energies

MDPI

Article

A Meta-Modeling Power Consumption Forecasting Approach Combining Client Similarity and Causality

Dimitrios Kontogiannis [1,*], Dimitrios Bargiotas [1], Aspassia Daskalopulu [1] and Lefteri H. Tsoukalas [2]

[1] Department of Electrical and Computer Engineering, School of Engineering, University of Thessaly, 38221 Volos, Greece; bargiotas@uth.gr (D.B.); aspassia@uth.gr (A.D.)
[2] AI Systems Lab, School of Nuclear Engineering, Purdue University, West Lafayette, IN 47907, USA; tsoukala@purdue.edu
* Correspondence: dimkonto@uth.gr

Abstract: Power forecasting models offer valuable insights on the electricity consumption patterns of clients, enabling the development of advanced strategies and applications aimed at energy saving, increased energy efficiency, and smart energy pricing. The data collection process for client consumption models is not always ideal and the resulting datasets often lead to compromises in the implementation of forecasting models, as well as suboptimal performance, due to several challenges. Therefore, combinations of elements that highlight relationships between clients need to be investigated in order to achieve more accurate consumption predictions. In this study, we exploited the combined effects of client similarity and causality, and developed a power consumption forecasting model that utilizes ensembles of long short-term memory (LSTM) networks. Our novel approach enables the derivation of different representations of the predicted consumption based on feature sets influenced by similarity and causality metrics. The resulting representations were used to train a meta-model, based on a multi-layer perceptron (MLP), in order to combine the results of the LSTM ensembles optimally. This combinatorial approach achieved better overall performance and yielded lower mean absolute percentage error when compared to the standalone LSTM ensembles that do not include similarity and causality. Additional experiments indicated that the combination of similarity and causality resulted in more performant models when compared to implementations utilizing only one element on the same model structure.

Keywords: power forecasting; energy; machine learning; neural networks; artificial intelligence; data analysis; feature engineering; ensemble neural networks; meta-modeling

Citation: Kontogiannis, D.; Bargiotas, D.; Daskalopulu, A.; Tsoukalas, L.H. A Meta-Modeling Power Consumption Forecasting Approach Combining Client Similarity and Causality. *Energies* **2021**, *14*, 6088. https://doi.org/10.3390/en14196088

Academic Editor: Yungcheol Byun

Received: 24 August 2021
Accepted: 21 September 2021
Published: 24 September 2021

1. Introduction

Data analysis and forecasting models are the cornerstones of research in the energy sector since they enable the development of sophisticated applications and strategies that optimize the flow of energy on the grid and improve the quality of life of electricity consumers. Modern data-driven approaches rely on the collection and processing of client information regarding their power consumption, socio-demographic features, and various external factors, such as weather variables, in order to examine consumption patterns and make accurate predictions. Forecasting models focused on the prediction of power consumption provide meaningful insights that can be utilized by electricity providers in order to monitor and control the demand efficiently, while being able to detect and avoid irregular events. On a larger scale, power forecasting models allow the providers to construct load profiles of buildings for months in advance and, based on that estimate, to validate energy meter readings and cluster buildings into groups, contributing towards more intelligent regional planning approaches. Additionally, smart pricing strategies can be implemented in an attempt to adjust electricity tariffs dynamically, based on client behavior. Price forecasting techniques complement electricity consumption models in that spectrum, with significant contributions towards the efficient execution

of regression tasks [1–3]. Furthermore, electricity forecasts can benefit each consumer individually due to the development of applications that allow clients to monitor and reschedule their daily tasks flexibly in order to gain additional control over the billing process. Consequently, there is a growing interest for the development of accurate and robust power forecasting models that are able to extract useful information from the underlying patterns and relationships of the collected energy data [4–6].

Energy data used in the design of forecasting models is commonly structured in the form of time series, where records consist of relevant features indexed in time order. Classical time series forecasting methods such as autoregression (AR) [7], moving-average (MA) [8], and autoregressive moving average (ARMA) are often used to predict the next time step in a univariate sequence modeled as a linear function of information extracted from previous time steps. Moreover, the autoregressive integrated moving average (ARIMA) method and its extensions [9] utilize differencing in the observations of previous time steps. Vector autoregression (VAR) models constitute a generalization of AR models since they support multivariate time series. A similar generalization is observed in vector autoregression moving-average models (VARMA) [10]. Additionally, simple exponential smoothing (SES) [11] and Holt Winter's exponential smoothing (HWES) [12] model the next time step as an exponentially weighted linear function of past observations. Traditional methods, such as AR, MA, ARMA, VAR, VARMA, and SES, usually do not utilize the trends and seasonal patterns of the input sequence and, while their extensions and variants can integrate those elements to construct more sophisticated models, there are more limitations associated with those methods. The limitations of those statistical methods mainly revolve around the structure of the available data, the relationship between input and output variables, and the ability to support highly dimensional time series [13,14]. Since classical forecasting models do not support missing values, data imputation techniques need to be implemented, thus altering the original dataset. Furthermore, traditional models often generate predictions based on the assumption of a linear relationship between the input and output variables, thereby omitting the more complex non-linear patterns and trends. Lastly, these classical methods are observed to be more suitable on univariate sequences in terms of performance, rendering the design and generalization process for more complex environments more difficult. It is clear that due to the complexity, availability, and structure of many energy datasets, most traditional approaches would not suffice for the derivation of accurate predictions.

On the other side of the spectrum, advances in artificial intelligence and machine learning led to the development of more robust models, which are capable of discovering complex relationships between input and output features. Many different architectures involving neural networks, such as the multilayer perceptron (MLP) [15] and long short-term memory (LSTM) network [16], were used successfully in many time series forecasting tasks, achieving impressive performance [17]. These neural network models follow a black-box approach in the approximation of nonlinear functions. The multilayer perceptron finds frequent application in regression, classification, and fitness approximation tasks with an emphasis on learning to map the set of inputs to the set of outputs. Long short-term memory networks take advantage of the temporal data characteristics in order to extract insights from the order dependencies that could be present in a sequence. The suitability of machine learning methods for energy data processing is evident since these models are able to capture more complex patterns from highly dimensional data without the requirement of having an optimally structured dataset. However, there are still many challenges that limit the performance of these models, and ongoing research attempts to address them. The lack of data needed to train a model successfully in combination with potentially missing values could hinder the performance of neural networks due to overfitting [18], since the model would not have an adequate number of training examples in order to perform well when new data is tested. Additionally, feature engineering is crucial in the design of a machine learning model, since the inclusion or exclusion of certain variables and data transformations can have a great impact in the learning process.

Therefore, some forecasting tasks in the energy sector can have poor performance due to a suboptimal data collection process or limited data availability given the forecasting horizon and the expected output. Ongoing research in the field focuses on the introduction of novel methods and hybrid models that utilize a combination of feature engineering techniques, architectural changes, and mechanisms that optimize the training process, thus rendering neural network models more resilient to data abnormalities. Additionally, there is an ongoing effort towards creating more well-structured datasets, while minimizing data distortion and noise for energy forecasts [19,20].

There are several recent projects addressing forecasting and classification tasks with the use of data-driven methods that often utilize neural networks and feature engineering techniques. Choi and Lee [21] proposed a framework based on an LSTM ensemble and a weighted combination of predictions for time series forecasting, showing that combinatorial approaches that utilize the output of multiple neural networks can achieve better performance compared to other popular forecasting methods. Tian et al. [22] presented a hybrid architecture based on the combination of a LSTM and a convolutional neural network (CNN) for short term load forecasting, improving prediction stability for that forecasting horizon. Mujeeb et al. [23] used a deep LSTM network to create a new load forecasting scheme for big data in smart cities, showing the capabilities of deep neural networks on highly dimensional historic load and price data. Markovič et al. [24] reinforced the importance of optimal data aggregation by presenting a data-driven method for the classification of energy consumption patterns based on functional connectivity networks. Jin et al. [25] proposed an encoder-decoder model utilizing an attention mechanism in order to learn long data dependencies from the input sequence efficiently. Tian et al. [26] developed a forecasting model based on transfer learning, using the outputs of already trained models for the estimation of building consumption according to similarity measures. This project provided substantial motivation towards research on meta-modeling techniques that could improve the accuracy of the predicted smart meter readings. Chen et al. [27] proposed a time series forecasting model that explored the impact of Granger causality for stock index predictions. This work presented interesting ideas on the use of causality in prediction models and could be extended to the field of energy forecasts. Boersma [28] studied the correlation and impact of internal and external factors on the prediction of household consumption using an MLP network. This project highlighted the importance of feature engineering as well as time resolution in the derivation of accurate predictions. Emamian et al. [29] implemented a solar power forecasting model using an LSTM ensemble to aggregate the predicted output of each network, demonstrating that ensemble models can achieve higher accuracy and more reliable results than single neural network models. Guo et al. [30] combined energy consumption and environmental data in the development of an LSTM forecasting model. Their study suggested that decent forecasting performance can be achieved when a good quality dataset is available. Lastly, Tao et al. [31] proposed a hybrid short-term forecasting model using an LSTM network for photovoltaic power predictions in conjunction with a bias compensation LSTM in an attempt to improve the predictions based on the residual error. This project highlighted the more positive effects of meta-modeling in neural network design and showed that there is more useful information to be extracted from the predicted output.

In this study, we focused on the prediction of power consumption extracted from monthly energy meter readings for electricity clients. Since the energy meter data is often collected monthly for each household or building, and the data collection process is dependent on the policies of the electricity provider, it is common for the resulting dataset structure to be problematic for most modern forecasting models. Insights and predictions are commonly based on patterns and trends extracted from recent years of consumption. Therefore, it is expected that a dataset containing monthly measurements might not have sufficient records for the training process of neural networks. Furthermore, due to different provider policies and the possibility of having a manual registration of the meter readings, the resulting datasets often contain missing or estimated values for clients that do not have

any electric energy meters installed. Consequently, machine learning models trained on such data would probably overfit or exhibit poor performance on both training and test sets. The main goals of this study were: to develop a combinatorial neural network model that manages to outperform the standard single network forecasting approach, while avoiding overfitting; and to demonstrate the impact of feature engineering in the implementation of a meta-modeling technique. The proposed model examined the impact of similarity and causality among clients in an LSTM ensemble architecture in order to derive the base, similar, and causal representations of the predicted output based on changes of the input feature set. Following this step, a multilayer perceptron was used to aggregate the predicted results, in order to discover the optimal combination of those representations that could be used to predict the actual power consumption more accurately, by formulating a meta-model for stacked generalization. This project aimed to stimulate further research in the design of models that do not rely on well-structured datasets, but rather explore the inclusion of potentially helpful features that express relationships between client time series, in order to improve the base performance of models that would otherwise be considered suboptimal. Our work contributes towards the study of influential features and the discovery of patterns within the communities of electricity clients. Additionally, the examination of combinatorial forecasting approaches, similar to the ones presented in this paper, help in the presentation of more complex ideas and greatly expand research knowledge through the investigation of alternative models.

In Section 2, we analyze the main methods utilized in the implementation of the proposed model and proceed to provide the forecasting problem framing, with appropriate references to the dataset and performance metrics used as a case study in order to test the performance of this approach. In Section 3, we discuss the results of our experiments and evaluate the performance of our model. Finally, in Section 4, we highlight the impact of the experimental results and address the advantages as well as the challenges of this approach. Additionally, we outline some ideas for further testing and improvement of this method for future research projects.

2. Materials and Methods

2.1. Long Short-Term Memory Networks

Long short-term memory networks [32] are a class of recurrent neural networks (RNN) that can identify long-term dependencies among the input features. LSTM networks are valuable tools for time series forecasting tasks, since they can perform well, when the duration between time lags of a given sequence is unknown. Additionally, LSTMs manage to preserve gradients throughout the computation, solving one of the main issues of RNNs, where the gradients would vanish during the training process. The structure of an LSTM consists of units known as the LSTM cells. Each cell contains a set of gates that can adjust the current cell state by adding or removing information at a given time step. The cell state is transferred from one unit to the next, where further adjustments occur. The input gate at time step t determines which values will be updated and stored in the cell state. Additionally, the output gate determines which parts of the current cell state will be transferred to the output of the cell, leading to the next unit. The following formulae describe the input gate i_t and the output gate o_t at the time step t, where x_t is the sequence information at the current timestep, h_{t-1} is the output of the previous LSTM unit, w_i and w_o are the respective weights of those gates, b_i and b_o are the biases, and σ is the sigmoid function:

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \quad (1)$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \quad (2)$$

The forget gate of the LSTM cell at time step t is similarly formulated as f_t and determines the sequence information that will be removed from the cell state. Given its

corresponding weight w_f and bias b_f the results of the forget gate are computed with the formula:

$$f_t = \sigma\left(w_f[h_{t-1}, x_t] + b_f\right) \tag{3}$$

The candidate sequence information that could potentially be stored to the current cell state is expressed as d_t and, given the respective weights w_d and biases b_d, is formulated as:

$$d_t = tanh(w_d[h_{t-1}, x_t] + b_d) \tag{4}$$

The current cell state at time step t is expressed as c_t and is formulated as the combination of the previous cell state c_{t-1} and the candidate information given the adjustments determined by the forget gate and the input gate. The following formula shows this combination:

$$c_t = f_t * c_{t-1} + i_t * d_t \tag{5}$$

Finally, the information that reaches the output of the cell at time step t is expressed as h_t and is given by the following formula:

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

An overview of the LSTM cell is presented in Figure 1 where each symbol corresponds to the respective symbols explained in Formulae (1)–(6) and the multiplication as well as the addition blocks connect the terms of each formula in this diagram. LSTM networks are trained with back propagation through time and gradient descent [33].

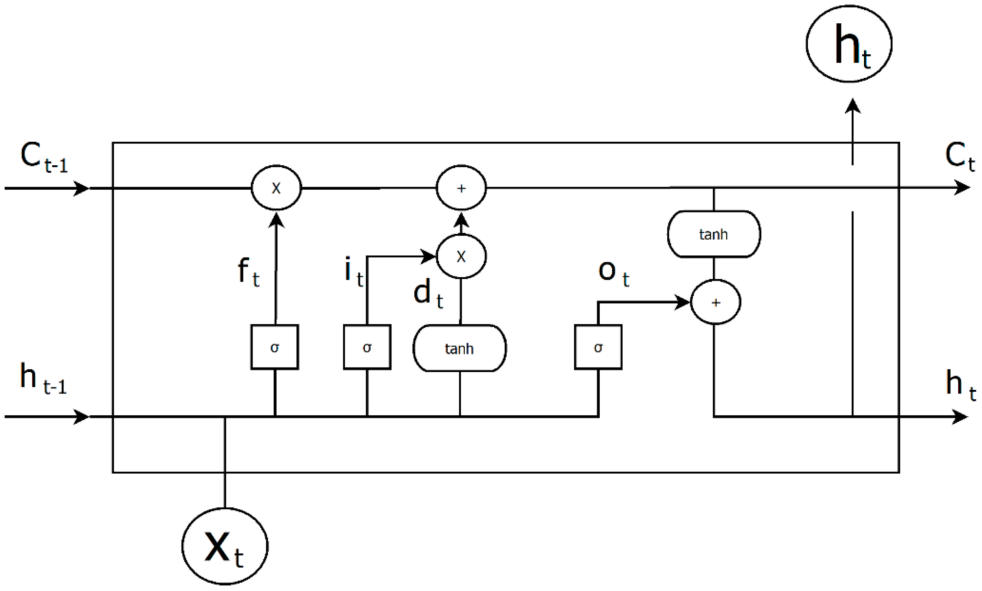

Figure 1. LSTM cell structure.

In the literature, several experiments were conducted with different LSTM variants, including a variable number of units and hidden layers as well as custom training loops for sequence forecasting. However, it is evident that, while changes in the structural parameters of an LSTM can boost model performance and achieve faster training time through faster convergence, stable and reliable results are derived from the aggregation of multiple LSTMs and the construction of ensemble models [34]. Therefore, for the purposes of this study, an LSTM ensemble was considered for the forecasting experiments and the weighted average of the ensemble members was used for each representation of the predicted output. It is worth mentioning that ensemble models can also yield small performance boosts compared to the standalone LSTM, but in this project, we focused

more on the stability and reproducibility that ensembles can ensure. Figure 2 presents the general ensemble LSTM structure.

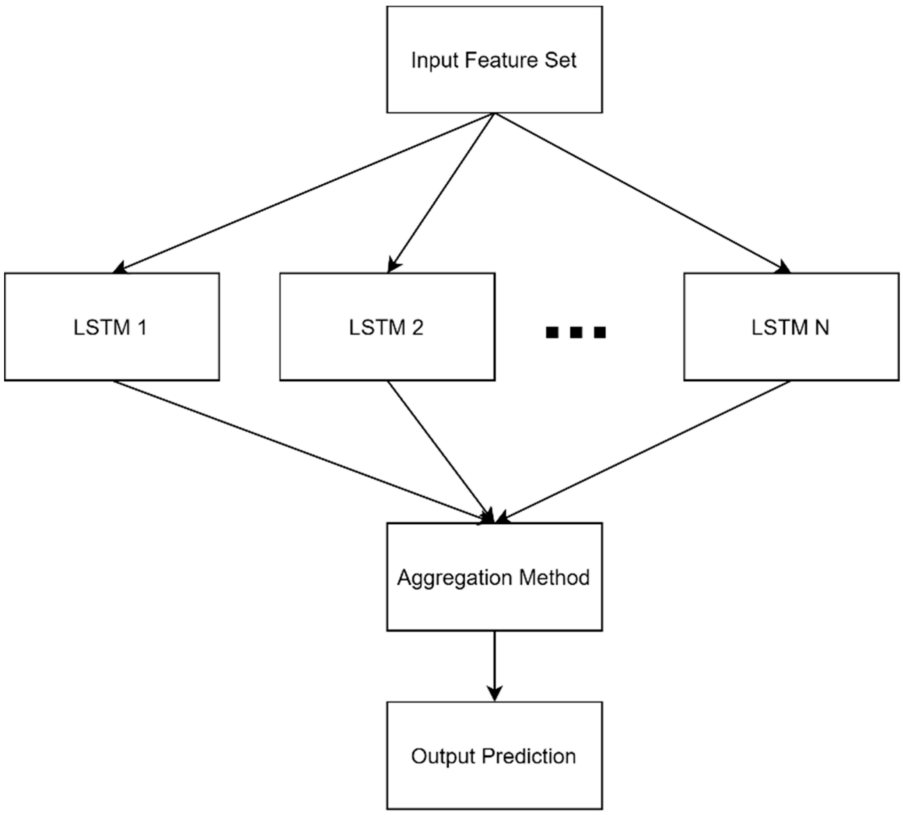

Figure 2. General LSTM ensemble structure.

2.2. Multi-Layer Perceptron

The multi-layer perceptron is a class of feed forward artificial neural networks (ANN) that extends the perceptron learning algorithm. The structure of MLP shares the same simple characteristics of ANNs, where the main units that form the network are neurons arranged in layers. These neural networks consist of the input layer, a set of hidden layers, and an output layer [35]. As the information flows from the input nodes to the output, weights and biases in the intermediate layers are adjusted to minimize the values of an error function that determines the performance of the model. Multi-layer perceptron networks are valuable function approximators that solve problems stochastically and yield decent performance in supervised learning tasks, such as classification and regression. For this study we consider a simple MLP with one hidden layer that approximates the function $f: R^D \rightarrow R^L$, where D denotes the size of an input vector x, and L denotes the size of the output vector $f(x)$. Formula (7) denotes the computation of output for a simple MLP network described in matrix notation, showing that from the input layer to the hidden layer, weight matrix $W^{(1)}$ and bias vector $b^{(1)}$ are adjusted and the result passes through an activation function s as the output of the hidden layer. Following this step, weight matrix $W^{(2)}$ and bias vector $b^{(2)}$ are adjusted as the information flows from the hidden layer to the output layer, with the result passing through an activation function G:

$$f(x) = G\left(b^{(2)} + W^{(2)}\left(s(b^{(1)} + W^{(1)}x)\right)\right) \quad (7)$$

The training process of MLPs can vary greatly based on the design of the model, but the most common method is back propagation with gradient descent for the calculation of weights. Time series models and applications that handle energy data often utilize MLPs for univariate and multivariate regression tasks. Alternatively, MLP networks can classify

load profiles as well as other variables that could group clients into distinct categories. For the purposes of this study, multi-layer perceptron was used as a meta-modeling prediction approximator that aggregates the results of LSTM ensembles and learns to predict the expected output through stacked generalization, in the spirit of [36]. Figure 3 presents the simple MLP structure discussed in this section.

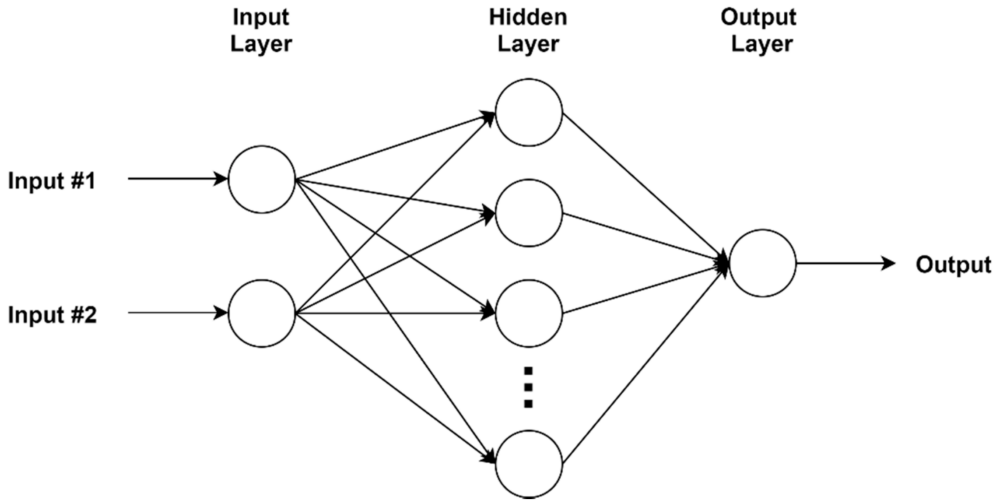

Figure 3. MLP with a single hidden layer.

2.3. Influential Community Factors

Feature engineering techniques [37] are useful because they contribute to the discovery of relationships and patterns between input features. Additionally, those methods assist towards an insightful ranking of features derived from the results of metrics and algorithms, leading to the inclusion of the most impactful features or exclusion of the least beneficial data records. Consequently, it is evident that the role of feature engineering techniques is crucial in the design of performant models for most machine learning tasks. In this study, we focused on the development of a forecasting model that utilizes the power consumption data of clients. In the literature, studies involving external variables, such as temperature and price, are common in this class of forecasting models and the focus is shifted towards the impact of additional data on a specific load profile through time. While the inspection of external variables is beneficial in the development of accurate forecasting models, we should also consider the discovery of interrelationships among the load profiles of clients and the overall community impact when selecting features that are extracted from a wider pool of consumers. The exploration of this approach could lead to the design of performant models after the investigation of features that discover associations between the power consumption of buildings. These associations can be useful when the data collection process is not ideal and external variables are not available. Additionally, models based on influential community features show the relative evolution of power consumption patterns, which is worth monitoring when electricity providers, as well as customers, want to estimate electricity tariffs and ensure that energy meters function properly. For the purposes of this study, we explored the effect of two influential community elements, namely similarity and statistical causality.

2.3.1. Similarity

Similarity metrics [38] quantify the structural closeness of features and rank them in order to find the ones most similar to the given input. In power forecasting tasks, where the similarity between the power consumption time series of clients is considered, the main goal is to create client associations by finding the most similar power consumption time series within the community, given the time series of one client. Since power consumption time series can vary in length or have missing values due to irregularities that occur

during the data collection process, we can easily observe that conventional distance metrics that assume optimal time series alignment, such as Euclidian distance, could produce a pessimistic dissimilarity measure due to the absence of a symmetrical point-by-point match of the time series or, in some cases, misinterpret the similarity of some time series. Therefore, we chose to examine the soft dynamic time warping (soft-DTW) algorithm [39] for this project. Soft-DTW is a differentiable loss function for time series and constitutes an extension of the dynamic time warping algorithm [40] for the computation of the best time series alignment through a dynamic programming approach. As a similarity measure, soft-DTW considers all alignment matrices of two time series and produces a score that encapsulates the soft-minimum of the distribution of all costs spanned by all possible alignments. This method yields decent performance in classification and regression tasks involving time series and is considered a useful metric that can serve different purposes in the design and training of a neural network model. In detail, for the comparison of two time series x and y with respective lengths n and m, given the cost matrix $\Delta(x, y)$, the inner product of Δ, with an alignment matrix A as $\langle A, \Delta(x, y)\rangle$, and the proposed generalized operator min^{γ} with the non-negative smoothing parameter γ, soft-DTW is computed with the formula:

$$softDTW(x, y) := min^{\gamma}\{\langle A, \Delta(x, y)\rangle, A \in A_{n,m}\} \quad (8)$$

2.3.2. Statistical Causality

Causality [41] usually refers to the abstract concept that defines a relationship between two variables, where the influence of one can partially justify the value of the other. Typically, when causality is present there are covert dependencies between those variables and discovering them can be useful in the construction of forecasting models. When variables have a simple structure and include descriptive labels, it can be easy to distinguish the causality between them through intuition and logical reasoning, but this is not the case for more complex data. Time series data are usually collected from complex and dynamic systems and due to their structure, the detection, quantification, and interpretation of causality are challenging tasks. The relationship of cause and effect in time series could describe partial dependencies of values on the same time step as well as changes caused to the values of one sequence due to the effect of past observations of another.

Statistical causality methods, such as Granger causality [42], attempt to determine the forecasting potency of one series with regards to another. The derivation of this predictive causality is a useful tool that could complement similarity measures and feature correlations when data analysis is performed. In the scenario of power consumption forecasting the role of statistical causality is twofold. First, models that rely on lagged observations of consumption can distinguish the most influential lags for prediction by eliminating the observations that fail the statistical causality tests. Second, in the scope of a wider client pool, the forecasting potency of a lagged observation that belongs to one client with regards to future consumption of a different client can enable data augmentation due to the significance of the underlying patterns that led to this causal relationship.

In this project, we utilized the Granger causality test to infer the predictive potency of power consumption time series. The Granger causality test is a bottom-up process, where the null hypothesis states that lagged values of a variable x do not explain the variation in variable y, hence x does not Granger-cause y. The p values of the chi-square and F distributions are compared to the desired statistical significance and the results can be interpreted with the following formula:

$$result = \begin{cases} Reject\ Null\ Hypothesis, & p < 0.05 \\ Accept\ Null\ Hypothesis, & p \geq 0.05 \end{cases} \quad (9)$$

2.4. Problem Framing and Proposed Design

This study focused on the design and implementation of a power consumption forecasting model for monthly consumption predictions based on the collection of energy

meter readings from a set of clients. This model attempted to integrate neural network architectures, feature engineering techniques, and a meta-modeling process in order to address the main challenges of this prediction horizon, as well as the difficulties that could arise due to a non-ideal data collection process.

Neural network models processing monthly client data with the common sliding window approach [43] often have insufficient observations for training, resulting in models that either overfit or have poor performance. Additionally, difficulties in the data collection process can result in unbalanced datasets with missing values that can affect model performance. It is also worth noting that changes in the original dataset addressing these problems could lead to the introduction of unnecessary noise, resulting in the misinterpretation of certain data patterns. Despite all the challenges mentioned above, research should not be limited to good quality datasets because data availability cannot be guaranteed for all machine learning tasks. Furthermore, the investigation of alternative approaches that could boost model performance in a non-ideal setting is interesting since those contributions shift the emphasis towards more robust structures that overcome data limitations. Our proposed approach maintained the original dataset following the common sliding window approach for predictions, while it introduced a revised model structure that can improve model performance under non-ideal conditions.

Following the sliding window approach using an LSTM network, the client dataset underwent the preprocessing phase, where the consumption dataset was clustered by client and the data of the client whose consumption was to be predicted was selected. The consumption of the client was organized into different columns representing lagged observations of the consumption at time $t-1$, $t-2, \ldots, t-n$ derived from the original time series shifted in time. The prediction of the next month was denoted as the next column at time t, which is the target output variable. The preprocessed dataset was split into a training and validation set, and the data was scaled appropriately based on the distribution of each feature, using standardization when the distribution was normal or using normalization otherwise. The scaled features were reshaped in the form of $[samples, timesteps, features]$ and passed to the LSTM network for training. Once the model was trained, the performance of the model was evaluated on new data from the test set and an error metric was used to determine the divergence of predictions from the actual consumption values. Figure 4 presents the standard design for a forecasting model utilizing an LSTM network and this common sliding window method that usually underperforms due to the challenges mentioned above.

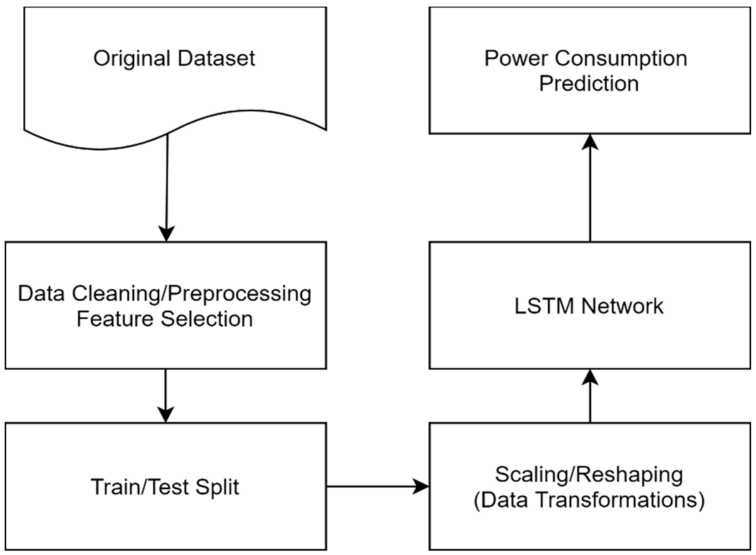

Figure 4. Standard design of power forecasting model using an LSTM network.

We extended the previously described method by introducing several modifications and new components aiming at a performance boost: instead of a standalone LSTM network, an LSTM network ensemble of n members was considered in order to derive more reliable predictions. Each member of the LSTM ensemble features an early stopping mechanism [44] that effectively stops the training process when the validation loss of the model stops decreasing. This mechanism prevents overfitting and selects the epoch where the model would achieve the best predictive performance on unseen data. The prediction of the LSTM ensemble is the aggregate prediction of the members derived from their weighted average, where the weights are determined using grid search [45]. This process was used on the original feature set of lagged observations to derive the base representation of the predicted consumption. Following this step, two different feature sets were constructed based on the influential community factors of similarity and causality. The first feature set contained the original lagged observations, as well as lagged observations of other clients at the same time steps, determined by their similarity ranking based on the previously described soft-DTW method. The second feature set contained the original lagged observations and lagged observations of other clients at the same time steps based on their predictive potency as determined by a Granger causality test when the targeted effect was the monthly consumption at time t of the original client in the training set. Those two feature sets passed through the LSTM ensemble and produced the similar and causal representations of the predicted consumption, respectively. Since the effects of causality and similarity are not strictly predetermined to be positive, we implement a meta-modeling technique that aimed to aggregate the three representations in order to create a model that discovered a weighted combination of those representations. Intuitively, this combinatorial model used an MLP as a meta-learner used for stacked generalization [46], thus creating an ensemble of ensembles that was expected to yield improved performance when compared to single layer LSTM ensembles. Figure 5 presents the design of the proposed approach.

2.5. Case Study and Experiments

In this section, we present a case study used to test our proposed approach. The dataset used for our experiments contained monthly power consumption data of clients located in seven municipalities of Nariño, Colombia from December 2010 to May 2016, and it is freely and publicly available [47]. The data was collected and registered by workers of the company Centrales Eléctricas de Nariño (CEDENAR) after manual inspection of electric energy meters installed at the building of each individual client. The consumption measurements were obtained from the monthly readings of those meters in kWh. The only exception to this manual procedure happened in the case where a client did not have an energy meter installed. In this situation, the estimated consumption derived from the installed electric load of the connected appliances was used. Additionally, the dataset contains socio-demographic features such as area, municipality, use, and stratum that further describe each client. The key index that uniquely identified each client is a code that includes a concatenation of the socio-demographic characteristics. According to the authors of the paper that introduces the dataset, the data was processed and ready for direct use in the implementation and testing of forecasting models. Furthermore, time series data for each client can be extracted when the observations are clustered by the unique code identifier. After further inspection, we deduced that this dataset was suitable for testing since the pool of clients is sufficiently diverse, containing clients that live in rural and urban areas, while using electricity in different environments ranging from residential to industrial and special. Furthermore, the feature of power consumption values is equally diverse, ranging from 1.009 to 305,687.4 kWh. Therefore, the exploration of influential community factors, such as similarity and causality, for the individual power consumption forecasts could be interesting as the dataset includes clients that satisfy a wider spectrum of consumption scenarios.

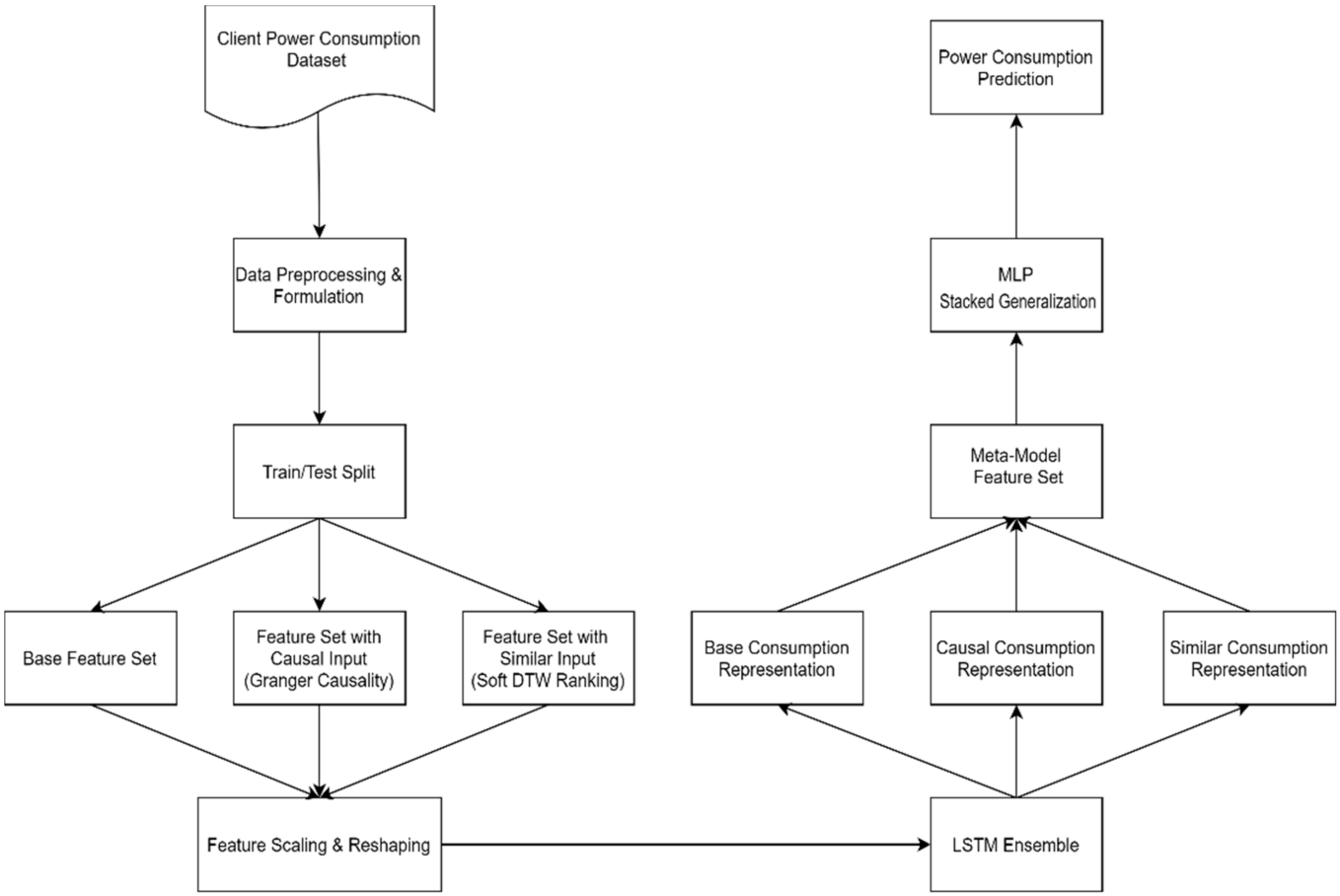

Figure 5. Proposed model design utilizing three LSTM ensemble sub-models in the development of a meta-model based on an MLP network.

Firstly, further inspection of the dataset was conducted and additional preprocessing was necessary for the extraction of the consumption time series for each client. Clients were clustered by code with the requirement that each date index contains one consumption measurement for that month. Consequently, 90 clients were detected and formed a new time series dataset. The resulting dataset fit the non-ideal scenario we wished to explore in this project, since it contains several missing values, possibly due to the manual registration process. Additionally, in terms of data shape, each user does not have more than 65 consumption observations associated to the corresponding months of data collection. Therefore, the possibility of having poor performance when training neural network models on this dataset was high. Initial testing was conducted on single layer LSTM networks containing 100 units and using the Adam optimizer in the prediction of the next monthly consumption value based on three previous months as input. The training set consisted of the first four years of data and the test set contained the remaining months. The result of the initial simple model, previously presented in Figure 4, exhibited overfitting, thus confirming our intuition to implement the extensions that we proposed in order to stabilize the model and improve its performance.

The problem formulation using lagged observations remained the same in order to have a fair comparison of the modified models. The first modification was the implementation of an early stopping mechanism that attempted to stop training the network when the error metrics derived from the evaluation of the model on validation data after each epoch stop decreasing. The initial number of epochs was set to 4000 and after continuous monitoring of the error metrics from consecutive executions the patience interval, which determines the number of epochs after no improvement to the loss function was detected, was set at 170 epochs, preventing overfitting. This patience interval remained

proportionately small when compared to the total training epochs and provided a sufficient window that allowed the improvement of the model. However, due to the decreased training epochs, the model yielded suboptimal performance. Therefore, the replacement of the single LSTM network with an LSTM ensemble of n members yielded more stable and reproducible training results and minor performance improvements. For the purposes of this study, the ensemble contained two members in order to balance execution time and stability benefit. The iterative increase of ensemble members only increased the execution time of our experiments, hence the choice of two ensemble members was appropriate for this dataset. Since the number of ensemble members was a parameter that depended on the dataset and the model structure, future research is encouraged to perform similar experimentation in order to establish the benefit of a larger ensemble before finalizing the model. The output prediction of the ensemble LSTM was the weighted average prediction of the members using grid search.

Taking this approach one step further, we explored the effects of similarity and causality among clients by forming two additional models utilizing the same LSTM ensemble structure as the base model. The first model focused on similarity and contained a modified feature set, where lagged observations from the most similar clients were included alongside the base client. A soft-DTW ranking was used to determine the top three closest clients that had the lowest distance scores. Intuitively, the number of the most similar clients should remain small when compared to the total number of clients in order to strengthen the similarity between the features used in the model. Our experiments indicated that three clients were sufficient in the construction of a feature set that includes power consumption values with a high likelihood of corresponding to the same electricity usage type. Generally, the number of the most similar clients is selected based on the dataset, with an emphasis on creating a small set of similar clients, reinforcing cohesion between the members of the set. The second model extended the base feature set by including lagged observations of power consumption after inspecting all other clients and selecting the columns of lags that satisfied the previously discussed predictive efficacy criterion, by rejecting the null hypothesis of a Granger causality test when that column was tested against the targeted output consumption of the training set for the main client. Each feature in every model was normalized or standardized based on the Shapiro-Wilk statistical test [48] before training.

Since all three LSTM ensemble models share those performance hurdles due to data limitations and the implementation of early stopping, the investigation of a combinatorial approach was interesting due to the variety of feature sets. Therefore, a meta-learner was developed, utilizing a single hidden layer MLP network with 100 neurons. The activation function was the rectified linear unit (ReLU) and the optimizer was Adam. Moreover, 4000 was the selected number of epochs for training and the same early stopping mechanism was utilized in order to prevent overfitting. The meta-learner used the output predictions of the three LSTM ensemble models in order to discover the best weighted combination and predict power consumption more accurately. Experiments for the comparison of those models focused on the prediction of the power consumption of a client for the next 14 months. The comparison considered the performance of each standalone LSTM ensemble using the base, causal, and similar feature set, respectively, as well as combinatorial models utilizing the meta-learner for the pairwise stacked generalization of the ensembles. Finally, the combinatorial model that utilized all three LSTM ensembles was examined and the results are presented in the following section.

The experiments presented in this study were implemented in Python 3.8.8, using the packages pandas 1.2.3, numpy 1.19.2, scikit-learn 0.24.1, tensorflow 2.3.0, keras 2.4.3, statsmodels 0.12.2 and matplotlib 3.3.4. It is worth mentioning that any model parameters not mentioned in this section follow the default values of those packages. The models and experiments were executed on a desktop computer with an AMD Ryzen 1700X processor, 8 gigabytes of RAM, and a NVIDIA 1080Ti graphics processor. The code of this study, containing the implementation of this power consumption forecasting approach, is publicly available on GitHub [49].

2.6. Performance Metrics

In this section, we present an overview of the performance metrics used in the evaluation of the neural network models in order to explain their intended usage in our experiments. Firstly, mean absolute error (MAE) [50] is a common loss function for the comparison of predicted and observed values, providing a more natural measure of average error. Given the predicted values y_i and real values x_i in a set of n samples, the mean absolute error is computed by the formula:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{10}$$

In this study, we used MAE as the loss function for the training of our neural network models since it is a simple measure that we can use to monitor how the divergence of predicted values from the real values decreased after every epoch. Second, the mean absolute percentage error (MAPE) [51] is one of the most common model evaluation functions that evaluate the prediction accuracy. This performance metric is often used in the evaluation of regression and time series models due to its intuitive interpretation of relative error practically and theoretically. Given the same information used to calculate MAE, MAPE is computed by the following formula:

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - x_i}{x_i} \right| \tag{11}$$

Since MAPE is a percentage error metric, it is scale independent and therefore constitutes the primary performance metric used to evaluate the models presented in this study. Finally, root mean squared error (RMSE) [52] is the standard deviation of prediction errors and, since error values are squared before they are averaged, this metric focuses on the impact of large errors. In this project, this metric was used as a secondary indicator with attention given only to the relative decrease of the value denoting the improved performance of the model. This method is affected by the scale of the data and, given the information used to compute the previously described metrics, RMSE is computed by the formula:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - x_i)^2}{n}} \tag{12}$$

3. Results

In this section, we present an overview of the experimental results through figures and error metrics that are based on the findings of the case study in order to evaluate the combinatorial model described in this project. The experiments consisted of the random selection of clients and the construction of individual forecasting models utilizing the base feature set of lagged consumption observations, all pairwise combinations of the base feature set, and the additional columns from the exploration of similarity and causality, as well as the final combinatorial model, which utilizes all three sub-models for stacked generalization. Furthermore, for the clear and concise demonstration of the results, we provide the representative comparison of those models for the predictions of 14 months of power comparison for an individual client. It is worth mentioning that the relative boost in performance following this method was maintained when different clients were selected from the dataset, following the same behavior for standalone models, sub-model pairs, as well as the combinatorial model. Additionally, the error metrics were derived as averages from 10 consecutive executions. Since the changes in the error metric values were miniscule, we found that 10 iterative executions were sufficient in the consolidation of measurements.

First, in Table 1 we list the values of MAPE and RMSE for all the models considered in this comparison. We can observe from the values of MAPE that, while the standalone models exhibited fair, but not optimal results given the dataset structure and the implementation of early stopping, the sub-model pairs contributed towards a more accurate meta-model. Moreover, the meta-model that utilized the base, similar, and causal sub-models performed better than all other models in this comparison, showing that the combination of many different models based on varying feature sets can result in a performance boost. The secondary performance metric values of RMSE showed a considerable decreasing trend when we transitioned from the standalone models to pairs of sub-models and, finally, to the three-component meta-model. The values of RMSE were justified due to the range of power consumption values in the dataset and we mainly focused on the decreasing trend in order to determine the improvement. Table 1 labels the standalone LSTM ensemble models as base, causal, and similar depending on the feature sets used. The meta-models utilizing pairs of LSTM ensembles are labeled as base-causal, base-similar, and causal-similar. The final combinatorial model using all sub-model ensembles is labeled as base-causal-similar.

Table 1. Performance comparison of standalone models, sub-model pairs, and combinatorial meta-model.

Model	MAPE	RMSE	MAE
Base	15.62	8485.73	5865.11
Causal	20.37	9749.18	7465.28
Similar	18.06	8984.84	6595.78
Base-Causal	6.36	6333.37	2739.73
Base-Similar	6.28	3635.15	1915.17
Causal-Similar	8.62	4595.11	2747.87
Base-Causal-Similar	3.49	1697.14	1122.30

Second, Figure 6 presents a direct comparison of the actual and predicted values of power consumption for the targeted output of 14 months between the standalone LSTM ensemble models and the final combinatorial meta-model utilizing an MLP. Through this comparison it is clear that no standalone model could get accurate predictions when consumption values show sudden valleys and peaks, such as the areas between data points 3 and 5, as well as data points 8 and 12. The standalone models managed to capture the decreasing and increasing patterns later in time, producing an outcome that seems to be shifted, distorting the result. Additionally, Figure 7 presents a direct comparison between the meta-models created by the combination of LSTM ensemble pairs and the meta-model that utilized all three LSTM ensembles. The inspection of this figure could lead to some interesting assumptions since the involvement of the base LSTM ensemble resulted in meta-models that could adapt better to sudden decrease in consumption. Similarly, the involvement of the component of similarity led to models that could capture the sudden increase in consumption. While this behavior could be situational to each model for each individual client, it shows that the combination of sub-models utilizing influential community characteristics could lead to a better fit in the regions where simpler standalone models would not be able to adapt that well. It is evident that the involvement of all three sub-models led to the development of the most accurate meta-model.

Finally, for completeness, we present the graph that shows the training history of the final meta-model in Figure 8. In this graph, we observe that the loss function MAE kept decreasing for the training and validation set. The initial training epochs were set to 4000, but the model stopped training after 3500 epochs due to an early-stopping mechanism that prevented overfitting.

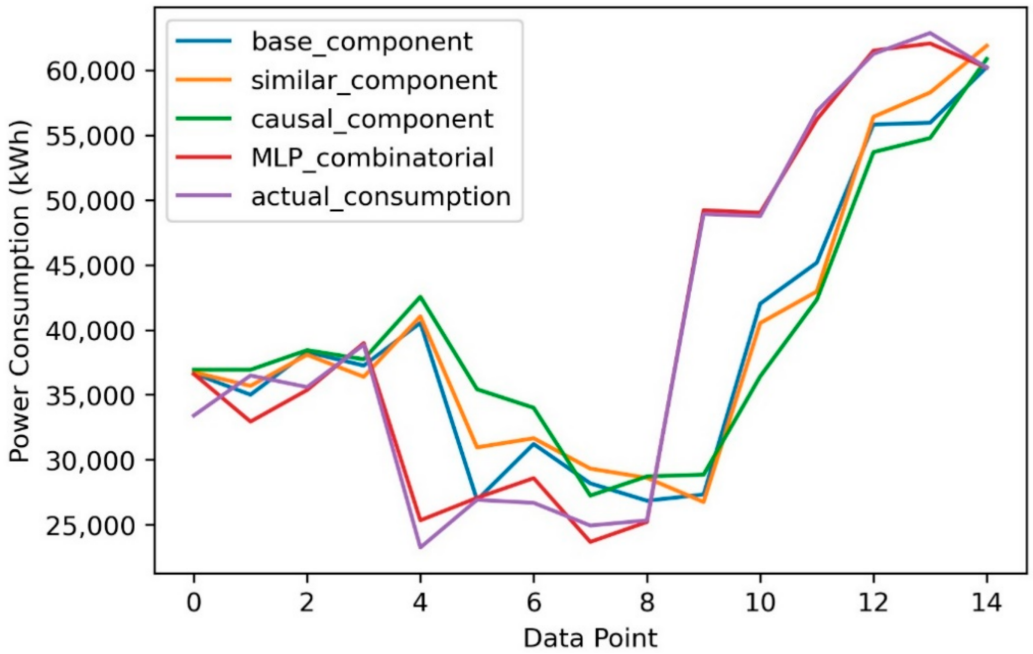

Figure 6. Comparison of the predicted values between the standalone models featuring one component and the final meta-model.

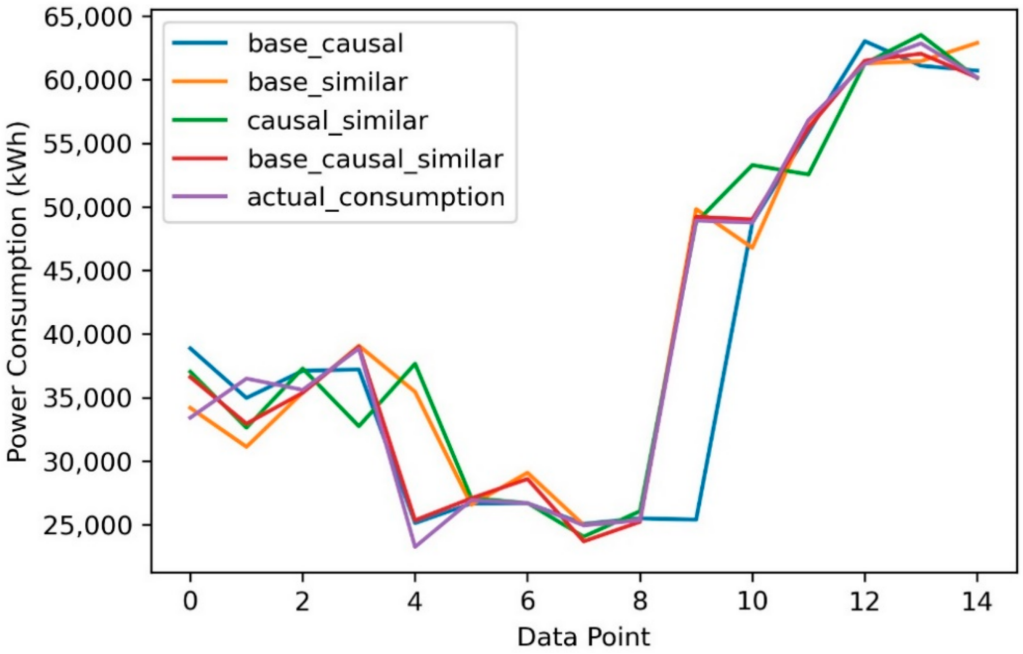

Figure 7. Comparison of the predicted values between the sub-model pairs and the final meta-model.

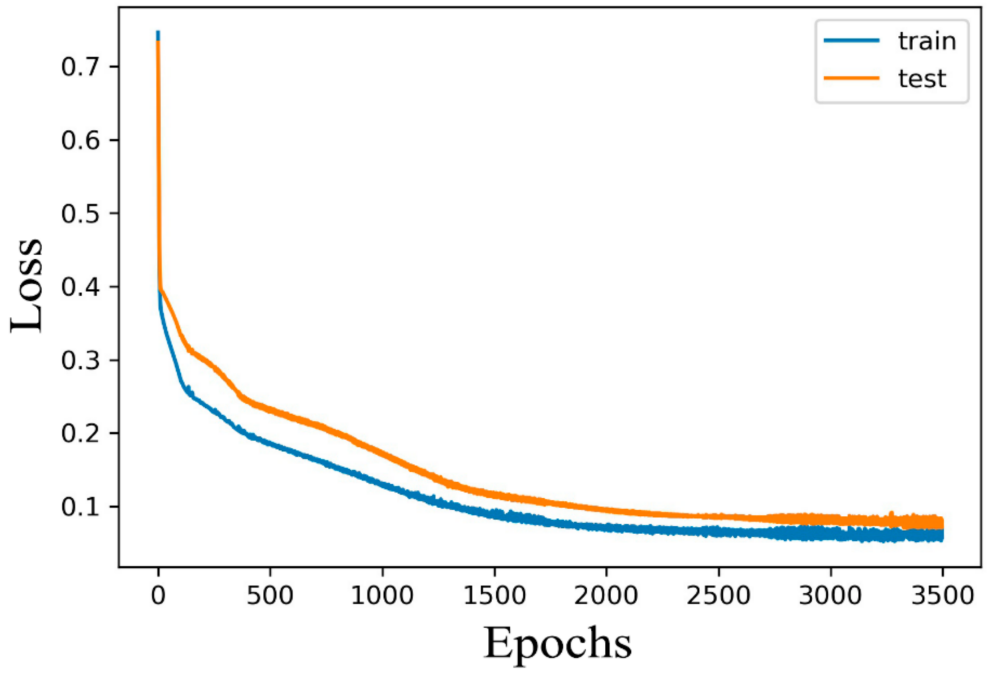

Figure 8. Training history showing the loss function of MAE for the final meta-model utilizing similarity and causality.

4. Discussion

This work explored the impact of similarity and causality in the development of a combinatorial power consumption forecasting model for electricity clients based on neural networks. Since the proposed model focused on a more realistic approach that addressed the main challenges in neural network model design, a case study was carried out using a dataset that was derived from a non-ideal data collection process. The research findings showed that, while the standard LSTM network, which only utilized lagged observations of the main client, could overfit and exhibit suboptimal performance, the development of meta-models based on combinations of feature sets that were influenced by the similarity and causality could achieve a better and more stable performance. In detail, the LSTM ensemble model utilizing only the lagged observations of the client had a MAPE of 15.62 and was outperformed by the meta-models, which utilized pairs of LSTM ensemble sub-models. In those experiments, the meta-model that utilized the output of the LSTM ensembles with the base and similar feature sets yielded the highest pairwise performance with a MAPE value of 6.28. In conclusion, the final meta-model that utilized the outputs of LSTM ensembles, which included the base feature sets as well as feature sets influenced by similarity and causality, yielded the highest performance when compared to all other models, achieving a MAPE of 3.49.

The process of designing this meta-model, as well as the results of this study, contribute greatly towards the introduction and development of more robust and complex combinatorial models that address the current challenges of forecasting model design and are more resilient towards the available dataset structure. This novel forecasting approach presents ideas that mitigate important hindrances in the performance of LSTM models and investigate the potential benefits of influential community factors, assisting in the implementation of performant models when the available data and the prediction horizon are far from ideal. Our project hopes to encourage further work in this field since it was observed that the consideration of many different feature sets can achieve better aggregated results. It is important to note that related work in this field shows that the standalone concepts of similarity and causality can be effective in the prediction of energy data in various horizons [53,54], but to the best of our knowledge, there are not many available experiments that consider the combination of the two on either short-term or long-term

predictions given a group of electricity clients regardless of data structure. Therefore, this work attempted to fill this research gap by providing useful insights given the scenario described in the case study. Future work on this class of meta-models could explore many different aspects, which were not available in the current dataset; for instance, it would be interesting to study the inclusion of more detailed features, such as occupancy and appliance information, in order to reinforce the results of similarity and causality tests. It would also be interesting to explore the performance of the model in a more ideal setting, where the available dataset contains consumption data from a much wider pool of clients, without missing values, in order to inspect how the model behaves with big data in a more ideal configuration. Finally, from the perspective of training performance and execution time, future work could parallelize this model and execute it on multiple graphics processors in order to inspect the improvements of the multithrea.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available in a publicly accessible repository. The data used in this study are openly available in [Mendeley Data] at [10.17632/xbt7scz5ny.3], reference number [55] under a Creative Commons Attribution 4.0 International license [https://creativecommons.org/licenses/by/4.0/ (accessed on 17 September 2021)]. The dataset was processed as the input for the design and performance assessment of the power consumption forecasting approach described in this article.

Acknowledgments: This work is supported in part by ONR under grant no N00014-18-1-2278 and by a GS-Gives Grant to AI Systems Lab (AISL) of Purdue University.

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

References

1. Alamaniotis, M.; Ikonomopoulos, A.; Bargiotas, D.; Tsoukalas, L.; Alamaniotis, A. Day-ahead Electricity Price Forecasting using Optimized Multiple-Regression of Relevance Vector Machines. In Proceedings of the 8th Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion (MEDPOWER 2012), Cagliari, Italy, 1–3 October 2012. [CrossRef]
2. Alamaniotis, M.; Bargiotas, D.; Bourbakis, N.; Tsoukalas, L. Genetic Optimal Regression of Relevance Vector Machines for Electricity Pricing Signal Forecasting in Smart Grids. *IEEE Trans. Smart Grid* **2015**, *6*, 2997–3005. [CrossRef]
3. Alamaniotis, M.; Bargiotas, D.; Tsoukalas, L.H. Towards smart energy systems: Application of kernel machine regression for medium term electricity load forecasting. *SpringerPlus* **2016**, *5*, 1–15. [CrossRef] [PubMed]
4. Hwang, J.; Suh, D.; Otto, M.-O. Forecasting Electricity Consumption in Commercial Buildings Using a Machine Learning Approach. *Energies* **2020**, *13*, 5885. [CrossRef]
5. Mir, A.; Alghassab, M.; Ullah, K.; Khan, Z.; Lu, Y.; Imran, M. A Review of Electricity Demand Forecasting in Low and Middle Income Countries: The Demand Determinants and Horizons. *Sustainability* **2020**, *12*, 5931. [CrossRef]
6. Kontogiannis, D.; Bargiotas, D.; Daskalopulu, A. Fuzzy Control System for Smart Energy Management in Residential Buildings Based on Environmental Data. *Energies* **2021**, *14*, 752. [CrossRef]
7. Dalal, M.; Li, A.; Taori, R. Autoregressive Models: What Are They Good For? *arXiv* **2019**, arXiv:1910.07737.
8. Johnston, F.R.; Boyland, E.J.; Meadows, M.; Shale, E. Some properties of a simple moving average when applied to forecasting a time series. *J. Oper. Res. Soc.* **1999**, *50*, 1267–1271. [CrossRef]
9. Advanced Time Series Analysis with ARMA and ARIMA, Medium. 2021. Available online: https://towardsdatascience.com/advanced-time-series-analysis-with-arma-and-arima-a7d9b589ed6d (accessed on 9 July 2021).
10. Luetkepohl, H. Forecasting with VARMA Models. In *Handbook of Economic Forecasting*, 1st ed.; Elliott, G., Granger, C., Timmerman, A., Eds.; Elsevier: Amsterdam, The Netherlands, 2006; Chapter 6; Volume 1, pp. 287–325.
11. Ostertagová, E.; Ostertag, O. Forecasting using simple exponential smoothing method. *Acta Electrotech. et Inform.* **2012**, *12*, 62–66. [CrossRef]

12. Holt-Winters Exponential Smoothing, Medium. 2021. Available online: https://towardsdatascience.com/holt-winters-exponential-smoothing-d703072c0572 (accessed on 9 July 2021).
13. An Overview of Time Series Forecasting Models Part 1: Classical Time Series Forecasting Models, Medium. 2021. Available online: https://shaileydash.medium.com/an-overview-of-time-series-forecasting-models-part-1-classical-time-series-forecasting-models-2d877de76e0f (accessed on 9 July 2021).
14. Mitrea, C.A.; Lee, C.K.M.; Wu, Z. A Comparison between Neural Networks and Traditional Forecasting Methods: A Case Study. *Int. J. Eng. Bus. Manag.* **2009**, *1*, 11. [CrossRef]
15. Shiblee, M.; Kalra, P.; Chandra, B. Time Series Prediction with Multilayer Perceptron (MLP): A New Generalized Error Based Approach. *Adv. Neuro-Inf. Process.* **2009**, 37–44. [CrossRef]
16. Khodabakhsh, A.; Ari, I.; Bakır, M.; Alagoz, S.M. Forecasting Multivariate Time-Series Data Using LSTM and Mini-Batches. In Proceedings of the 7th International Conference on Contemporary Issues in Data Science, Zanjan, Iran, 6–8 March 2019; Spinger: Cham, Switzerland, 2019; pp. 121–129. [CrossRef]
17. Kontogiannis, D.; Bargiotas, D.; Daskalopulu, A. Minutely Active Power Forecasting Models Using Neural Networks. *Sustainability* **2020**, *12*, 3177. [CrossRef]
18. López de Prado, M. Overfitting: Causes and Solutions (Seminar Slides). 2020. Available online: http://dx.doi.org/10.2139/ssrn.3544431 (accessed on 9 July 2021).
19. Jayalakshmi, N.; Shankar, R.; Subramaniam, U.; Baranilingesan, I.; Karthick, A.; Stalin, B.; Rahim, R.; Ghosh, A. Novel Multi-Time Scale Deep Learning Algorithm for Solar Irradiance Forecasting. *Energies* **2021**, *14*, 2404. [CrossRef]
20. Dai, W.; Yoshigoe, K.; Parsley, W. Improving Data Quality Through Deep Learning and Statistical Models. *Adv. Intell. Syst. Comput.* **2017**, 515–522. [CrossRef]
21. Choi, J.Y.; Lee, B. Combining LSTM Network Ensemble via Adaptive Weighting for Improved Time Series Forecasting. *Math. Probl. Eng.* **2018**, *2018*, 1–8. [CrossRef]
22. Tian, C.; Ma, J.; Zhang, C.; Zhan, P. A Deep Neural Network Model for Short-Term Load Forecast Based on Long Short-Term Memory Network and Convolutional Neural Network. *Energies* **2018**, *11*, 3493. [CrossRef]
23. Mujeeb, S.; Javaid, N.; Ilahi, M.; Wadud, Z.; Ishmanov, F.; Afzal, M.K. Deep Long Short-Term Memory: A New Price and Load Forecasting Scheme for Big Data in Smart Cities. *Sustainability* **2019**, *11*, 987. [CrossRef]
24. Markovič, R.; Gosak, M.; Grubelnik, V.; Marhl, M.; Virtič, P. Data-driven classification of residential energy consumption patterns by means of functional connectivity networks. *Appl. Energy* **2019**, *242*, 506–515. [CrossRef]
25. Jin, X.-B.; Zheng, W.-Z.; Kong, J.-L.; Wang, X.-Y.; Bai, Y.-T.; Su, T.-L.; Lin, S. Deep-Learning Forecasting Method for Electric Power Load via Attention-Based Encoder-Decoder with Bayesian Optimization. *Energies* **2021**, *14*, 1596. [CrossRef]
26. Tian, Y.; Sehovac, L.; Grolinger, K. Similarity-Based Chained Transfer Learning for Energy Forecasting with Big Data. *IEEE Access* **2019**, *7*, 139895–139908. [CrossRef]
27. Chen, T.-L.; Cheng, C.-H.; Liu, J.-W. A Causal Time-Series Model Based on Multilayer Perceptron Regression for Forecasting Taiwan Stock Index. *Int. J. Inf. Technol. Decis. Mak.* **2019**, *18*, 1967–1987. [CrossRef]
28. Boersma, K. Using Influencing Factors and Multilayer Perceptrons for Energy Demand Prediction. 2019. Available online: http://essay.utwente.nl/78789/ (accessed on 10 July 2021).
29. Emamian, M.; Milimonfared, J.; Aghaei, M.; Hosseini, R. Solar Power Forecasting with Lstm Network Ensemble, Researchgate. 2019. Available online: https://www.researchgate.net/publication/337494650 (accessed on 10 July 2021).
30. Guo, L.; Wang, L.; Chen, H. Electrical Load Forecasting Based on LSTM Neural Networks. *BDECE* **2019**, 107–111. [CrossRef]
31. Tao, C.; Lu, J.; Lang, J.; Peng, X.; Cheng, K.; Duan, S. Short-Term Forecasting of Photovoltaic Power Generation Based on Feature Selection and Bias Compensation–LSTM Network. *Energies* **2021**, *14*, 3086. [CrossRef]
32. Understanding LSTM Networks-Colah's blog, Colah.github.io. 2021. Available online: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (accessed on 10 July 2021).
33. An Introduction to Gradient Descent and Backpropagation, Medium. 2021. Available online: https://towardsdatascience.com/an-introduction-to-gradient-descent-and-backpropagation-81648bdb19b2 (accessed on 10 July 2021).
34. Kamal, I.M.; Bae, H.; Sunghyun, S.; Yun, H. DERN: Deep Ensemble Learning Model for Short- and Long-Term Prediction of Baltic Dry Index. *Appl. Sci.* **2020**, *10*, 1504. [CrossRef]
35. Ramchoun, H.; Idrissi, M.A.J.; Ghanou, Y.; Ettaouil, M. Multilayer Perceptron. In Proceedings of the 2nd international Conference on Big Data, Cloud and Applications, New York, NY, USA, 29–30 March 2017. [CrossRef]
36. Wolpert, D. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
37. Rohrhofer, F.; Saha, S.; Cataldo, S.; Geiger, B.; Linden, W.; Boeri, L. Importance of feature engineering and database selection in a machine learning model: A case study on carbon crystal structures. *arXiv* **2021**, arXiv:2102.00191.
38. Cassisi, C.; Montalto, P.; Aliotta, M.; Cannata, A.; Pulvirenti, A.C.A.A. Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining. In *Advances in Data Mining Knowledge Discovery and Applications*; IntechOpen: London, UK, 2012; [CrossRef]
39. Cuturi, M.; Blondel, M. Soft-DTW: A Differentiable Loss Function for Time-Series. *arXiv* **2017**, arXiv:1703.01541.
40. Time Series Similarity Using Dynamic Time Warping -Explained, Medium. 2021. Available online: https://medium.com/walmartglobaltech/time-series-similarity-using-dynamic-time-warping-explained-9d09119e48ec (accessed on 10 July 2021).

41. Inferring Causality in Time Series Data, Medium. 2021. Available online: https://towardsdatascience.com/inferring-causality-in-time-series-data-b8b75fe52c46 (accessed on 10 July 2021).
42. Amornbunchornvej, C.; Zheleva, E.; Berger-Wolf, T. Variable-Lag Granger Causality for Time Series Analysis. In Proceedings of the 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Washington, DC, USA, 5–8 October 2019. [CrossRef]
43. Brownlee, J. Time Series Forecasting as Supervised Learning, Machine Learning Mastery. 2021. Available online: https://machinelearningmastery.com/time-series-forecasting-supervised-learning/ (accessed on 10 July 2021).
44. Introduction to Early Stopping: An Effective Tool to Regularize Neural Nets, Medium. 2021. Available online: https://towardsdatascience.com/early-stopping-a-cool-strategy-to-regularize-neural-networks-bfdeca6d722e (accessed on 10 July 2021).
45. Liashchynskyi, P.; Liashchynskyi, P. Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS. *arXiv* **2019**, arXiv:1912.06059.
46. Brownlee, J. Stacking Ensemble Machine Learning with Python, Machine Learning Mastery. Available online: https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/ (accessed on 10 July 2021).
47. Parraga-Alava, J.; Moncayo-Nacaza, J.D.; Revelo-Fuelagán, J.; Rosero-Montalvo, P.D.; Anaya-Isaza, A.; Peluffo-Ordóñez, D.H. A data set for electric power consumption forecasting based on socio-demographic features: Data from an area of southern Colombia. *Data Brief* **2020**, *29*, 105246. [CrossRef]
48. Shapiro, S.S.; Wilk, M.B. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **1965**, *52*, 591. [CrossRef]
49. Dimkonto/Client-Power-Consumption-Forecasting, GitHub. 2021. Available online: https://github.com/dimkonto/Client-Power-Consumption-Forecasting (accessed on 13 July 2021).
50. Fürnkranz, J.; Chan, P.K.; Craw, S.; Sammut, C.; Uther, W.; Ratnaparkhi, A.; Jin, X.; Han, J.; Yang, Y.; Morik, K.; et al. Mean Absolute Error. In *Encyclopedia of Machine Learning*; Springer: Boston, MA, USA, 2011; p. 652. [CrossRef]
51. De Myttenaere, A.; Golden, B.; Le Grand, B.; Rossi, F. Mean Absolute Percentage Error for regression models. *Neurocomputing* **2016**, *192*, 38–48. [CrossRef]
52. Hyndman, R.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [CrossRef]
53. Dudek, G. Pattern similarity-based methods for short-term load forecasting—Part 1: Principles. *Appl. Soft Comput.* **2015**, *37*, 277–287. [CrossRef]
54. Cordova, J.; Sriram, L.M.K.; Kocatepe, A.; Zhou, Y.; Ozguven, E.E.; Arghandeh, R. Combined Electricity and Traffic Short-Term Load Forecasting Using Bundled Causality Engine. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 3448–3458. [CrossRef]
55. Parraga-Alava, J. PCSTCOL: Power Consumption Data from an Area of Southern Colombia. *Mendeley Data* **2020**. [CrossRef]