LSTM Modelling for Energy Prediction in the Current Global Context

VALERIA VEVERITA, University of Twente, The Netherlands

February 2022 marks the beginning of the Russian-Ukrainian conflict which targeted the lives of people that are even outside of the conflict. The energy sector is one of the major targets of geo-political events. The enforcement of energy sanctions by the western world caused economical changes which impacted the social models when it comes to energy consumption and raises questions about energy security. The new changes in the energy sector require the reevaluation of energy prediction models and the inclusion of socio-economic factors when predicting energy consumption models. This research aims to identify the socio-economic effect on the energy consumption prediction for the Netherlands and the Republic of Moldova. Two predictive models, the Long-Short Term Memory (LSTM) and Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors (SARIMAX) models, will be assessed and compared for the energy prediction in the context of the current geo-political events. The accuracy of the models will be reevaluated by taking into consideration socio-economic factors such as energy price, inflation rate, world context factor, and weather factors. The result indicate that the energy forecasting for the Republic of Moldova is more dependant on socio-economic factors compared to the Netherlands when using the LSTM predictive model. On the contrary, for the seasonal energy prediction both countries display dependency on the socio-economic and weather factors.

Additional Key Words and Phrases: Energy Prediction, Machine Learning, LSTM, SARIMA

1 INTRODUCTION

Energy is the foundational resource of modern society which ensures its stability and development. After the Russian military attack on Ukraine starting from February 2022, the emerging issue of energy security became a primary socio-economic concern worldwide and a timely research issue. Based on the Asia Pacific Energy Research Centre, there are four criteria [11], such as availability, accessibility, affordability and acceptability, which facilitate the assessment of energy security. After the energy sanctions imposed by the US and the EU, the energy prices are projected to experience a substantial increase, as it is predicted that the European crude oil consumer price would increase by 6.47%, the refined oil price would increase by 7.39% and gas price would increase by 1.16% [9]. Subsequently, due to the discrepancy between the escalation of energy prices and alteration in income levels, it is estimated that the energy consumption would decrease by 1.75% [9]. The military conflict proved the dependency of energy on the geo-political events and raised international awareness regarding energy security in the current socio-economic context. The volatility of the energy market directly impacted the global economy, which subsequently led to changes in social models when it comes to energy consumption.

Energy prediction models can contribute to ensuring energy security, and combating the violation of accessibility, affordability and acceptability. Forecasting energy production enables energy suppliers to determine the optimal quantity of energy required, which can help them identify and plan the available resources for localized production and rely less on energy import, thereby facilitating price stabilization. Additionally, energy forecasting can facilitate the transition to sustainable energy production which subsequently will lead to energy independence. Despite the fact that energy prediction is not a novel subject, the current geo-political event has introduced new challenges in the realm of energy prediction at the country level. The first challenge lies in the recentness of the conflict and the lack of systematic studies regarding the effect of the event on the energy sector. By taking into consideration that the conflict remains ongoing, it is not possible to determine the causal factors that contribute to changes in energy consumption. Therefore, in the scope of this paper it is only feasible to hypothesize and statistically infer the correlation between the exogenous factors related to the conflict and the energy sector. The second challenge is related to the human factor and the difficulty to predict and quantify social behaviors in crisis situations.

In light of the current context, it is important to reevaluate the efficiency of energy prediction models by taking into consideration socio-economic factors such as energy price, inflation rate, and world context. This project will focus on the assessment of the Long-Short Term Memory (LSTM) model comparative to the Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors (SARIMAX) model in the context of the current geo-political event. This project is limited in scope to the analysis of two countries: the Republic of Moldova and the Netherlands. The following countries were chosen due to their high dependence on Russian energy supplies, as the Netherlands was 58.9% reliant on the fossil fuels imported from Russia in 2021 [3], while the Republic of Moldova imported almost 80% of fossil fields in 2018 [1].

2 RESEARCH QUESTION

This paper addresses the research question and subquestions defined below in order to assess the performance of the LSTM model for energy consumption prediction in the defined context.

RQ: What is the accuracy of the LSTM model in predicting energy consumption in the current global context for the Netherlands and the Republic of Moldova?

RQ1: Which of the following exogenous factors (weather, energy price, inflation rate, world context) are correlated with energy consumption?
RQ2: How universal is the LSTM model with the identified exogenous features in predicting energy consumption in the Netherlands and the Republic of Moldova?

TScIT 39, July, 2023, Enschede, The Netherlands

RQ3: How does the machine learning (LSTM) model compare to the statistical (SARIMA) model with the identified features?

3 RELATED WORK

In the current energy sector there are multiple models, both statistical and/or machine learning models, to predict energy consumption with high efficiency and accuracy. Statistical models, such as time-series (e.g. SARIMA) [12] and regression models (e.g. linear regression) [10], apply statistical techniques in order to identify trends in the energy consumption time series and derive its predictions. However, machine learning models are considered to achieve higher predictive accuracy than statistical models, due to their ability to analyze a greater amount of complex data and identify complex patterns. Currently, artificial neural network (ANN) and Support Vector Machine (SVM) models [4] are widely used in forecasting, and show highly accurate and robust predictions in energy forecasting as well. Despite the numerous studies in the energy prediction sector, the topic of forecasting still remains a highly researched topic and new variations of models are appearing, such as Long Short-Term Memory (LSTM) networks which is a variation of neural networks and is commonly used in energy prediction due to its high performance compared to popular models such as Auto-Regressive Moving Average Model (ARMA), Auto-Regressive Fractionally Integrated Moving Average Model (ARFIMA) and Back Propagation Neural Network (BPNN).[24]

In 2020, Cuoto et al. [8], discuss the impact of the corona virus pandemic on the energy sector. The paper aims to optimize the predictive energy consumption model in the pandemic context, in order to assist energy providers to estimate the optimal energy demand that they need to supply. The paper mentions the use of deep neural network to forecast the energy consumption, which yields a mean absolute percentage error (MAPE) of 1.17%. Haoxiang et al. [13], proposes the use of random forest (RF) to forecast energy demand in office building in Shanghai. RF is a classifier which in the scope of the paper was used to predict consumption in a random office building without any historical data from the tested buildings. The model reached a mean absolute percentage error of 12% for predicting lighting energy consumption. On the other hand, the energy required for HVAC (Heating, Ventilation, and Air Conditioning) reached a significantly larger error value of 58%, due to the dependency on weather and season. Luo et al.[16] have proposed a genetic algorithm enhanced adaptive deep neural network (DNN) predictive model accompanied by feature extraction in order to improve the accuracy and effectiveness of building energy consumption prediction. The overall MAPE of the proposed algorithm is 1.43%, which compared to the regular genetic algorithm and DNN model, showcases an improvement of 15.9% (MAPE) in testing cases. In 2021, Kumar Dubey et al. [12] compared statistical and machine learning models, and it identifies that the LSTM network performs significantly better than the SARIMA model in the context of energy consumption forecasting. Additionally, the article remarks that the performance of the network is directly proportional to the number of lags and input data. LSTM network has also been compared with 12 data-driven models, of which 7 were shallow learning, 2 deep learning and 3 heuristic methods, and based on the comparison it was determined that LSTM network performs better in short term-prediction (1h ahead), while XGBoost should be considered for long-term prediction (24h ahead). [25] Nivethitha et al. [20] proposes a kCNN-LSTM model, where k stands for clustering and CNN is Convolutional Neural Network, and the model is trained on historical energy consumption data. The model is compared to a statistical model (ARIMA) and four neural network models, which are outperformed by the proposed model by approximately 10% based on the mean squared error (MSE).

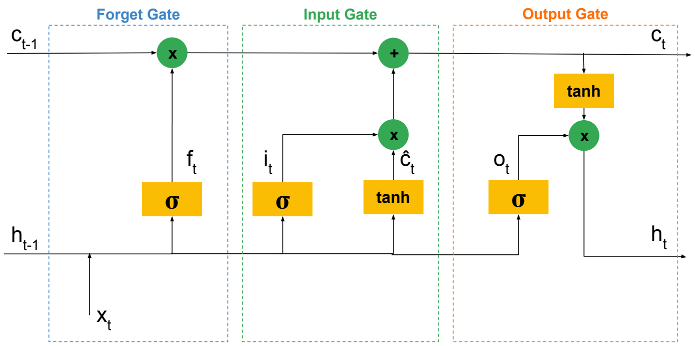


Fig. 1. Long-Short Term Memory cell

In 2022, Razak et. al. compares machine learning techniques, such as Deep Neural Network (DNN), Artificial Neural Network (ANN), Gradient Boosting (GB) etc., in order to determine which model has a greater performance when predicting building energy consumption. The research has determined that the most efficient predictive model is DNN, which showcases a MAE result of 0.92. On the other hand, the model has the greatest training time of 5.2 seconds. Tasarruf et al. [6] proposes a hybrid forecasting model which consists of Prophet model, ARIMA model, LSTM model and Back Propagation Neural Networm (BPNN). The aim of the model was to encounter both for the linear and non-linear trends in energy consumption. However, the proposed model has the lowest metric parameters in comparison to the standalone models.

4 BACKGROUND

4.1 BACKGROUND: LSTM Model

Long-Short Term Memory (LSTM) [6, 15] is a variation of the RNN network, and it tackles the problem of vanishing gradients that occur with the original RNN network. Gradient problems refer to the back-propagation of the output layer to the input layer, meanwhile calculating the error gradients which are later used to update the weights and biases in the model. Vanishing refers to the gradients that rapidly decrease and approach a value of 0, which subsequently prevent weights from updating and stagnate the learning mechanism. LSTM solves vanish problems by introducing a forget mechanism which is accomplished by the forget gates. The LSTM model is used in time series forecasting as it captures long-term dependencies in sequential data.

The LSTM model consists of three gates depicted in Fig.1: forget gate, input gate and output gate. Forget gate determines percentage wise how much of the long-term memory c_{t-1} should be remembered and propagated forward, based on the previous hidden state h_{t-1}, also called short term memory and the input data x_t. Initially, the hidden state and the new input data are inputted to the gate for the weight multiplication, after which the sigmoid activation function (σ) transforms the inputs into a value between 0 and 1. These values are further used to determine the long-term memory retention, as 0 factor indicated that data should be forgotten, while 1 factors indicates that data should be fully preserved. This factor is represented in (1).

$$f_t = \sigma(U_f x_t + V_f h_{t-1} + b_f) \tag{1}$$

Where U and V are weight matrices and b is the bias value for the corresponding gates. The the cell state is multiplied with f_t factor, to determine what percentage of the long-term memory should be remembered.

Input gate uses the input and short-term memory to compute the potential long-term memory value, which is subsequently ran through a tanh activation function, which will result in a vector ranging from -1 to 1. The computation of the potential long-term memory value is represented in (2). Additionally, the input gate computes the factors which determined how much of the new potential long-term memory should be remembered. The computation of this factors is represented in (3). Lastly, at the input gate, the final long-term memory value c_t is computed, as depicted by (4).

$$\hat{c}_t = tanh(U_c x_t + V_c h_{t-1} + b_c) \tag{2}$$

$$i_t = \sigma(U_i x_t + V_i h_{t-1} + b_i) \tag{3}$$

$$c_t = c_{t-1} \times f_t + \hat{c}_t \times i_t \tag{4}$$

Lastly, the output gate computes the potential short-term memory value. In the first step the previous short-term memory value and the input data are converged using the sigmoid activation function, as shown in (5). Lastly, the newly computed long-term memory value c_t is passed through a tanh activation function and multiplied by o_t, to compute the current short-term memory value h_t, as shown in (6).

$$o_t = \sigma(U_o x_t + V_o h_{t-1} + b_o) \tag{5}$$

$$h_t = o_t \times tanh(c_t) \tag{6}$$

4.2 BACKGROUND: SARIMAX Model

Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors (SARIMAX) [2, 12] model is a variation of ARIMA model, that is a statistical analysis model which is implemented in time-series forecasting. The foundational model composes three parts: the auto-regressive order (AR) denoted by p, the integrated order (I) denoted by d, and the moving average order (MA) denoted by q. In the model p represents the number of lagged observations which equals the number of lags in the partial auto-correlation (PAC) that crosses the limit set. The PAC is defined in (7)

$$PAC = \frac{cov(y_i, y_{i-h} | y_{i-1}, ..., y_{i-h+1})}{\sqrt{var(y_i | y_{i-1}, ..., y_{i-h+1}) var(y_{i-h} | y_{i-1}, ..., y_{i-h+1})}} \tag{7}$$

Where: y_i is the response variable, and y_{i-1}, ..., y_{i-h+1}, and y_{i-h} are the predictor variables of the h^{th} order. The integrated order d denotes the differencing order required to achieve stationary, which refers to a statistical property where the mean and variance are constant over time. The last term q represents the number lagged predictive errors which equals the number of lags in the autocorrelation (AC) which crosses the limit set. The AC is defined in (8)

$$AC = \frac{\sum_{i=1}^{n-k}(y_i - \hat{y})(yi + k - \hat{y})}{\sum_{i=1}^{n}(y_i - \hat{y})^2} \tag{8}$$

Where: y is the mean of the time series, k is the lag, and n is the complete series value.

The SARIMAX model accounts for the seasonality in data denoted by the terms P, D, Q, s, where P, D and Q represent the seasonal AR, I, and MA orders, while s denotes the period number in a season. Based on the described terms the SARIMAX model is denoted as SARIMAX(p, d, q)(P, D, Q)s, and has the following specification in (9):

$$\phi(B)\phi(B^s)\nabla_s^D \nabla^d x_t = \Theta(B)\Theta(B^s)\varepsilon_t \tag{9}$$

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p \tag{10}$$

$$\phi(B^s) = 1 - \phi_1 B^s - \phi_2 B^{2s} - ... - \phi_P B^{Ps} \tag{11}$$

$$\Theta(B) = 1 - \Theta_1 B - \Theta_2 B^2 - ... - \Theta_q B^q \tag{12}$$

$$\Theta(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - ... - \Theta_Q B^{Qs} \tag{13}$$

$$\nabla^d = (1 - B)^d \tag{14}$$

$$\nabla^d = (1 - B^s)^d \tag{15}$$

Where: x_t is the time series data at period t, B is the back-shift operator, ε_t is a series of independent and identically distributed random variables. Here (10) and (11) represent the AR non-seasonal and seasonal parameters of order p and P, while (12) and (13) represent the MS non-seasonal and seasonal parameters of order q and Q subsequently. Lastly, (14) and (15) refer to non-seasonal and seasonal difference components.

5 METHODS

5.1 Dataset

In the global context we are assuming multiple factors that are impacting the energy consumption such as endogenous variables (Actual Total Load that is measured in mega watts [MW]), and exogenous variables such as weather and socio-economic factors. Additionally, the geographical scope of the research extends to the Netherlands and Republic of Moldova, the data will be collected for the following countries.

Visual Crossing Weather [7] dataset provides access to weather data such as temperature (°C), feels-like temperature (°C), dew (°C), humidity (%), precipitation (mm), pressure (hPa) and cloud cover (%) for both geographical territories. The socio-economic factors are quantified by the energy prices (EUR/MWh), inflation and the world context. The world context is defined by three events: pre-covid period (1), Covid period (2), Covid and Russian-Ukrainian conflict period (3) and Russian-Ukranian conflict period (4). For the Netherlands, the energy prices were collected from entsoe [23] which is

Table 1. Summary of the statistical values for energy consumption, weather and socio-economic variables: the Netherlands

Variable	Range	Mean	Standard Deviation	PCC
Total Load	[5383.25, 17704.25]	11924.24	1912.884	1
Temperature	[-7.2, 36.9]	10.63	6.03	-0.33
Apparent Temperature	[-10.7, 35.3]	9.42	7.15	-0.34
Dew	[-12.3, 19.7]	6.95	5.08	-0.24
Humidity	[16.46, 100]	80.11	15.94	0.19
Precipitation	[0, 26.04]	0.084	0.675	0.019
Wind speed	[0, 60.3]	13.85	7.51	0.12
Pressure	[971.4, 1046.3]	1015.9	10.85	-0.069
Cloud Cover	[0, 100]	61.52	37.07	0.0915
Price	[-195.41, 871]	128.11	118.76	0.098
Inflation	[0.7, 14.5]	4.92	4.04	-0.16
World Context	[1, 4]	2.72	0.98	-0.16

Table 2. Summary of the statistical values for energy consumption, weather and socio-economic variables: the Republic of Moldova

Variable	Range	Mean	Standard Deviation	PCC
Total Load	[326, 1074]	665.64	143.80	1
Temperature	[-16.10, 36]	11.35	9.45	-0.05
Apparent Temperature	[-22.50, 38.60]	9.98	10.83	-0.08
Dew	[-19, 23.50]	4.99	7.74	-0.18
Humidity	[13.93, 100]	69.21	20.97	-0.18
Precipitation	[0, 201.61]	0.07	1.56	0.01
Wind speed	[0, 53.60]	12.45	7.02	0.18
Pressure	[988, 1044.40]	1016.39	7.93	0.08
Cloud Cover	[0, 100]	60.75	37.52	0.11
Price	[58.96, 228.53]	115.04	66.08	-0.05
Inflation	[100.22, 134.62]	113.83	12.25	-0.08
World Context	[1, 4]	2.71	0.98	-0.11

a central collection of electricity generation, transportation, and consumption data. Inflation is defined by the Consumer Price Index (CPI) (year-on-year %change), which measures the change in urban consumer prices compared to last year for a market basket of good and services. The inflation information for the Netherlands is collected from Statistics Netherlands [17]. The socio-economic contextual data (energy prices and inflation) for the Republic of Moldova was collected from Statistica Moldova [5, 18]. It is important to mention that in the Netherlands the energy prices rate are changing hourly while in the Republic of Moldova it changes periodically at the decision of ANRE (National Agency for Energy Regulation in the Republic of Moldova) in accordance with the market energy prices, energy consumption, etc. The data regarding the energy consumption was collected from entsoe [23] both for the Republic of Moldova and the Netherlands. The energy data per country was extracted from for 1 175 days (from 2020-02-11 00:00 until 2023-04-30 23:00) which resulted in 28 200 data samples per variable for a 1-hour resolution.

The dataset for the Netherlands was complete and there were no missing values or time stamps. For the Republic of Moldova there were 416 non-numerical entries (NaN), which were replaced using the scikit-learn k-Nearest Neighbours (KNN) Imputer [19], that computes the missing values based on the nearest defined numerical neighbours. In total 12 variables were collected, one endogenous and 11 exogenous for each country which are described in Table 1 for the Netherlands and Table 2 for the Republic of Moldova.

5.2 Feature Selection

This section explores the weather and socio-economic features relation in terms of power consumption, as addressed by the first research sub-question. In this research there are 11 exogenous features, 8 are weather factors and 3 are socio-economic factors, in the scope of two countries. Feature selection is used to improve model performance and accuracy, reduce data dimensionality and future data collection process. The feature selection process can be done either using a variable ranking method (correlation coefficient, mutual information) or a nested subset selection method.

The Pearson Correlation Coefficient (PCC) defined in (18) is used to determine the linear correlation between the defined variables and identify potential factors that influence the energy demand profiles. Positive PCC indicates correlation that tends to move in the same direction, therefore if one variable increase the correlated variable increases as well. On the other hand, negative PCC indicates indirect correlation, where the correlated variables follow opposite directions. Fig 2 and Fig 3 depict the correlation matrix between all the factors with the corresponding PCC values. The PCC values denote that in the Republic of Moldova weather (except for precipitation) and socio-economic variables have a similar correlation in regards to the energy consumption, while in the Netherlands the weather factors (except for the precipitation) have a greater correlation with the total load. However, in both scenarios the correlation values of all factors with the total load is relatively low compared to the maximum PCC (0.8), which can result due to the necessary energy demand on a country level which is more uniform and independent from the defined exogenous factors. Additionally, the transition to renewable energy can influence energy consumption patterns.

5.3 Long-Short Term Memory Network

The proposed network has two LSTM layers and one Dense layer. The first layer has 200 hidden units, the second layer has 100 hidden units and the third layer has 20 hidden units. At each LSTM layer, a dropout layer has been added with a 20% rate to prevent overfitting. To build and train the models, Keras Deep Learning Library [14] is used. Based on the feature selection procedure the following parameters (Actual Total Load, Price, Inflation, Temperature, Apparent

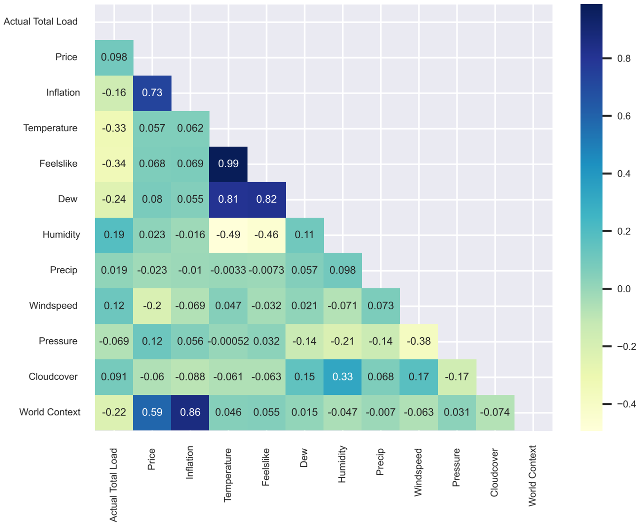


Fig. 2. Correlation heat map of energy consumption with exogenous factors for the Netherlands

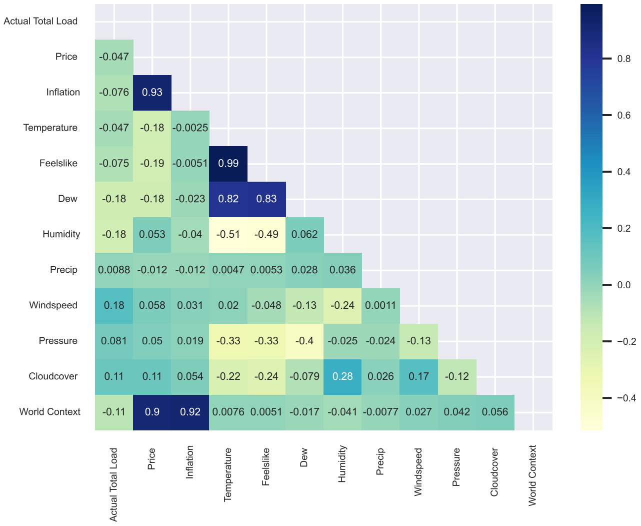


Fig. 3. Correlation heat map of energy consumption with exogenous factors for the Republic of Moldova

Temperature, Dew Point, Humidity, Wind speed, and World Context) to train the models. Each model makes use of Adam optimizer and they are trained with a learning rate 0.001 and a clip value of 0.5. Each model is trained for 30 epochs, with a batch size of 32. The training data consists of 27 480 entries, which represents values for 1 145 days (from 2020-02-11 00:00 until 2023-03-31 23:00). There are three test datasets for three testing scenarios (one day-ahead, one week-ahead and one month-ahead prediction).

5.4 SARIMAX

SARIMAX is composed of seasonal and non-seasonal components, which subsequently are dependent on the correlated orders. The non-seasonal component consists of the orders p, d, and q, while the seasonal component consists of the orders P, D, Q, and s. The main aim in building a SARIMAX model is to identify the defined orders.

- **s:** In the ACF plot of the Netherlands (Fig. 4a) and the Republic of Moldova (Fig. 4c) there is one peak every 24 lags (hours), therefore the seasonal period s for both countries is 24.
- **d:** The first step in building the model is determining if the time series is non-stationary, which means that the time series has a varying trend mean over time or seasonality. This can be verified using the Augmented Dickey-Fuller test, presented by the statsmodels library [21]. Based on the Augmented Dickey-Fuller, the p-value of $4.04e^{-21}$ for the Netherlands and $6.68e^{-19}$ for the Republic of Moldova is lower than the significance level of 5%, and hence we can reject the null hypothesis and take that the series is stationary, subsequently, the difference order d and D is 0.
- **p and P:** The p and P values are equal to the lags in the PACF (Fig. 4b, Fig. 4d) which crosses the limit set significantly, In the case of both countries there is a significant drop after the first lag, subsequently p and P values are equal to 1.
- **q and Q:** The q value is equal to the lags in the ACF (Fig. 4a, Fig. 4c) which crosses the limit set significantly. For the Republic of Moldova and the Netherlands, the first 5 lags are the most prominent, subsequently q is equal to 5 and Q is 1.

The defined SARIMAX (5, 0, 1) × (1, 0, 1, 24) model is applied for the energy forecast both for the Netherlands and the Republic of Moldova. The SARIMAX model was built and trained using python statsmodel library [22].

5.5 Prediction Evaluation Metrics

To assess the performance of the predictive LSTM model three assessment errors are used to evaluate the error between the actual energy load and the predicted energy load. The Root Mean Square Error (RMSE) defined in (16) is used to determine the deviation magnitude between the real and predicted values.

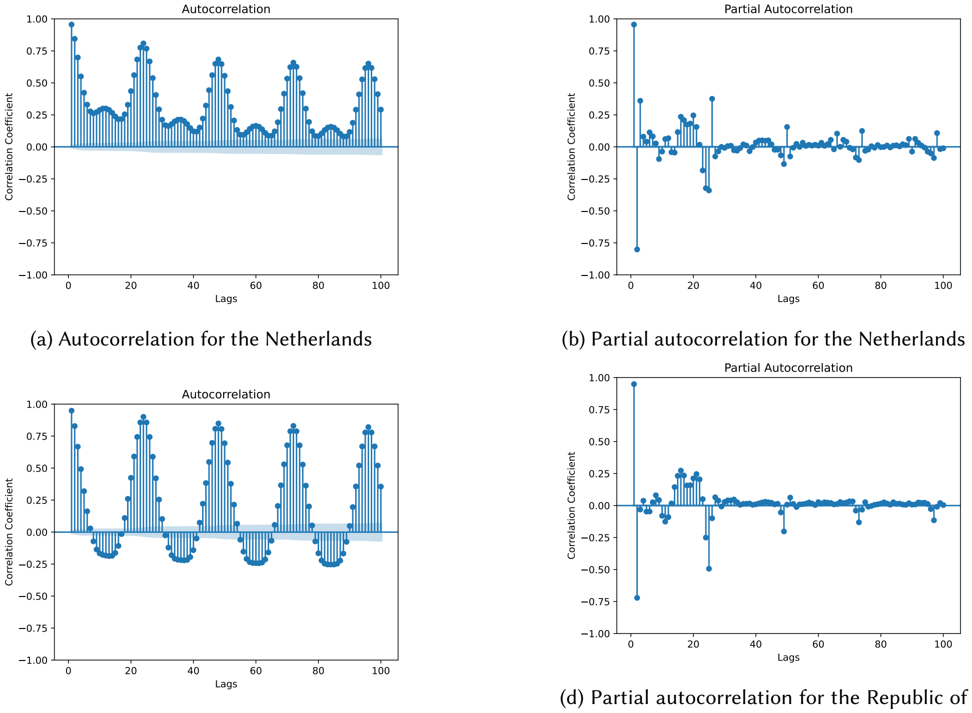
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{16}$$

Where: y_i represents the observed values, $\hat{y}_i$ represents the predicted values and n represents the number of observations.

The Normalized Root Mean Square Error (NRMSE) defined in (17) is used to determine the deviation magnitude between the predicted and measured value defined in percentage unit. NRMSE is used to determine the efficiency of the predictive model in both countries, as the energy consumption magnitude differs by country.

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{(y_{max} - y_{min})} * 100 \tag{17}$$

Where: y_{min} and y_{max} represent the min and max values of the observed values.



(a) Autocorrelation for the Netherlands

(b) Partial autocorrelation for the Netherlands

(c) Autocorrelation for the Republic of Moldova

(d) Partial autocorrelation for the Republic of Moldova

Fig. 4. Aucorrelation and partial autocorrelation of the actual total load

The Pearson Correlation Coefficient (PCC) defined in (18) is used to determine the similarity between the predicted and measured values.

$$PCC = \frac{\sum_{i=1}^{n} |(y - \mu_y)(\hat{y} - \mu_{\hat{y}})|}{\sigma_y * \sigma_{\hat{y}}} \tag{18}$$

Where: $\mu_{\hat{y}}$ and μ_y represents the mean of the predicted and observed dataset, and $\sigma_{\hat{y}}$ and σ_y represents the variance of the predicted and observed dataset respectively.

6 RESULTS

There are 4 different test scenarios which are defined based on the influencing factors. Figure 3 presents the list of scenarios which are further used to assess model efficiency.

Table 3. Simulation parameters

	Variables	Time Horizon	Resolution
Scenario 1	Energy data	1 day	1 hour
Scenario 2	Energy data + Socio-economic factors	1 day	1 hour
Scenario 3	Energy data + Weather factors	1 day	1 hour
Scenario 4	Energy data + Socio-economic + Weather	1 day	1 hour

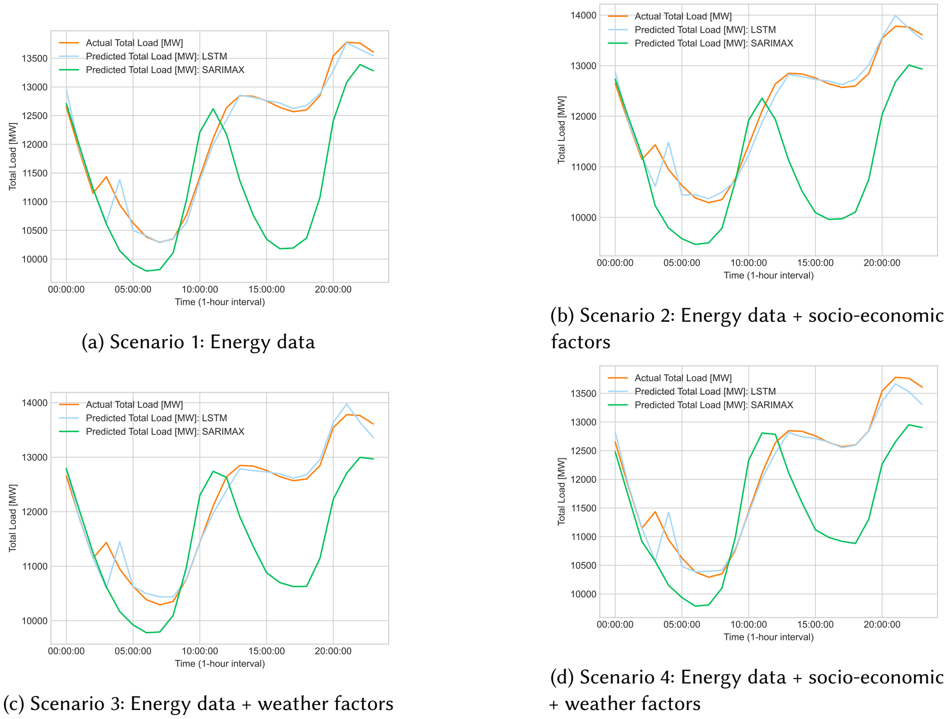
This set of scenarios defines a set of experiments to perform the forecast with hourly resolution for the total load of the Netherlands and the Republic of Moldova. The results for the Netherlands are summarized in Table 4 and for the Republic of Moldova in Table 5

6.1 Results for the Netherlands

Fig 5a, Fig 5b, Fig 5c and Fig 5d show the predictive similarity of the model, which implies that the socio-economic and weather factors do not have a great impact on the energy consumption prediction in the Netherlands. For the LSTM model, the accuracy metrics show that scenario 3, where the model was trained based on historical load data and weather data, performs the best, with an NRMSE accuracy of 5.88%. For the SARIMAX model, the accuracy metrics show that scenario 4, where the model was trained based on historical load, weather and socio-economic data, performs the best, with an NRMSE accuracy of 28.43%.

6.2 Results for the Republic of Moldova

Similar to the case of the Netherlands, the predictive figure also show a similarity in the predictive models. However, based on Fig 6d the predicted total load showcases a greater matching accuracy with the actual total load compared to the other figures. For the LSTM model, the accuracy metrics show that scenario 4, where the model was trained based on historical load data combined with the weather and socio-economic data, performs the best, with an NRMSE accuracy of 6.11%. For the SARIMAX model, the accuracy metrics show that scenario 4, where the model was trained based



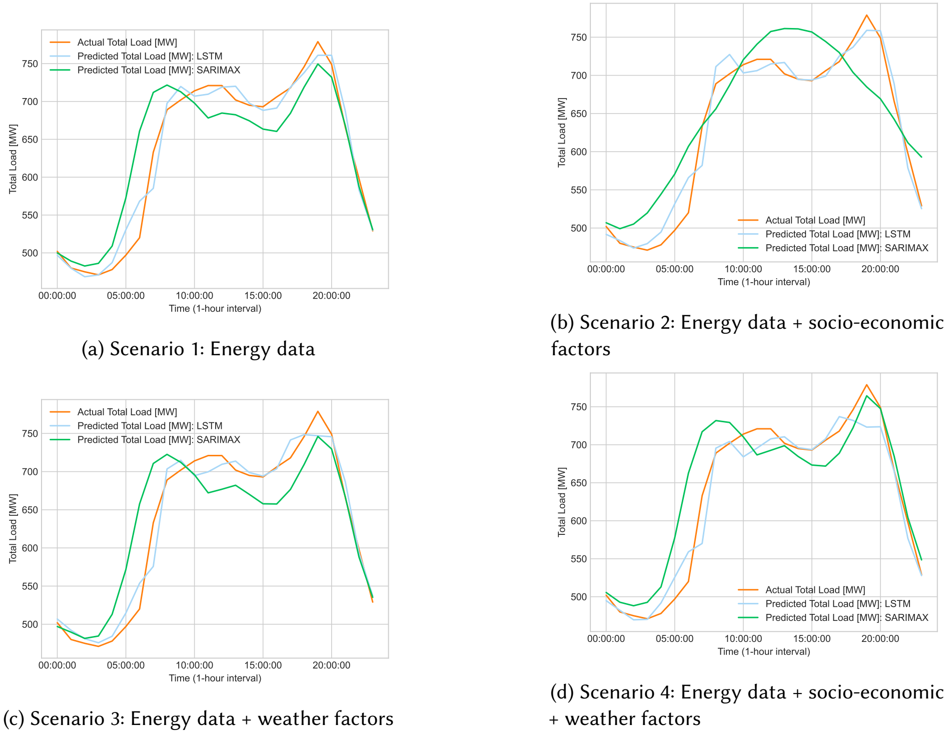
(a) Scenario 1: Energy data

(b) Scenario 2: Energy data + socio-economic factors

(c) Scenario 3: Energy data + weather factors

(d) Scenario 4: Energy data + socio-economic + weather factors

Fig. 5. Energy consumption prediction for the Netherlands for one day ahead using the LSTM and SARIMAX models for the defined scenarios



(a) Scenario 1: Energy data

(b) Scenario 2: Energy data + socio-economic factors

(c) Scenario 3: Energy data + weather factors

(d) Scenario 4: Energy data + socio-economic + weather factors

Fig. 6. Energy consumption prediction for the Republic of Moldova for one day ahead using the LSTM and SARIMAX models for the defined scenarios

Table 4. Evaluation metrics for the Netherlands

Model	Test Scenario	RMSE [MW]	NRMSE [%]	PCC
LSTM	Scenario 1	232.62	6.66	0.96
	Scenario 2	228.62	6.55	0.94
	Scenario 3	205.72	5.88	0.98
	Scenario 4	214	6.15	0.94
SARIMAX	Scenario 1	1255.30	35.98	0.65
	Scenario 2	1456.93	41.79	0.64
	Scenario 3	1089.57	31.23	0.72
	Scenario 4	991.89	28.43	0.78

Table 5. Evaluation metrics for the Republic of Moldova

Model	Test Scenario	RMSE [MW]	NRMSE [%]	PCC
LSTM	Scenario 1	21.83	7.09	0.89
	Scenario 2	20.91	6.78	0.92
	Scenario 3	22.67	7.36	0.87
	Scenario 4	18.81	6.11	0.94
SARIMAX	Scenario 1	43.12	14.00	0.91
	Scenario 2	49.80	16.17	0.89
	Scenario 3	44.37	14.41	0.90
	Scenario 4	43.10	13.99	0.92

on historical load, weather and socio-economic data, performs the best, with an NRMSE accuracy of 13.99%.

7 DISCUSSION

This paper presented a comparative study between the machine learning model LSTM and statistical model SARIMAX in forecasting energy consumption. For this paper two countries were considered, the Netherlands and the Republic of Moldova, and four different cases as described in Table 3. This study resulted in the following observations:

(1) For the Netherlands, the weather factors such as temperature, apparent temperature, dew and socio-economic factors such as inflation and world context, show prominent inverse correlation with energy consumption. On the other hand humidity, wind speed and price show direct correlation with energy consumption. For the Republic of Moldova, apparent temperature, dew, humidity, inflation and world context show prominent inverse correlation with energy consumption, while wind speed, pressure and cloud cover have a direct correlation with energy consumption.

(2) Based on the Pearson correlation coefficient, in the Netherlands the energy consumption is more dependant on external factors compared to the Republic of Moldova, where apparent temperature has the highest correlation of -0.34 with energy consumption, compared to the Republic of Moldova, where wind speed has the highest correlation of 0.18 with energy consumption.

(3) LSTM model is found to be have a higher predictive accuracy in comparison to SARIMAX for the Netherlands and the Republic of Moldova in all of the defined scenarios.

(4) For the Netherlands, the LSTM model results in a higher predictive accuracy in scenario 3, which indicates that weather features contribute more to energy forecasting. Conversely to this, the energy consumption in the Republic of Moldova is more dependant on weather and socio-economic features.

(5) The SARIMAX model results in a higher predictive accuracy in scenario 4 for both countries, which implies that the seasonal energy consumption is dependant both on weather and socio-economic features.

(6) The SARIMAX model displays higher predictive accuracy in the case of the Republic of Moldova compared to the Netherlands, where the best NRMSE of 28.43% in case of the Netherlands is greater by a factor of 2.03 than the NRMSE of 13.99% in case of the Republic of Moldova. This implies a greater seasonality in the energy consumption in the Republic of Moldova, which subsequently indicates that in the Republic of Moldova the energy demand is higher in seasonal sectors such as residential services and agriculture, while in the Netherlands the energy demand is higher in sectors that are not seasonal dependant such as industry and transport.

8 CONCLUSION

In this paper a comparison between a machine learning model (LSTM) and a statistical model (SARIMAX) to predict the energy consumption in the Netherlands and the Republic of Moldova based on weather and socio-economic factors. The comparison was performed based on three performance metrics, RMSE, NRMSE, and PCC. The prediction capabilities of the two models are analyzed under four scenarios. The obtained results indicate that the presented machine learning model has a significantly higher performance compared to the statistical predictive model, SARIMAX, for both countries. In the case of the energy prediction for the Netherlands, the LSTM models showcases a higher performance when the historical load and weather data is used for training. For the Republic of Moldova, the best performance is observed in the scenario when all the defined features (historical load, weather, socio-ecnomic) are considered for the model training. This observation dismisses the universality of the model in the context of the two countries, as for each country different feature selection results in different performances. Additionally, we can imply that feature selection is dependent on the context, as for the Republic of Moldova, the identified socio-economic features contribute to the model performance, while the Netherlands is independent from the influence of the socio-economic factors. The discrepancy in the countries' energy profiles can be explained by the geographical factors, as the Republic of Moldova is closer to Russian-Ukrainian conflict, which implies higher energy dependency on economic and political factors.

9 ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Elena Mocanu, for guiding me through this academic journey by providing expertise and support. I am also thankful to my parents, Aliona and

Ghenadie Veverita, for the undivided support and assistance in the data collection.

REFERENCES

[1] 2020. Moldova energy profile. (2020). https://www.iea.org/reports/moldova-energy-profile

[2] 2021. Comparative Analysis of Different Univariate Forecasting Methods in Modelling and Predicting the Romanian Unemployment Rate for the Period 2021–2022. *Entropy* 23 (2021), 325. Issue 3.

[3] 2023. National Reliance on Russian Fossil Fuel Imports. (2023). https://www.iea.org/reports/national-reliance-on-russian-fossil-fuel-imports

[4] A.S. Ahmad, M.Y. Hassan, M.P. Abdullah, H.A. Rahman, F. Hussin, H. Abdullah, and R. Saidur. 2014. A review on applications of ANN and SVM for building electrical energy consumption forecasting. *Renewable and Sustainable Energy Reviews* 33 (2014), 102–109. https://doi.org/10.1016/j.rser.2014.01.069

[5] ANRE. 2023. Electrical Energy. (2023). https://anre.md/energie-electrica-3-290

[6] Tasarruf Bashir, Chen Haoyong, Muhammad Faizan Tahir, and Zhu Liqiang. 2022. Short term electricity load forecasting using hybrid prophet-LSTM model optimized by BPNN. *Energy Reports* 8 (2022), 1678–1686. https://doi.org/10.1016/j.egyr.2021.12.067

[7] Visual Crossing Corporation. 2023. Visual Crossing Weather (2020-2023). (2023). https://www.visualcrossing.com/

[8] P. A. J. Couto, C. A. F. Rocha, F. P. Monteiro, S. C. A. Monteiro, M. E. L. Tostes, U. H. Bezerra, L. S. Soares, and E. C. S. Silva. 2020. Adaptive RNA Model for Very Short Energy Forecast Validated in the New Coronavirus Pandemic Context. In *2020 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, Vol. 4. 1–6. https://doi.org/10.1109/ROPEC50909.2020.9258775

[9] Lianbiao Cui, Suyun Yue, Xuan-Hoa Nghiem, and Mei Duan. 2023. Exploring the risk and economic vulnerability of global energy supply chain interruption in the context of Russo-Ukrainian war. *Resources Policy* 81 (2023). https://doi.org/10.1016/j.resourpol.2023.103373

[10] Nelson Fumo and M.A. Rafe Biswas. 2015. Regression analysis for prediction of residential energy consumption. *Renewable and Sustainable Energy Reviews* 47 (2015), 332–343. https://doi.org/10.1016/j.rser.2015.03.035

[11] Narumon Intharak. 2007. A Quest for Energy Security in the 21st Century. (2007). http://doi.acm.org/10.1145/1219092.1219093

[12] Ashutosh Kumar Dubey, Abhishek Kumar, Vicente García-Díaz, Arpit Kumar Sharma, and Kishan Kanhaiya. 2021. Study and analysis of SARIMA and LSTM in forecasting time series data. *Sustainable Energy Technologies and Assessments* 47 (2021), 101474. https://doi.org/10.1016/j.seta.2021.101474

[13] Haoxiang Li, Qi Zhou, Jing Tian, and Xiaoyu Lin. 2020. Energy Demand Forecasting for an Office Building Based on Random Forests. In *2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2)*. 29–32. https://doi.org/10.1109/EI250167.2020.9347021

[14] Keras Deep Learning Library. 2022. Long-Short Term Memory. (2022). https://www.tensorflow.org/api_docs/python/tf/keras/layers/LSTM

[15] Chujie Lu, Sihui Li, and Zhengjun Lu. 2022. Building energy prediction using artificial neural networks: A literature survey. *Energy and Buildings* 262 (2022), 111718. https://doi.org/10.1016/j.enbuild.2021.111718

[16] X.J. Luo, Lukumon O. Oyedele, Anuoluwapo O. Ajayi, Olugbenga O. Akinade, Hakeem A. Owolabi, and Ashraf Ahmed. 2020. Feature extraction and genetic algorithm enhanced adaptive deep neural network for energy consumption prediction in buildings. *Renewable and Sustainable Energy Reviews* 131 (2020), 109980. https://doi.org/10.1016/j.rser.2020.109980

[17] Statistics Netherlands. 2023. Annual rate of change CPI. (2023). https://www.cbs.nl/en-gb/figures/detail/70936eng

[18] Government of the Republic of Moldova. 2023. Statistica Moldovei. (2023). https://statistica.gov.md/en

[19] Sklearn. 2023. KNNImputer. (2023). https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html

[20] Nivethitha Somu, Gauthama Raman M R, and Krithi Ramamritham. 2021. A deep learning framework for building energy consumption forecast. *Renewable and Sustainable Energy Reviews* 137 (2021), 110591. https://doi.org/10.1016/j.rser.2020.110591

[21] Statsmodels. 2023. Augmented Dickey-Fuller unit root test. (2023). https://www.statsmodels.org/dev/generated/statsmodels.tsa.stattools.adfuller.html

[22] Statsmodels. 2023. Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors model. (2023). https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html

[23] ENTSO-E Transparency. 2023. Electricity Market Transparency. (2023). https://transparency.entsoe.eu/dashboard/show

[24] Jian Qi Wang, Yu Du, and Jing Wang. 2020. LSTM based long-term energy consumption prediction with periodicity. *Energy* 197 (2020), 117197. https://doi.org/10.1016/j.energy.2020.117197

[25] Zhe Wang, Tianzhen Hong, and Mary Ann Piette. 2020. Building thermal load prediction through shallow machine learning and deep learning. *Applied Energy* 263 (2020), 114683. https://doi.org/10.1016/j.apenergy.2020.114683