Forecasting Household Electrical Energy Consumption with AI Hybrid Model Approach

Janssen Mitchellano Hamaziah
Undergraduate Student
Binus University
Jakarta, Indonesia
janssen.hamaziah@binus.ac.id

Louis
Undergraduate Student
Binus University
Jakarta, Indonesia
louis004@binus.ac.id

Alfi Yusrotis Zakiyyah
School of Computer Science
Binus University
Jakarta, Indonesia
Alfi.zakiyyah@binus.edu

Meiliana
School of Computer Science
Binus University
Jakarta, Indonesia
Meiliana@binus.edu

***Abstract*— Forecasting household electrical energy consumption plays a critical role in energy policy planning, resource management, and sustainable development. This paper proposes a hybrid approach to forecast electrical energy consumption, leveraging the capabilities of advanced algorithms to improve prediction accuracy and robustness. The study explores the application of various machine learning techniques, including regression methods such as Linear Regression, XGBoost Regression, Artificial Neural Network Regression. The research evaluates the proposed hybrid models using real-world energy consumption datasets, comparing their performance against other forecasting methods. Furthermore, the paper discusses the implications of accurate energy consumption forecasting for energy policy formulation, demand-side management, and renewable energy integration. Through empirical analysis and validation, this research contributes to the advancement of predictive modeling techniques in the domain of non-renewable energy consumption forecasting, facilitating informed decision-making and strategic planning for energy stakeholders and policymakers.**

Keywords*— *electrical consumption energy, energy prediction, machine learning, XGBoost, Artificial Neural Networks, hybrid model

I. Introduction

First, why is electricity consumption forecasting important? Underestimating the energy demand can lead to huge consequences such as energy supply shortages and forced power outages that affect productivity and economic growth. On the other hand, overestimating the demand of energy can lead to overinvestment in generator capacity, financial distress, and higher electricity prices. Accurate forecast of electricity demand informs the energy policymakers, power utilities, development professionals and private investors about investment decisions on power generation and supporting network infrastructure, meanwhile an inaccurate forecast can lead to dire social and economic consequences [1].

Then, what is a Hybrid AI model? In theory, a hybrid machine learning model is a model consisting of traditional machine learning algorithms and blend them with advanced AI techniques, resulting in a powerful model to tackle complex tasks with remarkable efficiency and accuracy. Key advantages of using hybrid model offers greater accuracy, versatility on data handling, efficiency on complex scenarios and improved learning model [2].

In this research, we will discuss the prediction of electricity consumption in accordance with the demands of society to help the government to maintain existing non-renewable resources. Non-renewable energy sources, including fossil fuels like coal, oil, and natural gas, continue to serve as primary contributors to global energy consumption despite the increasing efforts toward renewable energy adoption. Accurate forecasting of the consumption is essential for effective energy policy formulation, resource allocation, and sustainable development planning. Due to a significant boost on AI by Artificial Neural Network, there is growing interest in leveraging machine learning techniques to enhance the accuracy and reliability of non-renewable energy consumption forecasts. components, incorporating the applicable criteria that follow. With that, it will make it easier for the government to maintain what resources need to be spent. Forecasting energy consumption has a significant role in saving energy, reducing power generation costs, and improving social and economic benefits [3, 4]. Many researchers are using machine learning technology to develop energy consumption forecasting algorithms to help on energy policy, resource management, and sustainable development [5, 6, 7, 8, 9]. It is also essential to increase power demand and energy

load forecasting accuracy for the power system's stable and efficient operation. Non-renewable sources such as coal, natural gas, fossil fuels, nuclear, etc., cannot be regenerated in a brief time, meaning the consumption of non-renewable sources consumption rate far exceeds their regeneration rate. Since there are already models for forecasting energy consumption, this research aims to try experimenting with a hybrid model combining both XGBoost and ANN (Artificial Neural Network) methods for energy consumption forecasting. Since open-source dataset is difficult to get and the time given is within this semester, we decided to use the dataset of household energy consumption shown later. In this paper, we would like to try to minimize the expenditure of non-renewable energy and only use what is needed.

II. Related Works

The forecast of energy consumption that related to non-renewable energy has been a topic of research for many years. Research on the consumption of non-renewable energy is something that can be said to be important because in everyday life we need electricity which is needed from those non-renewable energy sources. The prediction is done to find out the number of daily needs needed by the community. To predict the demand for energy we will use the AI / ML approach. This prediction aims to estimate the needs received by the community.

Matos, et al. [10] decided to train three models, which are Artificial Neural Networks (ANN), Gradient Boost Decision Tree, and XGBoost, and will use the model that has the best performance for their requested data of the city of Bath, UK. They found that XGBoost performs the best among the others by comparing training time, prediction time, and some error measurements, such as Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Weighted Absolute Percentage Error (WAPE), and R^2 Error, and decided to use the model for energy consumption forecasting. Originally, this research aims to help reduce the bi-hourly tariffs, to optimize energy management, so that the cost is lower on some specific tariffs, which can provide higher savings for the community.

However, Khan et al. [11] felt that traditional prediction models are not good enough to forecast energy consumption and proposed a hybrid model consisting of Multilayer Perceptron (MLP), Support Vector Regression (SVR), and CatBoost. Surprisingly, the performance of the hybrid model works better than XGBoost for the dataset gained from Jeju Energy Corporation (JEC). The research was conducted to help JEC plan to change non-renewable energy to renewable sources, and to execute the plan, they will need to forecast the energy consumption on Jeju Island by integrating both renewable and non-renewable energy consumption and then trying to fulfill the forecasted demand by replacing the non-renewable energy generator to renewable energy generator.

From both journals, we would like to try to propose a new hybrid model that we find interesting to experiment with. Seeing the advances of AI technology on Artificial Neural Networks, and the performance of XGBoost on these journals, we would like to combine both ANN and XGBoost Regression to see whether this model performs better or worse than the other model.

III. Methodology

In this research, the method we use is a hybrid model that combines the XGBoost model with the ANN model. This research will experiment whether the hybrid method can have better speed and accuracy. However, before training and testing the model, we need to do feature engineering, EDA (Exploratory Data Analysis) and data pre-processing. The dataset in this research we took from Kaggle https://www.kaggle.com/datasets/uciml/electric-power-consumption-data-set. The following is a more detailed explanation:

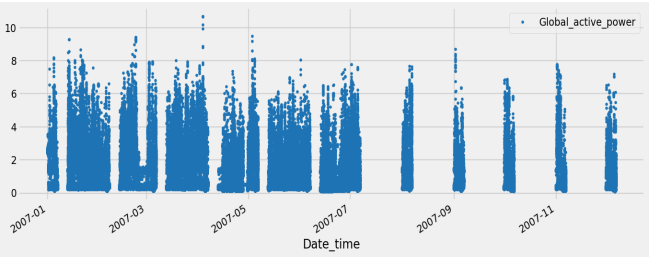

Fig. 1 Graph of global active power (in kilowatt) in a household per day

- *Feature Engineering*

Feature engineering stage is to transform raw data into relevant data that can be used for modeling. Feature selection consists of feature selection, feature transformation, feature creation and data cleaning. This will affect the speed performance and accuracy of the model to avoid overfitting. In this research we combined the date and time column as ‘date time’ column and changed the format to “date-time format.” This step is

important to make sure the data have a time series sequence. Because time series models need to have a proper date-time index and sequence pattern. This feature engineering process aimed to make the raw data to be relevant data. For the 'Global active power' columns we change to numerical format. This step format ensures that the data can be processed mathematically. In the feature engineering step we create a two-column variable as weekday and weekend. From data of those two columns, we can see that weekends consume more electrical energy than weekdays. This makes the data more explainable and relevant than only raw data. Because the model training needs to know that at weekends electrical consumption is slightly greater than normal day. The column needs to be separated by year, quarter, month, and day to determine date and time. After determining the time column, we sorted by ascending order to make the dataset sequential and time series. Make sure that the dataset used for training is relevant because irrelevant data influences model performance and accuracy. In this research we had tried to train the model, without declaring the weekdays and weekends variable and that really affects the performance model. The error rate is high because of irrelevant data. RMSE gets 0.608 by using hybrid models because of irrelevant data and some missing part on preprocessing. But after fixing the feature engineering and data preprocessing, we get 0.0272 for RMSE. From the description above we can conclude that feature engineering takes a significant role in this topic.

- *Exploratory Data Analysis (EDA)*

In the EDA stage, what needs to be done is to understand the data, such as patterns and anomalies. From understanding the data, we can make a hypothesis and analyze it. There are several EDA techniques performed in this research such as data visualization, pattern understanding, handling missing values, variable transformation, anomaly detection and hypothesis testing. EDA is for understanding the data before doing data preprocessing and modeling. Because of different data needs different ways to handle it, it is important to understand the pattern of dataset. In EDA we must make sure there are no null values. In our research dataset there are 3771 missing values in each submetering. The null values later will be handled in data preprocessing. And the most crucial part is to know whether the data had an anomaly data or not. If the data had anomaly data, data need to be normalized or standardized in preprocessing data. Pattern Understanding is a stage where interesting patterns are found in the data, such as the correlation of each variable and the effect on the distribution of the data.

From the dataset there are variable variables that have their respective roles. column variables as follows:

I. Date and time: Where the data is an independent variable that becomes a prediction benchmark for energy consumption. And later, the two variables will be combined to one column and change the format to 'date-time' format.
II. Global active power is an electric power system concept that refers to the total amount of global or overall active electricity from household electrical appliances. In the data, the unit for global active power is (KWh). Global active power is the most vital component because it reflects the total active power and data to predict energy consumption. OBJ This variable's values are shown in Fig. 1.
III. Submetering (1, 2, 3): Where there are class divisions like submetering 1 for the living room, submetering 2 for the kitchen, and 3 for the bathroom. So, the class is the main place in a household.

- *Data Preprocessing*

Data preprocessing is a crucial step where we process data before it is sent to be trained in the model. Data preprocessing in this research uses feature scaling techniques, remove unnecessary variables, transform variables, split datasets into train sets and test sets shown in Fig. 2.

Fig. 2 Graph of the data splitted to a train-test split

- *Modelling (Hybrid Model)*

So, in this modeling stage, we will train the model using XGBOOST and combined with ANN model. The range of window or moving average in this research we take 50 days (about 1 and a half months) before. And from that we can append and create x train, y train, x test, and y test. Firstly, we train using XGboost Model, and extract from the train evaluation. From that we combined the true x train and the feature extraction.

From these combining results, the x train is used to train ANN model. This research will experiment whether combining these two models can produce a better level of accuracy. So, after training with a combination of two models we do validation using RMSE, MSE, MAE, and R^2 to see the level of error rate and accuracy.

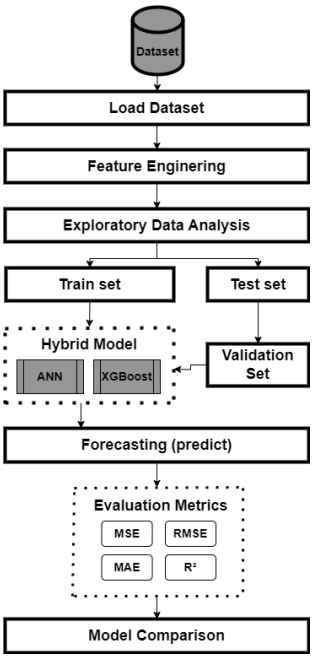

Fig. 3 Structure of the proposed hybrid model (XGBoost + ANN)

The model comparison we use will have the following:

A. *Polynomial Regression*

In this research we try to compare the polynomial regression model with our proposed hybrid model. Polynomial regression model is a machine learning model that is superior enough to perform regression. However, in the sources we read it is said that the polynomial model is not the best choice for performing time series-based regression. That is because polynomial regression models only capture temporal patterns and have difficulty understanding time correlation. But in our research this time, Polynomial regression produces a good RMSE with a score of 0.02811. According to our observations, this is due to the feature engineering process and supporting data preprocessing that helps make the data easier to process by the model. As a simple example, dividing the weekend and weekdays labels can help the polynomial regression model to read time patterns. However, for this research, the XGBoost model is slightly superior to the polynomial model with an RMSE score of 0.0277.

B. *XGBoost Regression*

The XGBoost model is one of the most powerful models. XGBoost is also very suitable for predicting regression time series. Basically, XGBoost was created to overcome overfitting machine learning models. The XGBoost model also has little chance of overfitting. Thus, we think this model can capture the complex data patterns in our dataset. XGBoost also has the advantage of being able to determine which features are useful and which are not, which can be used to make decisions by the XGBoost model. And when compared to other deep learning models XGBoost has speed in handling large and complex datasets and of course has accuracy that is not less good than Deep learning. In this research we use RMSE, MAE, and R^2 to determine how good this model is. The RMSE and MAE scores obtained by XGBoost are 0.0277 and 0.0102.

C. ANN (Artificial Neural Network.)

ANN model is one of the deep learning models that is powerful enough to perform a time series-based regression. This model is one of the models that has a simple structure compared to other deep-learning models such as LSTM or GRU. However, in this research, ANN is superior in processing data compared to the LSTM model. ANN also has an advantage in linearity so that this model can also be said to be flexible. When compared to other models compared in this research (except hybrid models), ANN has the lowest RMSE and the highest R^2. Which can be said to be good. The RMSE and R^2 of this model are 0.0273 and 0.9394. However, for the MAE value, XGBoost is superior with a value of 0.0102.

D. LSTM (Long Short-Term Memory)

Basically, the model we propose is a powerful model. And in this research, we tried to combine these two models (XGBoost + ANN). But we are interested in comparing ANN with LSTM model initially. The LSTM model has the advantage of being able to look at previous nodes and having long-term memory. LSTM is a model to improve the weakness of the RNN model which is vanishing gradient. According to some sources we read, LSTM is a good model in processing time series data. But in our research this time we are more interested in using the ANN model than LSTM. The reason is simple where ANN has a simpler structure than LSTM, so it will take faster time than LSTM. Not only that, ANN tends not to experience overfitting when compared to the LSTM model. We think that if we hybrid model using LSTM and XGBoost, then some of the shortcomings of the XGBoost model will be covered by the LSTM model that reads patterns well. So that the

model can read patterns better during training. This is where the risk of overfitting begins. Therefore, we use a simpler ANN rather than a more complex LSTM. And it should be noted that the difference in ANN testing error rate is smaller than LSTM, it is because the data we use counts per minute where if using the LSTM model, the model will overfit the pattern with the training model. But in this case LSTM does not experience overfitting, it is just that the ANN model is superior for this case.

E. Hybrid Model

The following steps are the steps of building this hybrid model:

- Feature Engineering: where data will be extracted or processed to give relevant data before training the model to prevent overfitting.
- Training The Hybrid Model:
  a. Train XGBoost: the training model using XGBoost techniques to ensure good prediction performance. After training use the train data to predict the train set. This prediction is used as a supporting feature for ANN. Data from the prediction results are integrated into the data set.
  b. Feature extraction and concatenate the train set. After extracting the model train set prediction from XGBoost model, then concatenate with the true train set. The train combination will be used to train the ANN model
  c. Train ANN: The training set is supported by the predictions made by XGBoost. Ensemble integration: combining the two prediction models above to make the better prediction
  d. XGBoost : Basically, XGBoost is summing multiple decision trees to boost the model from getting underfitting. This Model is quite powerful for forecasting time series data. In the formula above K stands for number of trees.

The ANN model in this research used ReLU activation function and the last unit used linear activation. The error rate for this model is low and gives good accuracy. So, we conclude that this model is good for real world applications. But the aim of this research is to make a better version of this topic. Because electrical energy consumption is playing a critical role for the economics and the government planning for a better future.

IV. Result And Discussion

The evaluation metrics we used in this research are RMSE, MSE, MAE, and R^2 to compare our proposed model to other models.

Models	RMSE	MAE	MSE	R^2
Polynomial Regression	**0.02811**	**0.0109**	**0.0008**	**0.9363**
XGBoost	**0.0277**	**0.0102**	**0.0008**	**0.9380**
ANN	**0.0273**	**0.0109**	**0.0007**	**0.9394**
LSTM	**0.0285**	**0.0124**	**0.0008**	**0.9341**
Hybrid Model (XGBoost + ANN)	**0.0272**	**0.0105**	**0.0007**	**0.9402**

TABLE I Model Comparison of Several Regression Model to The Proposed Model

As shown in table 1, the proposed model performs the best among the other models. XGBoost is shown to have a better performance in MAE to ANN, but ANN has a better performance in the rest of the evaluation metrics than XGBoost. We conclude that the hybrid model is performing better as both XGBoost and ANN are covering both their disadvantages and excel.

V. Conclusion

Forecasting energy consumption is a crucial problem to be overseen. This problem allows the planning of distributing the energy, controlling the cost of energy generating, and capturing the generated energy to a storage and its distribution. Forecasting energy consumption can also be used to plan to reduce non-renewable sources generators and boost the renewable sources generator without shutting down the entire city or town.

Several machine learning models have been developed to overcome this problem, such as SVR, Moving Average, Grey, ARIMA, etc. Not only machine learning models are used, but deep learning models are also developed to overcome this problem, such as the use of LSTM, ANN, MLP and so on. Since this problem has surfaced and received several attention, hybrid models are also developed to fit on this problem. With our pique of interest in developing a hybrid model to overcome this problem, we proposed a hybrid method of combining both ANN and XGBoost. As a matter of fact, both ANN and XGBoost are already independently powerful in solving this case and our proposed method is just our piqued interest in combining both powerful models.

To conclude this research, our proposed model has a low error rate and high accuracy and for comparison with other models, the hybrid model is superior, in terms of accuracy, though it only differs slightly to other models. The models that we combined both have their own advantages and disadvantages and combining them will complement the disadvantages of each model. In this research, we show that the hybrid model is feasible to be used in real-world applications but must also be considered for its compatibility with existing data.

References

[1] J. S. J. d. Wit, A. Kochnakyan, and V. Foster, "Forecasting Electricity Demand: Why are electricity-demand forecasts important?," Saylor Academy. https://learn.saylor.org/mod/book/view.php?id=61397#:~:text=Underestimating%20demand%20results%20in%20supply (accessed Jun. 19, 2024).

[2] "What is Hybrid Machine Learning?," Polymer. https://www.polymersearch.com/glossary/hybrid-machine-learning

[3] S. Ozturk and F. Ozturk, "Forecasting Energy Consumption of Turkey by Arima Model," Journal of Asian Scientific Research, vol. 8, no. 2, pp. 52–60, 2018, doi: https://doi.org/10.18488/journal.2.2018.82.52.60.

[4] F. Lu, F. Ma, and S. Hu, "Does energy consumption play a key role? Re-evaluating the energy consumption-economic growth nexus from GDP growth rates forecasting," Energy Economics, vol. 129, pp. 107268–107268, Jan. 2024, doi: https://doi.org/10.1016/j.eneco.2023.107268.

[5] victorchawsukho, "PowerConsumption_w_ML_n_BusinessView," Kaggle.com, 2024. https://www.kaggle.com/code/victorchawsukho/powerconsumption-w-ml-n-businessview (accessed Jun. 19, 2024).

[6] B. Zohuri, F. Behgounia, and Z. Z. Nezam, "Artificial Intelligence Integration with Energy Sources (Renewable and Non-renewable) -David Publishing Company," www.davidpublisher.com. https://www.davidpublisher.com/index.php/Home/Article/index?id=44556.html (accessed Jun. 19, 2024).

[7] A. Hussein and M. Awad, "Time series forecasting of electricity consumption using hybrid model of recurrent neural networks and genetic algorithms," Deleted Journal, pp. 100004–100004, Mar. 2024, doi: https://doi.org/10.1016/j.meaene.2024.100004.

[8] Wang, X., Wang, H., Bhandari, B. et al. AI-Empowered Methods for Smart Energy Consumption: A Review of Load Forecasting, Anomaly Detection and Demand Response. Int. J. of Precis. Eng. and Manuf.-Green Tech. 11, 963–993 (2024). https://doi.org/10.1007/s40684-023-00537-0

[9] H. Yang, M. Ran, and C. Zhuang, "Prediction of Building Electricity Consumption Based on Joinpoint−Multiple Linear Regression," Energies, vol. 15, no. 22, p. 8543, Jan. 2022, doi: https://doi.org/10.3390/en15228543.

[10] M. Matos, J. Almeida, P. Gonçalves, F. Baldo, F. J. Braz, and P. C. Bartolomeu, "A Machine Learning-Based Electricity Consumption Forecast and Management System for Renewable Energy Communities." Energies, vol. 17, no. 3, pp. 630-654, 2024, doi: 10.3390/en17030630. [Online]. Available: https://www.mdpi.com/1996-1073/17/3/630.

[11] P. W. Khan, Y.-C. Byun, S.-J. Lee, D.-H. Kang, J.-Y. Kang, and H.-S. Park, "Machine Learning-Based Approach to Predict Energy Consumption of Renewable and Nonrenewable Power Sources." Energies, vol. 13, no. 18, pp. 4870-4885, 2020, doi: 10.3390/en13184870. [Online]. Available: https://www.mdpi.com/1996-1073/13/18/4870.