



Incorporating causality in energy consumption forecasting using deep neural networks

Kshitij Sharma¹ · Yogesh K. Dwivedi^{2,3}  · Bhimaraya Metri⁴

Accepted: 22 June 2022
© The Author(s) 2022

Abstract

Forecasting energy demand has been a critical process in various decision support systems regarding consumption planning, distribution strategies, and energy policies. Traditionally, forecasting energy consumption or demand methods included trend analyses, regression, and auto-regression. With advancements in machine learning methods, algorithms such as support vector machines, artificial neural networks, and random forests became prevalent. In recent times, with an unprecedented improvement in computing capabilities, deep learning algorithms are increasingly used to forecast energy consumption/demand. In this contribution, a relatively novel approach is employed to use long-term memory. Weather data was used to forecast the energy consumption from three datasets, with an additional piece of information in the deep learning architecture. This additional information carries the causal relationships between the weather indicators and energy consumption. This architecture with the causal information is termed as entangled long short term memory. The results show that the entangled long short term memory outperforms the state-of-the-art deep learning architecture (bidirectional long short term memory). The theoretical and practical implications of these results are discussed in terms of decision-making and energy management systems.

Keywords Deep neural networks · Energy consumption · Forecasting · Machine learning

✉ Yogesh K. Dwivedi
y.k.dwivedi@swansea.ac.uk

Kshitij Sharma
kshitij.sharma@ntnu.no

Bhimaraya Metri
director@iimnagpur.ac.in

¹ Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway

² Emerging Markets Research Centre (EMaRC), School of Management, Swansea University, Room #323, Bay Campus, Fabian Bay, Swansea SA1 8EN, Wales, UK

³ Department of Management, Symbiosis Institute of Business Management, Pune & Symbiosis International (Deemed University), Pune, Maharashtra, India

⁴ Indian Institute of Management Nagpur, Nagpur, India

1 Introduction

Time-series forecasting is one of the most important quantitative models in which historical observations of the same variable are collected and analysed to develop a model that captures the underlying data generating process. Then the model is used to predict the future. In operation research (OR), such methods are used for volatility prediction (Vidal & Kristjanpoller, 2020), risk assessment (Du et al., 2020), supply chain management (Pacella & Papadia, 2021), price forecasting (Demir et al., 2021), demand and supply forecasting (Chen et al., 2020; Chen & Lu, 2021; Du et al., 2021; van Steenbergen & Mes, 2020; Zhang et al., 2021) and so on. There are multiple methods used in time series analysis that can be divided into three broad categories: 1) classical statistical modelling and forecasting (e.g., ARIMA, GARCH); 2) machine learning and deep learning; and 3) hybrid approaches. This paper used a novel method called "Entangled Recurrent Neural Network" (E-RNNs) for time series prediction (proposed by Yoon & van der Schaar, 2017). Thus, contributing to the deep learning-based methods for time series prediction in OR. As an application of the method, this paper forecasts the energy demands in the different regions.

Energy demand forecasting is becoming increasingly crucial for planning for energy consumption, formulation of distribution strategies, and recommending modern energy policies (Bhattacharyya & Timilsina, 2009). Energy is one of the most vital resources for developing and sustaining that development in any country (Suganthi & Samuel, 2012). This is true for all the sectors that are social, economic, and environmental. Energy is also a crucial part of multiple industries such as production, agriculture, health, and education. Therefore, it is important to have an efficient energy demand management system for allocating the available resources properly and effectively. Energy demand and supply can also be seen as the macro supply chain operations, with the energy companies and distributors as the main body. For example, Bagchi et al. (2022) showed how energy usage could be employed to achieve a better energy mix to, in turn, obtain the much-desired decoupling of growth and sustainable development in developing countries. Moreover, because of the widespread impact of energy consumption, the decision-making in this sector might involve multiple stakeholders, a multitude of uncertainty, and risk inflicting sources (Dorsman et al., 2021). The decision-making in the energy sector can also be different based on the time frame (long, medium, and short term) and area scopes (country, region, city, buildings). Such decisions can also affect the financial aspects of the organizations because the energy investments have capital intensive nature. Furthermore, with the increasing world population (that is estimated at 8.5 billion by 2030; United Nations, 2015), the energy demand will also increase. This increase in demand can only be met by using efficient energy planning and sustainable energy policies, both of which depend on a reliable energy demand forecasting system. Energy demand and supply can also be seen as the macro supply chain operations, with the energy companies and distributors as the main body. Therefore, accurate forecasting of the energy demand does not only have societal impacts but also can have financial impacts. Therefore, accurate forecasting of energy demand or consumption is the desired outcome of the applied methods. One of the most prominent states of the art methods of forecasting the energy demand is using recurrent neural networks.

Recurrent neural networks (RNNs) are the most used deep architectures for solving time-series-based prediction and/or regression tasks. They have been applied in a variety of domains. For example, speech recognition (Zhang et al., 2020), traffic forecasting (Zhang et al., 2020), language translation (Vathsala & Holi, 2020), and risk scoring (Clements et al., 2020). There are two main variants of RNNs used in the state-of-the-art research, such as

standard RNNs (Rumelhart et al., 1988) and bi-directional RNNs (Schuster & Paliwal, 1997). Standard RNNs can propagate the error in only one direction and, therefore, cannot use future inputs for current prediction. On the other hand, using causal predictions with bi-directional RNNs converts them into standard RNNs (Yoon & van der Schaar, 2017). Yoon and van der Schaar (2017) addressed this limitation of Bi-directional RNNs for causal prediction by proposing a novel RNN architecture called Entangled RNN (E-RNN). By stacking an additional forward hidden layer on top of the Bi-RNN structure, the causal prediction of E-RNN is dependent on all the previous backward hidden states. E-RNN can be used in a plethora of applications, ranging from medicine to finance. Importantly, E-RNN can be combined with various state-of-the-art RNN techniques such as multilayer (Parlos et al., 1994), dropout (Srivastava et al., 2014), LSTM (Hochreiter & Schmidhuber, 1997), and GRU (Chung et al., 2015) and leads to performance gains, without the need for any additional assumptions. Entangled Recurrent Neural Networks have been used in multiple domains such as face image sequencing (Oh et al., 2021), load forecasting (Sriram et al., 2018), and other power system applications (Sriram, 2020). However, as aforementioned, to the best of the authors' knowledge, E-RNNs have not been used in the domain of OR, especially in energy demand forecasting.

In this contribution, the relation between the weather conditions that are already used for forecasting purposes in energy consumption is used (Ghalekhondabi et al., 2017). The causal link between energy consumption and weather conditions might improve upon the state-of-the-art forecasting accuracies. Once the demand forecasting based on the proposed causal relation and using E-RNNs is obtained, the E-RNNs' performance is compared against the other state-of-the-art RNN techniques such as multilayer multivariate bi-bidirectional LSTM (Chung et al., 2015; Hochreiter & Schmidhuber, 1997; Parlos et al., 1994; Srivastava et al., 2014). Specifically, this paper addresses the following research question:

How much does the causal information between the weather conditions and the energy demand improves the energy demand forecasting using deep learning networks?

To address the aforementioned research question, the proposed method is applied to three datasets where additional information about the causal relations is used. These causal relations are between the weather conditions and the energy consumption in the training phase of the algorithm to forecast the energy consumption. These results will be compared against a multivariate bidirectional LSTM using just the weather conditions to forecast the energy consumption. Such a comparison will establish the role of causal relations in the forecasting of energy consumption. The current research in this field supports the use of deep learning methods for forecasting, but there is little to no knowledge about the role played by the factors causing the energy use in the different geopolitical landscapes such as a whole country, a region within a country, cities, and a standalone building. The causal relation also changes based on how wide the area has been considered for exploration. Therefore, it is necessary to study and understand the importance of the causal relationship between weather factors and energy consumption, especially when it comes to time-series prediction and forecasting.

The remainder of the paper is organized as follows. The second section presents an overview of deep learning applications in OR, the role of energy demand forecasting in OR, and deep learning applications for energy forecasting. The third section presents the methodology, including the description of the datasets brief introduction to LSTMs, and the proposed method. The fourth section presents the results, and the fifth section discusses the results and presents the implications. Finally, the sixth section concludes the paper.

2 Related work

2.1 Deep learning applications in operation research

Over the past few years, innovations in business analytics and operations management are becoming necessary factors for the success of the ventures (Lim et al., 2013; Mortenson et al., 2015; Ranyard et al., 2015). During these years, deep learning has been used in a variety of operations research applications. For example, supply chain management (Kilimci et al., 2019; Pacella & Papadia, 2021; Punia et al., 2020), understanding/predicting financial risk behaviour (Geng et al., 2015; Kim et al., 2020a, 2020b; Xu & He, 2020) price and price movement forecasting (Sen & Mehtab, 2021; Shahi et al., 2020), fault diagnosis (Kumar et al., 2016), maintenance prediction (Kumar et al., 2018) and asset management for maintenance prediction (Chen et al., 2021) and sustainability performance prediction (Rajesh, 2020). In the following paragraphs, a brief overview is provided about each of these sub-fields and how deep learning methods are applied within the sub-fields.

Organizations use supply chain forecasting as part of their supply chain management to fulfil the requirements of short-term and long-run aggregate forecasting. Such forecasting aids in the decision-making at strategic and tactical levels. In recent times applications of and improvements upon existing deep learning networks have been frequently utilized for this purpose. For example, Kilimci et al. (2019) used a decision integration strategy empowered using a combination of support vector regression and a deep learning network to forecast the demand in the Turkish market. Pacella and Papadia (2021) used LSTMs to forecast the demand of products in the supply chain to form a better and more accurate basis for the respective replenishment systems. Xu and He (2020) used a deep belief network to forecast the financial credit risk, as a crucial part of the supply chain finance, to maintain the sustainable profit growth for financial organizations. Along with demand forecasting, sales prediction is also important in the supply chain management to make sure that the stores do not sell overstock, avoid understocking, reduce losses and minimize waste. In this vein, Husna et al. (2021) used LSTM and convolutional neural networks to forecast the sales of grocery stores. Other examples of sales forecasting in the different sectors include fashion retail (Giri et al., 2019; Loureiro et al., 2018), e-commerce (Pan & Zhou, 2020; Qi et al., 2019), pharmaceuticals (Chang et al., 2017; Ferreira et al., 2018) to mention a few.

There are several examples where improvements upon existing deep learning methods were proposed. For example, Punia et al. (2020) proposed a novel and improved cross-temporal deep learning architecture to forecast all levels (e.g., individual stock units, product groups, online and offline channels) of a retail supply chain. In another example, Kegenbekov and Jackson (2021) used an adaptive deep reinforcement learning method to achieve synchronization between the inbound and outbound flows in an organization. In all these examples mentioned above, the models incorporating deep learning methods outperform various traditional methods. These traditional methods include basic machine learning algorithms (Husna et al., 2021; Kilimci et al., 2019; Xu & He, 2020), statistical methods (Pacella & Papadia, 2021; Xu & He, 2020), and sometimes even pre-existing deep learning methods (Punia et al., 2020). Another facet of deep learning applications for supply chain management is for verifying, generating, and augmenting supply chain maps. For this purpose, Wichmann et al. (2020) used natural language processing and LSTM to automatically extract the buyer–seller relations from texts and then use the features to generate basic supply chain maps and verify/augment existing ones. More recently, Guan and Yu (2021) used the features from the

resource distribution allocation index and a deep learning network to inform the design of the supply chain resource distribution allocation model.

For predicting financial risk behaviour, deep learning has been used at various stages of an artificial intelligence pipeline. In other words, deep learning has been used not only at the prediction stage of a process but also at earlier stages as well, such as during the feature extraction and selection phases. For example, Kim et al., (2020a, 2020b) used deep learning to extract features from structured data for retail traders to identify/predict risk-related behaviour. Kim et al., (2020a, 2020b) showed that the features extracted using deep learning methods provided better prediction outcomes than the traditional methods. Similarly, Geng et al. (2015) used deep neural networks to extract features from the data collected about customers to predict the bankruptcy chances of the customers. Another financial risk venture for deep learning algorithms is in the direction of volatility. Liu (2019) showed that the LSTM-RNN-based methods outperformed one of the most popular techniques (GARCH, Pérez-Cruz et al., 2003) for volatility prediction. Similar results and trends were reported by Xiong et al. (2015) in the case of volatility prediction and comparison with GARCH. Chatzis et al. (2018) used deep learning methods to forecast the stock market crisis and showed the prevalence of such methods over the classical methods. Along with similar trends, Moews et al. (2019) reported better performances shown by deep learning models when compared to traditional machine learning models. Eachempati et al. (2021) have also shown that the deep neural networks outperform the traditional methods of predicting the accounting disclosure, another type of financial behaviour that has gained considerable attraction from the deep learning applications (Almagtome, 2021). Another aspect of risk behaviour that was detected using such techniques is the detection of fraudulent reviews for online marketing websites (Kumar et al., 2022a, 2022b). Kumar et al., (2022a, 2022b) examined the different pre-processing and feature engineering techniques to extract both reviews and review-centric features and showed that unifying these features in ML classifiers resulted in better detection of fraudulent reviews than the contemporary methods. Furthermore, using deep learning techniques and text mining Huang et al. (2022) provided a more accurate estimate of financial distress for beneficiaries and investors than classical machine learning algorithms.

Concerning the stock market prices, deep learning has attracted a multitude of efforts in the direction of stock price forecasting. To compare the different deep learning models such as LSTM and GRU, Shahi et al. (2020) showed that there was no significant difference while using the models, but when the authors added additional sentiment data to forecast the stock prices, the performance was significantly increased. The high predictive performance of deep and extreme machine learning algorithms for stock price prediction was also reported by Balaji et al. (2018), Sen and Mehtab (2021), and Liu et al. (2022). Furthermore, Wang and Fan (2021) show that by incorporating complex non-linear relations into the architecture of the deep learning networks, one can achieve high stock price prediction capacity. Sirignano and Cont (2019) used a deep temporal network and trained it on all the stock data that they obtained and showed that this model outperformed individually trained models. Li and Pan (2022) proposed an ensemble deep learning model for predicting stock prices using the current affairs of the companies.) used a generative adversarial network for stock market price movement prediction. Another set of efforts used sentiment analysis of Twitter data for stock price prediction (Darapaneni et al., 2022; Jing et al., 2021; Mohan et al., 2019; Rao & Srivastava, 2012; Shivaprasad and Shetty, 2017; Swathi et al., 2022; Yusof et al., 2018). Sirignano and Cont (2019) further claim that the "general" model can also be used for "transfer" learning purposes. A sub-application of stock price prediction is predicting the stock price movements and price formulation. For example, Tsantekidis et al. (2017) used

stochastic deep learning networks to forecast the price movement from the large-scale high-frequency data. In the same vein, Zhao and Chen (2021) integrated ARIMA, convolutional neural networks, and long-short term memory to detect non-linear temporal patterns and predict the stock price movement. For deeper reviews on the applications of the deep learning algorithms in the domain of finance, see the surveys done by Ozbayoglu et al. (2020) and Sezer et al. (2020).

2.2 Energy demand/consumption forecasting

Energy demand forecasting/prediction is not a new problem. However, as mentioned in the introduction that it has become, over the years, an important problem to solve from an OR perspective. In the following, a few examples are described of how the past researchers have addressed the problem of energy demand/consumption forecasting/prediction. For comprehensive and in-depth reviews, see, Ghalekhondabi et al. (2017), Islam et al. (2020), and Suganthi and Samuel (2012). The examples covered in this brief overview, methods to forecast/predict the energy demand/consumption, are from basic time-series, regression and econometric analysis, ARMA and GARCH models, basic machine learning, and deep learning methods.

Time series models are concerned with trend analysis, Markov models, and spectrum analysis. Ediger and Tatlıdil (2002) analysed the cyclic patterns in the energy consumption data to forecast the energy demands in Turkey. Aydin (2014) used distribution analysis (t- and F-distributions) to forecast the demand for energy from fossil fuels globally. Aydin (2015) used a similar analysis to model the trends of coal-based energy demands in countries such as India, the United States, Japan, South Africa, and Thailand. Farajian et al. (2018) used the Box-Jenkins method for trend analysis to provide the agriculture energy demands in Iran over a 24-year period. Morakinyo et al. (2019) used the trend analysis to predict the energy consumption on the extreme weather days to delineate the effect of the extremely hot or extremely cold days. Tian et al. (2022) conducted a similar analysis as Morakinyo et al. (2019) with additional regional climate models to predict the energy consumption during the extreme weather days in various regions of Canada.

Efforts from a regression analysis point of view have used different variables to predict the energy demands of households, buildings, and regions. For example, Harold et al. (2017) used the income elasticity of households and quantile regressions to predict the energy consumption, with the aim of informing the use of constant mean elasticity for policy purposes. Maaouane et al. (2021) used the import, export, and energy efficiency measures information to predict the energy demands in the industrial sector. Catalina et al. (2013) used the global heat loss coefficient of buildings, the indoor set point, the sol–air temperature difference, and the south equivalent surface to predict the heating energy consumption. Many studies have considered whether factors as independent variables in regression analysis to predict the energy demands. For example, Braun et al. (2014) used humidity and temperature to predict the energy demands of supermarkets in the UK. Fumo and Biswas (2015) used indoor and outdoor temperature and solar radiation to predict the energy consumption of residential buildings. Tso and Guan (2014) also predicted the residential energy demands using the house size, housing type, heating requirement, and amount of air-conditioning use and a multiple regression model.

Econometric models use the correlations between the energy demand and the macro-economic variables to predict/forecast the energy demands/consumption. One of the methods for incorporating the macro-economic variables in the analysis is to use the causal analysis.

For example, Kayhan et al. (2010) used the causality between economic growth and energy consumption to forecast the energy demands in Romanian households. In another study, Sentürk and Sataf (2015) used the GDP and other socio-econometric information from the World Economic Forum for seven countries (Turkey, Kazakhstan, Azerbaijan, Kyrgyzstan, Uzbekistan, Turkmenistan, and Tajikistan) to predict the overall energy consumption in those countries. Ozturk et al. (2010) used similar information to analyse the causality between the GDP and the energy consumption for 51 countries and showed that for the low-income countries, the GDP Granger caused the energy consumption; while for the middle-income countries, there was a bi-directional causality. Other examples of studies using economic growth to cointegrate/predict energy consumption at a large scale could be found in Sentürk and Sataf (2015).

The next category of tools used is the models with auto-regressive components, that is, auto-regressive moving average (ARMA), auto-regressive integrated moving average (ARIMA), and generalized autoregressive conditional heteroskedasticity (GARCH). ARIMA and/or ARMA models are used to extract the historical trends in the time-series data and use this information for forecasting purposes (Erdogdu, 2007; Ho & Xie, 1998; Huang & Shih, 2003; Vo et al, 2021; Wang et al., 2019). The only difference between the two models is that, unlike ARMA models, the ARIMA models can be used only with stationary time-series (Valipour et al., 2013). On the other hand, GARCH models have similar functionality as ARMA/ARIMA models with one key difference (Bauwens et al., 2006; Engle, 2001). While ARMA/ARIMA models utilize the conditional mean of the time series to forecast the values, GARCH models use the conditional variances in the time series to perform similar forecasts. These models are especially useful when there is heterogeneity in the time series (Bauwens et al., 2006; Engle, 2001). Examples of studies using ARMA/ARIMA models for predicting the future energy demands include Li and Li (2017) predicting the future energy consumption in the Shandong province in China; Ozturk and Ozturk (2018) predicting the coal, oil, natural gas, and renewable energy consumption in Turkey. Furthermore, Eerdogdu (2007) also used cointegration analysis with ARIMA to predict total energy consumption in Turkey, while) used ARIMA models to predict agricultural loads at small scales. Other similar efforts include predicting energy consumption in Morocco (Kafazi et al., 2016), Ghana (Sarkodie, 2017), Afghanistan (Mitkov et al., 2019), India, China, and the USA (Jiang et al., 2018), and Middle Africa (Wang et al., 2018). In the case of using GARCH for prediction purposes in the energy sector, the primary use cases of this method if predicting the volatility of the energy market and load forecasting. For example, Efimova and Serletis (2014) use the extreme weather conditions, geopolitical tensions, and de-regularised markets along with GARCH models to forecast the volatility in the energy market; while Ergen and Rizvanoglu (2016) used GARCH and the historical changes due to weather and demand abnormalities to forecast the volatility in the natural gas energy demand. Fałdziński et al. (2020) used GARCH and SVM to forecast the volatility in the demands of multiple energy sources (e.g., oil, gas, gasoline). Concerning load forecasting, Hor et al. (2006) used GARCH not only for the load forecast daily but also to estimate the maximum daily demand with high accuracy. Similarly, Iwafune et al. (2014) used GARCH to forecast a building's energy load in short-term usage. For a comparison of time-series-based methods for energy demand forecasting, see Okawu et al. (2019).

Another category of studies to forecast the energy demands is concerned with the basic machine learning algorithms. For example, Eseye and Lehtonen (2020) used Artificial Neural Networks (ANN) and Support Vector Machines (SVM) with the weather, occupancy, and heat requirement data, to predict the energy consumption for residential buildings in Finland. Pelka (2021) also used SVM and ARIMA models to predict the mid-term energy

consumption in European countries. Wu and Shen (2018) use a swarm optimization-based SVM to predict the natural gas consumption. Johannesen et al. (2019) compared different machine learning algorithms for different time units and concluded that random forest regressors were the best performing for the long-term forecasting problems, while K-Nearest neighbour based regressors were the most efficient for the short term forecasts. Ahmad and Chen (2019) used Random Forest with non-linear auto-regression techniques to predict the energy demands during the different seasons and with the aim of having a more accurate grid-based distribution scheme than before. Wang et al. (2018) also used Random Forests to predict short-term energy demands for building in Florida.) compared future energy predictions using the ARIMA and Random Forest models, and the results show that Random forest was better performing for the long-term prediction between the two models. Other examples of using basic machine learning algorithms include Murat and Ceylan (2006) using ANN to predict the energy demand in the transport sector; Kankal and Uzlu (2017) also using ANN for long-term energy demand forecast in Turkey; Ferlito et al. (2015) using ANN to forecast a building's energy consumption; Lu et al. (2021) using SVM for forecasting the energy consumption in the USA; Ghazal et al. (2022) using IoT data and fusion of SVM algorithms to predict the industrial energy consumption and; Jana and Ghosh (2022) using discrete wavelet transform and ensemble machine learning algorithms to forecast natural gas prices and demand. Forouzandeh et al. (2022) also used ensemble machine learning algorithms to predict room energy demand. For a comprehensive review of studies using artificial neural networks and support vector machines for energy demand forecasting, see Ahmad et al. (2014).

Finally, the most recent set of studies (as is clear from the advancement of the methods and computing power in the last decade) concern the methods related to deep learning for forecasting the energy requirements in various sectors (e.g., household, industries), at various time scales (e.g., short-term, mid-term, long-term), and for the different area-scope (e.g., buildings, regions, countries). For example, Hrnjica and Mehr (2020) used the time-series decomposition and Recurrent Neural Networks (RNN) to predict the energy demand in Northern Nicosia, Cyprus. Real et al. (2020) used a combination of Convolutional Neural Networks (CNN) and ANN for load forecasting in French grids. Ishaq and Kwon (2021) used an Ensemble of deep network architectures to forecast short-term energy demands for local Korean buildings. Somu and Ramamritham (2021) also predicted the future energy for a local building using combinations of CNN and LSTM. Al Khafaf et al. (2019) used LSTM to forecast energy demands in Victoria, Australia. Kim and Cho (2019a, 2019b) also used LSTM and an auto-encoder for short-term forecasting of power demands in households. There have been a few studies where the traditional machine learning algorithm and deep learning algorithms were compared in terms of their forecasting accuracies. For example, Paterakis et al. (2017) compared multiple layer perceptron against Random forest, SVM, and other regressors; Ağbulut (2022) compared deep neural networks against SVM to predict the energy demands for the transport sector; Bakay and Ağbulut (2021) compared deep neural networks against SVM and ANN to forecast electricity demands in Turkey; Shirzadi et al. (2021) compared LSTMs against SVM and random forest to forecast long-term power demands in Ontario, Canada. In all these examples, the deep learning algorithms outperformed the basic machine learning algorithms. Another example of using deep networks in the energy sector is to perform the assets management for the electrical grid companies (Kala et al., 2020), where the authors showed that their proposed algorithm involves the faster regional convolutional neural networks outperformed the human-coding efforts for asset management. For a comprehensive review of deep learning for energy systems and building energy, please see, Forootan et al. (2022) and Ardabili et al. (2022), respectively.

From the studies reported in this section, there is a lot of potential in using deep learning methods to forecast energy consumption as a process in Operations Research. This contribution aims to improve upon the existing deep learning algorithms to predict future energy demands/consumption by incorporating the causal relation between the weather information and the energy consumption over different area-scope (e.g., buildings, regions, countries).

3 Methodology

3.1 Granger causality

Granger causality (Granger, 1969) tests for the ability of one time series to predict another one – in the present case, whether information flow provides sufficient information to predict 1) user focus size, 2) cognitive load, and 3) user attention flow. Granger causality investigates bi-directional, simultaneous, and continuous relationships and has been employed in several studies in OR (e.g., Ghouali et al., 2014; Mian & Liang, 2010; Tang & Chrsquo, 2011; Zhang & Xu, 2015). The basic definition of Granger causality has two assumptions (Granger, 1969). First, it assumes that the cause occurs prior to the effect. Second, the cause contains information about the effect that is more important than the history of the effect itself. Although Granger causality is defined for linear and stationary time-series contexts, variations for non-linear (Ancona et al., 2004; Chen et al., 2004) and non-stationary (Ding et al., 2000; Hesse et al., 2003) data exist.

The main idea behind Granger's definition of causality is that if the lag (past values) of variable one predicts the current value of variable two in a better manner than the lags (past values) of the variable two itself, it can be inferred that variable one causes variable two. To arrive at such an inference, there is a simple method to be followed. Considering the case of two variables, X and Y . To determine whether X Granger causes Y or the other way around, two models are created. The first model predicts the current value of Y using the past values of Y , while the second model predicts the current value of Y using the past values of X . Then, the quality of the prediction for both models is compared; if the second model outperforms the first model, it can be inferred that X Granger causes Y .

To conduct the data analysis, there are a number of statistical steps. First is data treatment, which is to divide the dataset comprising of weather information and energy consumption into 48 h windows for further analyses. Then the stationary nature of the time series is tested: a Ljung-Box test is used to determine whether there are significant non-zero correlation coefficients at lags 1–15. Small p-values suggest that the time series data is stationary. Further, the optimum value for the 'lag' is identified as the number of previous data points considered for modelling the causality. The value is identified based on the Akaike information criterion (AIC) value of the model. Different models are created with different values of lag that must be considered for the Granger causality consideration and select the model with the lowest AIC value.

Next, Granger Causality (Granger, 1969) is tested to examine the causality between the different variable pairs (humidity – energy consumption; information flow – energy consumption; information flow – energy consumption; wind speed – energy consumption). As aforementioned, the basic principle of Granger causality is to compare two models to test whether x causes y . The first model predicts the value of y at time t using the previous n values of y . The second model predicts the value of y at time t using the previous n values of both x and y . The comparison of the two models can tell whether the history of x contains

more information about y than the history of y itself. If this is the case, then it can be said that x *Granger causes* y .

$$y(t) = \sum_{j=1}^p \alpha_{11j} x(t-j) + \alpha_{12j} y(t-j) + \varepsilon_{1t}$$

$$y(t) = \sum_{j=1}^p \alpha_{22j} y(t-j) + \varepsilon_{2t}$$

where p model order, maximum lag included in the model, α coefficients matrix, the contribution of each lag value to the predicted value, ε residual, prediction error.

One might argue about the choice of the method to analyse the causality between the different pairs of measurements. This paper uses the definition of causality provided by Granger. There are three other methods that could be used to show the causality between different variables: 1) Structured Equation Modelling (SEM, (Edwards & Bagozzi, 2000)) 2) Cross-convergent mapping (CCM, (Sugihara et al., 2012)) and 3) conducting an intervention experiment where the hypothesized cause is controlled and the hypothesized 'effect' is measured (Shadish et al., 2002). SEM does not necessarily contain the information required to consider a causal relationship. Statistically speaking, testing an SEM is not a test for causality. There is a certain mathematical formulation under which SEM can be used for causal inference (Steyer, 2013; Steyer et al., 2002); however, the solutions are not available commercially. Bollen and Pearl (2013) provide a detailed account describing how SEM should not be used for modelling causal relations between variables. The second method, that is, CCM, is useful only in the cases where the time series is stationary (i.e., the mean and variance of the variable do not change over time) and non-linear (i.e., there is no autocorrelation in the time series). Eye-tracking data is stationary (as revealed by the Ljung-Box test) but auto-correlated (where users look at current time instances vastly depending on where they were looking at previous instances). Therefore CCM is not an adequate method for such data. In the case of identifying causal relations between two variables through an experimental or pseudo-experimental setup, such setups are typically costly or require an extensive duration to identify the cause-effect relationship between the two variables in question (Chambliss & Schutt, 2018). Moreover, it has also been shown that for longer time-series data, the Granger causality outperforms other contemporary methods (Zou & Feng, 2009).

In this contribution, four casualties for each dataset (and sub-datasets) are computed for a time window of 48 h with a one-hour shift between two consecutive windows: (1) pressure "Granger causing" demand; (2) wind speed "Granger causing" demand; (3) temperature "Granger causing" demand, and (4) humidity "Granger causing" demand. Once the F-values for all four causal relations are available over time, this is used as additional information for forecasting using entangled LSTM.

3.2 Entangled LSTM

3.2.1 LSTM

A single LSTM cell is comprised of four components:

1. Forget gate (f): this is a neural network with a sigmoid activation function. This gate is responsible for what information is propagated to the next time step and what information is discarded. Depending on the previously hidden state h_{t-1} and the current input x_t , the

forget gate assigns a value between zero and one to every element in the previous cell state C_{t-1} . For all the elements that are assigned a value of one, the information is retained, and for all the elements that are assigned a value of zero are discarded. For all the elements that are assigned a value between zero and one, the value decides how much information is to be retained.

2. Input gate (I): this is also a neural network that uses a sigmoid activation function. To make the decision about what new information is to be stored in the cell state (explained next), there are two different operations:
 - a. The input gate decides which values will be updated.
 - b. Using a tanh activation function, a set of candidate values is created (\tilde{C}_t). Once \tilde{C}_t and I_t are computed, the input given to the cell state can be decided.
3. Cell State (C_t): this functions as the memory of the LSTM. Due to the cell states, LSTMs usually outperform basic recurrent neural networks. For every time window, the C_{t-1} is combined with the forget gate, and it is determined what information is propagated to the next time step and what information is discarded. The retained information is then combined with I_t and \tilde{C}_t to create the new cell state that will be the new memory of the LSTM.
4. Output gate (O): this is another neural network that uses a sigmoid activation function. The cell state computed from the previous step is passed through a hyperbolic function (tanh), and this creates the cell values that are filtered between -1 and 1.

The schema in Fig. 1 shows the various gates and their arrangement;

For each gate, the following are the variables (set of weights and biases):

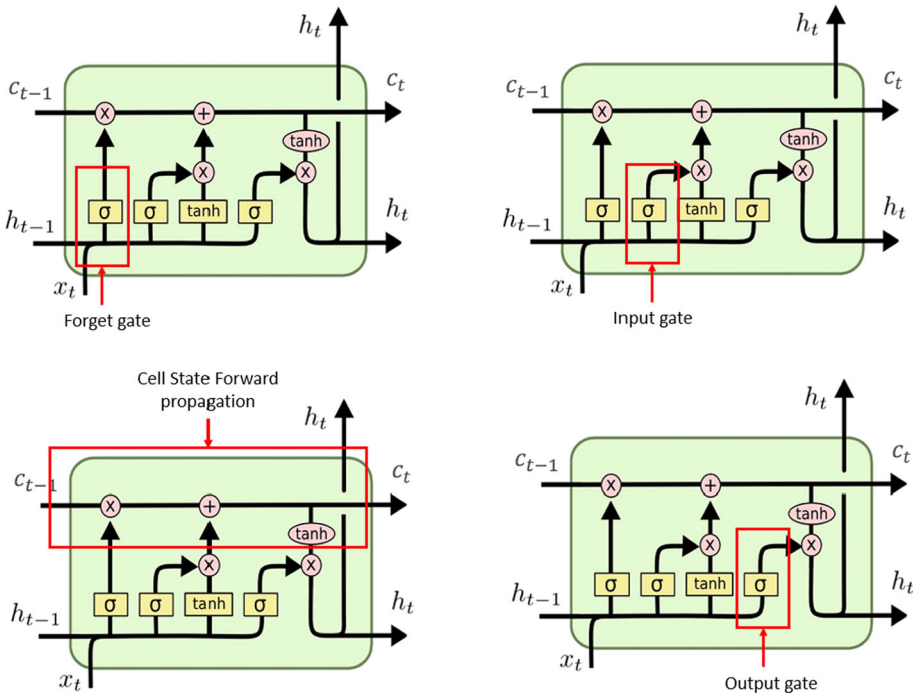


Fig. 1 One LSTM cell with all the components marked with red boxes. (Color figure online)

1. W_f and b_f are the Forget gate weight and bias, respectively.
2. W_i and b_i are the Input gate weight and bias, respectively.
3. W_c and b_c are the Candidate cell state weight and bias, respectively.
4. W_o and b_o are the Output gate weight and bias, respectively.
5. W_v and b_v are the Weight and bias associated with the Softmax layer, respectively.
6. f_t , i_t , C_t , and o_t are the Output of the activation functions for the forget, input, cell, and output gates, respectively.
7. a_f , a_i , a_c , and a_o are the Input to the activation functions to the forget, input, cell, and output gates, respectively.
8. CF is the cost function with respect to which the derivatives are calculated.

For each gate, the following equations show how the activation of each gate is calculated.

1. Forget gate: $a_f = W_f \cdot Z_t + b_f$ $f_t = \text{sigmoid}(a_f)$
2. Input gate: $a_i = W_i \cdot Z_t + b_i$ $i_t = \text{sigmoid}(a_i)$
3. Input gate: $a_c = W_c \cdot Z_t + b_c$ $\tilde{C}_t = \text{tanh}(a_c)$
4. Cell state: $C_t = f_t \otimes C_{t-1} \oplus i_t \otimes \tilde{C}_t$
5. Hidden state: $h_t = o_t \otimes \text{tanh}(C_t)$
6. Output Eq. 1: $V_t = W_v \cdot h_t + b_t$
7. Output Eq. 2: $\hat{y}_t = \text{softmax}(V_t)$
8. Output gate: $a_o = W_o \cdot Z_t + b_o$ $o_t = \text{sigmoid}(a_o)$

The outcome of all the gates is derived in a similar manner, and here the forget gate calculations are explained as an example. There is a fixed path from the activation of the forget gate to the cost function that is shown as the following:

$$f_t \rightarrow C_t \rightarrow h_t \rightarrow CF$$

The following is how the derivative is calculated for the cost function with respect to the forget gate.

$$\frac{dCF}{df_t} = \frac{dCF}{dh_t} * \frac{dh_t}{dC_t} * \frac{dC_t}{df_t}$$

All the derivatives of the cost function with respect to the cell state and hidden state are calculated in the same manner. Input to each LSTM is the previous cell state and the concatenated previous hidden state and current input. For simplicity $[h_{t-1}, x_t] Z_t$

$$\begin{aligned} \frac{dCF}{dC_{t-1}} &= \frac{dCF}{dC_t} * \frac{dC_t}{dC_{t-1}} = \frac{dCF}{dC_t} * f_t \\ \frac{dCF}{dZ_t} &= \frac{dCF}{da_f} * \frac{da_f}{dZ_t} + \frac{dCF}{da_i} * \frac{da_i}{dZ_t} + \frac{dCF}{da_o} * \frac{da_o}{dZ_t} + \frac{dCF}{da_c} * \frac{da_c}{dZ_t} \\ &= W_f^T * \frac{da_f}{dZ_t} + W_i^T * \frac{da_i}{dZ_t} + W_o^T * \frac{da_o}{dZ_t} + W_c^T * \frac{da_c}{dZ_t} \end{aligned}$$

The forget gate:

$$\frac{dCF}{df_t} = \frac{dCF}{dh_t} * \frac{dh_t}{dC_t} * \frac{dC_t}{df_t}$$

But,

$$\frac{dCF}{dC_t} = \frac{dCF}{dh_t} * \frac{dh_t}{dC_t}$$

So,

$$\begin{aligned}\frac{dCF}{df_t} &= \frac{dCF}{dC_t} * \frac{dC_t}{df_t} \\ \frac{dCF}{df_t} &= \frac{dCF}{dC_t} * \frac{d(f_t \otimes C_{t-1} \oplus i_t \otimes \hat{C}_t)}{df_t} \\ \frac{dCF}{df_t} &= \frac{dCF}{dC_t} * C_t\end{aligned}$$

And,

$$\begin{aligned}\frac{dCF}{da_f} &= \frac{dCF}{df_t} * \frac{df_t}{da_f} \\ \frac{dCF}{da_f} &= \frac{dCF}{dC_t} * C_{t-1} * f_t(1 - f_t)\end{aligned}$$

The other derivatives can be computed with respect to the inputs and outputs of the cost function. For example, for the input gate,

$$\begin{aligned}\frac{dCF}{di_t} &= \frac{dCF}{dC_t} * \hat{C}_t \\ \frac{dCF}{da_i} &= \frac{dCF}{dC_t} * \hat{C}_t * i_t(1 - i_t)\end{aligned}$$

For the Cell State:

$$\begin{aligned}\frac{dCF}{dC_t} &= \frac{dCF}{dh_t} * O_t * (1 - \tanh^2(C_t)) \\ \frac{dCF}{d\hat{C}_t} &= \frac{dC_t}{d\hat{C}_t} * i_t \\ \frac{dCF}{da_c} &= \frac{dCF}{dC_t} * i_t * (1 - \hat{C}_t^2)\end{aligned}$$

Output gate:

$$\begin{aligned}\frac{dCF}{dO_t} &= \frac{dCF}{dh_t} * \tanh(C_t) \\ \frac{dCF}{da_o} &= \frac{dCF}{dh_t} * \tanh(C_t) * O_t(1 - O_t)\end{aligned}$$

and finally, the hidden state:

$$\frac{dCF}{dh_t} = \frac{dCF}{dV_t} * \frac{d(W_V * h_t)}{dh_t} = W_V^T * \frac{dCF}{dV_t}$$

The weights for all the gates.

Forget gate:

$$\frac{dCF}{dW_f} = \frac{dCF}{da_i} * \frac{da_f}{dW_f}$$

Input gate:

$$\frac{dCF}{dW_i} = \frac{dCF}{da_i} * \frac{da_i}{dW_i}$$

Output:

$$\frac{dCF}{dW_V} = \frac{dCF}{dV_t} * \frac{dV_t}{dW_V}$$

Output gate:

$$\frac{dCF}{dW_o} = \frac{dCF}{da_o} * \frac{da_o}{dW_o}$$

3.2.2 Bi-directional-LSTM

Bi-directional LSTM, or Bi-LSTM, functions on a very simple principle. In Bi-LSTMs, there are just two LSTMs put together. The first LSTM runs on the exact input sequence that is provided by the dataset. The second LSTM runs in the reversed order of the input sequence as the first LSTM. This improves the LSTM by using both the temporal directions: past and future. The forward layer is responsible for the positive time (forward states or the future), and the backward state is responsible for the negative time (backward state or the past). It has been shown in various contexts that Bi-LSTM outperforms the simple LSTM. For example, Sun et al. (2018) reported better prediction from using Bi-LSTM as compared to LSTM while predicting the blood glucose levels. Shahid et al. (2020) also reported better prediction quality from Bi-LSTM than LSTM while predicting the COVID-19 infections. In terms of energy consumption prediction, Le et al. (2019) showed better performance of Bi-LSTM than regular LSTM. Moreover, Kim and Cho (2019a, 2019b) also show better energy consumption forecasting in specific regions while using Bi-LSTMs than regular LSTMs, whereas Ma et al. (2020) showed similar results while predicting the future energy consumption of individual buildings. Other examples where Bidirectional LSTMs outperformed LSTM in time-series forecasting/prediction tasks include crop detection (Crisóstomo et al., 2020), text mining (Alzaidy et al., 2019), news classification (Li et al., 2002), human activity classification (Shrestha et al., 2020) and sequence tagging (Huang et al., 2015).

The main reason for Bi-LSTM outperforming the simple LSTM can be attributed to the fact that by using two hidden states for each time step, the information from the past and the future is preserved, which in turn provides a better approximation of the time series and encodes the contexts in a better manner than just a forward layer. Therefore, Bi-LSTMs provide better forecasting performance than regular LSTMs. One of the key operations in the Bi-LSTMs is the merging of the two layers, that is, forward and backward layers. This operation is necessary because without merging, it will not be possible to combine the outputs of these layers since they function independently of each other. There are four primary ways of merging the output of these two layers. (1) Sum: The outputs are added together. (2) Multiply: The outputs are multiplied together. (3) Concatenation: The outputs are concatenated together (the default), providing double the number of outputs to the next layer. (4) Average: The average of the outputs is taken. The sum and multiplication artificially increase the variance of the outputs, while the average reduces them. At the same time, the concatenation maintains the original variances of the outputs of the forward and backward layers. Therefore most of the contributions use the concatenation operation to merge the outputs of the forward and backward layers in the Bi-LSTMs.

3.2.3 Entangled-LSTM

In the Entangled-LSTM, an additional layer containing the causal information is stacked on top of the backward layer of Bi-LSTM. This layer is a traditional LSTM layer where the positive direction of time (i.e., the future) is maintained. It can be seen in Fig. 2 that the forward hidden layer that is stacked on top of the backward hidden layer is used for propagating the backward hidden state to the current output. The following shows the update process of the hidden and current output states:

$$hf_t = \tanh(bf + Wf * hf_{t-1} + Uf * x_t)$$

$$hb_\tau = \begin{cases} \tanh(bb + Wb * hb_t + Ub * x_t) & \text{if } \tau = t \\ \tanh(bb + Wb * hb_{\tau+1} + Ub * x_\tau) & \text{otherwise} \end{cases}$$

$$h_t = \tanh(b + W * h_{t-1} + U[hf_t; hb_t])$$

$$O_t = c + Vh_t$$

where h, hf, and HB are the hidden states in simple, forward, and backward layers, respectively; b, bf and bb are the biases in the simple, forward, and backward layers, respectively; W, Wf, Wb, U, Uf, Ub are the weights for the respective networks.

The outcome of the output layer is dependent on the propagated hidden state, the current forward and backward states. The propagated hidden state is dependent on the previous propagated hidden state, which is dependent on three states: 1) the previous propagated hidden state, 2) the previous forward hidden state, and 3) the previous backward hidden state. This chain of dependencies shows that the output at any time t is dependent on the entire input, forward, backward and hidden states. Whereas, in the Bi-LSTM, the output at time t is dependent only on the entire input and forward states only. Figure 2 shows the difference between the Bi-LSTM and entangled-LSTM schematically.

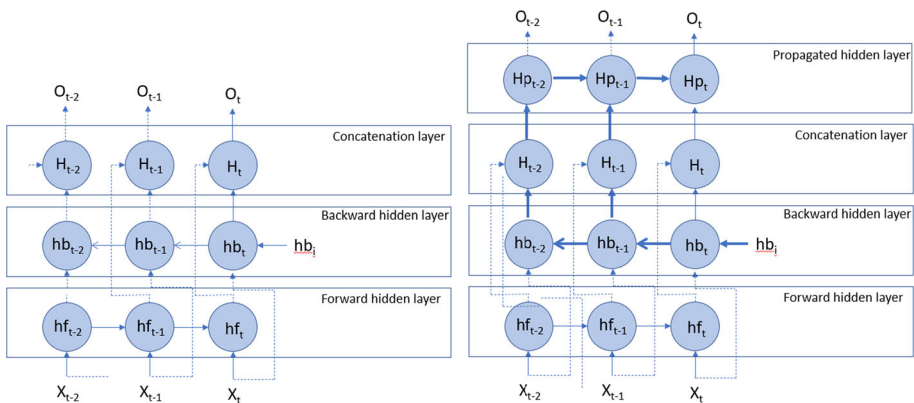


Fig. 2 Bidirectional LSTM (left) and entangled LSTM (right). Dash line: Independent connection to o_t , Thick blue line: New connections by Entangled-LSTM. (Color figure online)

3.3 Datasets and pre-processing

This paper used three datasets available online. Following is a brief description of the datasets and how they have been pre-processed to be used in this paper. These three datasets cover three different area scopes: the first dataset (Spain) is a country-wide dataset, the second dataset (Paraguay) is a region-specific dataset, and the third dataset (France) is for a specific building. For the convenience of expression, datasets 1, 2, and 3 are referred to as Spain, Paraguay, and France datasets, respectively, in the rest of this paper. The three datasets contain different amounts of data in terms of their time span. As reported next, the first dataset had 48 months of data, and the second and the third datasets have 3 and 29 months as the total duration of the recorded data, respectively. Therefore, for maintaining equal grounds for comparison across the three datasets, only the first 29 months of data from the first two datasets were used.

3.3.1 Dataset1: Spain

This dataset¹ contains four years (a total of 48 months) of electrical consumption, generation, pricing, and weather data for Spain. Consumption and generation data were retrieved from ENTSOE, a public portal for Transmission Service Operator (TSO) data. Settlement prices were obtained from the Spanish TSO Red Electric España. Weather data was purchased as part of a personal project from the Open Weather API for the five largest cities in Spain and made public on Kaggle. The dataset contains the following sources of energy: biomass, coal/lignite, coal gas, natural gas, coal, oil, shale oil, peat, geothermal, hydro, sea, nuclear, and other renewable. This detailed information was not present in all the datasets. Therefore, these diverse data sources were aggregated to reflect the total demand for every hour across four years. This dataset also contains hourly weather parameters from the five largest cities in Spain that are Madrid, Barcelona, Valencia, Seville, and Bilbao. Once again, for the analysis and consistency across the datasets, humidity, temperature, pressure, and wind speed were used as the weather parameters. Further, there was no specific energy data for the separate cities; therefore, for the purpose of this contribution, the granger causality was computed (explained in the next subsection) between the weather data from all the cities separately and virtually divided this dataset into five sub-datasets, one for each city.

3.3.2 Dataset2: Paraguay

This dataset² contains the electricity consumption and the meteorological data of the Alto Parana region in Paraguay. Both datasets are from January 2017 to December 2020, a total of 36 months of a time period. The weather data contains temperature (Celsius), relative humidity (percentage), wind speed (km/h, kilometres per hour), and atmospheric pressure (hPa, hectopascal) at the station level with a frequency of every three hours. To be consistent with the other two datasets, the weather data was extrapolated to represent hourly data. A simple smoothing function was used to extrapolate the weather data. The window size for the smoothing function was 24 data points (3 days) with a shift of one data point between two consecutive windows. The electricity consumption data was recorded from 55 feeders in 15 substations in an hourly fashion. Once again, to maintain consistency, the data from the 55 feeders were combined into one by aggregating the consumed amount from all the substations. This was done because there was only one weather station form where the data

¹ <https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather>.

² <https://data.mendeley.com/datasets/hzfwzsk8f/4>.

was gathered, and there was no specific location provided for the weather station. Another way to process this dataset was similar to dataset1, where five sets of causal relations were computed (one for each city). However, this would not have been possible here because even computing 14 causalities would be cumbersome; and because the data is from a region and not a country, therefore aggregating the electricity consumption is the better choice.

3.3.3 Dataset3: France

This dataset³ contains the energy consumption and weather data from one Challenger building in Guyancourt, France. The dataset has 29 months of high-frequency energy consumption data, with the recording frequency being every 10 min. The energy consumption includes Heating and cooling, electrical consumption (indoor and outdoor comfort units), Lighting, plug load, blinds, sanitary consumption, air handling unit consumption, and total consumption. For maintaining consistency across the three datasets, the total energy consumption was used, which is the aggregate of all the individual energy consumptions. The data was aggregated in terms of temporal frequency. The energy consumption data were recorded every 10 min; therefore, to compute the hourly consumption, the data from one hour (six or fewer values in certain hours) was added. The weather data included daily degrees during the days, hourly humidity, hourly temperature, and daily sunshine hours. There was no pressure or wind speed data. However, the exact coordinates of the building were available. Therefore, it was possible to extract the missing information (hourly pressure and hourly wind speed) from online resources (e.g., scrapping certain web pages and some freely available data). It was possible to extract the missing information for the whole period represented in the dataset.

3.4 Training and testing setup

To train, validate, and test the Bi-LSTM and Entangled-LSTM, the data from the first 29 months of the three datasets were used. This was done to have an accurate comparison among the three datasets because 29 months was the lowest of the durations across them. The input of the data for the Bi-LSTM contains the batch size (48 h, i.e., two days), number of time steps (120 h, i.e., five days), and the hidden size. The input data is then fed into three "stacked" layers of LSTM cells (of 50 lengths for the hidden size), and the LSTM network is shown as unrolled over all the time steps. The term "unrolled" means that the feedback loop of an LSTM cell is not shown. The loop is used for keeping the information persisting within the recurrent network. The output from these unrolled cells is the same as the input (batch size, number of time steps, hidden size). This output data is passed to the time distributed layer, which is the set of inputs that the model will learn from to predict the input data coming after. Finally, the output layer has a softmax activation applied to it. This output is compared to the training data for each batch, and the error and gradient are then backpropagated. The Entangled LSTM is created in the same manner as the Bi-LSTM except for one difference. In Entangled-LSTM, there is one additional layer containing the F-value of the causal relationship of humidity, pressure, temperature, and wind speed with energy consumption/demand. This additional layer is also trained with backpropagation.

For training and validating both the models, 26 months of hourly data were used, and for the testing, the remaining three months of data were used. Hyndman and Koehler (2006) and

³ <https://zenodo.org/record/3445332#.Ybcj6n3MJ24>.

Davydenko and Fildes (2013) have provided overviews of the metrics that could be used to evaluate the forecasts. In this contribution, the following three error metrics are used:

Mean Absolute Error (MAE): this is the mean of the absolute difference between the original and the predicted values.

$$MAE = \frac{1}{n} \sum |e(t)|$$

Root Mean Squared Error (RMSE): this is the square root of the mean of the squared difference between the original and the predicted values.

$$RMSE = \sqrt{\frac{1}{n} \sum e(t)^2}$$

Mean Absolute Percentage Error (MAPE): this is the mean of the absolute error when the error is reported as a ratio of the original values.

$$MAPE = \frac{100}{n} \sum \frac{|e(t)|}{o(t)}$$

where

$$e(t) = orig(t) - pred(t)$$

In all the above formulae, $e(t)$ is the error for time t , $orig(t)$ is the original value at time t , $pred(t)$ is the predicted value at time t , and n is the number of data points used in the test set. Among the three metrics, MAE and RMSE are the two most user error metrics. But they are scale-dependent, which indicates a requirement for an additional scale-invariant evaluation metric. Therefore, the MAPE is used. MAPE is also good for comparing the error rates among different datasets because of its scale-invariant nature. In this paper, only the performances of the bidirectional-multivariate-LSTM and the causal-LSTM (which by extension is also bidirectional and multivariate) are compared.

4 Results

4.1 Simulation results

First of all, the results with simulated data will be presented. The purpose of this set of results is to show that if there are two-time series where one time series is perfectly causing the other time series, which of the two LSTM architectures (Bi-LSTM or Entangled-LSTM) would perform better. For this purpose, two time series were generated, where the time series one is perfectly causing the time series two, and the task is to forecast the values for the second time series. Next, both the LSTM architectures were trained and tested for the simulated dataset. Table 1 contains the outcome of the comparison. The entangled LSTM clearly outperforms the Bidirectional LSTM. This shows the theoretical confirmation of the proposed method. That is, adding the causal information in perfect causal conditions will provide better forecasting than the model without the causal information. It is clear that when one-time series "perfectly" causes another time series, the performance of Entangled LSTM is better than the Bi-directional LSTM.

Table 1 The testing performance from the two LSTM architectures for the simulated data where the causality is established

| Bidirectional LSTM | | | | Entangled LSTM | | | |
|--------------------|------|------|------|----------------|------|------|------|
| Dataset | MAE | RMSE | MAPE | Dataset | MAE | RMSE | MAPE |
| Simulated | 0.15 | 0.12 | 0.08 | Simulated | 0.09 | 0.06 | 0.02 |

The MAE and RMSE are scale-dependent, and the MAPE is scale-independent. Therefore, for the purpose of understandability, the time series were normalized between 0 and 1

4.2 Comparison of multivariate-bidirectional and multivariate entangled LSTMs

The two LSTM architectures (bidirectional and entangled) are compared for the three datasets based on the different metrics (MAE, RMSE, MAPE). Figures 3, 4, and 5 show the training and validation losses for Spain (different sub-datasets based on the cities), Paraguay, and France datasets. It can be clearly observed from the losses that both the architectures are not overfitting on any of the datasets. Moreover, the training losses (the blue curves in all the figures) fluctuate in the range of 0.05 to 0.10 before eventually stabilizing. On the other hand, the validation losses (the orange curves in all the figures) fluctuate in a smaller range (0.02 to 0.04) and stabilize. None of the plots (in Figs. 3, 4, and 5) have any alarming differences between the training and validation losses. Therefore, it can be concluded that there was no overfitting of the data in any of the three cases.

As shown in Table 2, the entangled LSTM outperforms the bidirectional LSTM for all three datasets. The minimum improvement is for the Paraguay dataset (MAPE difference is 2.07%), and the maximum improvement is for the France dataset (MAPE difference is 7.73%). The highest difference between the two architectures is for the France dataset. This could be because of the reason that in the France dataset, the weather information is the most accurate for the geographical location. Therefore, the causality computed between the weather indicators (pressure, temperature, humidity, wind speed) and the energy consumption of the building (the France dataset has the data from an individual building) captures the causal relation that is closest to the reality among all the datasets. On the other hand, the Paraguay dataset having the minimum improvement shows that the data from the single weather station in the whole region does not exemplify the weather conditions that are in general covered by the 14 substations from which the data was collected. Another reason for the Paraguay dataset corresponding to the minimum improvement could be the original frequency of the weather parameters. In the Paraguay dataset, the original weather parameters were recorded every three hours, and the weather data was extrapolated to obtain the hourly frequencies for all the four weather parameters used in the paper. This could be the reason why the variability in the weather data in the Paraguay dataset was lower than in the other two datasets, and therefore, the causation with the energy data was not as informative as it was in the other two cases.

Another important aspect that can be observed is the differences between the bidirectional LSTM and entangled LSTM for the Sub-datasets of the Spain dataset. These sub-datasets are treated as independent datasets, and indeed there is no specific trend revealed from the presented forecasting and analysis. The order of the cities in terms of size of the population is Madrid, Barcelona, Valencia, Seville, and Bilbao, while their respective improvements from bidirectional to entangled LSTM are 2.86%, 2.41%, 6.91%, 4.04%, 2.65%. As a post-hoc prediction, the energy consumption data was divided in a way that the proportion of the energy consumed reflected the ratio of these cities' populations to the Spanish population.

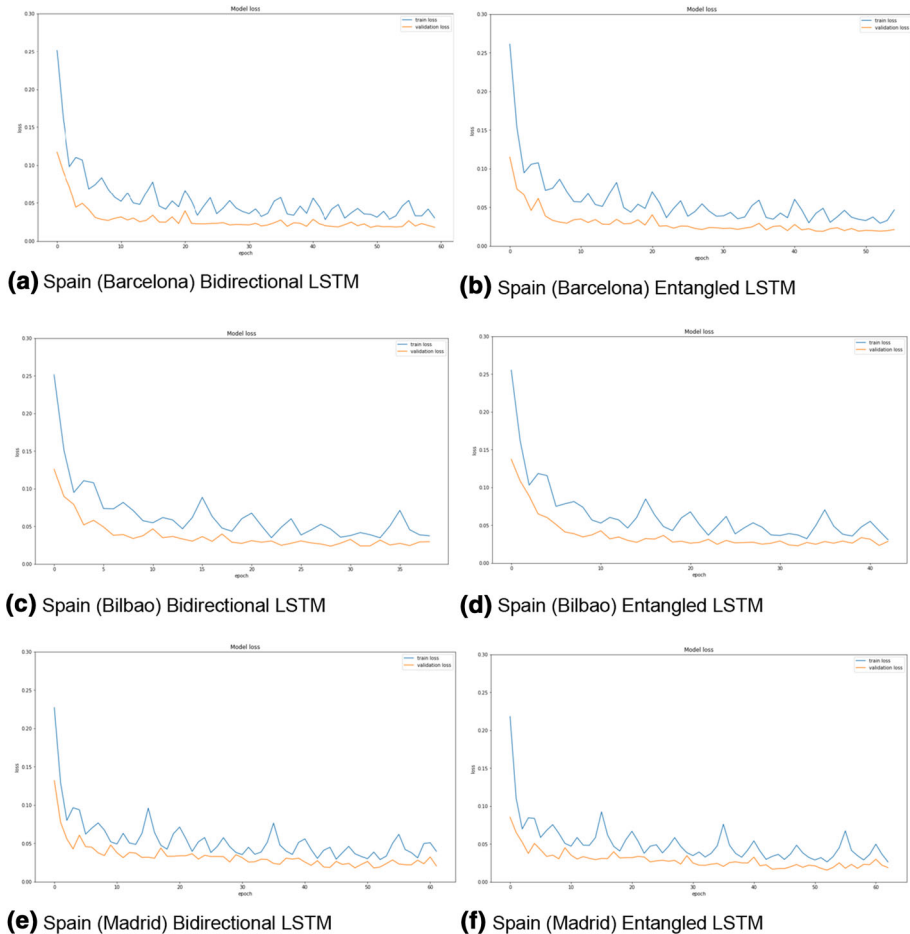


Fig. 3 The training (blue curves) and validation (orange curves) losses from the first dataset (Spain dataset and the five sub-datasets). (Color figure online)

There was no significant difference in the results with either of the two methods. This shows that including the causal information is even more important in cases like the Spain dataset. In the following, an explanation is provided. Looking at the MAPE of the two methods, Valencia has the worst forecasting performance, and Bilbao has the best MAPE, looking at the bidirectional LSTM. This would indicate that adding the causal information plays a role in improving the forecasts. The MAPE for Valencia with entangled LSTM was cut down to half of what it was with bidirectional LSTM, while the MAPE for Bilbao was also reduced by almost a quarter with entangled LSTM.

In a nutshell, it is evident that adding the causal information for forecasting the energy consumption/demand improves the forecasting accuracy. This is shown across all three datasets. The range in MAPE for the entangled LSTM is 2.31% which, considering the variation in the three datasets, is neither alarming nor significant, especially because the three datasets cover large variations in terms of their area scope. As aforementioned, the Spain dataset covers national consumption while Paraguay and the France dataset cover regional and individual

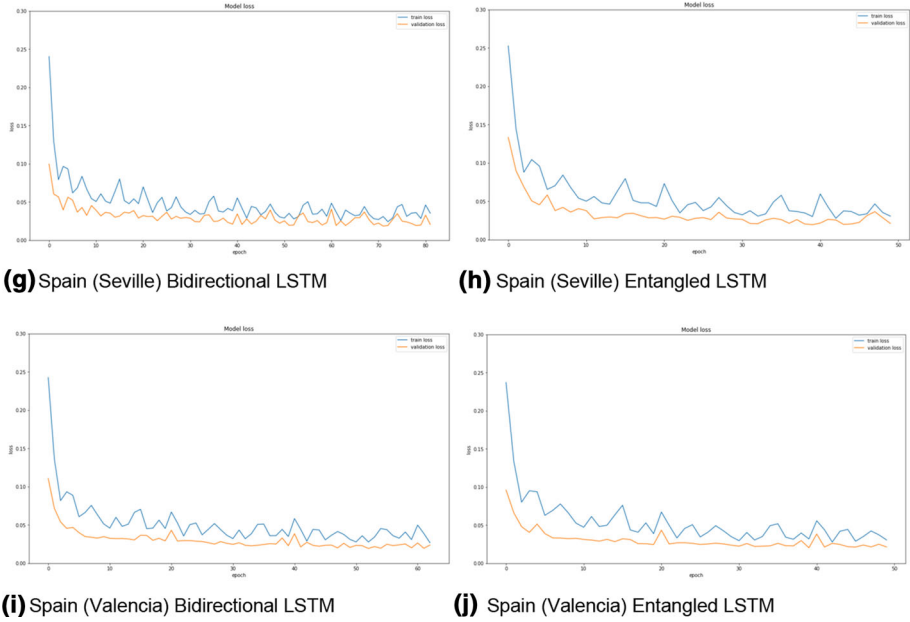


Fig. 3 continued

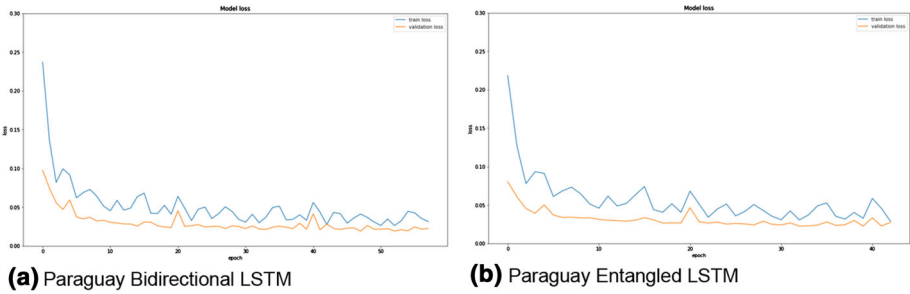


Fig. 4 The training (blue curves) and validation (orange curves) losses from the first dataset (Paraguay dataset). (Color figure online)

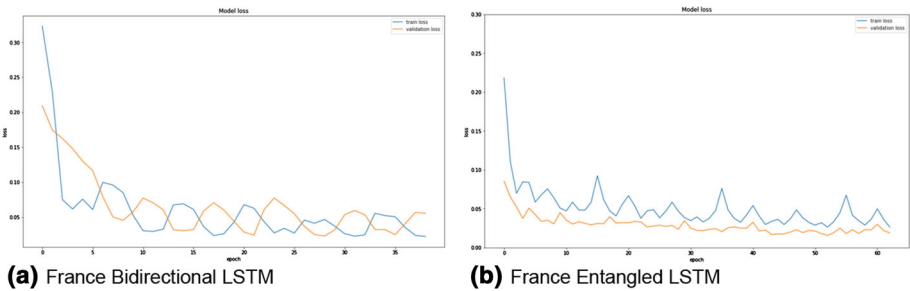


Fig. 5 The training (blue curves) and validation (orange curves) losses from the third dataset (France dataset). (Color figure online)

Table 2 The testing performance from the two LSTM architectures. As it is evident that the MAE and RMSE are scale-dependent, and the MAPE is scale-independent

| Bidirectional LSTM | | | | Entangled LSTM | | | |
|--------------------|---------|----------|-------|-------------------|---------|---------|-------|
| Dataset | MAE | RMSE | MAPE | Dataset | MAE | RMSE | MAPE |
| Spain (Barcelona) | 3358.60 | 3929.88 | 12.55 | Spain (Barcelona) | 2692.17 | 3317.99 | 10.14 |
| Spain (Bilbao) | 3278.86 | 3873.20 | 12.49 | Spain (Bilbao) | 2462.59 | 2840.11 | 9.84 |
| Spain (Madrid) | 3583.12 | 4243.87 | 13.31 | Spain (Madrid) | 2848.33 | 3252.92 | 10.45 |
| Spain (Seville) | 3472.90 | 3917.47 | 12.89 | Spain (Seville) | 2283.42 | 2708.95 | 8.83 |
| Spain (Valencia) | 3885.11 | 4523.322 | 15.05 | Spain (Valencia) | 2117.02 | 2473.69 | 8.14 |
| Paraguay | 2651.78 | 3099.92 | 10.36 | Paraguay | 2101.28 | 2798.55 | 8.29 |
| France | 1962.67 | 2311.56 | 17.80 | France | 1066.35 | 1454.34 | 10.07 |

building consumption, respectively. This, combined With the proof that none of the models have overfitted in the training and validation phases, the generalizability of this method can be assumed.

4.3 Comparison of univariate-bidirectional and univariate entangled LSTMs

The results from the previous multivariate forecasting show that using Granger's definition of causality and modelling the causal relationship between the two-time series can provide better forecasting results than simply using one or more time series to forecast another time series. Next, the comparison of univariate forecasting using bidirectional and Entangled LSTMs, was performed. For comparing the univariate forecasting quality, only MAPE was used because the other two metrics (i.e., MAE and RMSE) are scale-dependent and will follow the same trend as the MAPE. There are two key aspects that can be observed in Table 3. First, the univariate results are underperforming when compared to the multivariate results. This depicts the importance of considering the multiple weather features. Second, for all the univariate results, the Entangled LSTM outperforms the Bidirectional LSTM. This is another proof that incorporating causal relation between the weather features and the energy consumption is beneficial for the forecasting of energy consumption. From the univariate forecasting, it is also clear that the temperature is the most important weather feature for forecasting. Using the temperature, in both the Entangled and Bidirectional LSTMs, the MAPE is the lowest for all the datasets (and data-subsets). In some cases, temperature features marginally outperform all the other features; however, it emerges as the best feature to be used in the forecasting, nonetheless. In summary, With these univariate forecasting results, it is shown that it is important not only to include multivariability in the LSTM but also causal relationships.

Table 3 The univariate testing performance, MAPE, from the two LSTM architectures using temperature, humidity, wind speed, and pressure, separately

| Dataset | Bidirectional LSTM | | | | Entangled LSTM | | | |
|-------------------|--------------------|-------|-------|-------|----------------|-------|-------|-------|
| | Temp | Humid | WS | Pres | Temp | Humid | WS | Pres |
| Spain (Barcelona) | 11.25 | 16.87 | 17.76 | 14.22 | 10.10 | 15.85 | 17.76 | 12.62 |
| Spain (Bilbao) | 14.07 | 18.70 | 13.46 | 14.28 | 11.45 | 13.63 | 13.46 | 12.37 |
| Spain (Madrid) | 15.19 | 18.10 | 15.36 | 16.39 | 13.23 | 14.49 | 15.36 | 14.79 |
| Spain (Seville) | 14.41 | 15.90 | 19.14 | 15.44 | 12.02 | 12.41 | 19.14 | 12.08 |
| Spain (Valencia) | 15.76 | 16.23 | 18.03 | 17.29 | 11.84 | 14.25 | 18.03 | 12.18 |
| Paraguay | 14.03 | 16.51 | 16.80 | 14.96 | 16.18 | 15.48 | 16.80 | 17.83 |
| France | 20.16 | 21.24 | 22.98 | 20.45 | 20.07 | 20.62 | 22.98 | 22.78 |

4.4 Comparing early predictions

While the Entangled LSTM outperforms the Bidirectional LSTM in all the comparisons, it is also important to understand "how much data is needed to obtain reliable forecasts?" To answer this question, less data was used to forecast than what was available in the given datasets. For example, the results in Sect. 4.2 are based on training the algorithms using the data from 29 months. In different experiments, the same forecasts were obtained using half data (14.5 months), third data (9.67 months), fourth data (7.25 months), sixth data (4.85 months), and eighth data (3.63 months). The purpose was to know at what proportion of the data the forecasting accuracy decreases to a level, as compared to the results from Sect. 4.2, that it ceases to be potentially useful. In all these cases, both the entangled and bidirectional LSTMs were compared. These results were also compared using the MAPE values of the forecasting performance. Figure 6 shows the results. In all the cases (i.e., all the

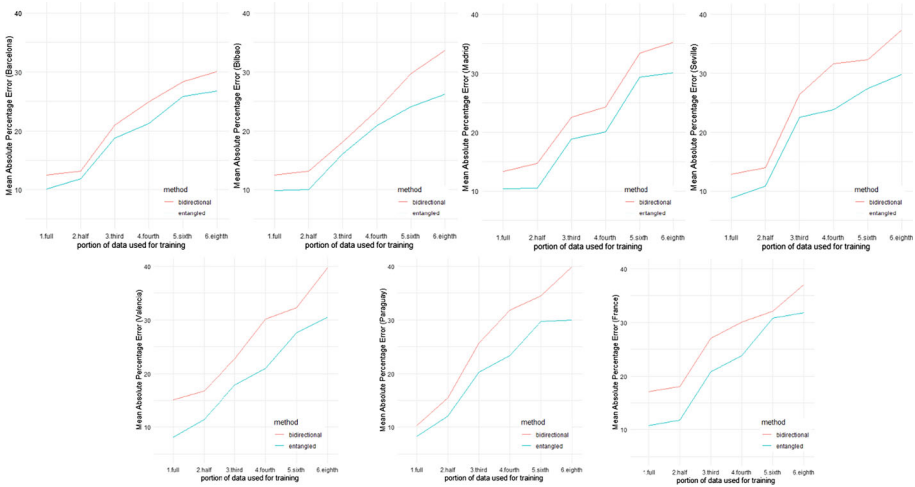


Fig. 6 Results (MAPE values) for the early prediction experiments

datasets and the data subsets), it is observed that when the data is reduced from half of the original training data length to a third of the data length, a considerable increment in the MAPE occurs. Moreover, with subsequent reductions in the data, further increments in the MAPE values can be seen for all the datasets. In all these experiments, there are two key takeaways. First, as aforementioned, given these three datasets, half of the dataset (14.5 months) is sufficient to obtain similar forecasting performance as with the full data (29 months). Second, in early predictions, as is seen with univariate and multivariate forecasting, the entangled LSTM outperforms the bidirectional LSTM. Another proof for the initial hypothesis is that incorporating the causal information in the deep network for forecasting the energy demand yields better results than simply using the weather features in a multivariate LSTM.

4.5 Comparing generalizability across datasets

Finally, the two algorithms are compared on the scale of generalizability. To perform such a comparison, cross-training testing was used. In other words, the training of the forecasting algorithm was carried out using one dataset, and the trained model was tested on another dataset. For example, training on the Spain dataset (all cities combined) and testing on France and Paraguay datasets. In this manner, both the algorithms were tested if they could use the learned model to predict in an unseen environment. The algorithm is deemed to be pseudo-generalizable if the testing accuracy on another dataset is comparable to the testing accuracy on the same dataset. For example, in the case of training on Spain and testing on Paraguay datasets, it can be observed from Table 4 that both LSTM models produce similar MAPE as when they were tested on the Spain dataset. When Bidirectional LSTM is trained on the Spain dataset, the testing MAPE for the Spain dataset is 13.32, while on the Paraguay dataset, the MAPE was 16.21. On the other hand, when the entangled LSTM is trained on the Spain dataset, the testing MAPE for the Spain dataset is 9.28, while on the Paraguay dataset, the MAPE is 11.21. In both cases, the algorithms can be considered generalizable. This difference is greater in the opposite case (i.e., when the training is done on the Paraguay dataset). However, the Entangled LSTM seems to be more generalizable

Table 4 Results (MAPE values) from cross-training testing (i.e., training on one dataset and testing on another one)

| Bidirectional LSTM | Testing dataset | | |
|--------------------|-----------------|----------|--------|
| | Spain | Paraguay | France |
| Training dataset | | | |
| Spain | 13.32 | 16.21 | 23.91 |
| Paraguay | 16.32 | 10.36 | 24.53 |
| France | 26.23 | 26.56 | 17.80 |
| Entangled LSTM | Testing dataset | | |
| | Spain | Paraguay | France |
| Training dataset | | | |
| Spain | 9.28 | 11.12 | 18.93 |
| Paraguay | 11.92 | 8.29 | 19.21 |
| France | 21.53 | 19.24 | 10.07 |

than the Bidirectional LSTM. In the end, both algorithms do not seem to generalize for the France dataset. It is shown that when the algorithms are trained using the country-wide dataset (Spain), they generalize on the region-based dataset (Paraguay) but not on a building-specific dataset (France). Moreover, the region-based dataset does not generalize for the building-based dataset; also, the algorithms do not generalize or generalize to a smaller extent when they are trained on a smaller area and tested on a larger area (i.e., trained on region or building based data and tested on country or region based data, respectively). These results can be explained by the fact that the larger the area in the trained dataset, the higher the variance in the model, and therefore they generalize on specific cases, but the opposite is not possible. Another aspect worth mentioning about the results from Table 4 is that the MAPE differences are lower in Entangled LSTM than those for the bidirectional LSTM. This can be explained based on the fact that the causal relationship between the weather and the energy consumption across countries might be more generalizable than the temporal nature of the weather conditions across different countries. This might be why the Entangle LSTM achieves better generalizability than the Bidirectional LSTM.

5 Discussion

5.1 Implications for theory

Causal analysis has lately become one of the major and important components for improving modern processes, and a few examples include Capability Maturity Model Integration (CMMI), Six Sigma, and Lean. Incorporating the causal analysis is also becoming more important for the organizations than before to obtain high levels of process maturation. This is evident because of two facts. First, CMMI has assigned the Causal Analysis and Resolution at the Maturity Level five. Second, Six Sigma programs also often embed causal analysis in a curriculum that is of more challenging statistical levels. Identifying causal relationships in the data helps the systems to predict future events in a more robust manner than those done using just the correlations and regressions (Kleinberg, 2013). One of the key goals of causal analysis is to aid organizations in making better decisions (Zheng et al., 2020). Causal relations can provide better support in decision-making due to the following two reasons. First, by analysing causalities in a particular data set, one can not only gain a deeper understanding of the relationship between pairs of time series and provide better predictions, but one can also analyse the effects of certain past events or decisions on the future predictions and outcomes. Second, based on the known causal relations, one can forecast the effect before it takes place and simulates the effects of certain decision-making activities. Encoding the intricacies of the relationships in the given datasets using the causal analysis might be the root cause of the results that were obtained. This contribution shows how to improve the time series forecasting using by adding the causal information (entangled LSTM) to the basic forecasting model (bidirectional LSTM).

Furthermore, it is also backed by a few theoretical frameworks that the causal information should be used in the decision making. One of the most prominent ones is the Expected Utility Theory (von Neumann & Morgenstern, 1944; Savage, 1954). This theory has two assumptions, and incorporating the causal relation between the weather information and the energy demands shows how these two assumptions are satisfied. Especially because the results show an improvement in forecasting capabilities when the causal information is incorporated into the deep learning network as compared to when the causal information is absent from the

forecasting problem. The first assumption is that each outcome has a corresponding utility to the decision-maker. In the presented case, this is the amount of energy consumed. The second assumption is that each outcome is assigned a probability. That is, each outcome is uncertain because there is always a lack of knowledge and evidence for an outcome to take place given a particular event or action. The theory also dictates that the outcome should maximize the expected utility. This lack of knowledge can be taken care of by using the causal relationship between the events and the outcomes. In this case, the energy consumption is the outcome, and the weather information creates the event. Therefore, it is intuitive to consider the causal relationship between energy consumption and weather data to effectively predict energy consumption. The results confirm the theory, where the entangled LSTM produces better forecasts than the bidirectional LSTM in all the cases, by showing that the power of causality in the predictor makes up for the lack of knowledge about the relation between an event (weather condition) and outcome (energy consumption). The inclusion of causal information is also supported by the causal decision theory (Joyce, 1999; Lewis, 1981; Maher, 1987; Nozick, 1993; Skyrms, 1982), which extends the Expected Utility Theory by dictating that knowing the outcomes of one must be aware of the causal relationship between an event and outcomes.

5.2 Implications for practice

Deep learning architectures encode the representation of the input data at multiple levels of abstraction using their multiple processing layers (LeCun et al., 2015). These encodings then can and are used to generate better predictions and forecasts than the other basic machine learning and statistical methods (Husna et al., 2021; Kilimci et al., 2019; Xu & He, 2020). It has been shown that predictive analysis brings competitive advantages to organizations (Ransbotham et al., 2017), and by extension, using deep learning to obtain better predictions can improve the advantageous positions for these organizations. Most of the predictions and/or forecasts aid in the decision making for different operations such as supply chain management (Husna et al., 2021; Pacella & Papadia, 2021; Punia et al., 2020), digital marketing (Pan & Zhou, 2020; Qi et al., 2019), and financial decision making (Abu-Mostafa & Atiya, 1996; Geng et al., 2015; Kim et al., 2020a, 2020b; Xu & He, 2020). With the current availability of the data in huge quantities, traditional machine learning algorithms tend to saturate their training and risk overfitting and specificity to one case. On the other hand, because of their complex structures and multiple weights to be trained, deep learning architectures can handle this large amount of data in a manner that is beneficial for various operations in organizations. Another advantage of deep learning algorithms is that they do not need extensive pre-processing and feature engineering because such algorithms are known to function well with noisy and unstructured or semi-structured datasets. Most of the contributions cited in the related work section of this paper do not use many pre-processing and/or feature engineering schemes. This also gives an additional advantage to the organizations in faster decision-making. Two other virtues of deep learning algorithms that help organizations to invest less time in data-driven decision-making are: 1) the less requirement of human intervention during the training phase of deep learning algorithms, and 2) the support for parallel and distributed processing. The first one refers to the self-learning capabilities of the deep learning algorithms using the error-backpropagation through its multiple layers and therefore requiring less human intervention as compared to the traditional machine learning algorithms. Moreover, the second advantage is mostly due to the advancements in

computing technology that allows the deep networks to be trained at scale and thus proving to be a big aid in fast data-driven decision support systems. With this contribution, by improving upon the state-of-the-art bidirectional LSTM networks mainly because of the use of causal information, the case of using deep learning architectures for operations research is emphasized.

The terms of the presented results (MAPE from entangled-LSTM), which are in the range of 0.08 and 0.11, stand comparable to some of the state-of-the-art contributions (Hrnjica & Mehr, 2020 report in the range 0.06—0.11 and Al Khafaf et al., 2019 report an average MAPE of 0.16); and better than others (Ishaq & Kwon, 2021: 0.35 and Somu & Ramamritham, 2021: 0.26); only Real et al. (2020) have reported a better MAPE range of 0.02—0.06. Considering that the better prediction of future energy consumption might lead to improving energy management processes and systems, this contribution also has certain implications for energy management systems. Efficient energy management is becoming a necessary process both at the supplier (i.e., smart grids) and consumer levels (i.e., energy management in smart homes). A smart grid energy management system contains multiple modules, among which the load and demand forecasting modules are also included (Chen et al., 2011). Effective forecasting systems can also support the real-time energy management systems in creating efficient load-balancing, operating routines, and minimizing operational costs (Luna et al., 2017). Moreover, better energy consumption forecasting can optimize the peak shavings for the utility grids and maximize the revenue for the grid (Shen et al., 2016). Finally, using highly accurate forecasts, it could be possible to minimize the power peaks and fluctuations while the grids are exchanging energy with each other or with the main grid (Arcos-Aviles et al., 2017). In terms of the consumer side of the energy supply chain, better energy forecasts can lead to better planning for optimizing smart home appliances (Hossen et al., 2018). Better energy consumption forecasts can have a major impact on the home energy management systems, which in turn can have a huge impact on the energy conservation, reliability, economics, and efficiency of the energy usage (Zhou et al., 2016). Whenever there is an option to choose between more than one source of energy, an efficient and individualized energy consumption forecast can also enable smart home energy management to switch among the multiple energy sources in a cost-effective manner (Olatomiwa et al., 2016).

5.3 Limitations and future work

Our contribution extends state the art in energy demand forecasting by incorporating the causal relationships between the weather parameters and the energy consumption at three different levels, that is, country, a specific region, and individual building. However, there were certain issues that limited the extent of this work and simultaneously opened new avenues for exploration. For example, in this paper, all the causal information that was available in the data was added, that is, the pressure causes consumption/demand, the humidity causes consumption/demand, the temperature causes consumption/demand, and the wind speed causes consumption/demand. Although each of these causalities seems intuitive to be added to the forecasting model, not all might have the same amount of mutual information with the actual demand. Therefore, it is important to explore which ones or which combinations would provide the most appropriate amount of information for the desired increase in the forecasting capability. Furthermore, the datasets in Spain and Paraguay were limited by the amount of information provided. For example, the Spain dataset had the national energy demand, but the weather information was about the five largest cities in the country. It is safe to assume that the five largest cities can control a big proportion of energy consumed.

However, it is a limitation in terms of analyses performed in this paper. On the other hand, in the Paraguay datasets, this problem was inverted. That is, local energy consumption data was available, but the weather information was centralised. Once again, it is not a completely valid assumption that the weather parameters would remain the same across a big region, but it is safe to assume that feeders in the high to medium vicinity of the weather station would be parameterised better by the causal relationship in the aggregate energy consumption data. In the future, it should be aimed to obtain the data such that the geographical spread of the two data streams is better matched than the two datasets. Different sources of energy (available in the Spain dataset) and the different modalities of usage (available in the France dataset) were not considered, where the causal relationship among each of these could have with the weather indicators. This choice was made to have one common analysis across the three datasets to showcase the generalizability of the proposed approach. However, exploring the different causal relationships between various energy usages and modalities with the weather conditions might also provide a better forecast for individual cases. Finally, for consistency of analysis across the three datasets, only four parameters were used to indicate weather conditions, that is, humidity, pressure, temperature, and wind speed. In the future, rainfall, snowfall, and wind direction, among other additional weather parameters, can also be considered.

From the current results presented in this paper, several venues emerge that could bring novel knowledge in the field of forecasting energy demand or consumption. First, at the forecasting level, exploring the multivariate nature of the causal relationships between the weather conditions and the energy consumption could improve the forecasting performance; because the interaction effects between the weather conditions would also be exploited in such a manner. Second, on a higher level, implementing and controlling the energy production using such methods (on a small scale) would provide an opportunity to study the effectiveness of the forecasting algorithm in the real-life scenarios because, with the current contribution, the practical nature of such an improved consumption forecasting could not be studied. Third, as it is with any deep learning application, the transparency and explainability of the algorithms are not up to the standards in some other industry sectors; therefore, after knowing that causality plays a significant role in the forecasting processes, the explanation for "how the forecasting works" could be provided to users at various levels, such as customers, managers, and policymakers.

6 Conclusions

In a nutshell, this paper presents a deep learning method to forecast energy consumption using not only the weather data but also the causal relationship between the weather indicators and the energy consumption. For the causal modelling, the definition of causality between the two-time series provided by Granger was used. This method was applied to three freely available datasets and showed that in all the cases, that is, a country, a specific region, and an individual building, augmenting the forecasting model by the causal information also augments the forecasting performance. This contribution extends the state-of-the-art in four ways.

1. This paper proposes the inclusion of causal relations in the deep learning frameworks for forecasting the energy demand/consumption.
2. Extending the LSTM architecture by using the causal information about how weather conditions cause the changes in the energy consumption provides better forecasting results.

3. Using the causal relations within the LSTM framework also provides better early prediction. That is, it requires less amount of data to achieve a similar level of forecasting performance as it would have been required by a setup without the casual information.
4. By using cross-training–testing routines, this paper also shows the higher generalizability of the proposed method than the contemporary methods.

The theoretical and practical implications of the results are also provided, both of which indicate that including the causal information does not only confirm certain widely accepted theoretical frameworks but also provides better energy management opportunities both at the supplier (i.e., smart grids) and the consumer (i.e., smart homes) levels.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abu-Mostafa, Y. S., & Atiya, A. F. (1996). Introduction to Financial Forecasting. *Applied Intelligence*, 6(3), 205–213.
- Ağbulut, Ü. (2022). Forecasting of transportation-related energy demand and CO₂ emissions in Turkey with different machine learning algorithms. *Sustainable Production and Consumption*, 29, 141–157.
- Ahmad, A. S., Hassan, M. Y., Abdullah, M. P., Rahman, H. A., Hussin, F., Abdullah, H., & Saidur, R. (2014). A review on applications of ANN and SVM for building electrical energy consumption forecasting. *Renewable and Sustainable Energy Reviews*, 33, 102–109.
- Ahmad, T., & Chen, H. (2019). Nonlinear autoregressive and random forest approaches to forecasting electricity load for utility energy management systems. *Sustainable Cities and Society*, 45, 460–473.
- Al Khafaf, N., Jalili, M., & Sokolowski, P. (2019). Application of deep learning long short-term memory in energy demand forecasting. In *International conference on engineering applications of neural networks* (pp. 31–42). Springer, Cham.
- Almagtome, A. H. (2021). Artificial Intelligence Applications in Accounting and Financial Reporting Systems: An International Perspective. In *Handbook of Research on Applied AI for International Business and Marketing Applications* (pp. 540–558). IGI Global.
- Alzaidy, R., Caragea, C., & Giles, C. L. (2019). Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents. In *The world wide web conference* (pp. 2551–2557).
- Ancona, N., Marinazzo, D., & Stramaglia, S. (2004). Radial basis function approach to nonlinear Granger causality of time series. *Physical Review E*, 70(5), 056221.
- Arcos-Aviles, D., Pascual, J., Guinjoan, F., Marroyo, L., Sanchis, P., & Marietta, M. P. (2017). Low complexity energy management strategy for grid profile smoothing of a residential grid-connected microgrid using generation and demand forecasting. *Applied Energy*, 205, 69–84.
- Ardabili, S. F., Abdilalizadeh, L., Mako, C., Torok, B., & Mosavi, A. (2022). Systematic review of deep learning and machine learning for building energy.
- Aydin, G. Ö. K. H. A. N. (2014). Production modeling in the oil and natural gas industry: An application of trend analysis. *Petroleum Science and Technology*, 32(5), 555–564.
- Aydin, G. Ö. K. H. A. N. (2015). The Application of trend analysis for coal demand modeling. *Energy Sources, Part B: Economics, Planning, and Policy*, 10(2), 183–191.
- Bagchi, P., Sahu, S. K., Kumar, A., & Tan, K. H. (2022). Analysis of carbon productivity for firms in the manufacturing sector of India. *Technological Forecasting and Social Change*, 178, 121606.
- Bakay, M. S., & Ağbulut, Ü. (2021). Electricity production based forecasting of greenhouse gas emissions in Turkey with deep learning, support vector machine and artificial neural network algorithms. *Journal of Cleaner Production*, 285, 125324.
- Balaji, A. J., Ram, D. H., & Nair, B. B. (2018). Applicability of deep learning models for stock price forecasting an empirical study on BANKEX data. *Procedia Computer Science*, 143, 947–953.

- Bauwens, L., Laurent, S., & Rombouts, J. V. (2006). Multivariate GARCH models: A survey. *Journal of Applied Econometrics*, 21(1), 79–109.
- Bhattacharyya, S. C., & Timilsina, G. R. (2009). Energy demand models for policy formulation: a comparative study of energy demand models. World Bank Policy Research Working Paper, (4866).
- Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research* (pp. 301–328). Springer, Dordrecht.
- Braun, M. R., Altan, H., & Beck, S. B. M. (2014). Using regression analysis to predict the future energy consumption of a supermarket in the UK. *Applied Energy*, 130, 305–313.
- Catalina, T., Iordache, V., & Caracaleanu, B. (2013). Multiple regression model for fast prediction of the heating energy demand. *Energy and Buildings*, 57, 302–312.
- Chambliss, D. F., & Schutt, R. K. (2018). *Making sense of the social world: Methods of investigation*. Thousand Oaks: Sage Publications.
- Chang, O., Naranjo, I., Guerron, C., Criollo, D., Guerron, J., & Mosquera, G. (2017). A deep learning algorithm to forecast sales of pharmaceutical products. no. August.
- Chatzis, S. P., Siakoulis, V., Petropoulos, A., Stavroulakis, E., & Vlachogiannakis, N. (2018). Forecasting stock market crisis events using deep and statistical machine learning techniques. *Expert Systems with Applications*, 112, 353–371.
- Chen, H., Rossi, R. A., Mahadik, K., & Eldardiry, H. (2020). A context integrated relational spatio-temporal model for demand and supply forecasting. *arXiv preprint arXiv:2009.12469*.
- Chen, J., Lim, C. P., Tan, K. H., Govindan, K., & Kumar, A. (2021). Artificial intelligence-based human-centric decision support framework: an application to predictive maintenance in asset management under pandemic environments. *Annals of Operations Research*, 1–24.
- Chen, C., Duan, S., Cai, T., Liu, B., & Hu, G. (2011). Smart energy management system for optimal microgrid economic operation. *IET Renewable Power Generation*, 5(3), 258–267.
- Chen, I. F., & Lu, C. J. (2021). Demand forecasting for multichannel fashion retailers by integrating clustering and machine learning algorithms. *Processes*, 9(9), 1578.
- Chen, Y., Rangarajan, G., Feng, J., & Ding, M. (2004). Analyzing multiple nonlinear time series with extended Granger causality. *Physics Letters A*, 324(1), 26–35.
- Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2015). Gated feedback recurrent neural networks. In *International conference on machine learning* (pp. 2067–2075).
- Clements, J. M., Xu, D., Yousefi, N., & Efimov, D. (2020). Sequential deep learning for credit risk monitoring with tabular financial data. *arXiv preprint arXiv:2012.15330*.
- Crisóstomo de Castro Filho, H., Abílio de Carvalho Júnior, O., Ferreira de Carvalho, O. L., Pozzobon de Bem, P., dos Santos de Moura, R., Olino de Albuquerque, A., & Trancoso Gomes, R. A. (2020). Rice crop detection using LSTM, Bi-LSTM, and machine learning models from sentinel-1 time series. *Remote Sensing*, 12(16), 2655.
- Darapaneni, N., Paduri, A. R., Sharma, H., Manjrekar, M., Hindlekar, N., Bhagat, P., & Agarwal, Y. (2022). Stock price prediction using sentiment analysis and deep learning for Indian markets. *arXiv preprint arXiv:2204.05783*.
- Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. *International Journal of Forecasting*, 29(3), 510–522.
- Del Real, A. J., Dorado, F., & Durán, J. (2020). Energy demand forecasting using deep learning: Applications for the French grid. *Energies*, 13(9), 2242.
- Demir, S., Mincev, K., Kok, K., & Paterakis, N. G. (2021). Data augmentation for time series regression: Applying transformations, autoencoders and adversarial networks to electricity price forecasting. *Applied Energy*, 304, 117695.
- Ding, M., Bressler, S. L., Yang, W., & Liang, H. (2000). Short-window spectral analysis of cortical event-related potentials by adaptive multivariate autoregressive modeling: Data preprocessing, model validation, and variability assessment. *Biological Cybernetics*, 83(1), 35–45.
- Dorsman, A. B., Atici, K. B., Ulucan, A., & Karan, M. B. (2021). Introduction: Applied operations research and financial modeling in energy. In *Applied operations research and financial modelling in energy* (pp. 1–6). Springer, Cham.
- Du, B., Zhou, Q., Guo, J., Guo, S., & Wang, L. (2021). Deep learning with long short-term memory neural networks combining wavelet transform and principal component analysis for daily urban water demand forecasting. *Expert Systems with Applications*, 171, 114571.
- Du, Z., Ge, L., Ng, A. H. M., Zhu, Q., Horgan, F. G., & Zhang, Q. (2020). Risk assessment for tailings dams in Brumadinho of Brazil using InSAR time series approach. *Science of the Total Environment*, 717, 137125.
- Eachempati, P., Srivastava, P. R., Kumar, A., Tan, K. H., & Gupta, S. (2021). Validating the impact of accounting disclosures on stock market: A deep neural network approach. *Technological Forecasting and Social Change*, 170, 120903.

- Ediger, V. Ş, & Tatlıdil, H. (2002). Forecasting the primary energy demand in Turkey and analysis of cyclic patterns. *Energy Conversion and Management*, 43(4), 473–487.
- Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, 5(2), 155.
- Efimova, O., & Serletis, A. (2014). Energy markets volatility modelling using GARCH. *Energy Economics*, 43, 264–273.
- Engle, R. (2001). GARCH 101: The use of ARCH/GARCH models in applied econometrics. *Journal of Economic Perspectives*, 15(4), 157–168.
- Erdogdu, E. (2007). Electricity demand analysis using cointegration and ARIMA modelling: A case study of Turkey. *Energy Policy*, 35(2), 1129–1146.
- Ergen, I., & Rizvanoghlu, I. (2016). Asymmetric impacts of fundamentals on the natural gas futures volatility: An augmented GARCH approach. *Energy Economics*, 56, 64–74.
- Eseye, A. T., & Lehtonen, M. (2020). Short-term forecasting of heat demand of buildings for efficient and optimal energy management based on integrated machine learning models. *IEEE Transactions on Industrial Informatics*, 16(12), 7743–7755.
- Fałdziński, M., Fiszeder, P., & Orzeszko, W. (2020). Forecasting volatility of energy commodities: Comparison of GARCH models with support vector regression. *Energies*, 14(1), 1–1.
- Farajian, L., Moghaddasi, R., & Hosseini, S. (2018). Agricultural energy demand modeling in Iran: Approaching to a more sustainable situation. *Energy Reports*, 4, 260–265.
- Ferlito, S., Atrigna, M., Graditi, G., De Vito, S., Salvato, M., Buonanno, A., & Di Francia, G. (2015). Predictive models for building's energy consumption: An Artificial Neural Network (ANN) approach. In *2015 xviii aiseem annual conference* (pp. 1–4). IEEE.
- Ferreira, R., Braga, M., & Alves, V. (2018, March). Forecast in the pharmaceutical area—statistic models vs deep learning. In *World conference on information systems and technologies* (pp. 165–175). Springer, Cham.
- Forootan, M. M., Larki, I., Zahedi, R., & Ahmadi, A. (2022). Machine learning and deep learning in energy systems: A review. *Sustainability*, 14(8), 4832.
- Forouzandeh, N., Zomorodian, Z. S., Shaghaghian, Z., & Tahsildoost, M. (2022). Room energy demand and thermal comfort predictions in early stages of design based on the machine learning methods. *Intelligent Buildings International*, 1–18.
- Fumo, N., & Biswas, M. R. (2015). Regression analysis for prediction of residential energy consumption. *Renewable and Sustainable Energy Reviews*, 47, 332–343.
- Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research*, 241(1), 236–247.
- Ghalekhondabi, I., Ardjmand, E., Weckman, G. R., & Young, W. A. (2017). An overview of energy demand forecasting methods published in 2005–2015. *Energy Systems*, 8(2), 411–447.
- Ghazal, T. M., Noreen, S., Said, R. A., Khan, M. A., Siddiqui, S. Y., Abbas, S., & Ahmad, M. (2022). Energy demand forecasting using fused machine learning approaches.
- Ghouali, S., Feham, M., & Ghouali, Y. Z. (2014). The direction of information between cardiorespiratory hemodynamic signals: test analysis using granger causality. *GSTF Journal of Mathematics, Statistics & Operations Research*, 2(2).
- Giri, C., Thomassey, S., Balkow, J., & Zeng, X. (2019). Forecasting new apparel sales using deep learning and nonlinear neural network regression. In *2019 International Conference on Engineering, Science, and Industrial Applications (ICESI)* (pp. 1–6). IEEE.
- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424–438.
- Guan, Y., & Yu, L. (2021). Design of supply chain resource distribution allocation model based on deep learning. In *International conference on multimedia technology and enhanced learning* (pp. 321–332). Springer, Cham.
- Harold, J., Cullinan, J., & Lyons, S. (2017). The income elasticity of household energy demand: A quantile regression analysis. *Applied Economics*, 49(54), 5570–5578.
- Hesse, W., Möller, E., Arnold, M., & Schack, B. (2003). The use of time-variant EEG Granger causality for inspecting directed interdependencies of neural assemblies. *Journal of Neuroscience Methods*, 124(1), 27–44.
- Ho, S. L., & Xie, M. (1998). The use of ARIMA models for reliability forecasting and analysis. *Computers & Industrial Engineering*, 35(1–2), 213–216.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Hor, C. L., Watson, S. J., & Majithia, S. (2006). Daily load forecasting and maximum demand estimation using ARIMA and GARCH. In *2006 International conference on probabilistic methods applied to power systems* (pp. 1–6). IEEE.

- Hossen, T., Nair, A. S., Noghianian, S., & Ranganathan, P. (2018). Optimal operation of smart home appliances using deep learning. In *2018 North American Power Symposium (NAPS)* (pp. 1–6). IEEE.
- Hrnjica, B., & Mehr, A. D. (2020). Energy demand forecasting using deep learning. In *Smart cities performativity, cognition, & security* (pp. 71–104). Springer, Cham.
- Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint. <http://arXiv:1508.01991>.
- Huang, B., Yao, X., Luo, Y., & Li, J. (2022). Improving financial distress prediction using textual sentiment of annual reports. *Annals of Operations Research*, 1–28.
- Huang, S. J., & Shih, K. R. (2003). Short-term load forecasting via ARMA model identification including non-Gaussian process considerations. *IEEE Transactions on Power Systems*, 18(2), 673–679.
- Husna, A., Amin, S. H., & Shah, B. (2021). Demand forecasting in supply chain management using different deep learning methods. In *Demand forecasting and order planning in supply chains and humanitarian logistics* (pp. 140–170). IGI Global.
- Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688.
- Ishaq, M., & Kwon, S. (2021). Short-term energy forecasting framework using an ensemble deep learning approach. *IEEE Access*, 9, 94262–94271.
- Islam, M. A., Che, H. S., Hasanuzzaman, M., & Rahim, N. A. (2020). Energy demand forecasting. In *Energy for sustainable development* (pp. 105–123). Academic Press.
- Iwafune, Y., Yagita, Y., Ikegami, T., & Ogimoto, K. (2014). Short-term forecasting of residential building load for distributed energy management. In *2014 IEEE international energy conference (ENERGYCON)* (pp. 1197–1204). IEEE.
- Jana, R. K., & Ghosh, I. (2022). A residual driven ensemble machine learning approach for forecasting natural gas prices: Analyses for pre-and during-COVID-19 phases. *Annals of Operations Research*, 1–22.
- Jiang, F., Yang, X., & Li, S. (2018). Comparison of forecasting India's energy demand using an MGM, ARIMA model, MGM-ARIMA model, and BP neural network model. *Sustainability*, 10(7), 2225.
- Jing, N., Wu, Z., & Wang, H. (2021). A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Systems with Applications*, 178, 115019.
- Johannesen, N. J., Kolhe, M., & Goodwin, M. (2019). Relative evaluation of regression tools for urban area electrical energy demand forecasting. *Journal of Cleaner Production*, 218, 555–564.
- Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge University Press.
- Kafazi, I., Bannari, R., & Abouabdellah, A. (2016). Modeling and forecasting energy demand. In *2016 international renewable and sustainable energy conference (IRSEC)* (pp. 746–750). IEEE.
- Kala, J. R., Kre, D. M., Gnassou, A. N. G., Kala, J. R. K., Akpablin, Y. M. A., & Coulibaly, T. (2020). Assets management on electrical grid using Faster-RCNN. *Annals of Operations Research*, 1–14.
- Kankal, M., & Uzlu, E. (2017). Neural network approach with teaching–learning-based optimization for modeling and forecasting long-term electric energy demand in Turkey. *Neural Computing and Applications*, 28(1), 737–747.
- Kayhan, S., Adiguzel, U., Bayat, T., & Lebe, F. (2010). Causality relationship between real GDP and electricity consumption in Romania (2001). *Romanian Journal of Economic Forecasting*, 169.
- Kegenbekov, Z., & Jackson, I. (2021). Adaptive supply chain: Demand-supply synchronization using deep reinforcement learning. *Algorithms*, 14(8), 240.
- Kilimci, Z. H., Akyuz, A. O., Uysal, M., Akyokus, S., Uysal, M. O., Atak Bulbul, B., & Ekemis, M. A. (2019). An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain. *Complexity*, 2019.
- Kim, A., Yang, Y., Lessmann, S., Ma, T., Sung, M. C., & Johnson, J. E. (2020a). Can deep learning predict risky retail investors? A case study in financial risk behavior forecasting. *European Journal of Operational Research*, 283(1), 217–234.
- Kim, J. Y., & Cho, S. B. (2019a). Electric energy consumption prediction by deep learning with state explainable autoencoder. *Energies*, 12(4), 739.
- Kim, M., Ryu, J., Cha, D., & Sim, M. K. (2020b). Stock price prediction using sentiment analysis: From. *The Journal of Society for e-Business Studies*, 25(4), 61–75.
- Kim, T. Y., & Cho, S. B. (2019b). Predicting residential energy consumption using CNN-LSTM neural networks. *Energy*, 182, 72–81.
- Kleinberg, S. (2013). *Causality, probability, and time*. Cambridge University Press.
- Kumar, A., Gopal, R. D., Shankar, R., & Tan, K. H. (2022b). Fraudulent review detection model focusing on emotional expressions and explicit aspects: Investigating the potential of feature engineering. *Decision Support Systems*, 113728.

- Kumar, A., Alsadoon, A., Prasad, P. W. C., Abdullah, S., Rashid, T. A., Pham, D. T. H., & Nguyen, T. Q. V. (2022a). Generative adversarial network (GAN) and enhanced root mean square error (ERMSE): Deep learning for stock price movement prediction. *Multimedia Tools and Applications*, 81(3), 3995–4013.
- Kumar, A., Shankar, R., Choudhary, A., & Thakur, L. S. (2016). A big data MapReduce framework for fault diagnosis in cloud-based manufacturing. *International Journal of Production Research*, 54(23), 7060–7073.
- Kumar, A., Shankar, R., & Thakur, L. S. (2018). A big data driven sustainable manufacturing framework for condition-based maintenance prediction. *Journal of Computational Science*, 27, 428–439.
- Le, T., Vo, M. T., Vo, B., Hwang, E., Rho, S., & Baik, S. W. (2019). Improving electric energy consumption prediction using CNN and Bi-LSTM. *Applied Sciences*, 9(20), 4237.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy*, 59(1), 5–30.
- Li, S., & Li, R. (2017). Comparison of forecasting energy consumption in Shandong, China Using the ARIMA model, GM model, and ARIMA-GM model. *Sustainability*, 9(7), 1181.
- Li, T., Li, Q., Zhu, S., & Ogihara, M. (2002). A survey on wavelet applications in data mining. *ACM SIGKDD Explorations Newsletter*, 4(2), 49–68.
- Li, Y., & Pan, Y. (2022). A novel ensemble deep learning model for stock prediction based on stock prices and news. *International Journal of Data Science and Analytics*, 13(2), 139–149.
- Lim, E. P., Chen, H., & Chen, G. (2013). Business intelligence and analytics: Research directions. *ACM Transactions on Management Information Systems (TMIS)*, 3(4), 1–10.
- Liu, Q., Tao, Z., Tse, Y., & Wang, C. (2022). Stock market prediction with deep learning: The case of China. *Finance Research Letters*, 46, 102209.
- Liu, Y. (2019). Novel volatility forecasting using deep learning—long short term memory recurrent neural networks. *Expert Systems with Applications*, 132, 99–109.
- Loureiro, A. L., Miguéis, V. L., & da Silva, L. F. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems*, 114, 81–93.
- Lu, H., Ma, X., & Ma, M. (2021). A hybrid multi-objective optimizer-based model for daily electricity demand prediction considering COVID-19. *Energy*, 219, 119568.
- Luna, A. C., Meng, L., Diaz, N. L., Graells, M., Vasquez, J. C., & Guerrero, J. M. (2017). Online energy management systems for microgrids: Experimental validation and assessment framework. *IEEE Transactions on Power Electronics*, 33(3), 2201–2215.
- Ma, J., Cheng, J. C., Jiang, F., Chen, W., Wang, M., & Zhai, C. (2020). A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data. *Energy and Buildings*, 216, 109941.
- Maouane, M., Zouggar, S., Krajačić, G., & Zahboune, H. (2021). Modelling industry energy demand using multiple linear regression analysis based on consumed quantity of goods. *Energy*, 225, 120270.
- Maher, P. (1987). Causality in the logic of decision. *Theory and Decision*, 22(2), 155–172.
- Mian, D. U., & Liang, G. U. (2010). Study on the causal relationship between proportion of blockholder and corporate performance—Based on panel-data granger causality tests of listed companies in China. *Forecasting*, 29(3), 50–54.
- Mitkov, A., Noorzad, N., Gabrovska-Evstatieva, K., & Mihailov, N. (2019). Forecasting the energy consumption in Afghanistan with the ARIMA model. In *2019 16th conference on electrical machines, drives and power systems (ELMA)* (pp. 1–4). IEEE.
- Moews, B., Herrmann, J. M., & Ibkunle, G. (2019). Lagged correlation-based deep learning for directional trend change prediction in financial time series. *Expert Systems with Applications*, 120, 197–206.
- Mohan, S., Mullanpudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019). Stock price prediction using news sentiment analysis. In *2019 IEEE fifth international conference on big data computing service and applications (BigDataService)* (pp. 205–208). IEEE.
- Morakinyo, T. E., Ren, C., Shi, Y., Lau, K. K. L., Tong, H. W., Choy, C. W., & Ng, E. (2019). Estimates of the impact of extreme heat events on cooling energy demand in Hong Kong. *Renewable Energy*, 142, 73–84.
- Mortenson, M. J., Doherty, N. F., & Robinson, S. (2015). Operational research from taylorism to terabytes: A research agenda for the analytics age. *European Journal of Operational Research*, 241(3), 583–595.
- Murat, Y. S., & Ceylan, H. (2006). Use of artificial neural networks for transport energy demand modeling. *Energy Policy*, 34(17), 3165–3172.
- Noureen, S., Atique, S., Roy, V., & Bayne, S. (2019a). Analysis and application of seasonal ARIMA model in energy demand forecasting: A case study of small scale agricultural load. In *2019a IEEE 62nd international midwest symposium on circuits and systems (MWSCAS)* (pp. 521–524). IEEE.
- Noureen, S., Atique, S., Roy, V., & Bayne, S. (2019b). A comparative forecasting analysis of arima model vs random forest algorithm for a case study of small-scale industrial load. *International Research Journal of Engineering and Technology*, 6(09), 1812–1821.

- Nozick, R. (1993). *The Nature of Rationality*. Princeton: Princeton University Press.
- Oh, G., Jeong, E., & Lim, S. (2021). Causal affect prediction model using a facial image sequence. arXiv preprint [arXiv:2107.03886](https://arxiv.org/abs/2107.03886).
- Okakwu, I. K., Oluwasogo, E. S., Ibhaze, A. E., & Imoize, A. L. (2019). A comparative study of time series analysis for forecasting energy demand in Nigeria. *Nigerian Journal of Technology*, 38(2), 465–469.
- Olatomiwa, L., Mekhilef, S., Ismail, M. S., & Moghavvemi, M. (2016). Energy management strategies in hybrid renewable energy systems: A review. *Renewable and Sustainable Energy Reviews*, 62, 821–835.
- Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications: A survey. *Applied Soft Computing*, 93, 106384.
- Ozturk, I., Aslan, A., & Kalyoncu, H. (2010). Energy consumption and economic growth relationship: Evidence from panel data for low and middle income countries. *Energy Policy*, 38(8), 4422–4428.
- Ozturk, S., & Ozturk, F. (2018). Forecasting energy consumption of Turkey by Arima model. *Journal of Asian Scientific Research*, 8(2), 52.
- Pacella, M., & Papadia, G. (2021). Evaluation of deep learning with long short-term memory networks for time series forecasting in supply chain management. *Procedia CIRP*, 99, 604–609.
- Pan, H., & Zhou, H. (2020). Study on convolutional neural network and its application in data mining and sales forecasting for E-commerce. *Electronic Commerce Research*, 20(2), 297–320.
- Parlos, A. G., Chong, K. T., & Atiya, A. F. (1994). Application of the recurrent multilayer perceptron in modeling complex process dynamics. *IEEE Transactions on Neural Networks*, 5(2), 255–266.
- Paterakis, N. G., Mocanu, E., Gibescu, M., Stappers, B., & van Alst, W. (2017). Deep learning versus traditional machine learning methods for aggregated energy demand prediction. In *2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)* (pp. 1–6). IEEE.
- Peřka, P. (2021). Pattern-based forecasting of monthly electricity demand using support vector machine. In *2021 International joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.
- Pérez-Cruz, F., Afonso-Rodríguez, J. A., & Giner, J. (2003). Estimating GARCH models using support vector machines. *Quantitative Finance*, 3(3), 163.
- Punia, S., Singh, S. P., & Madaan, J. K. (2020). A cross-temporal hierarchical framework and deep learning for supply chain forecasting. *Computers & Industrial Engineering*, 149, 106796.
- Qi, Y., Li, C., Deng, H., Cai, M., Qi, Y., & Deng, Y. (2019). A deep neural framework for sales forecasting in e-commerce. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 299–308).
- Rajesh, R. (2020). Sustainability performance predictions in supply chains: Grey and rough set theoretical approaches. *Annals of Operations Research*, 1–30.
- Ransbotham, S., Kiron, D., Gerbert, P., & Reeves, M. (2017). Reshaping business with artificial intelligence: Closing the gap between ambition and action. *MIT Sloan Management Review*, 59(1).
- Ranyard, J. C., Fildes, R., & Hu, T. I. (2015). Reassessing the scope of OR practice: The influences of problem structuring methods and the analytics movement. *European Journal of Operational Research*, 245(1), 1–13.
- Rao, T., & Srivastava, S. (2012). Analyzing stock market movements using twitter sentiment analysis.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive Modeling*, 5(3), 1.
- Sarkodie, S. A. (2017). Estimating Ghana's electricity consumption by 2030: An ARIMA forecast. *Energy Sources, Part B: Economics, Planning, and Policy*, 12(10), 936–944.
- Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.
- Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673–2681.
- Sen, J., & Mehtab, S. (2021). Accurate stock price forecasting using robust and optimized deep learning models. In *2021 International Conference on Intelligent Technologies (CONIT)* (pp. 1–9). IEEE.
- Sentürk, C., & Sataf, C. (2015). The determination of panel causality analysis on the relationship between economic growth and primary energy resources consumption of Turkey and Central Asian Turkish Republics. *Procedia-Social and Behavioral Sciences*, 195, 393–402.
- Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, 90, 106181.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.
- Shahi, T. B., Shrestha, A., Neupane, A., & Guo, W. (2020). Stock price forecasting with deep learning: A comparative study. *Mathematics*, 8(9), 1441.
- Shahid, F., Zameer, A., & Muneeb, M. (2020). Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons & Fractals*, 140, 110212.

- Shen, J., Jiang, C., Liu, Y., & Qian, J. (2016). A microgrid energy management system with demand response for providing grid peak shaving. *Electric Power Components and Systems*, 44(8), 843–852.
- Shirzadi, N., Nizami, A., Khazen, M., & Nik-Bakht, M. (2021). Medium-term regional electricity load forecasting through machine learning and deep learning. *Designs*, 5(2), 27.
- Shivaprasad, T. K., & Shetty, J. (2017). Sentiment analysis of product reviews: a review. In *2017 International conference on inventive communication and computational technologies (ICICCT)* (pp. 298–301). IEEE.
- Shrestha, A., Li, H., Le Kernec, J., & Fioranelli, F. (2020). Continuous human activity classification from FMCW radar with Bi-LSTM networks. *IEEE Sensors Journal*, 20(22), 13607–13619.
- Sirignano, J., & Cont, R. (2019). Universal features of price formation in financial markets: Perspectives from deep learning. *Quantitative Finance*, 19(9), 1449–1459.
- Skyrms, B. (1982). Causal decision theory. *The Journal of Philosophy*, 79(11), 695–711.
- Somu, N., MR, G. R., & Ramamritham, K. (2021). A deep learning framework for building energy consumption forecast. *Renewable and Sustainable Energy Reviews*, 137, 110591.
- Sriram, L. M. K. (2020). *Causality theory and advanced machine learning in power systems applications* (Doctoral dissertation, The Florida State University).
- Sriram, L. M. K., Gilanifar, M., Zhou, Y., Ozguven, E. E., & Arghandeh, R. (2018). Causal Markov Elman network for load forecasting in multinetwork systems. *IEEE Transactions on Industrial Electronics*, 66(2), 1434–1442.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1), 1929–1958.
- Steyer, R. (2013). *Wahrscheinlichkeit und Regression*. Springer-Verlag.
- Steyer, R., Nachtigall, C., Wüthrich-Martone, O., & Kraus, K. (2002). Causal regression models III: Covariates, conditional, and unconditional average causal effects. *Methods of Psychological Research Online*, 7(1), 41–68.
- Suganthi, L., & Samuel, A. A. (2012). Energy models for demand forecasting—A review. *Renewable and Sustainable Energy Reviews*, 16(2), 1223–1240.
- Sugihara, G., May, R., Ye, H., Hsieh, C. H., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, 338(6106), 496–500.
- Sun, Q., Jankovic, M. V., Bally, L., & Mougiakakou, S. G. (2018). Predicting blood glucose with an lstm and bi-lstm based deep neural network. In *2018 14th symposium on neural networks and applications (NEUREL)* (pp. 1–5). IEEE.
- Swathi, T., Kasiviswanath, N., & Rao, A. A. (2022). An optimal deep learning-based lstm for stock price prediction using twitter sentiment analysis. *Applied Intelligence*, 1–14.
- Tang, C. F., & Chrsquo, K. S. (2011). The Granger causality between health expenditure and income in Southeast Asia economies. *African Journal of Business Management*, 5(16), 6814–6824.
- Tian, C., Huang, G., Piowar, J. M., Yeh, S. C., Lu, C., Duan, R., & Ren, J. (2022). Stochastic RCM-driven cooling and heating energy demand analysis for residential building. *Renewable and Sustainable Energy Reviews*, 153, 111764.
- Tsantekidis, A., Passalis, N., Tefas, A., Kannianen, J., Gabbouj, M., & Iosifidis, A. (2017). Using deep learning to detect price change indications in financial markets. In *2017 25th European signal processing conference (EUSIPCO)* (pp. 2511–2515). IEEE.
- Tso, G. K., & Guan, J. (2014). A multilevel regression approach to understand effects of environment indicators and household features on residential energy consumption. *Energy*, 66, 722–731.
- Valipour, M., Banihabib, M. E., & Behbahani, S. M. R. (2013). Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. *Journal of Hydrology*, 476, 433–441.
- van Steenberg, R. M., & Mes, M. R. (2020). Forecasting demand profiles of new products. *Decision Support Systems*, 139, 113401.
- Vathsala, M. K., & Holi, G. (2020). RNN based machine translation and transliteration for Twitter data. *International Journal of Speech Technology*, 23(3), 499–504.
- Vidal, A., & Kristjanpoller, W. (2020). Gold volatility prediction using a CNN-LSTM approach. *Expert Systems with Applications*, 157, 113481.
- Vo, T. T., Le, P. H., Nguyen, N. T., Nguyen, T. L., & Do, N. H. (2021). Demand forecasting and inventory prediction for apparel product using the ARIMA and fuzzy EPQ model. *Journal of Engineering Science & Technology Review*, 14(2), 80–89.
- von Neumann, J., & Oskar, M. (1944). *Theo: of Games and Economic Behavior*. Princeton: Princeton University Press. Second edition, 1947; third edition, 1953. Section 3, chapter I, reprinted in Alfred N. Page. (1968). *Utility Theoov A Book of Readings*. New York: Wiley, pp. 215–233.
- Wang, L., Zhan, L., & Li, R. (2019). Prediction of the energy demand trend in middle Africa—a comparison of MGM, MECM, ARIMA and BP mod.

- Wang, G., & Fan, Y. (2021). Research on stock price forecasting model based on deep learning. In *2021 4th international conference on information systems and computer aided education* (pp. 2946–2948).
- Wang, Z., Wang, Y., Zeng, R., Srinivasan, R. S., & Ahrentzen, S. (2018). Random Forest based hourly building energy prediction. *Energy and Buildings*, *171*, 11–25.
- Wichmann, P., Brintrup, A., Baker, S., Woodall, P., & McFarlane, D. (2020). Extracting supply chain maps from news articles using deep neural networks. *International Journal of Production Research*, *58*(17), 5320–5336.
- Wu, Y. H., & Shen, H. (2018). Grey-related least squares support vector machine optimization model and its application in predicting natural gas consumption demand. *Journal of Computational and Applied Mathematics*, *338*, 212–220.
- Xiong, R., Nichols, E. P., & Shen, Y. (2015). Deep learning stock volatility with google domestic trends. arXiv preprint [arXiv:1512.04916](https://arxiv.org/abs/1512.04916).
- Xu, R. Z., & He, M. K. (2020). Application of deep learning neural network in online supply chain financial credit risk assessment. In *2020 international conference on computer information and big data applications (CIBDA)* (pp. 224–232). IEEE.
- Yoon, J., & van der Schaar, M. (2017). E-RNN: Entangled recurrent neural networks for causal prediction. In *Proc. ICML workshop principled approaches deep learn.* (pp. 1–5).
- Yusof, N. N., Mohamed, A., & Abdul-Rahman, S. (2018). A review of contextual information for context-based approach in sentiment analysis. *International Journal of Machine Learning and Computing*, *8*(4), 399–403.
- Zhang, L., & Xu, Y. A. O. (2015). An improved method of granger causality test and application on the stock market risk transmission. *Economic Computation & Economic Cybernetics Studies & Research*, *49*(2).
- Zhang, Q., Lu, H., Sak, H., Tripathi, A., McDermott, E., Koo, S., & Kumar, S. (2020). Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. In *ICASSP 2020–2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 7829–7833). IEEE.
- Zhang, Y., Li, G., Muskat, B., & Law, R. (2021). Tourism demand forecasting: A decomposed deep learning approach. *Journal of Travel Research*, *60*(5), 981–997.
- Zhao, Y., & Chen, Z. (2021). Forecasting stock price movement: New evidence from a novel hybrid deep learning model. *Journal of Asian Business and Economic Studies*.
- Zheng, M., Marsh, J. K., Nickerson, J. V., & Kleinberg, S. (2020). How causal information affects decisions. *Cognitive Research: Principles and Implications*, *5*(1), 1–24.
- Zhou, B., Li, W., Chan, K. W., Cao, Y., Kuang, Y., Liu, X., & Wang, X. (2016). Smart home energy management systems: Concept, configurations, and scheduling strategies. *Renewable and Sustainable Energy Reviews*, *61*, 30–40.
- Zou, C., & Feng, J. (2009). Granger causality vs. dynamic Bayesian network inference: A comparative study. *BMC Bioinformatics*, *10*(1), 1–17.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.