zhu, Y. Predict Oil Production with LSTM Neural Network. In Proceedings of the 9th International Conference on Computer Engineering and Networks, Changsha, China, 18–20 October 2019; Liu, Q., Liu, X., Li, L., Zhou, H., Zhao, H.-H., Eds.; Springer: Singapore, 2021; pp. 357–364.
25. Sagheer, A.; Kotb, M. Time Series Forecasting of Petroleum Production Using Deep LSTM Recurrent Networks. *Neurocomputing* **2019**, *323*, 203–213. [CrossRef]
26. Abbasimehr, H.; Paki, R. Improving Time Series Forecasting Using LSTM and Attention Models. *J. Ambient. Intell. Hum. Comput.* **2022**, *13*, 673–691. [CrossRef]
27. Wang, F.; Zai, Y.; Zhao, J.; Fang, S. *Field Application of Deep Learning for Flow Rate Prediction with Downhole Temperature and Pressure*; OnePetro: Richardson, TX, USA, 2021.
28. Kumar, I.; Tripathi, B.K.; Singh, A. Attention-Based LSTM Network-Assisted Time Series Forecasting Models for Petroleum Production. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106440. [CrossRef]
29. Abbasi, T.; Lim, K.H.; Yam, K.S. Predictive Maintenance of Oil and Gas Equipment Using Recurrent Neural Network. *IOP Conf. Ser. Mater. Sci. Eng.* **2019**, *495*, 012067. [CrossRef]
30. Dong, Y.; Zhang, Y.; Liu, F.; Cheng, X. Reservoir Production Prediction Model Based on a Stacked LSTM Network and Transfer Learning. *ACS Omega* **2021**, *6*, 34700–34711. [CrossRef] [PubMed]
31. Ma, X.; Hou, M.; Zhan, J.; Zhong, R. Enhancing Production Prediction in Shale Gas Reservoirs Using a Hybrid Gated Recurrent Unit and Multilayer Perceptron (GRU-MLP) Model. *Appl. Sci.* **2023**, *13*, 9827. [CrossRef]
32. Aljameel, S.S.; Alomari, D.M.; Alismail, S.; Khawaher, F.; Alkhudhair, A.A.; Aljubran, F.; Alzannan, R.M. An Anomaly Detection Model for Oil and Gas Pipelines Using Machine Learning. *Computation* **2022**, *10*, 138. [CrossRef]
33. Lian, Y.; Geng, Y.; Tian, T. Anomaly Detection Method for Multivariate Time Series Data of Oil and Gas Stations Based on Digital Twin and MTAD-GAN. *Appl. Sci.* **2023**, *13*, 1891. [CrossRef]
34. Jambol, D.D.; Sofoluwe, O.O.; Ukato, A.; Ochulor, O.J. Transforming Equipment Management in Oil and Gas with AI-Driven Predictive Maintenance. *Comput. Sci. IT Res. J.* **2024**, *5*, 1090–1112. [CrossRef]
35. Abdrakhmanov, I.; Kanin, E.; Boronin, S.; Burnaev, E.; Osiptsov, A. Development of Deep Transformer-Based Models for Long-Term Prediction of Transient Production of Oil Wells. In Proceedings of the 2021 SPE Russian Petroleum Technology Conference, Day 2 Wed, Virtual, 13 October 2021; p. D021S006R008.
36. Li, J.; Hu, D.; Chen, W.; Li, Y.; Zhang, M.; Peng, L. CNN-Based Volume Flow Rate Prediction of Oil–Gas–Water Three-Phase Intermittent Flow from Multiple Sensors. *Sensors* **2021**, *21*, 1245. [CrossRef] [PubMed]

37. Kapoor, S. Shivankurkapoor/Flowrate-Prediction 2024. Available online: https://github.com/shivankurkapoor/flowrate-prediction/ (accessed on 5 May 2024).
38. Yue, M.; Dai, Q.; Liao, H.; Liu, Y.; Fan, L.; Song, T. Prediction of ORF for Optimized CO_2 Flooding in Fractured Tight Oil Reservoirs via Machine Learning. *Energies* **2024**, *17*, 1303. [CrossRef]
39. Iskandar, U.P.; Kurihara, M. Time-Series Forecasting of a CO_2-EOR and CO_2 Storage Project Using a Data-Driven Approach. *Energies* **2022**, *15*, 4768. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.