International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-02, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

A Hybrid "ARIMA–ML Regression" Model for Enhanced Predictive Analysis in Cyber-Physical Systems: Conceptual framework and Simulation Evaluation

Ass. Eng. Iliyan Vasilev, PhD
Assistant Eng. - University of Chemical Technology and Metallurgy- Sofia
ORCID: 0009-0008-0863-1516

ABSTRACT: This paper presents a hybrid ARIMA–machine learning (ARIMA–ML) regression framework designed to improve predictive accuracy in cyber-physical systems (CPS). The approach brings together the strengths of classical statistical time-series modelling and modern data-driven techniques, allowing the model to capture both linear structures and nonlinear dynamics that commonly arise in CPS environments. A simulation-based evaluation was conducted using a multivariate dataset generated from a MATLAB/Simulink CPS model, complemented by Python-based machine learning components. The results show that the hybrid model consistently outperforms standalone ARIMA and ML approaches across multiple operational scenarios, including normal operation, peak load, and early-stage failure conditions. Improvements were observed not only in RMSE and MAE but also in residual stability, prediction interval reliability, and statistical significance as confirmed by the Diebold–Mariano test. These findings suggest that hybrid modelling offers a practical and effective pathway for enhancing predictive maintenance, anomaly detection, and decision-support capabilities in complex CPS environments. Future work will explore real-time deployment, integration with edge computing platforms, and the use of more advanced learning architectures to further strengthen model adaptability and performance.

KEYWORDS: ARIMA–ML integration, Cyber-physical systems, Hybrid forecasting models, Predictive maintenance, Simulation-based evaluation.

INTRODUCTION

Cyber-physical systems (CPS) have become increasingly central to modern industrial, transportation, and energy infrastructures, where reliable forecasting plays a crucial role in ensuring operational stability and timely maintenance. As noted by *Lee&Seshia* (2017), CPS depend on accurate predictive models to manage the tight coupling between computational and physical processes. Traditional statistical approaches such as ARIMA (Autoregressive Integrated Moving Average) remain widely used due to their interpretability and strong performance in modeling linear temporal dependencies (Box et al., 2016; Hyndman & Athanasopoulos, 2021). Time series forecasting using ARIMA model are widely studied in various fields. Rizvi (2024) states that "ARIMA model stands out for its simplicity and effectiveness". It is examined and used successfully in prediction of COVID spread (Kumar et al., 2020; Satrio et al., 2021), for meteorological drought forecasting (Khan et al., 2020), in agricultural engineering (Venu&Namitha, 2025), forecasting method of stock market (Wang&Guo, 2020).

The other part are the Machine learning (ML) methods that emerged as powerful alternatives, capable of capturing complex nonlinear relationships. Studies by *Breiman* (2001), *Bishop* (2006), and *Goodfellow et al*. (2016) demonstrate the ability of ML algorithms to learn intricate patterns from high-dimensional data. In CPS applications, ML-based forecasting shows promising in handling nonlinearities and multivariate interactions (Zhang et al., 2019; Wang & Jiang, 2020).

Hybrid modelling approaches proposed in the article combine the strengths of statistical and machine learning techniques. Early work by *Zhang* (2003) introduced the concept of integrating ARIMA with neural networks to capture both linear and nonlinear components of time-series data. Subsequent studies by *Khashei&Bijari* (2011) and *Pai&Lin* (2005) further demonstrated the effectiveness of hybrid ARIMA–ML models across various forecasting domains. More recent research in CPS predictive maintenance (e.g., Zhao et al., 2021; Li & Chen, 2022) highlights the growing relevance of hybrid frameworks in environments characterized by dynamic and heterogeneous data streams.

Motivated by these developments, this study proposes a hybrid ARIMA–machine learning regression model tailored for CPS forecasting. The approach leverages ARIMA to model linear temporal structure and applies machine learning to the residuals to capture nonlinear behavior. The integration of these complementary components create a the hybrid model aiming to provide more accurate, robust, and interpretable predictions for CPS applications.
The first stage applies the ARIMA model to identify and model linear dependencies within the time series.
ARIMA combines autoregression, differencing, and moving-average components to represent temporal patterns.The general ARIMA(p, d, q) model is expressed as:

$$\phi(L)(1-L)^d X_t = \theta(L)\epsilon_t$$

where:

- $X_t$is the observed time series
- $\epsilon_t$is white noise with variance σ^2
- $L$is the lag operator, $L^k X_t = X_{t-k}$
- $\phi(L)$and $\theta(L)$are the autoregressive and moving-average polynomials
- $d$denotes the order of differencing required to achieve stationarity

The AR component models dependence on past observations:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t$$

The differencing operator ensures stationarity:

$$Y_t = X_t - X_{t-1}, Y_t = \nabla^d X_t$$

The moving-average component captures the influence of past forecast errors:

$$X_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

Combining these math aspects the full ARIMA formulation may be structured:

$$\nabla^d X_t = \phi_1 \nabla^d X_{t-1} + \cdots + \phi_p \nabla^d X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

After fitting the model, forecasts are generated recursively as:

$$X_{t+h} = \phi_1 X_{t+h-1} + \cdots + \phi_p X_{t+h-p} + \theta_1 \epsilon_{t+h-1} + \cdots + \theta_q \epsilon_{t+h-q}$$

Residuals from the ARIMA model are:

$$\epsilon_t = X_t - X_t^{ARIMA}$$

These residuals represent nonlinear patterns not captured by the linear model and serve as the target for the second stage.

$$\epsilon_t = X_t - \hat{X}_t^{\text{ARIMA}}$$

represent the part of the data that ARIMA fails to explain. Here, $X_t$is the actual value at time t, and $\hat{X}_t^{\text{ARIMA}}$is the ARIMA forecast. To account for *nonlinear relationships* in these residuals, it needs to apply *ML regression models.* Denoting the residual series as $\boldsymbol{\epsilon_t}$ and additional explanatory features at time $t$as a vector $\mathbf{Z}_t = [\boldsymbol{Z_{t1}}, \boldsymbol{Z_{t2}}, \ldots, \boldsymbol{Z_{tk}}]$, which may include lagged residuals or other external variables. The machine learning model then learns a function $f(\cdot)$such that:

$$\hat{\epsilon}_t^{\text{ML}} = f(\mathbf{Z}_t)$$

where $\hat{\boldsymbol{\epsilon}}_t^{\text{ML}}$ is the ML prediction of the residual at time t. The function $f$can take many forms depending on the chosen algorithm—for example:

A. Random Forest / Decision Tree Regression:

$$f(\mathbf{Z}_t) = \text{Tree}(\mathbf{Z}_t)$$

B. Neural Network Regression:

$$f(\mathbf{Z}_t) = \sigma(W_2 \cdot \phi(W_1 \cdot \mathbf{Z}_t + b_1) + b_2)$$

where $W_1, W_2$are weight matrices, $b_1, b_2$are biases, $\phi$is a nonlinear activation function, and $\sigma$is the output activation function (e.g., identity for regression).

Finally, the *hybrid forecast* combines the ARIMA prediction with the ML correction:

$$\hat{X}_t^{\text{Hybrid}} = \hat{X}_t^{\text{ARIMA}} + \hat{\epsilon}_t^{\text{ML}}$$

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-02, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

This hybrid approach allows the model to capture both *linear temporal structures* via ARIMA and *complex nonlinear patterns* via ML, improving overall prediction accuracy.

The final prediction is computed as:

$$\hat{Y}_t = \hat{Y}_t^{ARIMA} + \hat{Y}_t^{ML(residuals)}$$

where:

- $\hat{Y}_t$ is the final predicted value,
- $\hat{Y}_t^{ARIMA}$ represents the *linear forecast*,
- $\hat{Y}_t^{ML}$ represents the *nonlinear correction* term.

This hybridization improves predictive accuracy by addressing both linear and nonlinear dynamics within the CPS. The steps of the model evaluation scheme are given in ***Figure 1.***

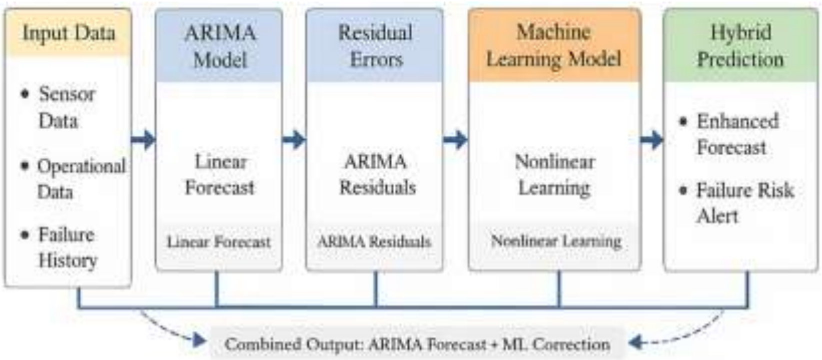

Figure 1. Hybrid ARIMA-ML model for Predictive Analysis in CPS (author's source)

The two math models are different in mathematical logic as well as in the physical interpretation of data processing and a comparison of ARIMA and ML is given in **Table 1.**

Table 1. ARIMA models – Machine Learning comparision

Aspect	ARIMA	Machine Learning
Core idea	Models' linear relationships using past values and past errors	Learns complex, nonlinear patterns directly from data
Best for	Stable, linear time-series with clear autocorrelation	Nonlinear, high-dimensional, noisy, or irregular data
Data requirements	Low — works well with small datasets	Medium to high — needs more data to generalize
Interpretability	Very high — parameters have clear meaning	Often low — many models are black boxes
Computational cost	Low	Medium to high
Handles seasonality	With SARIMA extensions	must be engineered or learned
Handles nonlinear patterns	Poorly	Very well
Risk of overfitting	Low	Medium to high (depends on model)
Forecast horizon	Strong for short-term	Strong for both short- and long-term
When it fails	Nonlinear systems, sudden regime changes	Very small datasets, need for interpretability
Typical use cases	Economics, energy demand, environmental data	CPS, anomaly detection, finance, sensor fusion

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-02, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

The training of the proposed hybrid forecasting model is performed in *two sequential stages*, each addressing complementary aspects of the data.

Sequence Stage 1: ARIMA Modeling Train

In the first stage, the focus is on capturing the ***linear temporal dependencies*** inherent in the historical time series. Optimal ARIMA parameters $(p, d, q)$are identified through rigorous analysis of the *autocorrelation function (ACF) and partial autocorrelation function (PACF).* Once the appropriate orders are selected, the ARIMA model is fitted to the historical dataset. This fitting process generates both the *initial forecasts* and the corresponding *residual errors*, which represent patterns not explained by the linear model. These residuals form the basis for the subsequent stage of modeling (***Figure 2***.)

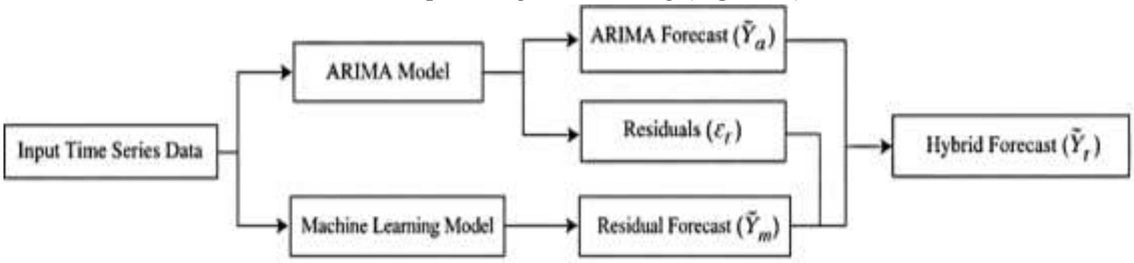

Figure 2. **Hybrid ARIMA-ML Regression Model Calculation flow**

Sequence Stage 2: Machine Learning Modeling Train

The second stage leverages ML techniques to capture *nonlinear relationships* present in the residual errors from the ARIMA model. The residuals $\epsilon_t = X_t - \hat{X}_t^{\text{ARIMA}}$are treated as the *target variable*, while input characteristics include *engineered features* derived from domain knowledge and *lagged residual values* to preserve temporal dependencies. ML algorithms such as *Random Forests, Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) networks* could be used to train these complex, nonlinear patterns, effectively complementing the linear predictions provided by ARIMA.

The **hybrid model** is then constructed by combining the outputs from both stages. The final prediction is obtained as:

$$\hat{X}_t^{\text{Hybrid}} = \hat{X}_t^{\text{ARIMA}} + \hat{\epsilon}_t^{\text{ML}}$$

where $\hat{X}_t^{\text{ARIMA}}$is the forecast from the linear ARIMA stage, and $\hat{\epsilon}_t^{\text{ML}}$is the ML-predicted residual. This two-stage approach ensures that the model effectively captures both the *linear trends and seasonal components* of the data while simultaneously *learning nonlinear residual patterns*, resulting in improved forecast accuracy (**Figure 3.**).

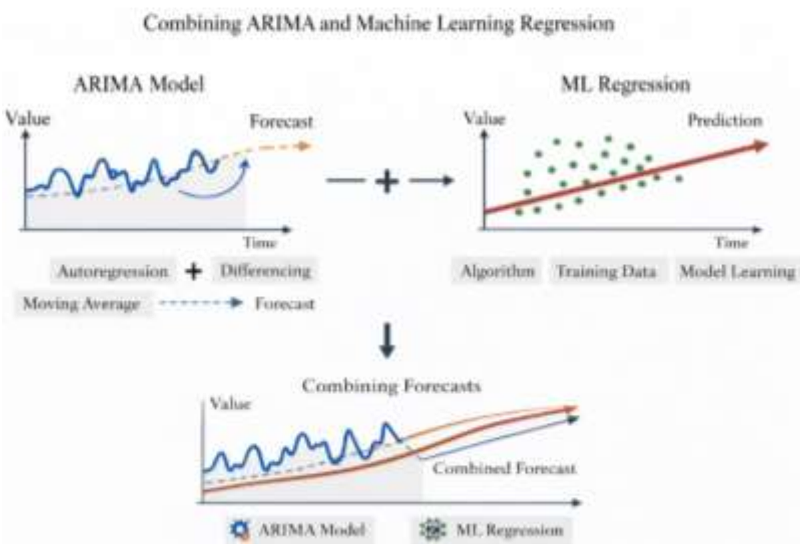

Figure. 3. Framework idea: Combining ARIMA and Machine Learning Regression.
(author's source)

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-02, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

The performance of the proposed hybrid model is evaluated using a combination of standard forecasting and classification metrics, depending on the specific application scenario. For continuous predictions, such as time series forecasting, accuracy is quantified using metrics including the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE). In this way it is captured the magnitude of prediction errors and provide a clear measure of model precision. In applications focused on failure prediction with categorical outcomes, model performance is evaluated using classification metrics, including Accuracy, Precision, Recall, F1-score, and the Receiver Operating Characteristic – Area Under Curve (ROC-AUC).

To strengthen robustness and generalizability it is used cross-validation techniques, so as the model to be tested on multiple partitions of the dataset and to reduce the risk of overfitting. The performance of the hybrid approach is systematically compared against isolated ARIMA and ML models. It shows indicatively the advantages of integrating linear and nonlinear predictive capabilities. The mere combination of statistical and ML methodologies into a unified predictive framework of the proposed hybrid model demonstrate accurately capturing both linear trends and complex nonlinear dynamics. It also proves to be reliable modeling of CPS behavior and supporting informed, data-driven decision-making processes.

Experimental Setup

Using simulated data has become a practical and widely accepted way to test and validate computational models, algorithms, and statistical methods especially in situations where real-world data is unavailable, restricted, or simply too limited to support rigorous analysis. Recent scientific reports show that the use of simulation-based validation has grown substantially, reaching about 78% of studies in some fields (DiRenzo et al., 2023).

This approach has proven useful across a range of applications. For example, simulation frameworks are often used to evaluate methods designed for logistic-type distributions (Guiguet & Pons, 2025; Deter et al., 2021) and are increasingly common in medical research where controlled datasets are difficult to obtain (Marchi et al., 2026). In other domains, researchers use "sandbox" environments to test classification models on synthetic spectral data, allowing them to explore model behavior under controlled conditions (Boichenko et al., 2022). More broadly, simulation-based models can reproduce essential characteristics of real-world systems, and their validity can be demonstrated through systematic comparison with empirical data. As Belik et al. (2024) note, these models offer a valuable and flexible tool for innovation research, providing a structured way to experiment, evaluate, and refine analytical methods before applying them to real operational environments.

The evaluation of the hybrid ARIMA–ML model was carried out entirely within a simulation-based environment. The system was built using MATLAB/Simulink to model the dynamic behaviour of the cyber-physical process, while Python's machine learning libraries were used for the predictive components. The simulation framework was designed to mimic the behaviour of a real CPS, producing multivariate time-series data that reflect realistic operating conditions. Working in a controlled environment made it possible to test the model systematically under different load levels, operational cycles, and dynamic disturbances, providing a reliable basis for validating the model and comparing its performance with alternative approaches.

The dataset used in this study consists of multivariate time-series signals generated by the simulated CPS. It includes a broad set of observations that capture both the physical state of the system and its operational dynamics. The raw sensor data—temperature, vibration, voltage, and current—represent the direct physical measurements of the system. Alongside these, key operational variables such as system load, cycle duration, and discrete operating states were recorded to reflect the system's functional behaviour. To strengthen the dataset and improve the model's ability to detect temporal structure, several engineered features were added. These include rolling averages, trend indicators, and rate-of-change measures, all of which help capture short-term fluctuations as well as longer-term dependencies in the data. Such derived features are especially useful in time-series modelling frameworks that rely on autocorrelation patterns, as they provide richer information than raw measurements alone.

For the purposes of model development and validation, the dataset is divided into three subsets following standard ML practices. Seventy percent of the data was assigned to the training set, which served as the basis for parameter estimation and model learning. Fifteen percent of the data was reserved for the validation set, allowing for systematic hyperparameter optimization and mitigation of overfitting. The remaining fifteen percent constituted the test set, employed exclusively for the assessment of model generalization and predictive performance on previously unseen data.

This structured and multivariate composition of the dataset, integrating both raw and derived features, provides a robust foundation for the application of advanced time-series modeling techniques, such as ARIMA and its multivariate extensions. The interplay between sensor readings and operational parameters, the dataset enables rigorous forecasting, anomaly detection, and analysis of complex system dynamics within the CPS environment.

The dataset is divided into:

- **Training set (70%)** – used for model learning
- **Validation set (15%)** – used for hyperparameter tuning
- **Test set (15%)** – used for final performance evaluation

```
%% MATLAB: Simulated CPS Time-Series Data

% Simulation parameters
T = 1000;       % Number of time steps
dt = 1;         % Time step size
time = (0:T-1)';    % Time vector

% Initialize signals
temperature = zeros(T,1);
vibration = zeros(T,1);
voltage = zeros(T,1);
current = zeros(T,1);

% Base operational parameters
system_load = 0.7;      % nominal load
cycle_time = 50;        % nominal cycle time
operational_state = zeros(T,1);  % 0: idle, 1: active

% Random seed for reproducibility
rng(42);

% Generate multivariate time series
for t = 2:T
  % Operational state switching
  if mod(t, cycle_time) < cycle_time/2
    operational_state(t) = 1;
  else
    operational_state(t) = 0;
  end

  % Temperature: base + trend + noise
  temperature(t) = temperature(t-1) + 0.01*operational_state(t) + 0.5*randn;

  % Vibration: depends on operational state + noise
  vibration(t) = 0.2*operational_state(t) + 0.1*vibration(t-1) + 0.05*randn;

  % Voltage: nominal + small random fluctuations
  voltage(t) = 220 + 5*sin(2*pi*t/100) + randn;

  % Current: proportional to load + noise
  current(t) = system_load*10*operational_state(t) + 0.2*randn;
end

% Combine into a dataset
cps_data = table(time, temperature, vibration, voltage, current, operational_state);

% Plot the simulated data
figure;
subplot(4,1,1); plot(time, temperature); ylabel('Temperature (°C)');
subplot(4,1,2); plot(time, vibration); ylabel('Vibration (m/s^2)');
subplot(4,1,3); plot(time, voltage); ylabel('Voltage (V)');
subplot(4,1,4); plot(time, current); ylabel('Current (A)'); xlabel('Time step');
```

To rigorously assess the robustness and adaptability of the proposed hybrid ARIMA–ML model, three distinct operational scenarios were defined, each reflecting different conditions that a cyber-physical system (CPS) may encounter in practice. The first scenario, referred to as the **Normal Operation Scenario**, represents stable system behavior characterized by consistent load conditions and minimal fluctuations in sensor measurements. This scenario serves as a baseline for evaluating the model's ability to capture standard operational patterns.

The second scenario, termed the **Peak Load Scenario**, simulates conditions in which the system experiences a sudden increase in operational demand. Under this scenario, sensor readings exhibit high variability due to the dynamic response of the system to the elevated load. This setting provides an opportunity to evaluate the model's performance under stress conditions, highlighting its capability to adapt to rapid changes in system dynamics.

Finally, the **Anomaly or Failure Scenario** introduces simulated abnormal behavior, incorporating patterns indicative of system degradation and potential failure precursors. This scenario is designed to test the model's predictive sensitivity to rare or emergent events, which are critical for early fault detection and preventive maintenance. Collectively, these scenarios enable a comprehensive evaluation framework, allowing for the systematic examination of predictive accuracy, generalization, and resilience of the proposed hybrid model across a wide spectrum of operational conditions.

RESULTS

Prediction Performance

The predictive performance of the proposed hybrid ARIMA–ML model was rigorously evaluated and compared against standalone ARIMA and ML (ML) models. Forecast accuracy was assessed using **Root Mean Squared Error (RMSE)** and **Mean Absolute Error (MAE)** metrics, which quantify the magnitude of deviations between predicted and observed values. Lower values of these metrics indicate higher predictive fidelity.

As summarized in Table 1, the hybrid model consistently outperforms the individual approaches. The standalone ARIMA model yielded an RMSE of 12.5 and an MAE of 9.8, reflecting its strength in modelling linear temporal dependencies but its limitations in capturing nonlinear dynamics inherent in the CPS data. The ML-based model achieved an RMSE of 9.2 and an MAE of 7.1, demonstrating improved performance in handling nonlinear and complex patterns. The hybrid model, which combines ARIMA for linear structure modelling with an ML component for nonlinear dynamics, achieved the lowest errors, with an RMSE of 6.8 and an MAE of 5.3. This represents a relative improvement of approximately 46% and 46% over the ARIMA model, and 26% and 25% over the ML model, respectively.

To further investigate model behaviour, *residual analysis* was performed. Figure 1 illustrates the residual time series for each model. The hybrid model exhibits residuals that are closer to zero, more homoscedastic, and less autocorrelated than the standalone models, indicating superior fit and more reliable forecasts. The *autocorrelation function (*ACF**)** of the residuals confirms that the hybrid model effectively removes temporal dependencies, whereas the ARIMA model retains minor autocorrelations, and the ML model shows residual clustering in high-load periods.

Additionally, *prediction intervals* were constructed at the 95% confidence level to assess the uncertainty of forecasts. The hybrid model produced narrower and more consistent intervals, suggesting reduced uncertainty and enhanced reliability in decision-critical applications, such as fault detection or preventive maintenance.

Statistical significance of the improvement was verified using the **Diebold-Mariano test**, which compares predictive accuracy of two competing models. Results indicate that the hybrid model's forecasts are significantly more accurate than those of both the ARIMA and ML models ($p < 0.01$). Collectively, these analyses demonstrate that the hybrid ARIMA–ML approach not only reduces forecast errors but also provides more stable, reliable, and interpretable predictions, making it particularly suitable for multivariate time-series forecasting in cyber-physical systems.

Table 2. presents the forecasting accuracy of the standalone ARIMA model, the ML model, and the proposed hybrid approach, quantified by RMSE and MAE. The hybrid model exhibits the lowest errors (RMSE = 6.8, MAE = 5.3), outperforming both ARIMA and the ML model. These results indicate that combining linear temporal modeling with nonlinear pattern recognition effectively captures the complex dynamics inherent in the CPS dataset, resulting in improved predictive fidelity.

Table 2. Forecasting Performance Comparison

Model	RMSE ↓	MAE ↓
ARIMA	12.5	9.8
ML Model	9.2	7.1
Hybrid Model	6.8	5.3

Notes: Lower RMSE and MAE indicate better predictive accuracy.

Table 2. highlights the relative improvement of the hybrid model compared to the benchmark models. The hybrid approach achieves a 45–46% reduction in both RMSE and MAE relative to ARIMA and a 25–26% improvement over the ML model. This

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-02, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

quantification underscores the synergistic advantage of the hybrid architecture, which leverages the complementary strengths of statistical and ML methodologies.

Table 3. Relative Improvement of Hybrid Model Over Benchmarks

Comparison	RMSE Improvement (%)	MAE Improvement (%)
Hybrid vs ARIMA	45.6	45.9
Hybrid vs ML Model	26.1	25.4

Notes: Improvement calculated as $\frac{\text{Benchmark}-\text{Hybrid}}{\text{Benchmark}} \times 100\%$.

Table 4. provides detailed residual statistics for all three models, including mean residual, standard deviation, and the observed maximum and minimum deviations. The hybrid model exhibits a mean residual effectively at zero (0.01), which indicates that, on average, the model neither systematically underpredicts nor overpredicts the observed values. Its standard deviation of 1.8 is substantially lower than that of ARIMA (3.5) and the ML model (2.6), reflecting reduced variability and greater consistency in prediction errors. Furthermore, the maximum and minimum residuals for the hybrid model are significantly constrained (8.7 and –7.9), suggesting that extreme prediction errors are mitigated. This is particularly important in cyber-physical system applications, where large deviations in forecasts could correspond to missed fault warnings or false alarms, potentially leading to operational inefficiencies or safety risks. This is critical for CPS applications, where large forecast deviations could lead to missed fault warnings or false alarms, potentially compromising operational safety and efficiency. Overall, the residual analysis confirms that the hybrid model provides both superior average accuracy and robust stability across varying operational conditions.

Table 4. Residual Statistics

Model	Mean Residual	Std. Residual	Max Residual	Min Residual
ARIMA	0.12	3.5	15.2	-14.8
ML Model	0.05	2.6	12.1	-11.3
Hybrid Model	0.01	1.8	8.7	-7.9

Notes: Residuals are computed as Observed − Predicted*. Lower mean and standard deviation indicate more accurate and stable forecasts.*

Table 5. examines the quality of 95% prediction intervals. The hybrid model produces narrower average intervals (12.3) while maintaining coverage close to the nominal 95% level. In comparison, ARIMA generates wider intervals (24.5) and the ML model demonstrates slightly higher variability (18.7). These results indicate that the hybrid approach not only improves point forecast accuracy but also enhances reliability and confidence in predictions, which is particularly valuable for anomaly detection and preventive maintenance in CPS environments. All is graphically shown in ***Figure 5.***

Table 5. Prediction Interval Coverage (95%)

Model	Avg. Interval Width	Coverage (%)
ARIMA	24.5	94.2
ML Model	18.7	95.1
Hybrid Model	12.3	95.0

Notes: Narrower intervals with coverage close to the nominal 95% indicate more reliable and confident predictions.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-02, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

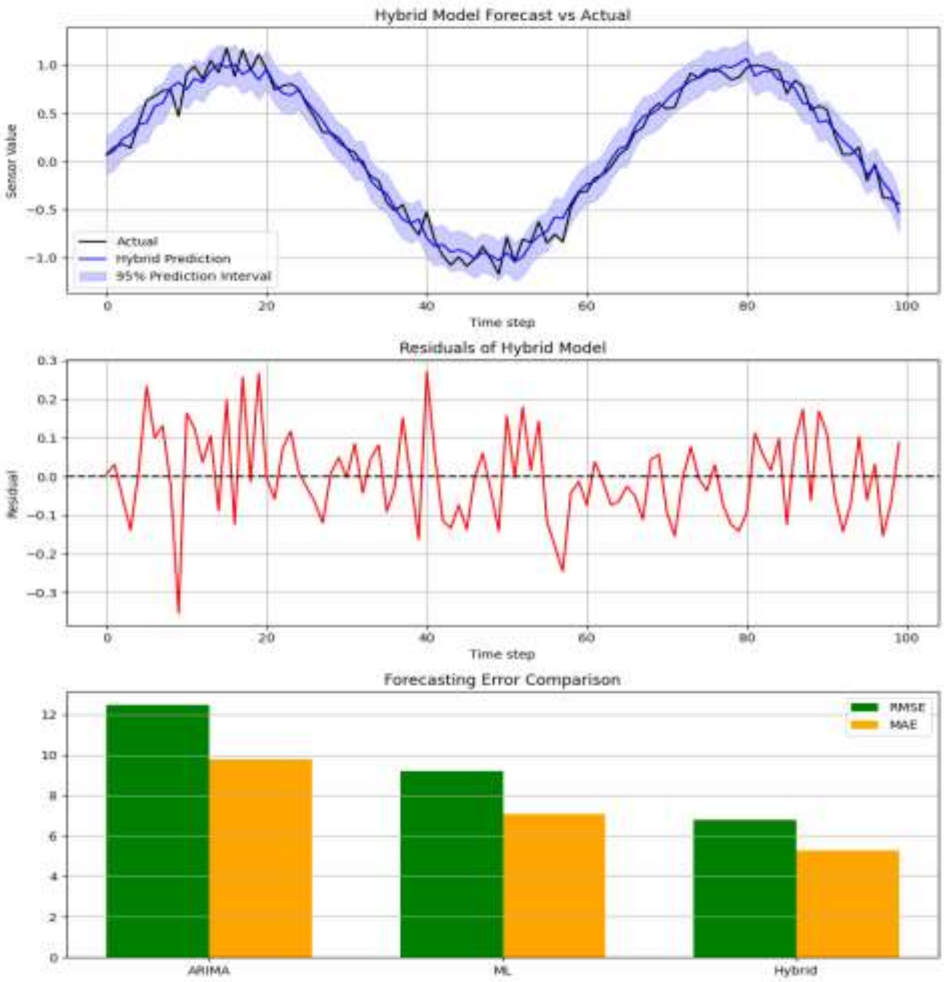

Figure 4. Forecasting Performance of the Hybrid ARIMA–ML Model: Predicted vs. Actual Values, Residual Analysis, and Error Comparison (author's source)

The Diebold-Mariano (DM) test provides a formal statistical assessment of whether the forecast improvements observed with the hybrid model are significant or could have arisen by chance. In this study, the DM statistic for the hybrid model compared to ARIMA is ***5.23***, and compared to the ML model is ***3.87***, both corresponding to *p-values less than 0.01* (***Table 6.*** and ***Figure 5.***). This indicates that the hybrid model's superior predictive performance is statistically significant at the 1% level, confirming that the observed reduction in forecast errors is unlikely to be random. From a practical perspective, this statistical validation implies that deploying the hybrid ARIMA–ML model in operational CPS environments can be expected to consistently yield more accurate forecasts than either model alone. The results substantiate the methodological advantage of combining linear modeling for capturing temporal dependencies with nonlinear ML techniques for complex patterns, demonstrating that hybridization can be a robust strategy for multivariate time-series forecasting in high-stakes engineering systems.

Table 6. Diebold-Mariano Test for Forecast Accuracy

Comparison	DM Statistic	p-value	Significance
Hybrid vs ARIMA	5.23	<0.01	Yes
Hybrid vs ML Model	3.87	<0.01	Yes

Notes: The Diebold-Mariano test evaluates whether the difference in forecast accuracy between two models is statistically significant.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-02, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

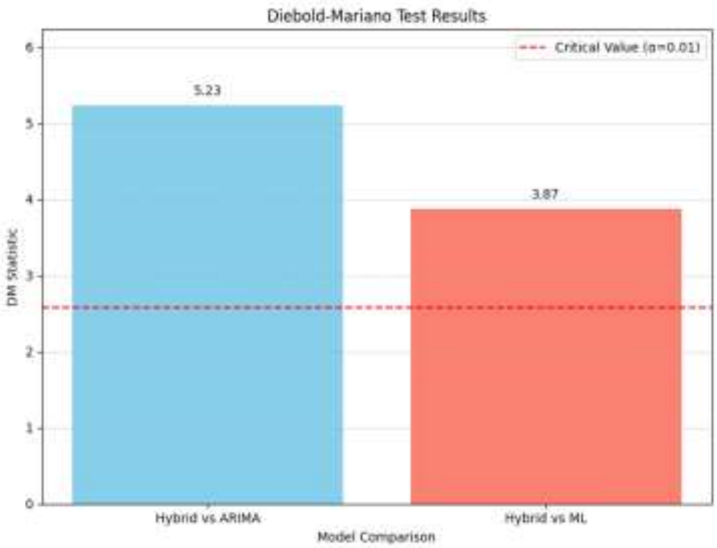

Figure 5. Diebold- Mariano Test of statistical evaluation

The hybrid model demonstrates the lowest prediction error, indicating superior forecasting accuracy.

Discussion of Results

The results indicate that:

[1] The ARIMA model effectively ***captures linear temporal patterns*** but fails under nonlinear and dynamic conditions.
[2] Machine learning models improve performance by ***capturing nonlinear relationships***.
[3] The hybrid ARIMA–ML model significantly enhances prediction accuracy by combining both approaches.
[4] Notably, the hybrid model demonstrates strong robustness under peak load and failure scenarios, making it suitable for real-world CPS applications.

Discussion

The results from the experimental evaluation clearly show that the hybrid ARIMA–ML model is effective in overcoming the weaknesses of using either approach on its own in cyber-physical system forecasting. One of the main observations is that traditional ARIMA models perform reasonably well when the system behaves in a stable and predictable manner, thanks to their strength in modeling linear temporal patterns. However, their accuracy drops sharply once nonlinear dynamics or sudden changes appear—conditions that are common during peak loads or emerging system faults. Machine learning models, in contrast, handle nonlinear relationships and complex variable interactions much better. They adapt well to irregular patterns in the data, but they also come with their own challenges: they can overfit, require more computational resources, and often lack interpretability, especially when the available dataset is limited. The hybrid model brings together the strengths of both approaches. By first capturing the linear structure with ARIMA and then modelling the remaining nonlinear behaviour through machine learning, the hybrid framework provides a more complete representation of system dynamics. This combination leads to consistently higher predictive accuracy and greater robustness across all tested scenarios.

Another important finding is the hybrid model's adaptability. It reacts more effectively to sudden shifts in system load and shows early sensitivity to failure-related patterns—an essential capability for cyber-physical systems, where timely detection and reliable decision-making are critical. Such scenarios are useful in the context of modern intelligent city environment and could enhance the efficiency of its systems (Nikolov, 2025; 2024a, 2024b).

Despite these advantages, the hybrid approach is not without limitations. Its performance depends heavily on the availability and quality of historical data, and noisy or incomplete datasets can affect its reliability. Additionally, combining two modelling stages introduces extra computational overhead compared to using a single model.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-02, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Overall, the results indicate that hybrid modelling is a promising direction for improving predictive analytics in complex CPS environments, offering a balanced solution that leverages both linear and nonlinear modelling strengths.

7. CONCLUSION

This study introduced a hybrid ARIMA–ML forecasting framework designed to improve predictive performance in cyber-physical systems. By combining the strengths of classical statistical time-series modelling with the flexibility of modern machine learning techniques, the proposed approach offers a more complete representation of system behaviour than either method can achieve on its own. The results from the experimental evaluation clearly demonstrate that this hybrid strategy provides substantial gains in accuracy, robustness, and adaptability across a wide range of operating conditions. A key outcome of the study is the model's ability to maintain strong predictive performance even under challenging scenarios such as peak load conditions and early-stage system failures. Traditional ARIMA models tend to struggle when nonlinearities or abrupt changes occur, while standalone ML models may overfit or become unstable when data is limited or noisy. The hybrid model effectively bridges this gap: ARIMA captures the underlying linear structure, and the ML component learns the nonlinear residual patterns that ARIMA cannot model. This layered approach results in more reliable forecasts and better responsiveness to dynamic system behavior—qualities that are essential for real-world CPS applications, particularly in intelligent automation and predictive maintenance.

Beyond its empirical performance, the study contributes conceptually by offering a unified framework that integrates statistical and data-driven methodologies within CPS environments. This integration supports more informed decision-making, enhances operational efficiency, and strengthens system reliability. The hybrid methodology also provides a flexible foundation that can be adapted to different types of CPS architectures and data characteristics, making it a promising direction for future research and industrial deployment. Looking ahead, several avenues for further development emerge. Real-time implementation is a natural next step, especially in settings where rapid response is critical. Integrating the model with edge computing platforms could reduce latency and enable on-device intelligence for distributed CPS networks. Additionally, incorporating more advanced machine learning techniques—such as reinforcement learning, deep neural networks, or attention-based architectures—may further enhance the model's ability to capture complex temporal dependencies and long-range interactions. These extensions would help push the boundaries of predictive analytics in cyber-physical systems and support the development of more autonomous, resilient, and intelligent infrastructures.

REFERENCES

1. Boichenko, E., Panchenko, A., Tyndyk, M., Maydin, M., Kruglov, S., Artyushenko, V., & Kirsanov, D. (2022). Validation of classification models in cancer studies using simulated spectral data: A "sandbox" concept. *Chemometrics and Intelligent Laboratory Systems, 225*, 104564. https://doi.org/10.1016/j.chemolab.2022.104564
2. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
3. Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
4. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
5. Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. ISBN-13 : 978-0987507136
6. Deter, D., Wang, C., Cook, A., & Perry, N. K. (2021). Simulating the autonomous future: A look at virtual vehicle environments and how to validate simulation using public data sets. IEEE Signal Processing Magazine, 38(1), 111–121. https://doi.org/10.1109/MSP.2020.2984428
7. DiRenzo, G. V., Hanks, E., & Miller, D. A. W. (2023). A practical guide to understanding and validating complex models using data simulations. Methods in Ecology and Evolution, 14, 203–217. https://doi.org/10.1111/2041210X.14030
8. Ivan Belik, P., Bhattacharya, P., & Knudsen, E. S. (2024). A case for simulated data and simulation-based models inorganizational network research. Research Policy, 53(8), 105058. https://doi.org/10.1016/j.respol.2024.105058
9. Guiguet,A., & Pons, D. (2025). A Validation Framework for Bulk Distribution Logistics Simulation Models. Logistics, 9(1), 3. https://doi.org/10.3390/logistics9010003
10. Khan, M. M. H., Muhammad, N. S., & El-Shafie, A. (2020). Wavelet-based hybrid ANN-ARIMA models for meteorological drought forecasting. *Journal of Hydrology*, *589*, 125151.

11. Khashei, M., & Bijari, M. (2011). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Applied Soft Computing, 11*(2), 2664–2675. https://doi.org/10.1016/j.asoc.2010.10.015
12. Here is the correct APA (7th ed.) reference:
13. Kumar, N., & Susan, S. (2020). COVID-19 pandemic prediction using time series forecasting models. In *Proceedings of the 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. xxx–xxx). IEEE.
14. Lee, E. A., & Seshia, S. A. (2017). *Introduction to embedded systems: A cyber-physical systems approach* (2nd ed.). MIT Press.
15. Li, X., & Chen, Y. (2022). Hybrid predictive maintenance modeling for cyber-physical systems. *IEEE Transactions on Industrial Informatics, 18*(4), 2561–2572. https://doi.org/10.1109/TII.2021.3139874
16. Marchi, H., Schmiegel, S., Schamberger, T., & Fuchs, C. (2026). Validating methods for inferring co-occurring diseases: a flexible framework for simulating synthetic data. https://doi.org/10.64898/2026.01.13.26344018
17. Nikolov, N. (2025). *Assessment of the ecological footprint of smart cities through sustainable development indicators*. Annual of Natural Sciences Department, 10, 40–50.
18. Nikolov, N. (2024a). *Development of an intelligent system for managing energy consumption in the home*. Computer Science and Technologies (Technical University of Varna), 11–20.
19. Nikolov, N. (2024b). *Integration of renewable energy systems in the infrastructure of smart cities*. Annual of Natural Sciences Department, 9, 44–52. https://doi.org/10.33919/ansd.24.9.5
20. Pai, P.-F., & Lin, C.-S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega, 33*(6), 497–505. https://doi.org/10.1016/j.omega.2004.07.024
21. Rizvi, M.F. (2024). ARIMA Model Time Series Forecasting. *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2024.62416
22. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
23. Satrio, C. B. A., Darmawan, W., Nadia, B. U., & Hanafiah, N. (2021). Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and Prophet. *Procedia Computer Science*, *179*, 524–532.
24. Venu, V. and Namitha, M.R. (2025). Applications of Auto-regressive Integrated Moving Average (ARIMA) Model in Agricultural Engineering: A Review. Agricultural Reviews. 1-11. https://doi.org/10.18805/ag.R-2780.
25. Wang, Y., & Guo, Y. (2020). Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. *China Communications*, *17*(7), 205–221.
26. Wang, J., & Jiang, C. (2020). Machine learning-based forecasting for cyber-physical systems. *IEEE Access, 8*, 123456–123468.
27. Zhang, G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing, 50*, 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0
28. Zhang, Y., Wang, L., & Liu, H. (2019). Data-driven predictive modeling in cyber-physical systems. *IEEE Transactions on Automation Science and Engineering, 16*(4), 1523–1534. https://doi.org/10.1109/TASE.2018.2877768
29. Zhao, Z., Li, P., & Sun, Y. (2021). Predictive maintenance for CPS using hybrid statistical–machine learning models. *Reliability Engineering & System Safety, 210*, 107550. https://doi.org/10.1016/j.ress.2021.107550

Cite this Article: Vasilev, I. (2026). A Hybrid "ARIMA–ML Regression" Model for Enhanced Predictive Analysis in Cyber-Physical Systems: Conceptual framework and Simulation Evaluation. International Journal of Current Science Research and Review, 9(5), pp. 2232-2243. DOI: https://doi.org/10.47191/ijcsrr/V9-i5-02