Forecasting of Biogas Production Using Advanced Time Series Algorithms

Nalluri Rishi Chaitanya Sri Prasad[1], Prabakaran Ganeshan[2], Karthik Rajendran[2 *]

[1] *Department of Computer Science and Engineering, School of Engineering and Sciences, SRM University-AP, Amaravati, Andhra Pradesh 522240, India.*

[2] *Department of Environmental Science and Engineering, School of Engineering and Sciences, SRM University-AP, Amaravati, Andhra Pradesh 522240, India.*

**Corresponding author Email: rajendran.k@srmap.edu.in*

Abstract

The anaerobic digestion (AD) process poses challenges in maintaining process stability and time series-based prediction and forecasting due to the intricate nature of the system. Process instability is a consequence of the unpredictability in the raw material received at the facility, as well as temperature fluctuations and pH changes resulting from microbiological processes. Consequently, it is necessary to implement constant monitoring and control measures for higher biogas production. The challenges associated with anaerobic digestion (AD) systems can be effectively addressed through the integration of advanced machine learning (ML) algorithms and industry 4.0 systems within biogas facilities. This integration holds the potential to enhance system efficiency and enable on-site control capabilities. Machine learning (ML) based solutions have the potential to enhance process performance in AD facilities, leading to improved system operation and maintenance. The present study focuses on advanced ML techniques, specifically time series algorithms (ARIMA and SARIMAX), have been employed to forecast the daily biogas production. These algorithms are trained to discern critical process parameters and forecast daily biogas production rates, measured in Liters. For forecasting 117 days of experimental data

used and identified ARIMA was best algorithm to forecast the daily production. This algorithm excelled not only in predicting biogas production but also in forecasting yield, resulting in a Root Mean Square Error (RMSE) of 3.26. Furthermore, a comparison between the forecasted values of both ARIMA and SARIMAX was conducted. The predictive ARIMA model underwent statistical validation with unknown data, resulting in a P-value is >0.05.

Keywords:

Biogas; Anaerobic digestion; Modelling and prediction; Forecasting; Time series; Machine learning

Highlights

- Forecast the biogas production with time series algorithms using ARIMA and SARIMAX.
- Feature Engineering and Hyperparameter tuning is done for better modelling.
- Predicted and forecasted biogas production with an RMSE of 3.26 and 24.02 respectively.
- By using this model, Biogas production can be forecasted for only the shorter period.

46 # List of Tables

50

List of Figures

List of Equations

List of Abbreviations

AD	Anaerobic Digestion
SRT	Sludge Retention Time
VS	Volatile Solids
TS	Total Solids
BOD	Biochemical Oxygen Demand
COD	Chemical Oxygen Demand
SS	Total Suspended Solids
TN	Total Nitrogen
TP	Total Phosphorus
VFA	Volatile Fatty Acids
HRT	Hydraulic retention time
DNN	Deep Neural Networks
ML	Machine Learning
ARIMA	Autoregressive integrated moving average
AR	Autoregressive
MA	Moving Average
I	Integrated
SARIMAX	Seasonal Autoregressive integrated moving average
RMSE	Root Mean Square Error
NRMSE	Normalized Root Mean Square Error
MAE	Mean Absolute Error
MW	Megawatt

CAGR	Compound Annual Growth Rate
LSTM	Long Short-Term Memory
DA-LSTM	Dual-stage-attention based long short-term memory
DA-LSTM-VSN	Dual-stage-attention based long short-term memory integrated with variable selection networks.
RF	Random Forest
KNN	K-Nearest Neighbors
SVM	Support Vector Machine
ANN	Artificial Neural Network
Xgboost	Extreme Gradient Boosting
SELU	Scaled Exponential Linear Units
RMSprop	Root Mean Squared Propagation
AIC	Akaike information criterion
BIC	Bayesian information criterion
ADF	augmented Dickey-Fuller test
PACF	Partial Autocorrelation Function
ACF	Autocorrelation Function
NAN	Not a Number

1. Introduction

Utilizing the promising approach like anaerobic digestion (AD) for the manufacturers achieve the production of biogas, guaranteeing improved yields. The process of anaerobic digestion involves microorganisms like acetogens, acidogens, and methanogens breaking down organic matter such as animal manure, wastewater, and food waste, in the absence of oxygen (Nguyen et al., 2019). AD is efficiently designed and free from risks, ensuring that it structures its processes in a manner that preserves human health. The outcomes of the anaerobic digestion process were, biogas (energy) and digestate (manure). The central primary focus revolves around biogas production. Biogas consists of 25% - 45% carbon dioxide and 50% - 70% methane and some other traces gas. The usage of biogas finds extensive application in the heating, electricity, and transportation sectors, serving as renewable fuels (Abanades et al., 2022).

In 2013, the worldwide recorded biogas production capacity reached approximately 14.173 MW, and this capacity increased to 18.505 MW by 2018, signifying a compound annual growth rate (CAGR) of 30.69%. By 2022, the capacity further increases to 21.512 MW, demonstrating a production rate increase of 16.24% between 2018 and 2022. Over the span of 10 years, from 2013 to 2022, the potential for biogas production escalation is estimated to be around 51.84%. This decade-long surge in biogas production exemplifies a substantial achievement in growth. The potential of biogas is widely spread for the utilization across diverse sectors. The distribution of biogas percentages across sectors indicates allocations of 45% in electricity, 35% in heating, and 10% in transport. In 2022, the global biogas market held a valuation of $71.59 billion, and this value escalated to $78.25 billion in 2023, showcasing a growth rate of 9.3% in CAGR. The projection for global market growth in 2027 stands at an estimated $102.7 billion, featuring a CAGR of 7.0%.

116 **Table 1.** Similar studies have been reported in literature.

Algorithm used	Parameter considered	Best Algorithm	Emphasis of the study	Model Evaluation	Author(s)
DA-LSTM-VSN, DA-LSTM, LSTM	Continuous: Qsludge, Qsludge, SRT, temperature. Discontinuous: VS/TS ratio, BOD, COD, SS, Total nitrogen (TN), and total phosphorus (TP)	DA-LSTM-VSN	Prediction of biogas production in anaerobic co-digestion of organic wastes using deep learning models	R^2 = 0.76 NRMSE = 0.09	(Jeong et al., 2021)
RF, KNN, SVM, ANN, Xgboost	%TS, %VS, temperature, pH, alkalinity, VFA, CH_4 and CO_2.	RF	Prediction of biogas production of industrial scale anaerobic digestion plant by machine learning algorithms	R^2 = 0.92 RMSE=1405.8	(Yildirim and Ozkaya, 2023)
DNN	COD, BOD, TSS, PH, T, Total Nitrogen (TN), Total Phosphorus (TP)	DNN	DNN model development of biogas production from an anaerobic wastewater treatment plant using Bayesian hyperparameter optimization	R^2 = 0.712	(Sadoune et al., 2023)
Extra Tree	TS%, VS, VFA mg/L, ALK mg/L, and VFA/ALK ratio	Extra Tree	Tree-Based Automated Machine Learning to Predict Biogas Production for Anaerobic Co-digestion of Organic Waste	R^2 = 0.72	(Wang et al., 2021)
KNN, ARIMA, Decoder-ANN, Temporal Fusion Transformer	Raw sludge dry matter load, Sludge loss on ignition, Hydraulic retention time, Overnight stay, Raw sludge dry matter, Temperature, Sludge dry matter load, public holiday, Time index, Ambient temperature,	Temporal Fusion Transformer	Machine learning for quantile regression of biogas production rates in anaerobic digesters	RMSE=246.1	(Sappl et al., 2023)

	Co-fermentation biowaste, Month, Sludge total, Sludge, Raw sludge, pH value, Sludge dry matter, Weekday				

117

118 In literature, the authors showed the comparison of biogas predictions using various
119 algorithms are represented in Table 1. (Jeong et al., 2021) They assembled time series data to
120 facilitate the training of diverse algorithms in deep learning. For both continuous and
121 discontinuous time series data, they utilized DA-LSTM-VSN as the optimal model for biogas
122 prediction, achieving a test R^2 of 0.76 and a normalized root mean square error (NRMSE) of
123 0.68. Trained DA-LSTM with continuous input features, resulting in a test NRMSE of 0.10,
124 while LSTM, which was also trained with continuous input features, resulted in a test NRMSE
125 of 0.14. This model exclusively focused on biogas prediction and did not venture into future
126 forecasting. (Yildirim and Ozkaya, 2023) Named a total of 8 input features to predict biogas
127 production, utilizing 5 machine learning algorithms. Among these, Random Forest (RF) stood
128 out as the top performer, attaining an R^2 of 0.92 and an RMSE of 1405.84. Although the R^2
129 value showcases satisfactory performance, the notable error rate indicates instances of high
130 model overfitting. Furthermore, the model restricted its application to singular sample
131 predictions, and it did not include provisions for forecasting. (Sadoune et al., 2023) Deep
132 Neural Networks (DNN) to train the model using 7 parameters. To enhance data modelling
133 through scaling, utilized three different scaling techniques: Minmax scaler, Robustscaler, and
134 Standardscaler. Among these techniques, Robustscaler emerged as the most effective for the
135 neural network, resulting in an R^2 of 0.712, considering a dense layer of 1 and a dropout rate
136 of 0.144, activation function (SELU), and optimizer (RMSprop). The model was limited to
137 singular sample predictions without delving into forecasting. (Wang et al., 2021) Extra Tree
138 Regressor algorithm with 5 parameters as a tree-based model for biogas prediction. Artificial

Neural Network (ANN) modelled for result comparison with the Extra Tree algorithm. The tree-based algorithm yielded an R^2 of 0.72, whereas the ANN algorithm resulted in an R^2 of 0.56. Their research mainly focus was solely on prediction, without incorporating biogas forecasting into the study. (Sappl et al., 2023) Total of 4 prediction algorithms for biogas forecasting, with one of them being a time series algorithm, namely ARIMA. All these algorithms underwent consideration and were trained with the dataset. Among them, the Temporal Fusion Transformers algorithm yielded the best result, achieving an RMSE of 246.17, outperforming the other algorithms. On the other hand, ARIMA generated a prediction result with an RMSE of 569.67, which is notably high compared to all the other algorithms. It's important to note that ARIMA solely focused on prediction and didn't involve forecasting biogas production daily. Currently, no published work has achieved the development of a time series model for forecasting biogas production with significantly low RMSE and MAE. This is a crucial concern as the demand for biogas is rapidly increasing and holds high significance for the future. Various sectors will actively utilize biogas in the upcoming days.

In this study focuses on biogas production, utilizing a dataset covering 117 experimental days. Leveraging the advanced algorithm of time series within the realm of machine learning techniques, the primary focus is on forecasting and predicting model outcomes. Prediction involves determining a solitary value for a given parameter or day, while forecasting entails the continuous projection of outputs in accordance with the data sequence. Hyperparameter tuning was conducted to select optimal parameters for AIC and BIC, crucial for model training. Matplotlib library was utilized for clear graph visualization. Through the comparison of model evaluation, particularly RMSE and MAE, the model was chosen for both prediction and data forecasting. If the model's performance falls short, it's retrained with varying AIC and BIC values. The entire methodology of this paper is depicted in Figure 2.

2. Methodology

2.1 Data Collection and pre-processing

The data was collected from previously published research article from Scopus database (Bautista Angeli et al., 2022). which was conducted experiment on a batch mode to produce daily biogas production for a period of 117 days. This comprehensive experiment incorporated several crucial operating and process parameters, including VS consumed, initial pH, and final pH, Temperature, HRT (hydraulic retention time), CH_4 mean production, , CH_4 content. To facilitate a comprehensive analysis of biogas yield production, the experiment timeline was segmented into three distinct periods based on these parameters. In the initial period, first 12 days, an initial pH was 6.8 and a final pH of 5.5 were measured and a CH_4 composition in produced biogas was measured at 44% (±6%). After the completion of initial period, period 2 extended up to 63 days, continuously from 13 day. In this stage an initial pH of 5.8, a final pH was 7.0, and a % CH_4 (% biogas) was 49% (±8%). The subsequent phase, period 3, covered the duration from the end of period 2 to till 117th day, comprising around 54 days. During this final stage, the experiment adhered to an initial pH of 7.0 and a final pH of 7.0, and a measured % CH_4 (% biogas) content was 52% (±3%), as detailed in Table 2 (Bautista Angeli et al., 2022).

Table 2. Analytical follow-up of the co-digestion in the reactor

	CH4 mean production (mL/ gVS)	VS consumed (%)	VS consumed (%)	Initial pH	Final pH
Period 1	94 (±21)	94 (±1)	44 (±6)	6.8	5.5
Period 2	353 (±115)	83 (±7)	49 (±8)	5.8	7.0
Period 3	321 (±39)	83 (±7)	52 (±3)	7.0	7.0

The dataset encompasses 117 days, facilitating a day-by-day examination of biogas yield production data. We undertook data cleaning to address null values or missing entries within the dataset, utilizing the front fill method to ensure a comprehensive dataset for biogas yield analysis. This cleaned dataset was harnessed for the application of machine learning (ML) techniques, specifically time series analysis, to forecast and predict the biogas production. Two distinct time series algorithms, namely ARIMA and SARIMAX, were employed to predict or forecast biogas production. To performance of these models can evaluate by using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), for ensure the robust model assessment. Figure 1 illustrates the relationship between biogas production with the number of days. This visualization provides a clear overview of the trend over time. In Figure 2 understanding of the entire methodology, in-depth for visually outlines the comprehensive process employed in this study.

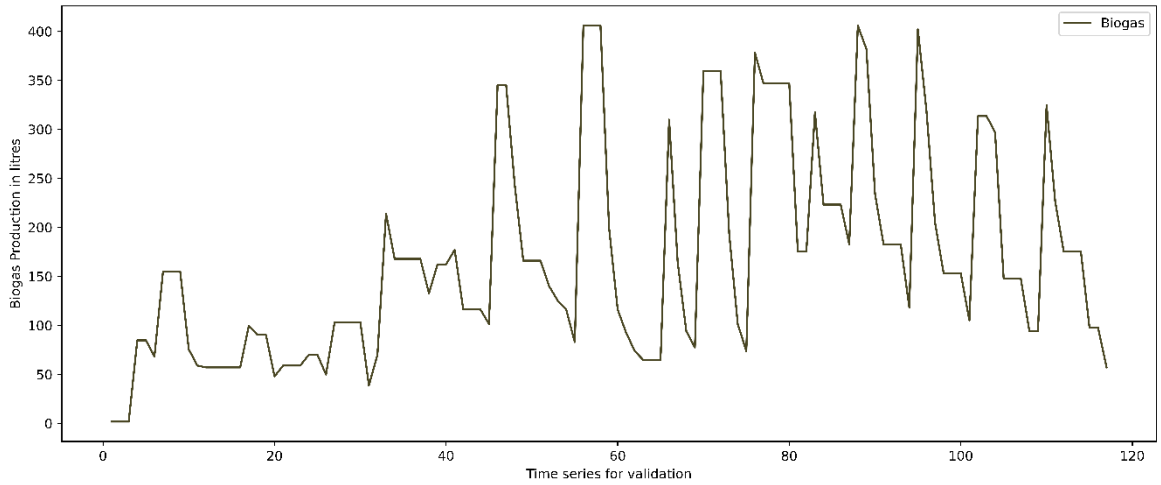


Figure 1. Biogas plot with respect to days

2.2 Time Series Algorithms

2.2.1 ARIMA

Autoregressive Integrated Moving Average (ARIMA) constitutes a generalized model derived from Autoregressive Moving Average (ARMA) in the realm of time series analysis. Application of algorithms are supposed to time-based data inputs predominantly called for time series algorithms (Brockwell and Davis, 2016a), with these being tailored to predict or forecast the patterns within the data. Often, such data exhibits non-stationary behavior, and is characterized by variations in statistical properties. To enable the effective use of time series models for forecasting, it's imperative to transform this data into a stationary form. Achieving stationarity ensures that the mean average of data points remains consistent throughout time. The augmented Dickey-Fuller test (ADF) serves as a reliable method to assess data stationarity. The p-value is equal to or less than 0.05, else not the data is deemed stationary. Conversely, when data lacks stationarity, differencing becomes a vital step to induce stationarity. This process is repeated iteratively until the data achieves the desired stationary state (Brockwell and Davis, 2016b).

The ARIMA framework contains three integral components that come into play: Autoregressive (AR), Integrated (I), and Moving Average (MA). The AR segment, denoted as "p" in Equation (1), embodies the autoregressive model. It signifies the number of observed lags in the model, commonly referred to as the lag order (Brockwell and Davis, 2016c).

AR(p) model:

$$y_t = c + \left(1 - \sum_{i=0}^{p} \varphi_i y_{t-i}\right) + \varepsilon_t \quad (1)$$

Where, φ_i = AR coefficients, y_t = Time series at time t, y_{t-i} = lagged values of time series, ε_t = error time at time t, c = constant.

The Moving Average (MA) component, denoted as "q" in Equation (2), represents the rolling window size, also referred to as the order of the moving average. This parameter defines the span of observations considered for calculating the moving average at a specific time point (Brockwell and Davis, 2016c).

MA(q) model:

$$y_t = c + \left(1 - \sum_{i=0}^{q} \theta_i \varepsilon_{t-i}\right) + \varepsilon_t \quad (2)$$

Where, y_t = Time series at time t, θ_i = MA coefficients, ε_{t-i} = residual error at a lagged period, ε_t = error time at time t, c = constant.

The selection of "p" and "q" values is derived from the Partial Autocorrelation Function (PACF) and Autocorrelation Function (ACF) plots from the data. These plots encompass the upper and lower boundaries. If the error bands cross these boundaries, they indicate time lags that significantly impact the time series. These crossing points dictate the values for "p" and "q". Specifically, "p" is determined from the PACF plots, and “q" is extracted from the ACF plots. Within the ARIMA framework, the "I" component signifies integration, denoted as "d" in Equation (3). This value represents the number of differencing operations applied to the data to achieve stationarity. Each difference operation reduces the data's non-stationary behavior, progressively rendering it suitable for time series analysis (Brockwell and Davis, 2016b).

$$L^d y_t = y_t - y_{t-d} \quad (3)$$

For the first differencing (d=1) of the data to make it stationary Equation (4),

$$y_t^1 = y_t - y_{t-1} \tag{4}$$

In some cases, second differencing (d=2) also needed to make data to stationary Equation (5)

$$\begin{aligned} y_t^2 &= y_t^1 - y_{t-1}^1 \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2} \end{aligned} \tag{5}$$

During this process, the application of differencing is pivotal to achieve data stationarity, enhancing the predictive capabilities for forecasting. ARIMA's effectiveness is notably pronounced in short-term prediction scenarios, relying primarily on past data for its forecasting abilities. The optimal selection of "p" and "q" values is determined through statistical analysis, specifically utilizing metrics like the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). These metrics gauge the model's fit by assessing the goodness of fit relative to the complexity of the model. Lower AIC and BIC values signify a better-fitting model. This approach, ARIMA constructs a predictive model capable of accurate forecasting based on the learned data. The ARIMA equation, denoted as ARIMA (p, d, q), is represented in Equation (6), encapsulating the interplay of autoregressive, differencing, and moving average components (Brockwell and Davis, 2016c).

ARIMA (p, d, q) model:

$$\left(1 - \sum_{i=0}^{p} \varphi_i L^i\right)(1 - L)^d y_t = \left(1 - \sum_{i=0}^{q} \theta_i L^i\right)\varepsilon_t \tag{6}$$

L = lag operator, d = differencing term.

2.2.2 SARIMAX

Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) extends from the foundation of Seasonal Autoregressive Integrated Moving Average (SARIMA) model. The inclusion of "X" in SARIMAX signifies its capability to handle exogenous variables. This algorithm is particularly effective for data displaying seasonality, making it a fitting choice for forecasting and prediction. Seasonal ARIMAX, denoted as SARIMA (p, d, q) (P, D, Q, s) in Equation (7), encompasses seasonal components in addition to the ARIMA parameters. The seasonal cycle length, represented as "s," typically corresponds to 12 months. The values "p," "d," and "q" are derived from the ACF and PACF plots of the ARIMA model (Brockwell and Davis, 2016c), indicated in Equations (1) & (2) & (3). Before modelling data need to achieve stationarity through differencing. Since SARIMAX is the algorithm with a seasonal cycle of 12 months requires a 12-month shift, differencing is performed iteratively to ensure stationarity. The number of times differencing is conducted is denoted as "D" in the SARIMAX model. The autoregressive model "P" in SARIMAX is determined by assessing the seasonal PACF plot, while the moving average "Q" is deduced from the seasonal ACF plot (Shumway and Stoffer, 2017).

SARIMA (p, d, q) (P, Q, D, s) model:

$$\left(1-\sum_{i=1}^{p}\varphi_i L^i\right)\left(1-\sum_{i=1}^{P}\Phi_i L^{is}\right)(1-L)^d(1-L^s)^D y_t = \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\left(1+\sum_{i=1}^{Q}\Theta_i L^{is}\right)X_t + \varepsilon_t \quad (7)$$

283 Φ_i = Seasonal AR coefficients, Θ_i = Seasonal MA coefficients, s= length of seasonal
284 period, X_t = exogenous variables.

285 2.3 Hyperparameter Optimization

286 Hyperparameter tuning is a technique for refining model performance and controlling the
287 learning process to train the model efficiently. It aids to identifying the optimal parameters,
288 leading to enhance the model performance. In the context of time series algorithm used for
289 data modelling and forecasting, finding the best autoregressive (p, P) and moving average (q,
290 Q) values, both for seasonal and non-seasonal components, holds utmost significance.
291 Optimized values of these parameters contribute to improved model fitting, enhancing
292 prediction and forecasting the outcomes. The selection process involves leveraging Akaike
293 Information Criterion (AIC) and Bayesian Information Criterion (BIC) techniques (Brockwell
294 and Davis, 2016a). These methods assist in identifying the most suitable AR (p, P) and MA (q,
295 Q) values for both ARIMA and SARIMAX models. The chosen values exhibit lower AIC and
296 BIC scores, aligning with better model performance. AIC and BIC values, calculated as per
297 Equation (8) and (9), carry no fixed range. Instead, they are influenced by the specific AR and
298 MA values derived from ACF and PACF plots. AIC and BIC scores need to be substantially
299 lower compared to alternatives, indicating the ideal choice of parameters for optimal
300 modelling.

301 Akaike information criterion (AIC):

$$AIC = 2k - 2\ln(L) \tag{8}$$

303 k = number of parameters in the model, L = max value of likelihood function of model

304 Bayesian information criterion (BIC):

$$BIC = k * \ln(n) - 2\ln(L) \tag{9}$$

306 n = number of observations.

307 2.4 Model Evaluation

308 Time series algorithms are utilized for both prediction and forecasting of the model.
309 These models were effectively implemented and adapted to the dataset. The model
310 performance was evaluated based on the metrics of Mean Absolute Error (MAE) and
311 Root Mean Square Error (RMSE) (James et al., 2013). MAE represents the average of
312 the absolute differences between the true and predicted values. As the difference
313 between the true and predicted values increases, the RMSE also increases, as
314 demonstrated in Equation (10).

315 $$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \quad (10)$$

316 Where, y_i = Actual values, $\hat{y}_i$ = Predicted values, n = Number of data points

317 Root Mean Square Error is employed to calculate the squared differences between the
318 actual and predicted values of the output variable. The RMSE provides insight into
319 the proximity of the actual data points to the predicted data points. A lower RMSE
320 signifies a minimal difference between the true and predicted values, as illustrated in
321 Equation (11).

322 $$RMSE = \frac{\sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}}{n} \quad (11)$$

323 y_i = Actual values, $\hat{y}_i$ = Predicted values, n = Number of data points

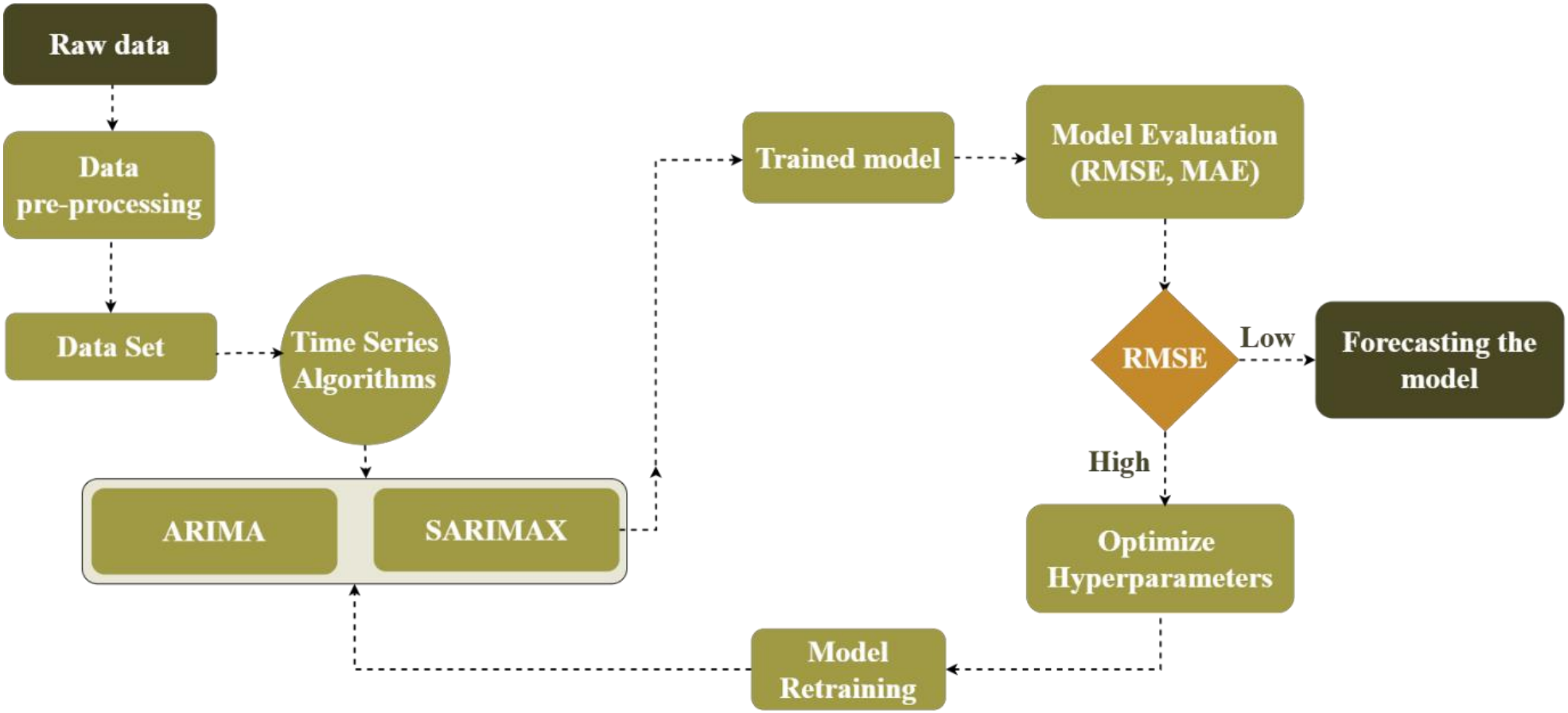


Figure 2. Overview of the Research Methodology Process.

3. Results and Discussion

3.1 Data Statistics before Modelling

The statistical analysis of the dataset involves 117 days of daily biogas production experiments. Throughout this period, the biogas produced ranged from a minimum of 1.86 to a maximum of 405.9 liters, with a mean production of 153.14 liters per day. The experiments were divided into three distinct periods, each characterized by specific parameters such as initial pH and final pH, which were detailed in Table 2. To address missing values, a front-fill method was employed. Additional insights into the data's distribution and variability can be assembled from the provided box plot in Figure 3.

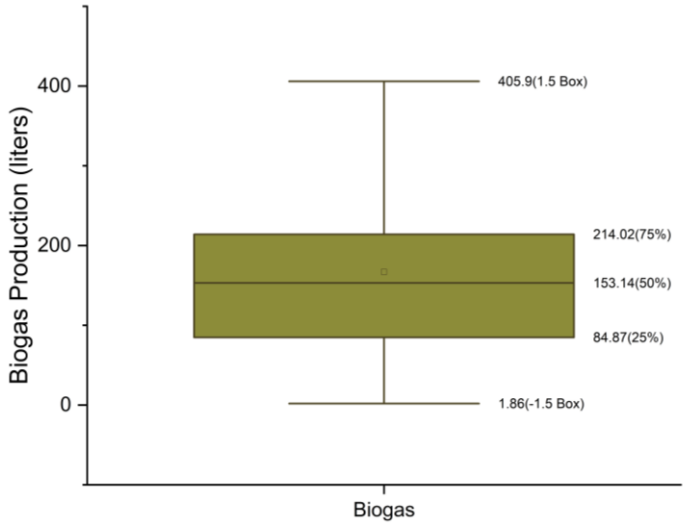


Figure 3. Statistics of the time series data

3.2 Seasonal Decomposition

Seasonal decomposition involves utilizing statistical analysis to dissect time series data into distinct components such as trend, seasonality, and residual patterns. These components offer valuable insights into the various variations present within the data, aiding in comprehensive analysis. Two primary models utilized for seasonal decomposition are Additive and Multiplicative (Cowpertwait and Metcalfe, 2009). Additive models are employed when the magnitude of seasonal data remains consistent throughout the entire period without notable fluctuations. This model aggregates the seasonal component, trend, and residual. The formulation of the additive model is presented as Equation (12).

$$y_t = T_t + S_t + R_t \tag{12}$$

The Multiplicative model comes into play when the magnitude of seasonal data exhibits changes throughout the time span. This model is employed in situations where the fluctuations in seasonal patterns are not consistent. It involves calculating the product of seasonal components, trends, and residuals. The formulation of the multiplicative model is represented as Equation (13).

$$y_t = T_t * S_t * R_t \tag{13}$$

T_t = Trend component, S_t = Seasonal component, R_t = Residual component.

In seasonal decomposition, the trend component signifies the smoothed trajectory of the time series data. This component offers insights into whether the time series exhibits an upward, downward, or steady trend. On the other hand, the Seasonal Component is utilized to identify recurring patterns that manifest at regular intervals

363 within the time series data. These patterns can be associated with specific days,
364 months, or years, and the periodicity can be adjusted based on analytical
365 requirements. The Residual Component captures the residual noise that remains once
366 the trend and seasonal patterns have been removed from the data. This segment
367 encompasses short-term fluctuations that cannot be explained by the underlying trend
368 and seasonality components.

369 We collected 117 daily experimental observations for the present time series data
370 concerning biogas production. Utilizing the multiplicative model for seasonal
371 decomposition was appropriate (Cowpertwait and Metcalfe, 2009), given the varying
372 nature of the data over time. Examining the decomposed components, the observed
373 component presented a normal distribution of the data. The trend component
374 depicted a discernible upward trajectory, indicating an increasing trend within the
375 time series. In terms of the seasonal component, recognizable patterns were evident,
376 showcasing periodicity in the data. The residual component, denoting the error, fell
377 within a range of 0.5 to 1.5, suggesting relatively minor discrepancies. The seasonal
378 decompose plot can be observed in Figure 4, visually encapsulating the outlined
379 components and their behavior.

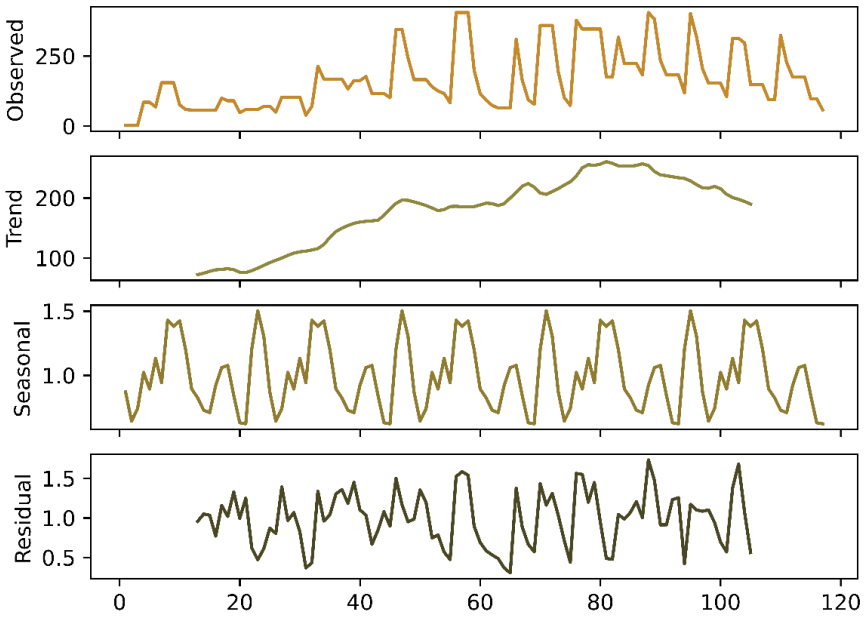


Figure 4. Seasonal decomposition of time series data.

3.3 Time Series Analysis

3.3.1 ADF (Augmented Dickey-Fuller test)

The Augmented Dickey-Fuller test (ADF) is utilized to determine whether the data exhibits stationarity. Stationarity ensures a constant average mean across all data points over time. In the ADF test, you calculate a p-value; if this value is less than or equal to 0.05, the data is considered stationary. To achieve stationarity, the technique of differencing is employed. In the context of predicting biogas using the ARIMA model, the ADF test initially yielded a p-value of 0.57, indicating non-stationarity of the data. The first difference was applied by shifting the data by 1. Subsequently, the ADF test for this first-differenced data revealed an exceptionally low p-value of 9.99e-07, confirming its stationarity. Transitioning to the Seasonal ARIMA (SARIMAX) model, which accounts for seasonality, a significant shift of 12 was required for

differencing due to its seasonal nature. Remarkably, after applying this seasonal difference, the SARIMAX model exhibited stationarity, supported by a p-value of 0.04.

3.3.2 ACF & PACF plots

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) serve as essential statistical analysis tools for enhancing time series models in the quest for improved prediction and forecasting accuracy. ACF plots play a pivotal role in determining the Moving Average (MA) model components, while PACF plots are instrumental in identifying the Autoregressive (AR) model components. These plots play a central role in the development of ARIMA and SARIMAX algorithms. Both ACF and PACF plots are accompanied by upper and lower bounds, which form a confidence interval. When the error bands within these plots intersect or cross over this confidence interval, it signifies the statistical significance of that lag. ACF plots encapsulate the correlation between a data point and its own past lags. These plots are particularly adept at unveiling patterns within the data. When ACF values exhibit a gradual decrease, it indicates the presence of a persistent pattern in the data. PACF plots, on the other hand, reveal the specific relationship between a data point and its immediately preceding value, effectively capturing the direct influence of past observations on the current point. The range of ACF and PACF values spans from -1 to 1. Where, -1 indicates a negative correlation, 0 signifies no correlation, and 1 represents a positive correlation (Shumway and Stoffer, 2017).

In the ARIMA (p, d, q) model, we posit an AR(p) value of 40 based on observations from the PACF plot. For the MA(q) component, we designate a value of 10, informed by insights gleaned from the ACF plot. As for the differencing aspect denoted by 'd',

a value of 1 is selected. This decision aligns with the need to attain data stationarity, which was achieved through a single differencing operation, as illustrated in Figure 5.

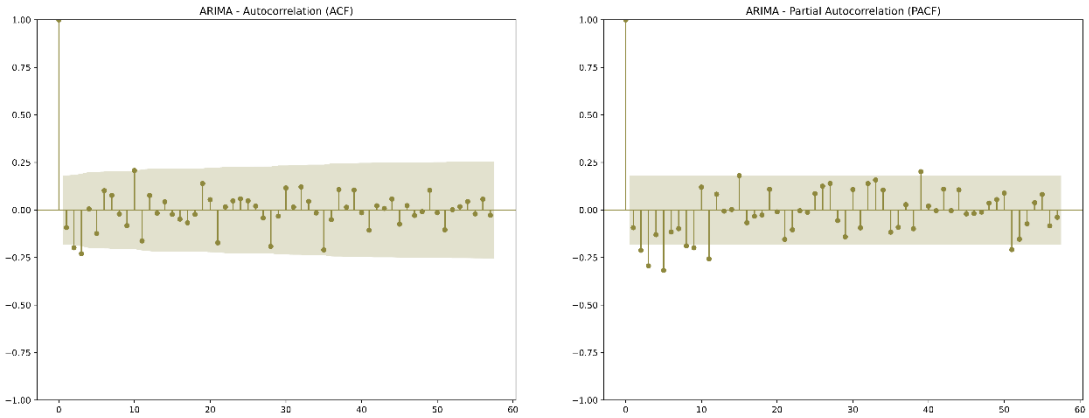


Figure 5. ACF and PACF plots of ARIMA

To build an accurate and efficient time series forecasting model, we also utilized the SARIMAX approach. The non-seasonal plots of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) guided us in determining the non-seasonal orders of differencing, represented as (2, 1, 2). For the seasonal component, we focused on the ACF and PACF plots after applying a seasonal shift of 12, corresponding to a yearly cycle. This allowed us to extract the seasonal orders of the SARIMAX model, which we established as (2, 1, 1, 12) Figure 6. For identifying the optimal parameters for our model, we relied on two key criteria: the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). By comparing different parameter combinations using these metrics, we were able to select the parameters that provided the best fit for the time series data. Overall, these comprehensive analyses of the non-seasonal and seasonal plots of ACF and PACF, coupled with the utilization of AIC and BIC, facilitated the development of a robust

time series forecasting model. This model, based on SARIMAX, ensures a precise and efficient prediction of future values, thereby enhancing our ability to make informed decisions based on the insights derived from the time series data (Brockwell and Davis, 2016a).

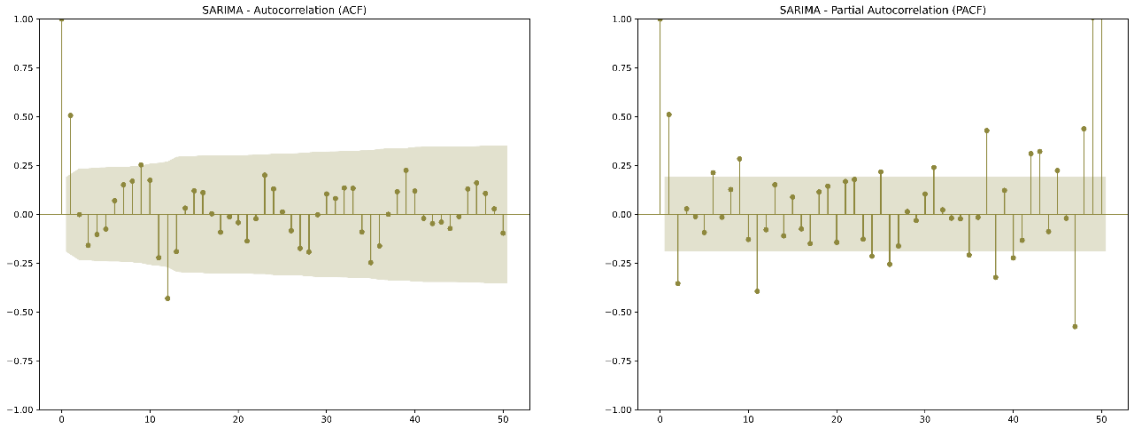


Figure 6. ACF and PACF plots of SARIMAX

3.3.3 Plot Diagnostics

The diagnostic analysis of the plotted statistics involves four main types of statistical graphs: Standardized residuals, Histogram plots, Normal Q-Q plot, and Correlogram. To assess the quality of the model's fit, various visualizations are utilized. Standardized residuals portray the discrepancies between actual values and predicted values. Calculating standardized residuals involves dividing normal residuals by the standard deviation of total residuals. This produces a standardized residual plot, illustrating the standardized differences in errors between actual and predicted values. Histogram plots showcase the distribution of residuals. The graph displays two curves: the normal distribution curve N (0,1) and the Kernel Density Estimate (KDE) curve. If the lines exhibit minimal disparity, the model is considered well-fitted. The KDE, a smoothed version of the histogram, generates a continuous representation

of data density. Particularly for ARIMA and SARIMAX models, a Gaussian kernel is employed for KDE plotting. Peaks in the KDE graph pinpoint areas of higher data density, while lower regions signify lower density. This visualization effectively communicates the data distribution's central tendency and variance. The Normal Q-Q plot gauges data distribution against the normal distribution. When data points align with the red reference line, it indicates normal residuals. Both the ARIMA (Figure 7) and SARIMAX (Figure 8) models closely adhere to this red line, signifying a nearly normal distribution of residuals.

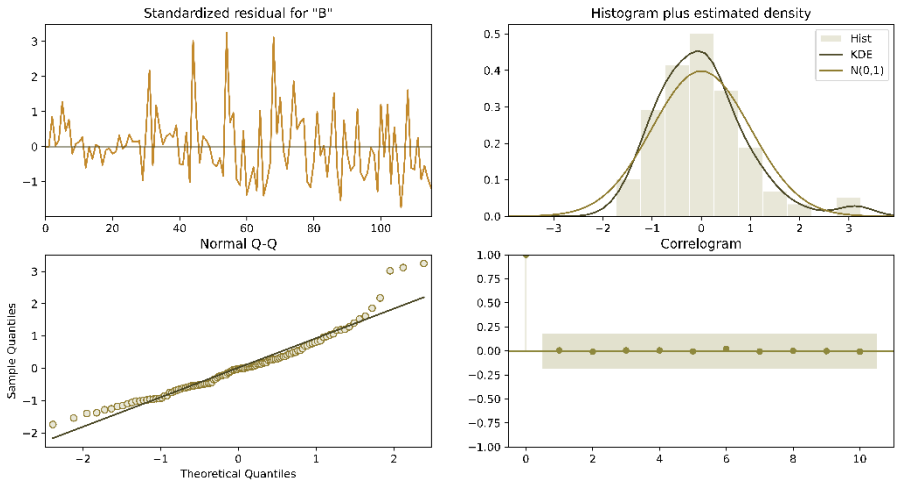


Figure 7. Plot diagnostics for ARIMA.

The Correlogram plot generates the Autocorrelation Function (ACF) plot using residuals instead of data. This plot reveals whether error bands fall within the confidence interval, indicating if residuals conform to a normal distribution. Consistency within the interval signifies normal distribution adherence, while deviations suggest potential data omission by the model. Notably, the lines of the ARIMA model fall insignificantly within the confidence interval, implying a lack of significance. Conversely, SARIMAX exhibits noticeable differences Figure 8. When

comparing both algorithms, it becomes evident that the ARIMA model predicts data more accurately than SARIMAX. Upon model evaluation, the ARIMA model demonstrates an RMSE of 3.26, whereas the RMSE for SARIMAX is 24.02. The superiority of the ARIMA model in this regard is apparent. These plot diagnostics leverage a variety of graphical tools to observe model residuals in time series analysis.

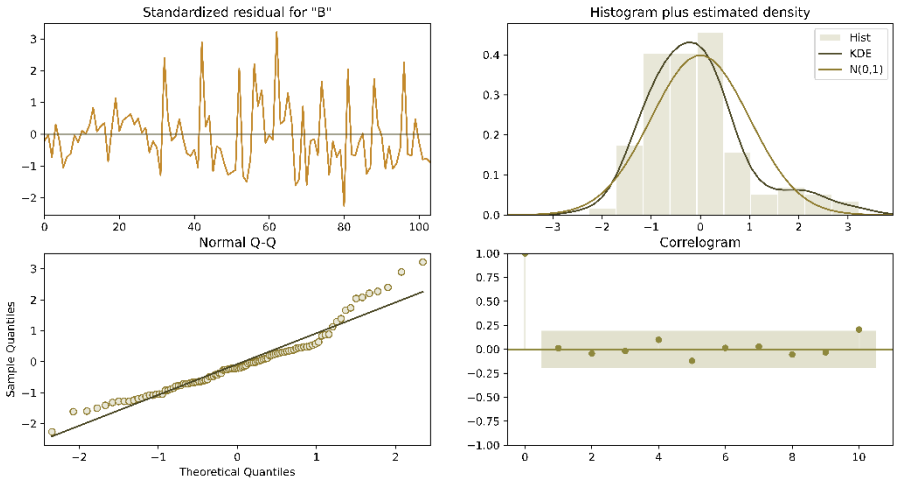


Figure 8. Plot diagnostics for SARIMAX

3.4 Comparison of Time Series algorithms

3.4.1 Prediction

We used time series analysis to predict the model by fitting two algorithms, namely ARIMA and SARIMAX, to the data. Both algorithms were applied to the output for prediction, specifically biogas production, with the p, d, q values determined from the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. After observing AIC and BIC values of 1402.4 and 1542 for the order (40, 1, 10), the ARIMA model was trained and found to be better fitted to the data. In the case of SARIMAX, the seasonal orders (P, D, Q) were interpreted from seasonal ACF and PACF plots, resulting in an order of (2, 1, 2) and (2, 1, 1, 12). The AIC and BIC values

were 1245.3 and 1266.5, respectively, for this SARIMAX configuration, which was then fitted and used for prediction. Both ARIMA and SARIMAX algorithms provided predictions that were well-fitted to the actual biogas production, as shown in Figure 9. The ARIMA model predicted daily biogas production with Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) values of 46.81 and 3.26 respectively. On the other hand, SARIMAX achieved MAE and RMSE values of 60.81 and 24.02.

Considering the time series modelling, both ARIMA and SARIMAX algorithms exhibited very low prediction errors for biogas production. This underscores the effectiveness of these algorithms in accurately forecasting biogas production based on the time series data.

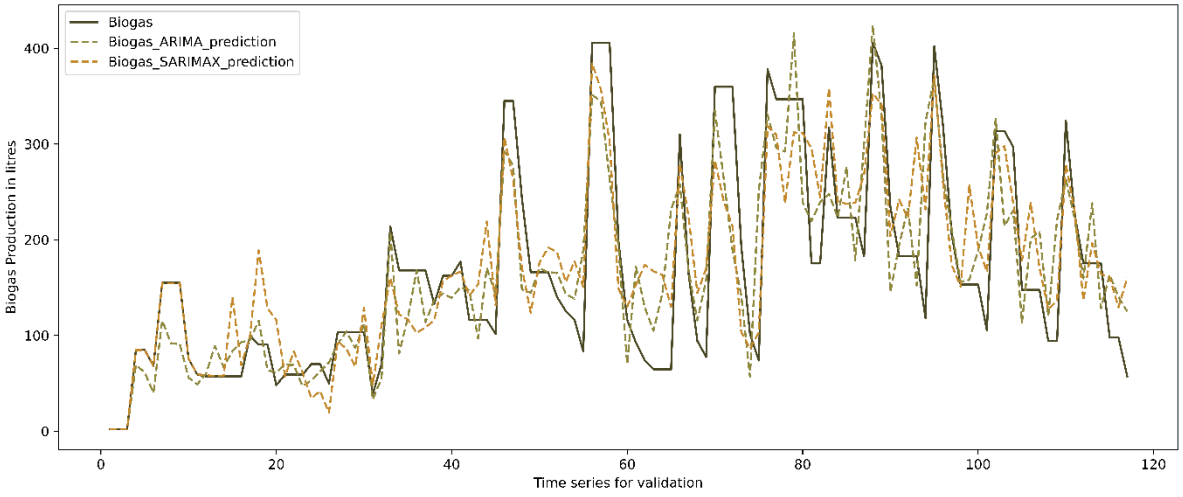


Figure 9. Prediction of Biogas production in liters

3.4.2 Forecasting

Forecasting involves predicting future unknown days by training the model using past data. This type of modelling allows us to anticipate whether outcomes will

increase or decrease in the future. After fitting the model with the time series algorithms, namely ARIMA and SARIMAX, the focus shifts to the forecasting phase. To execute the forecasting technique, we introduce unknown days representing a future period. These days are initialized with "Not a Number" (NAN) values to facilitate predictions for unfamiliar data. These NAN values are then combined with the existing dataset, creating an extended dataset. This extended dataset is subsequently utilized to train the time series algorithms, enabling them to forecast the future data points. In our dataset, information is available for 117 days, while the final dataset incorporates new days to facilitate forecasting. To define the range for forecasting, we designate start and end dates as day 77 and day 150. This range signifies that forecasting spans from the 77th day to the 150th day, though this range can be customized to suit our requirements. Both ARIMA and SARIMAX algorithms perform well and effectively forecast biogas production up to the 150th day. The forecasting results of ARIMA and SARIMAX are visually represented in Figure 10. This visual representation provides insights into the predicted trends and variations in biogas production over the forecasting period.

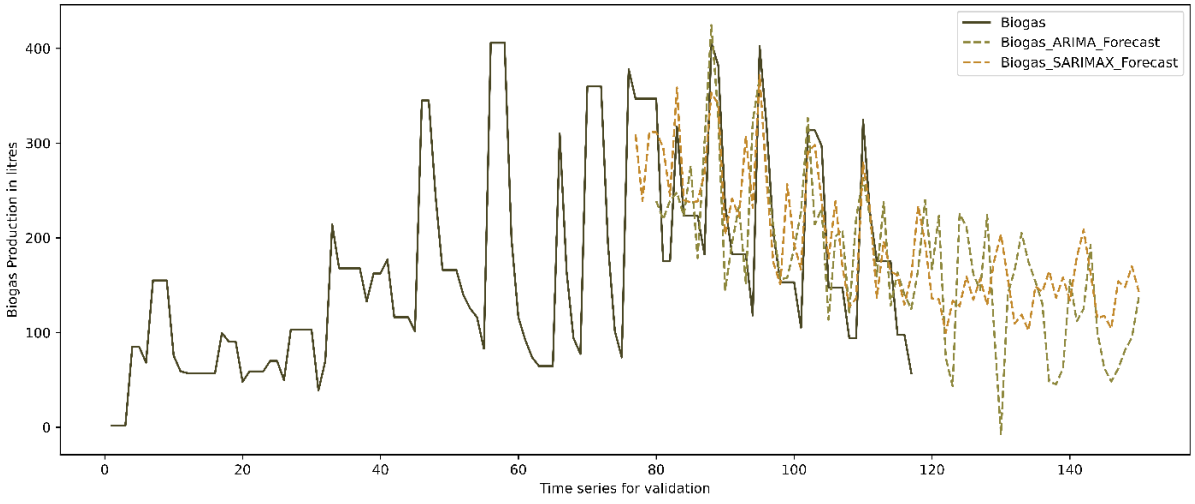


Figure 10. Prediction of Biogas production in liters

3.4.3 Testing model against new data

To evaluate the model using unfamiliar data, we examined five new data points corresponding to days when biogas experiments were conducted. These data points were utilized to assess the variance between the biogas production predicted by the time series model and the actual biogas generated. Predictions using both the ARIMA and SARIMAX algorithms, which yielded RMSE values of 3.26 and 24.02, respectively. When we fed this data into the model for prediction, it yielded a significantly lower error in comparison to the outcomes generated by the time series algorithm. All these values are presented in Table 3 for reference. To evaluate the forecasting results, both ARIMA and SARIMAX algorithms were employed to predict biogas production from the 77th day up to the 150th day (as depicted in Figure 10). These forecasts were made for the month.

Table 3. Difference between the predicted yield and the actual yield

Time Series Algorithm	Day	Observed Biogas	Forecasted Biogas	Deviation (%)
ARIMA	5	71.96	61.35	10.61
	13	83.03	89.03	-6
	19	77.49	63.54	13.95
	25	55.35	62.49	-7.14
	35	162.36	118.6	43.76
SARIMAX	5	71.96	84.59	-12.63
	13	83.03	57.61	25.42
	19	77.49	128.4	-50.91
	25	55.35	41.86	13.49
	35	162.36	116.59	45.77

3.4.4 Statistical Analysis

This technique is utilized to test whether a statistically significant difference exists between the two models, or it can also be employed to assess the statistical difference between the actual and predicted values. This testing process is essential because, for a model to align with the null hypothesis, the p-value must be greater than 0.05. This implies that the newly predicted values fall within the 95% confidence interval of the data distribution, signifying no significant difference between the actual and predicted values. If the p-value is less than 0.05, the null hypothesis is rejected in Favor of the alternative hypothesis. In such cases, the model is deemed to exhibit a significant difference between the predicted and actual values (James et al., 2023). Based on the data presented in Table 3, we extracted the disparity between the actual and predicted models to execute the p-test. The outcome revealed a p-value exceeding 0.05, signifying that both models demonstrate an insignificant difference in their predictive capabilities. This observation is consistent with the model comparison shown in Figure 11.

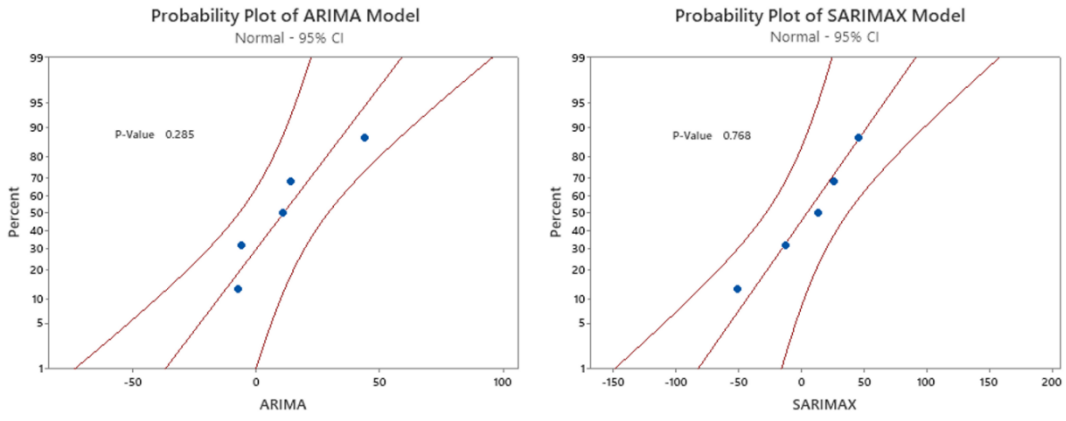
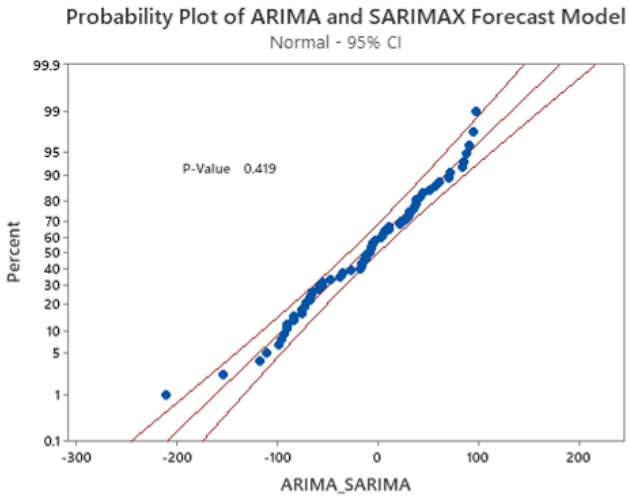


Figure 11. Probability plot for the ARIMA and SARIMAX prediction with new data.

554 For the forecasting model, we calculated the discrepancy between the ARIMA and
555 SARIMAX predictions from the 77th day to the 150th day. Subsequently, a paired t-
556 test was conducted between these two sets of predictions. The obtained p-value was
557 greater than 0.05, indicating that the forecasting model similarly suggests an absence
558 of significant difference between these two models, even in the context of biogas
559 production forecasting. This outcome is illustrated in Figure 12.



560

561 **Figure 12**. Probability plot for the ARIMA and SARIMAX Forecasting of
562 the biogas production.

563 ## 4. Conclusion

564 The culmination of this paper is the forecasting of biogas production through the
565 utilization of time series algorithms, specifically ARIMA and SARIMAX. The dataset
566 comprises 117 days of daily biogas production experimentation, organized into three
567 distinct periods. Throughout each period, key parameters such as CH_4 mean
568 production, VS consumed, CH_4 (% biogas), initial pH, and final pH are meticulously
569 upheld, as detailed in Table 2. To address missing data for certain days, feature
570 engineering techniques are employed. Additionally, hyperparameter tuning is

conducted to determine optimal values for ARIMA (p, d, q) and SARIMAX (p, d, q) (P, D, Q, s) parameters. This process is guided by the AIC and BIC values to ensure a well-fitted model for predictive and forecasting tasks. The identification of suitable parameter values is facilitated by examining the plots of the autocorrelation function (ACF) and partial autocorrelation function (PACF). Consequently, the dataset is fitted to the ARIMA and SARIMAX algorithms. The application of these models results in the prediction and forecasting of biogas production for future unknown days. The achieved RMSE values are 3.26 for ARIMA and 24.02 for SARIMAX, respectively. Subsequent statistical analysis of the new data confirms p-values greater than 0.05 for both prediction and forecasting, signifying the absence of significant differences between the two methods. Thus, these models are deemed viable for biogas production forecasting.

5. References

Abanades, S., Abbaspour, H., Ahmadi, A., Das, B., Ehyaei, M.A., Esmaeilion, F., El Haj Assad, M., Hajilounezhad, T., Jamali, D.H., Hmida, A., Ozgoli, H.A., Safari, S., AlShabi, M., Bani-Hani, E.H., 2022. A critical review of biogas production and usage with legislations framework across the globe. Int. J. Environ. Sci. Technol. 19, 3377–3400. https://doi.org/10.1007/s13762-021-03301-6

Bautista Angeli, J.R., LeFloc'h, T., Lakel, A., Lacarrière, B., Andres, Y., 2022. Anaerobic digestion of urban wastes: integration and benefits of a small-scale system. Environ. Technol. 43, 3414–3425. https://doi.org/10.1080/09593330.2021.1921857

Brockwell, P.J., Davis, R.A., 2016a. Modeling and Forecasting with ARMA Processes, in: Introduction to Time Series and Forecasting. Springer International Publishing, Cham, pp. 121–155. https://doi.org/10.1007/978-3-319-29854-2_5

Brockwell, P.J., Davis, R.A., 2016b. Stationary Processes, in: Introduction to Time Series and Forecasting. Springer International Publishing, Cham, pp. 39–71. https://doi.org/10.1007/978-3-319-29854-2_2

Brockwell, P.J., Davis, R.A., 2016c. ARMA Models, in: Introduction to Time Series and Forecasting. Springer International Publishing, Cham, pp. 73–96. https://doi.org/10.1007/978-3-319-29854-2_3

Cowpertwait, P.S.P., Metcalfe, A. V, 2009. Time Series Data, in: Introductory Time Series with R. Springer New York, New York, NY, pp. 1–25. https://doi.org/10.1007/978-0-387-88698-5_1

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. Statistical Learning, in: An Introduction to Statistical Learning: With Applications in R. Springer New York, New York, NY, pp. 15–57. https://doi.org/10.1007/978-1-4614-7138-7_2

James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J., 2023. Multiple Testing, in: An Introduction to Statistical Learning: With Applications in Python. Springer International Publishing, Cham, pp. 557–596. https://doi.org/10.1007/978-3-031-38747-0_13

Jeong, K., Abbas, A., Shin, J., Son, M., Kim, Y.M., Cho, K.H., 2021. Prediction of biogas production in anaerobic co-digestion of organic wastes using deep learning models. Water Res. 205, 117697. https://doi.org/https://doi.org/10.1016/j.watres.2021.117697

Nguyen, L.N., Nguyen, A.Q., Nghiem, L.D., 2019. Microbial Community in Anaerobic Digestion System: Progression in Microbial Ecology, in: Bui, X.-T., Chiemchaisri, C., Fujioka, T., Varjani, S. (Eds.), Water and Wastewater Treatment Technologies. Springer Singapore, Singapore, pp. 331–355. https://doi.org/10.1007/978-981-13-3259-3_15

Sadoune, H., Rihani, R., Marra, F.S., 2023. DNN model development of biogas production from an anaerobic wastewater treatment plant using Bayesian hyperparameter optimization. Chem. Eng. J. 471, 144671. https://doi.org/https://doi.org/10.1016/j.cej.2023.144671

Sappl, J., Harders, M., Rauch, W., 2023. Machine learning for quantile regression of biogas production rates in anaerobic digesters. Sci. Total Environ. 872, 161923.

https://doi.org/https://doi.org/10.1016/j.scitotenv.2023.161923

Shumway, R.H., Stoffer, D.S., 2017. ARIMA Models, in: Time Series Analysis and Its Applications: With R Examples. Springer International Publishing, Cham, pp. 75–163. https://doi.org/10.1007/978-3-319-52452-8_3

Wang, Y., Huntington, T., Scown, C.D., 2021. Tree-Based Automated Machine Learning to Predict Biogas Production for Anaerobic Co-digestion of Organic Waste. ACS Sustain. Chem. Eng. 9, 12990–13000. https://doi.org/10.1021/acssuschemeng.1c04612

Yildirim, O., Ozkaya, B., 2023. Prediction of biogas production of industrial scale anaerobic digestion plant by machine learning algorithms. Chemosphere 335, 138976. https://doi.org/https://doi.org/10.1016/j.chemosphere.2023.138976