Nowcasting of Lumber Futures Price with Google Trends Index Using Machine Learning and Deep Learning Models

Mingtao He
Wenying Li
Brian K. Via
Yaoqi Zhang

Abstract

Firms engaged in producing, processing, marketing, or using lumber and lumber products always invest in futures markets to reduce the risk of lumber price volatility. The accurate prediction of real-time prices can help companies and investors hedge risks and make correct market decisions. This paper explores whether Internet browsing habits can accurately nowcast the lumber futures price. The predictors are Google Trends index data related to lumber prices. This study offers a fresh perspective on nowcasting the lumber price accurately. The novel outlook of employing both machine learning and deep learning methods shows that despite the high predictive power of both the methods, on average, deep learning models can better capture trends and provide more accurate predictions than machine learning models. The artificial neural network model is the most competitive, followed by the recurrent neural network model.

Lumber futures have been traded at Chicago Mercantile Exchange since 1969 (Mehrotra and Carter 2017). Since the COVID-19 pandemic, the lumber futures price has experienced huge volatility. Figure 1 plots the daily opening price of lumber futures from May 3, 2011, to May 28, 2021. The opening price of lumber futures plummeted on April 1, 2020, then it returned to normal levels seen prior to the pandemic. After that, it continued to climb steeply and finally reached its highest point in 10 years on May 7, 2021, with $1,677 per thousand board feet (mbf). It was $425.9 per mbf on January 21, 2020, when the first COVID-19 case in the United States was confirmed (Sahu and Kumar 2020). The average opening price from May 2011 to January 2020 was $337 per mbf, while the average opening price from February 2020 to May 2021 was $698 per mbf. The unusual fluctuations exposed lumber futures products that were originally designed to hedge uncertainties to huge risks. Therefore, there is an urgent need to find a reliable method to predict the lumber futures price, which would help enterprises and investors hedge risks and make correct decisions in the market.

In recent decades, several lumber price prediction methods have been proposed, such as ordinary least-squares regression (Mehrotra and Carter 2017), vector autoregressive model (VAR) (Song 2006), autoregressive integrated moving average model (ARIMA) (Buongiorno and Balsiger 1977, Oliveira et al. 1977, Banaś and Utnik-Banaś 2021), seasonal autoregressive moving average model (SARIMA) (Banaś and Utnik-Banaś 2021), seasonal autoregressive moving average model with exogenous variables (SARIMAX) (Banaś and Utnik-Banaś 2021), forest simulation model (FORSIM) (Buongiorno et al. 1984), and sales & operations planning network model (Marier et al. 2014). Most of the literature on lumber price prediction is based on traditional statistical models (Marier et al. 2014), econometric models (Banaś and

The authors are, respectively, Ph.D. Candidate, School of Forestry and Wildlife Sci., Auburn Univ., Auburn, Alabama (mzh0097@auburn.edu); Assistant Professor, Dept. of Agric. Economics & Rural Sociology, Auburn Univ., Auburn, Alabama (wenying.li@auburn.edu [corresponding author]); Regions Professor, Forest Products Development Center (FPDC), School of Forestry and Wildlife Sci., Auburn Univ., Auburn, Alabama (brianvia@auburn.edu); and Alumni Professor, School of Forestry and Wildlife Sci., Auburn Univ., Auburn, Alabama (Zhangy3@auburn.edu). This paper was received for publication in October 2021. Article no. 21-00061.

Forest Prod. J. 72(1):11–20.
doi:10.13073/FPJ-D-21-00061

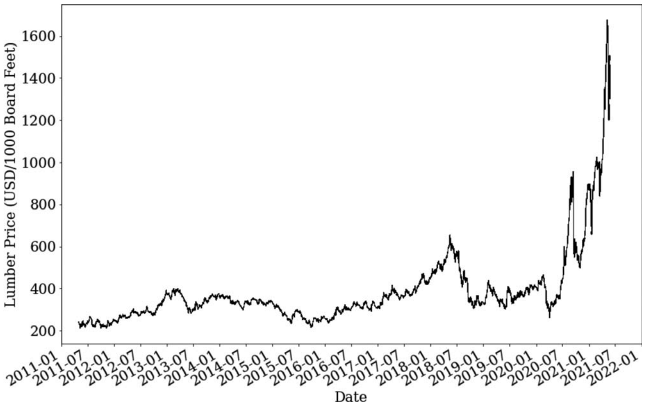

Figure 1.—The opening prices of lumber futures, United States, May 2011 to May 2021.

Utnik-Banaś 2021, Buongiorno and Balsiger 1977, Mehrotra and Carter 2017, Oliveira et al. 1977, Song 2006), or mathematical models (Buongiorno et al. 1984). So far, only one paper has used a recurrent neural networks model, which is a deep learning method to predict the closing price of lumber futures in the next few days using the price obtained from the previous few days (Verly Lopes et al. 2021).

In other domains, machine learning models and deep learning models were widely used for time series forecasting. A support vector machine (SVM) method was employed to forecast the daily electrical load (Singh and Mohapatra 2021) or wind speed (Gangwar et al. 2020). A random forest method was conducted in other studies to estimate poverty (Zhao et al. 2019) or the biomass weight of wheat (Zhou et al. 2016). XGBoost was run to forecast crude oil price (Gumus and Kiran 2017) or sales of the enterprise (Gurnani et al. 2017, Ji et al. 2019). Classification and regression tree (CART) was carried out to forecast precipitation (Choubin et al. 2018) or currency exchange rate (Haeri et al. 2015). And the deep learning models, including artificial neural network (ANN), recurrent neural network (RNN), and convolutional neural network (CNN), were applied to forecast construction material prices (Mir et al. 2021), photovoltaic power (Abdel-Nasser and Mahmoud 2019), gas demand (Su et al. 2019a), stock markets (Hoseinzade and Haratizadeh 2019), or river discharges (Awchi 2014). Overall, machine learning models and deep learning models have been widely employed to predict economic indicators, socioeconomic indicators, and science indicators. Machine learning models and deep learning models are statistical approaches. Compared to the traditional econometric models, they capture the hidden nonlinear characteristics among variables and provide more accurate predictions, while the econometric models are based on strict linear assumptions (Herrera et al. 2019) and might overfit the sample and yield forecasting error (Shobana and Umamaheswari 2021).

Some previous studies predicted the future lumber price based on the past values, which is an autoregressive technique (Song 2006). Other studies use some exogenous independent variables to predict lumber prices, such as the construction confidence index (Banaś and Utnik-Banaś 2021) and specific characteristics of the lumber supply chain (Marier et al. 2014). Models that include exogenous independent variables can produce good prediction results because the exogenous variables normally contain more information. However, none of these studies included public attention as an exogenous variable. Google is the most popular search engine in the United States. Google Trends is a publicly available service provided by Google. It provides access to aggregated information about different search queries and how those queries change over time. The Google Trends index is an index measuring the search volume of different queries over time. Users can use the Google Trends index to observe changes in the query volume of certain keywords over time and compare the query volume of different keywords over time. This provides an opportunity to capture the interest and concern of the public in real time without any cost. Therefore, Google Trends index is widely used to predict economic indicators and socioeconomic indicators, such as sales, unemployment, travel, consumer confidence (Choi and Varian 2012), consumer behavior (Carrière-Swallow and Labbé 2013), housing market (Dietzel 2016), the stock price (Hu et al. 2018), and so on.

This prospective study aims to use the Google Trends index of some keywords from the previous day to predict the next day's opening price of lumber futures. Nowcasting is the process of predicting the present, the very near future, or the very recent past value of an indicator based on real-time data (Banbura et al. 2010, Chumnumpan and Shi 2019). Nowcasting the opening price of lumber futures can help investors to take appropriate actions during the premarket trading hours between 8:00 a.m. to 9:30 a.m. Eastern each trading day. It would have a beneficial impact on hedging risks and expanding trade opportunities (Dungey et al. 2009). It would also be useful in helping enterprises navigate during normal and unusual times such as a pandemic. The statistical significance of the keywords of the Google Trends index will change over time. In other words, different factors have various effects on lumber futures price in different situations. The models can dynamically select the keyword variables in different time periods. As a result, the components of variables will change to capture dynamic trends of the real world. This study fills the gap in the literature by using machine learning and deep learning models to nowcast the lumber futures prices via Google Trends index.

This paper consists of five sections. The ''Data'' section briefly introduces the data. The ''Prediction Models'' section describes the models adopted in this study. The ''Results and Discussion'' section presents and discusses the results, and the ''Conclusion'' section concludes this study.

Data

Data collection

The Chicago Mercantile Exchange lumber futures price daily data were extracted from Investing.com. The dataset includes opening price, closing price, highest price, and lowest price of lumber futures. The data are from May 2011 to May 2021, with a total of 2,523 entries of data. The opening price of lumber futures is plotted in Figure 1.

The actual Google search requests for some lumber price–related keywords were then extracted from Google Trends index to match the same time series as the lumber price datasets. Keyword variables include 2 by 4 (a length of sawn wood 2 inches thick and 4 inches wide), BDFT (board foot), CLT (cross-laminated timber), commodity, DIY (do it yourself), fire, forest products association, forestry, hardwood, harvest, home building, home improvement, home renovation, invest, logging, logs, lumber futures, lumber price, lumber yard, MDF (medium density fiberboard), OSB (oriented strand board), plywood, sawmill, softwood, stock market, timber, and wood. Research has seen an effect on the lumber prices for a reduction in the quality of softwood lumber or in that case any lumber. Hence, more general keywords were included instead of the specific kinds of lumber. For example, the Southern pine and Douglas-fir lumber, which are the two most commercially important types of softwood lumber, have not changed in strength and stiffness over the last five decades (Miyamoto et al. 2018, França et al. 2021, Shmulsky et al. 2021), and thus they were not included in the keywords.

Google Trends index will standardize the data to a scale of 0 to 100 to represent the ''interest over time.'' But the scale of this data set will change if the same variable is colisted with other keywords or if the time range is changed. Therefore, it is important to always extract the same combination of words in the same time range during the modeling and prediction process to avoid restandardization of the same data set to different scales. However, the many years of daily keyword data cannot be downloaded directly from Google Trends. In order to avoid restandardization, application programming interface (API) was applied to extract Google Trends index data via *R*. The library ''gtrendsR'' on R was employed to extract the Google Trends index, and it retrieves the index via APIs. The descriptive statistics of opening price and closing price of lumber futures price and the whole Google Trends index of keywords is provided in Table 1.

Variable selection

To increase the model interpretability, remove redundant or irrelevant variables, and reduce overfitting, least absolute shrinkage and selection operator (LASSO) was first applied to perform independent variable selection (Fonti and Belitser 2017). The LASSO estimate can be written as

$$\hat{\beta} = arg \min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (1)$$

where $\lambda \geq 0$ is a constant parameter that controls the strength of regularization. The value of λ is directly proportional to the amount of regularization (Muthukrishnan and Rohini 2016, Fonti and Belitser 2017). In the LASSO process, the variables that have nonzero coefficients after the regularization are selected as part of the model (Fonti and Belitser 2017). As a result, the lumber futures closing price, and the Google Trends index of the four terms ''2 by 4,'' ''commodity,'' ''invest,'' and ''lumber futures'' were selected as the feature inputs of the models (Table 2). Figure 2 plots the daily Google Trends index of the above keywords from May 3, 2011, to May 28, 2021.

Sample splitting

Before building up the models, the dataset was divided into two subsets: a training set and a test set, which can avoid overfitting the models and improve the accuracy of the models (LeCun et al. 2015, Roelofs et al. 2019). The models will be trained on the training set, and the fitted models will be used to estimate the predicted value in the test set, which can provide an evaluation of the models. The different splitting rate of the data set is selected in respect to the object of characteristics of the studied subjects (Tao et al. 2020, Nguyen et al. 2021) and the sample size (Tai et al. 2019). In this study, considering that the lumber price does not fluctuate abnormally until the second half of 2020 and there are thousands of entries of samples, the splitting rate of the data set is determined to be 95 percent. The training set and the test set contain 95 and 5 percent of the total sample, respectively, which means the data of the first nine and a half years (May 3, 2011, to November 24, 2020) was used as the training set, and the data of the last six months (November 25, 2020, to May 28, 2021) will be used as the test set.

Prediction Models

Machine learning (ML) models and deep learning (DL) models have emerged with the advent of big data technology and gained in popularity as frontier prediction methods (Liakos et al. 2018). Machine learning models are the algorithms of providing machines the ability to optimize the performance without being strictly programmed (Schmidt et al. 2019, Kadam et al. 2020). Machine learning models include support vector machine (SVM), random forest, XGBoost, classification and regression trees (CART), and many more (Friedman et al. 2001). Deep learning models are defined as representation-learning algorithms composed of processing units organized in input, hidden layers, and output layers (LeCun et al. 2015, Shrestha and Mahmood 2019). Deep learning models include artificial neural network (ANN), recurrent neural network (RNN), and convolutional neural network (CNN) (Miotto et al. 2018).

Machine learning models

Support vector machine.—Support vector machine is an algorithm that maximizes a specific mathematical function based on a given data set (Noble 2006). SVM can be applied to time series prediction by introducing kernel functions (Pyo et al. 2017). In the SVM, the input vector *x* is mapped to the high-dimensional feature space using the nonlinear mapping function $\Phi(x)$ and run regression in the space (Wang et al. 2008). The SVM can be represented as the following equation:

$$\widehat{y_{\mathrm{SVM}}} = b + \sum_{i}^{n} w_i \Phi_i(x) \quad (2)$$

where $\widehat{y_{\mathrm{SVM}}}$ is the predicted value, parameters b and w_i can be estimated by minimizing the regularized risk function:

$$R(C) = C\frac{1}{n}\sum_{i=1}^{n} L_{\varepsilon}(y, \widehat{y_{\mathrm{SVM}}}) + \frac{1}{2}||w||^2 \quad (3)$$

where C is a regularization constant, y is the actual value, L_{ε} is the loss function, $(1/2)||w||^2$ is a measurement of function flatness. By introducing the kernel function K (x, y), Equation 2 can be transformed into the explicit form:

$$f_{\mathrm{SVM}}(x, \partial_i, \partial_i^*) = \sum_{i=1}^{n} (\partial_i - \partial_i^*) K(x, x_i) + b \quad (4)$$

where ∂_i and $\partial_i{}^*$ are the Lagrange multipliers which satisfy the condition: $\partial_i \times \partial_i{}^* = 0$, $\partial_i \geq 0$ and $\partial_i{}^* \geq 0$ (Choudhry and

Table 1.—Descriptive statistics of lumber price and Google trends index, United States, May 2011 to May 2021.

	Count	Mean	Std	Min	25%	50%	75%	Max
Google trends index								
2 by 4	2523	0.014	0.017	0.000	0.004	0.007	0.018	0.148
BDFT	2523	0.003	0.024	0.000	0.000	0.000	0.000	0.500
CLT	2523	0.185	0.093	0.000	0.111	0.174	0.249	0.500
Commodity	2523	0.060	0.027	0.007	0.040	0.057	0.073	0.260
DIY	2523	2.053	1.362	0.200	1.210	1.950	2.400	12.420
Fire	2523	31.476	9.235	7.360	25.440	30.680	36.580	100.000
Forest products association	2523	0.013	0.073	0.000	0.000	0.000	0.000	0.890
Forestry	2523	0.429	0.131	0.070	0.336	0.420	0.507	1.000
Hardwood	2523	0.082	0.034	0.020	0.060	0.080	0.100	0.210
Harvest	2523	0.357	0.181	0.050	0.240	0.300	0.450	2.240
Home building	2523	0.054	0.018	0.007	0.042	0.052	0.064	0.153
Home improvement	2523	0.041	0.060	0.000	0.010	0.020	0.030	0.400
Home renovation	2523	0.033	0.022	0.000	0.016	0.030	0.047	0.126
Invest	2523	1.523	0.694	0.304	1.035	1.382	1.849	6.000
Logging	2523	0.650	0.130	0.231	0.557	0.650	0.739	0.990
Logs	2523	0.042	0.021	0.010	0.020	0.040	0.060	0.140
Lumber futures	2523	0.0001	0.0007	0.0000	0.0000	0.0000	0.0000	0.0128
Lumber price	2523	0.001	0.004	0.000	0.000	0.000	0.001	0.044
Lumber yard	2523	0.008	0.007	0.000	0.003	0.006	0.011	0.046
MDF	2523	0.197	0.078	0.018	0.143	0.191	0.248	0.470
OSB	2523	0.226	0.123	0.000	0.140	0.211	0.291	1.000
Plywood	2523	0.777	0.273	0.207	0.577	0.740	0.918	2.000
Sawmill	2523	0.475	0.148	0.080	0.370	0.466	0.572	0.980
Softwood	2523	0.0002	0.0002	0.0000	0.0000	0.0002	0.0004	0.0028
Stock market	2523	0.707	2.003	0.020	0.150	0.280	0.540	35.000
Timber	2523	1.544	0.333	0.619	1.325	1.530	1.734	3.230
Wood	2523	12.518	4.947	3.630	9.180	11.760	15.360	37.600
Opening price of lumber futures	2523	384.8	186.1	211.9	292.5	336.6	390.5	1677.0
Closing price of lumber futures	2523	384.7	186.7	209.7	292.7	336.3	390.0	1686.0

Garg 2008, Wang et al. 2008). In this study, the $K\ (x, x_i)$ is the polynomial kernel function:

$$K(x,x_i) = (xx_i)^3 \tag{5}$$

where x_i is the sample in the training set (Choudhry and Garg 2008).

Random Forest.—Random forest is an algorithm that obtains the output by combining many decision trees to form forests (Breiman 2001). Specifically, it selects a bootstrap sample from the training set, which is selected randomly with replacement, and then obtains the optimal split point to split the node into two subtrees by minimizing mean squared error (MSE), which is called growing a random forest tree, T_m. After creation of M trees, the final output of random forest is defined as (Huang and Liu 2019, Peng et al. 2021, Yoon 2021):

$$\widehat{y_{RF}} = \frac{1}{M}\sum_{m=1}^{M} T_m(x) \tag{6}$$

XGBoost.—XGBoost is a regression tree algorithm, which is also called extreme gradient boosting. XGBoost is based on the gradient boosting decision tree algorithm and applies the addition of regularization terms to control the complexity of the model, which can prevent overfitting and improve the accuracy (Peng et al. 2019). As a result, the objective functions consist of two parts: training loss $L(\theta)$ and regularization $\Omega(\theta)$:

$$\mathrm{obj}(\theta) = L(\theta) + \Omega(\theta) \tag{7}$$

where θ is the parameter (Gurnani et al. 2017, Peng et al. 2019). The training loss is defined as:

$$L(\theta) = \sum_{i=1}^{n} (y_i - \widehat{y_{\mathrm{XGB}i}}) \tag{8}$$

where y_i is the actual value. In the XGBoost, each inner node represents the value of the attribute test, and the leaf node with values represents a decision (Xie and Zhang 2021). $\widehat{y_{\mathrm{XGB}i}}$ is the output, which is the sum of all predict values form M trees and can be written in the form:

$$\widehat{y_{\mathrm{XGB}i}} = \sum_{m=1}^{M} f_m(x_i), f_m \in F \tag{9}$$

where m is the number of trees, x_i is the ith training sample, f_m is the value for the mth tree in the functional space F (Peng et al. 2019, Xie and Zhang 2021).

The target function can be finally expressed as:

$$\mathrm{obj}(\theta) = \sum_{i=1}^{n} L(y_i, \widehat{y_{\mathrm{XGB}i}}) + \sum_{m=1}^{M} \Omega(f_m) \tag{10}$$

Classification and regression trees.—Classification and regression trees (CART) is a nonparametric statistical model, which is employed for classification problems or regression problems. If the output variable is continuous, the CART model will generate a regression tree. The CART tree is a hierarchical binary tree that is built up by splitting subsets of

Table 2.—LASSO results.

Variable	LASSO	Variable	LASSO	Variable	LASSO
Close	0.002241	Harvest	0	MDF	0
2 by 4	0.063729	Home building	0	OSB	0
BDFT	0	Home improvement	0	Plywood	0
CLT	0	Home renovation	0	Sawmill	0
Commodity	−0.35150	Invest	0.007146	Softwood	0
DIY	0	Logging	0	Stock market	0
Fire	0	Logs	0	Timber	0
Forest products association	0	Lumber futures	−8.96809	Wood	0
Forestry	0	Lumber price	0		
Hardwood	0	Lumber yard	0		

the data set by applying all output variables to generate two subnodes repeatedly. For determining the splitting, each predictor is evaluated to discover the best cut point, based on the least-squares deviation (LSD) impurity measure, $R(t)$ (Mahjoobi and Etemad-Shahidi 2008, Samadi et al. 2014):

$$R(t) = \frac{1}{N_{\omega}(t)} \sum_{i \in \omega} \omega_i f_i \Big(y_i - \bar{y}_{\mathrm{CART}}(t) \Big)^2 \tag{11}$$

where $N_{\omega}(t)$ is the weighted number of records at node t, ω_i is the value of the weighting field for record i, f_i is the value of the repeat field, y_i is the value of the target field, and $\bar{y}_{\mathrm{CART}}(t)$ is the mean of the output variable at node t.

Deep Learning Models

Artificial neural network model.—The artificial neural network (ANN) model connects the units called artificial neurons to generate complex networks (Kurbatsky et al. 2014, Su et al. 2019b). In each unit, there is an activation function, f, which applies the input variables, x_i, to generate the output value. The output of a unit conveyed to next unit as an input via a weighted connection. Given a unit, j, the output of this unit can be expressed as (Su et al. 2019b):

$$\widehat{y_{\mathrm{ANN}i}} = f_{\mathrm{ANN}} \left(\sum_{i=1}^{n} \omega_{ij} x_i + t_j \right) \tag{12}$$

where ω_{ij} is the connection weights and t_j is the bias term. The activation function, f_{ANN}, is rectified linear unit activation function in this study. The ANN model in this study is composed of an input layer, seven hidden layers, and an output layer. The output layer sums up the output of units from hidden layers. Different values of hyperparameter were tested, and the model with the best performance has a batch size 8, epochs 100, an optimizer of Adam, loss function of mean squared error, and one hidden layer with 64 units in this study.

Recurrent Neural Network.—Recurrent neural network (RNN) is a model of neural network. It applies the previous values of observations to calculate the future value by connecting the computational units from a directed circle (Selvin et al. 2017, Moghar and Hamiche 2020). However, the RNN confronts two problems: vanishing gradient and exploding gradient (Bouktif et al. 2018). As a result, long short-term memory (LSTM) was introduced to solve these problems in this study. The usually hidden layers were replaced with LSTM cells. The LSTM cells consist of input gate, forget gate, output gate, and cell state, which makes it possible to control the gradient flow and then overcome the vanishing and exploding gradient problems (Selvin et al. 2017, Bouktif et al. 2018). The LSTM cell can be expressed as (Bouktif et al. 2020):

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{13}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{14}$$

$$c_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{15}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{16}$$

$$h_t = o_t {*} \tanh(c_t) \tag{17}$$

where x_t is input vector at time t; h_{t-1} and h_t are output vector of hidden units at time $t - 1$ and time t, respectively; f_t, i_t, and o_t are forget, input, and output gate vector, respectively; c_t is the cell state vector; and $W*$ and $b*$ are the weight matrices and bias vector parameters of the LSTM unit, respectively. In this study, the RNN model is composed of an LSTM layer with 500 units and has epochs 50, batch size 9, an optimizer of Adam, and loss function of mean squared error. The activation function and recurrent activation function are hyperbolic tangent activation function and hard sigmoid activation function, respectively.

Convolutional neural network.—Convolutional neural network (CNN) is a class of feedforward neural networks, which can be effectively applied in image recognition, natural language processing, and time series data prediction (Lu et al. 2020). CNN consists of convolution layer, pooling layer, and fully connected layers. It extracts data features via the convolution layer and connects the units locally using the pooling layer, which reduces the redundant features (Chen et al. 2021). Then it converts the features in the previous layers to the final output using fully connected layers, which can be expressed as (Balaji et al. 2018):

$$\widehat{y_{\mathrm{CNN}i}}^{j} = f_{\mathrm{CNN}} \left(\sum_{k} \widehat{y_{\mathrm{CNN}k}}^{j-1} w_{k,i}^{j-1} \right) \tag{18}$$

where $\widehat{y_{\mathrm{CNN}i}}^{j}$ is the output value of unit i at the layer j,$\widehat{y_{\mathrm{CNN}k}}^{j-1}$ is the output value of unit k at the layer $j - 1$, f_{CNN} is the activation function. In this study, the activation function of CNN is rectified linear unit activation function. $w_{k,i}^{j-1}$ is the weight of the connection between unit k at layer $j - 1$ and unit i at layer j. In this study, the data is convoluted through a Conv-1D layer within 16 units, and then the max pooling layer. Next, the data are convoluted through another Conv-1D layer within 32 units, and then the global max

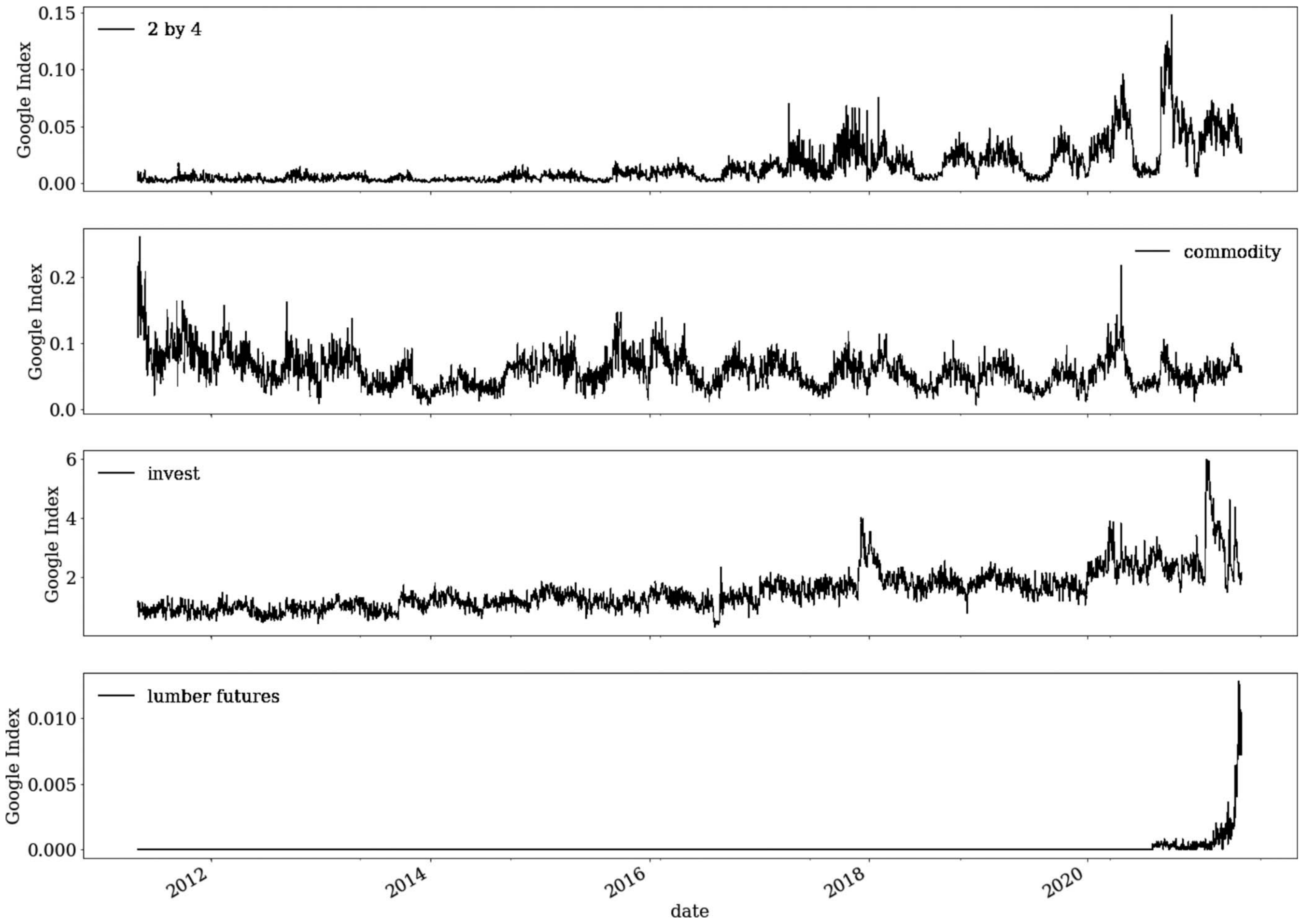

Figure 2.—Google trends index after LASSO, United States, May 2011 to May 2021.

pooling layer. The activation function is rectified linear unit. The CNN model has epochs 1500, an optimizer of Adam, and a loss function of mean squared error.

Evaluation of Models

To evaluate the performance of these models, the mean squared error (MSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and symmetric mean absolute percentage error (SMAPE) were used as the criteria. The measures are as follows:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y_i} - y_i)^2 \tag{19}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\widehat{y_i} - y_i| \tag{20}$$

$$\text{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{\widehat{y_i} - y_i}{y_i}\right| \tag{21}$$

$$\text{SMAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\frac{|\widehat{y_i} - y_i|}{(|\widehat{y_i}| + |y_i|)/2} \tag{22}$$

where N is the number of training set samples or test set samples, y_i is a real value at time t, and $\widehat{y_i}$ is the corresponding predicted value.

Results and Discussion

In this study, a baseline model was established, based on the naïve forecasting method, to provide the required point of comparison when evaluating all other models.[1] Naïve forecasting is the method in which actual values in the last period are simply taken as predicted values in this period. In the baseline model, the opening price at the previous time step $t - 1$ was used to be the predicted value at the time step t.[2]

[1] We have built up a multiple linear regression (MLR) model with the open price at time $t - 1$ according to the recommendations. The MSE, MAE, MAPE, and SMAPE of the MLR model are 2067.48, 30.65, 2.89, and 2.90 percent, respectively. Overall, the performance is slightly better than the naïve forecasting, but does not differ substantially. Therefore, we decide to use the naïve forecasting model, the common baseline method in the machine learning research field. This also follows the zero rule algorithm for the baseline method (Choudhary and Gianey 2017).

[2] Akaike information criterion (AIC), Bayesian information criterion (BIC), and Hannan-Quinn information criterion (HQIC) were employed to determine the lag order for the baseline model. Based on the selection criterion of the three models, the Lag 1 was selected because it has the smallest AIC, BIC, and HQIC values.

The prediction results of different models of the test set are shown in Figure 3, which contains 127 observations from November 25, 2020, to May 28, 2021. All four machine learning models and three deep learning models showed strong predictive ability because the predicted lumber prices are close to the actual prices.

Figure 3 shows that the random forest, XGBoost, CART, ANN, RNN, and CNN models can capture the trends and dynamics in the test set, while the SVM model fails to identify the pattern in the highest price interval, which makes the nowcasting less accurate. It should be noted that the actual lumber price in the test set is much higher than that in the training set. Most of the machine learning and deep learning models can still capture the trends and identify the pattern. This shows that the machine learning and deep learning models have the ability to extract hidden features among variables in high-dimensional and multivariate data sets in a complex and dynamic environment (Köksal et al. 2011, Wuest et al. 2016).

From the overall performance, the ANN model performs better than other models. There is a large overlap between predicted prices and actual prices, especially for the prediction of an abnormal trend of rapid growth from mid-March 2021 to early May 2021. Moreover, the ANN model provides significantly better predictions than the baseline model. Although the random forest, XGBoost, CART, and RNN models are inferior to ANN, the predicted prices of these models were highly consistent with the actual observations. SVM and CNN models have the weakest prediction effects among the machine learning and deep learning models, respectively, although predicted prices of these two models are also roughly close to the actual prices. The SVM model overestimates the lumber price from mid-March to early May significantly, and the CNN model cannot capture the trend of rapid growth very well, compared with the other two deep learning models. This result might be explained by the fact that the CNN model

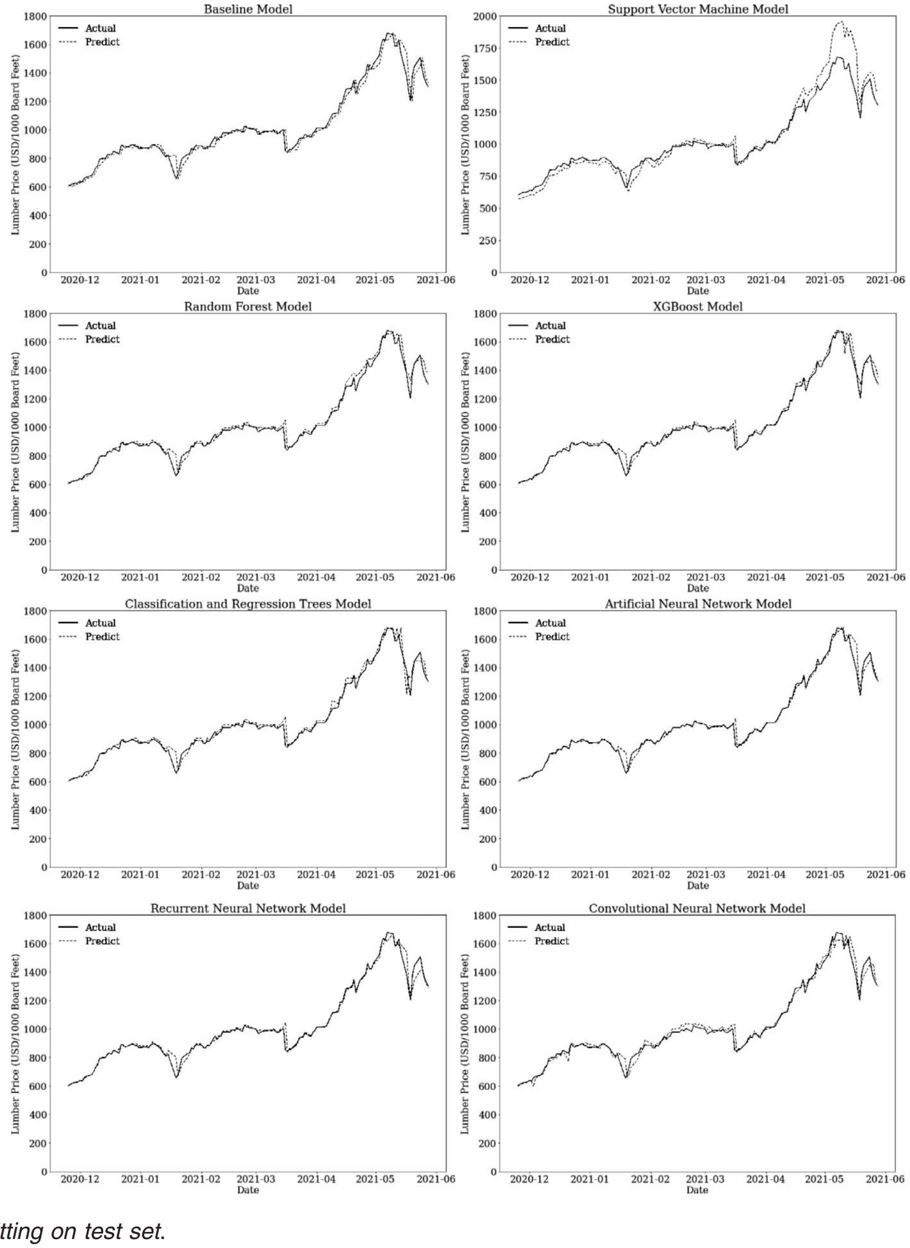

Figure 3.—Models fitting on test set.

does not depend on any information from previous observations to make a prediction (Selvin et al. 2017).

Figure 4 compares the average prediction performance between machine learning models, deep learning models, and the baseline model. Comparing the predictive performance of all seven models shows that the ANN model performs the best overall. The MSE, MAE, MAPE, and SMAPE of the test set are the lowest among these models. This may be explained by the good self-learning, self-adapting, and self-organizing ability of the ANN model, which can analyze the patterns and rules of observations through training (Su et al. 2019b). The RNN model is the second-best prediction performance model, which could be attributed to the good ability to use information from previous lags to predict the future values by RNN (Selvin et al. 2017). XGBoost gives more accurate predictions than other machine learning models, and it is also the third-best model among all seven models. ANN, RNN, XGBoost, random forest, CART, and CNN models provide more accurate results than the baseline model. In addition, the performance of the machine learning and deep learning models are generally better than traditional time series models. For example, Banaś and Utnik-Banaś (2021) forecasted round wood prices from 2019 Q1 to 2020 Q4 in Poland using ARIMA, SARIMA, and SARIMAX models, whose MAPE was 2.57, 2.20, and 1.75 percent on average, respectively. All the models except for SVM in this study have better performance than the ARIMA model. The ANN, RNN, XGBoost, random forest, and CART models in this study are better than the SARIMA model, and the ANN and RNN are better than the SARIMAX model.

Figures 3 and 4 show that, compared with machine learning models, deep learning models are, on average, more capable of capturing the trends and providing more

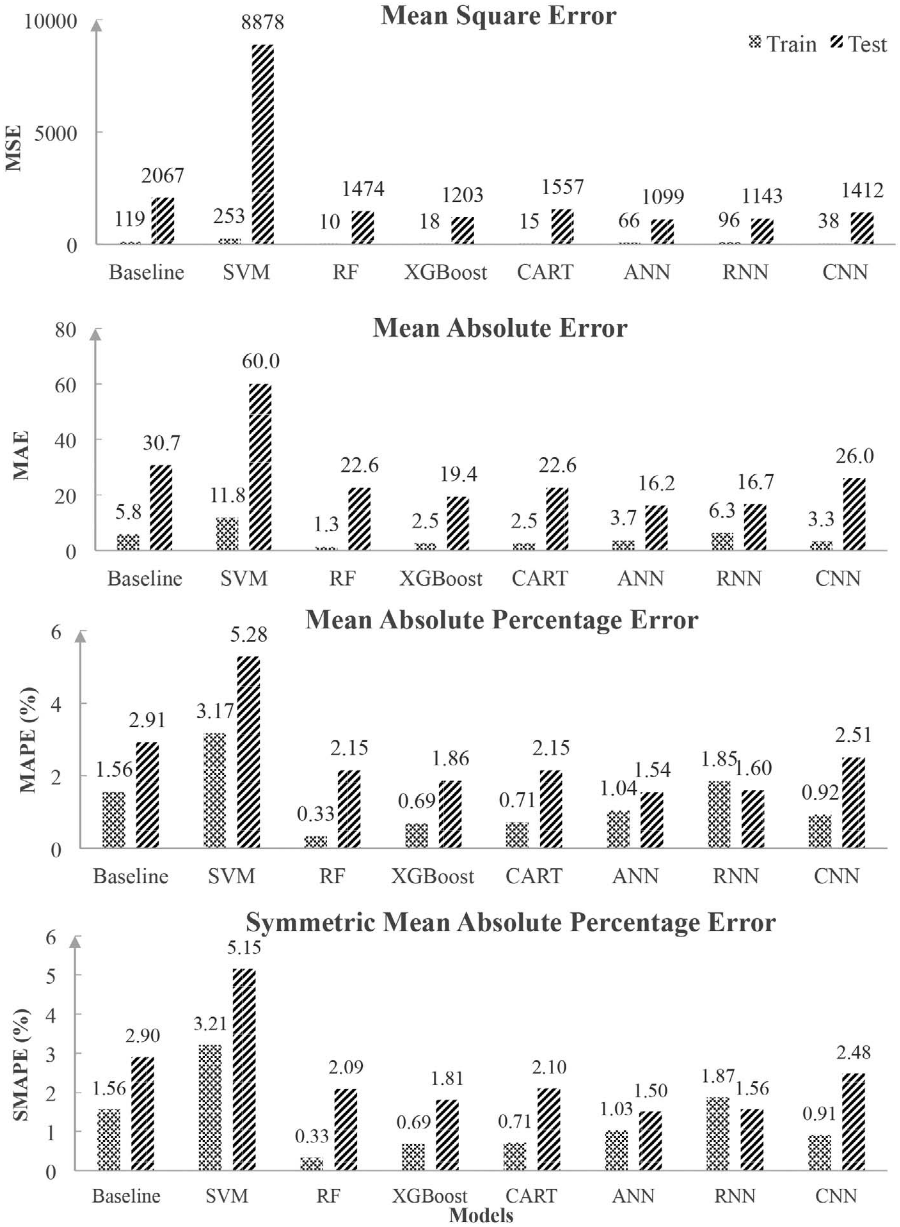

Figure 4.—Models evaluation.

accurate predictions. This may result from the better overfitting reduce ability of deep learning models. This can also be seen in Figure 4. The fitting performance of the three deep learning models to the training set is worse than that of the machine learning models.

Conclusions

This study describes a new approach for nowcasting the lumber futures price using Google Trends index through machine learning models (SVM, random forest, XGBoost, and CART) and deep learning models (ANN, RNN, and CNN). We show that deep learning models generally give more accurate predictions than machine learning models. Among the seven models, the ANN model provides the best performance, followed by the RNN model. The comparison with the baseline model shows that the random forest, XGBoost, CART, ANN, RNN, and CNN models provide more accurate predictions than the baseline model. Our findings also imply that the Google Trends index, which reflects the dynamic changes of the interest and attention from the public, can provide enough information to be good predictors in nowcasting lumber futures prices.

By using the prediction methods and Google Trends index, investors can take appropriate measures to hedge risks and make profits during premarket trading hours. The high predictive power of this approach implies that the big data models should be added to the toolbox of investors and policymakers to predict other economic variables. One probable criticism to these methods being applied to predict the lumber futures price followed by appropriate actions is that it might enhance the lumber futures market volatility and further lead to the invalidation of the forecasting.

Literature Cited

Abdel-Nasser, M. and K. Mahmoud. 2019. Accurate photovoltaic power forecasting models using deep LSTM-RNN. *Neural Comput. Appl.* 31(7):2727–2740.

Awchi, T. A. 2014. River discharges forecasting in northern Iraq using different ANN techniques. *Water Resour. Manag*. 28(3):801–814.

Balaji, A. J., D. H. Ram, and B. B. Nair. 2018. Applicability of deep learning models for stock price forecasting an empirical study on BANKEX data. *Procedia Computer Sci*. 143:947–953.

Banaś, J. and K. Utnik-Banaś. 2021. Evaluating a seasonal autoregressive moving average model with an exogenous variable for short-term timber price forecasting. *Forest Policy Econ*. 131:102564.

Banbura, M., D. Giannone, and L. Reichlin. 2010. Nowcasting. ECB Working Paper No. 1275.

Bouktif, S., A. Fiaz, A. Ouni, and M. A. Serhani. 2018. Optimal deep learning LSTM model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies* 11(7):1636.

Bouktif, S., A. Fiaz, A. Ouni, and M. A. Serhani. 2020. Multi-sequence LSTM-RNN deep learning and metaheuristics for electric load forecasting. *Energies* 13(2):391.

Breiman, L. 2001. Random forests. *Machine Learning* 45(1):5–32.

Buongiorno, J. and J. W. Balsiger. 1977. Quantitative analysis and forecasting of monthly prices of lumber and flooring products. *Agric. Syst*. 2(3):165–181.

Buongiorno, J., F. Mey Huang, and H. Spelter. 1984. Forecasting the price of lumber and plywood: Econometric model versus futures markets. *Forest Prod. J*. 34(7).

Carrière-Swallow, Y. and F. Labbé. 2013. Nowcasting with Google trends in an emerging market. *J. Forecast*. 32(4):289–298.

Chen, Y., R. Fang, T. Liang, Z. Sha, S. Li, Y. Yi, W. Zhou, and H. Song. 2021. Stock price forecast based on CNN-BiLSTM-ECA Model. *Scientific Programming* 2021:2446543.

Choi, H. and H. Varian. 2012. Predicting the present with Google Trends. *Economic Record* 88:2–9.

Choubin, B., G. Zehtabian, A. Azareh, E. Rafiei-Sardooi, F. Sajedi-Hosseini, and Ö. Kişi. 2018. Precipitation forecasting using classification and regression trees (CART) model: A comparative study of different approaches. *Environ. Earth Sci*. 77(8):1–13.

Choudhary, R. and H. K. Gianey. 2017. Comprehensive review on supervised machine learning algorithms. *In:* 2017 International Conference on Machine Learning and Data Science (MLDS), December 4–15, 2017, Noida, India; IEEE Computer Society's Conference Publishing Services (CPS), Piscataway, New Jersey. pp. 37–43.

Choudhry, R. and K. Garg. 2008. A hybrid machine learning system for stock market forecasting. *World Acad. Sci. Eng. Technol*. 39(3):315–318.

Chumnumpan, P. and X. Shi. 2019. Understanding new products' market performance using Google Trends. *Australasian Marketing J. (AMJ)* 27(2):91–103.

Dietzel, M. A. 2016. Sentiment-based predictions of housing market turning points with Google trends. *Int. J. Housing Markets Anal*. 9(1):108–136.

Dungey, M., L. Fakhrutdinova, and C. Goodhart. 2009. After-hours trading in equity futures markets. *J. Futures Markets: Futures, Options, Other Derivative Prod*. 29(2):114–136.

Fonti, V. and E. Belitser. 2017. Feature selection using lasso. VU Amsterdam Research Paper in Business Analytics, 30 pp.

França, F. J. N., R. Shmulsky, J. T. Ratcliff, B. Farber, C. A. Senalik, R. J. Ross, and R. D. Seale. 2021. Yellow pine small clear flexural properties across five decades. *Forest Prod. J*. 71(3):233–239.

Friedman, J., T. Hastie, and R. Tibshirani. 2001. The Elements of Statistical Learning. Springer Series in Statistics, New York.

Gangwar, S., V. Bali, and A. Kumar. 2020. Comparative analysis of wind speed forecasting using LSTM and SVM. *EAI Endorsed Trans. Scalable Inf. Syst*. 7(25):e1.

Gumus, M. and M. S. Kiran (Eds.). 2017. Crude oil price forecasting using XGBoost. *In:* 2017 International Conference on Computer Science and Engineering (UBMK). pp. 1100–1103.

Gurnani, M., Y. Korke, P. Shah, S. Udmale, V. Sambhe, and S. Bhirud. 2017. Forecasting of sales by using fusion of machine learning techniques. *In:* 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), February 24–26, 2017, Pune, India; IEEE Computer Society's Conference Publishing Services (CPS), Piscataway, New Jersey. pp. 93–101.

Haeri, A., S. M. Hatefi, and K. Rezaie. 2015. Forecasting about EUR/JPY exchange rate using hidden Markova model and CART classification algorithm. *J. Adv. Comput. Sci. Technol*. 4(1):84–89.

Herrera, G. P., M. Constantino, B. M. Tabak, H. Pistori, J.-J. Su, and A. Naranpanawa. 2019. Long-term forecast of energy commodities price using machine learning. *Energy* 179:214–221.

Hoseinzade, E. and S. Haratizadeh. 2019. CNNpred: CNN-based stock market prediction using a diverse set of variables. *Expert Syst. Appl*. 129:273–285.

Hu, H., L. Tang, S. Zhang, and H. Wang. 2018. Predicting the direction of stock markets using optimized neural networks with Google Trends. *Neurocomputing* 285:188–195.

Huang, S. and S. Liu. 2019. Machine learning on stock price movement forecast: The sample of the Taiwan stock exchange. *Int. J. Econ. Financial Issues* 9(2):189.

Ji, S., X. Wang, W. Zhao, and D. Guo. 2019. An application of a three-stage XGBoost-based model to sales forecasting of a cross-border E-commerce enterprise. *Math. Problems Eng*. 2019.

Kadam, V. S., S. Kanhere, and S. Mahindrakar. 2020. Regression techniques in machine learning & applications: A review. *Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET)* (10):826–830.

Köksal, G., I. Batmaz, and M. C. Testik. 2011. A review of data mining applications for quality improvement in manufacturing industry. *Expert Syst. Appl*. 38(10):13448–13467.

Kurbatsky, V. G., D. N. Sidorov, V. A. Spiryaev, and N. V. Tomin. 2014. Forecasting nonstationary time series based on Hilbert-Huang transform and machine learning. *Automation Remote Control* 75(5):922–934.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521(7553):436–444.

Liakos, K. G., P. Busato, D. Moshou, S. Pearson, and D. Bochtis. 2018. Machine learning in agriculture: A review. *Sensors* 18(8):2674.

Lu, W., J. Li, Y. Li, A. Sun, and J. Wang. 2020. A CNN-LSTM-based model to forecast stock prices. *Complexity* 2020.

Mahjoobi, J. and A. Etemad-Shahidi. 2008. An alternative approach for the prediction of significant wave heights based on classification and regression trees. *Appl. Ocean Res*. 30(3):172–177.

Marier, P., S. Bolduc, M. B. Ali, and J. Gaudreault (Eds.). 2014. S&OP network model for commodity lumber products. *In:* Proceedings of the 10th International Conference on Modeling, Optimization, and Simulation (MOSIM), November 5–7, 2014, Nancy, France; CIRRELT, Québec, Canada.

Mehrotra, S. N. and D. R. Carter. 2017. Forecasting performance of lumber futures prices. *Econ. Res. Int*. 2017:1–8.

Miotto, R., F. Wang, S. Wang, X. Jiang, and J. T. Dudley. 2018. Deep learning for healthcare: Review, opportunities and challenges. *Briefings Bioinformatics* 19(6):1236–1246.

Mir, M., H. D. Kabir, F. Nasirzadeh, and A. Khosravi. 2021. Neural network-based interval forecasting of construction material prices. *J. Building Eng*. 39:102288.

Miyamoto, B. T., K. Cheung, M. Clauson, and A. Sinha. 2018. Revisiting the compression parallel to grain design values of Douglas-fir. *Forest Prod. J*. 68(2):132–137.

Moghar, A. and M. Hamiche. 2020. Stock market prediction using LSTM recurrent neural network. *Procedia Computer Sci*. 170:1168–1173.

Muthukrishnan, R. and R. Rohini. 2016. LASSO: A feature selection technique in predictive modeling for machine learning. *In:* 2016 IEEE International Conference on Advances in Computer Applications (ICACA), October 24–24, 2016, Coimbatore, India; IEEE Computer Society's Conference Publishing Services (CPS), Piscataway, New Jersey. pp. 18–20.

Nguyen, Q. H., H.-B. Ly, L. S. Ho, N. Al-Ansari, H. van Le, Q. van Tran, I. Prakash, and B. T. Pham. 2021. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Math. Problems Eng*. 2021.

Noble, W. S. 2006. What is a support vector machine? *Nature Biotechnol*. 24(12):1565–1567.

Oliveira, R. A., J. Buongiorno, and A. M. Kmiotek. 1977. Time series forecasting models of lumber cash, futures, and basis prices. *Forest Sci*. 23(2):268–280.

Peng, L., L. Wang, X.-Y. Ai, and Y.-R. Zeng. 2021. Forecasting tourist arrivals via random forest and long short-term memory. *Cognitive Computation* 13(1):125–138.

Peng, Z., Q. Huang, and Y. Han. 2019. Model research on forecast of second-hand house price in Chengdu based on XGboost algorithm. *In:* 2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT), October 18–20, 2019, Jinan, China; IEEE Computer Society's Conference Publishing Services (CPS), Piscataway, New Jersey. pp. 168–172.

Pyo, S., J. Lee, M. Cha, and H. Jang. 2017. Predictability of machine learning techniques to forecast the trends of market index prices: Hypothesis testing for the Korean stock markets. *PloS One* 12(11):e0188107.

Roelofs, R., S. Fridovich-Keil, J. Miller, V. Shankar, M. Hardt, B. Recht, and L. Schmidt. 2019. A meta-analysis of overfitting in machine learning. *In:* Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS), December 8–14, 2019, Vancouver, Canada; Curran Associates, Inc., Red Hook, New York. pp. 9179–9189.

Sahu, K. K., and R. R. Kumar, 2020. Current perspective on pandemic of COVID-19 in the United States. *J. Family Med. Primary Care* 9(4):1784.

Samadi, M., E. Jabbari, and H. M. Azamathulla. 2014. Assessment of M5′ model tree and classification and regression trees for prediction of scour depth below free overfall spillways. *Neural Comput. Appl*. 24(2):357–366.

Schmidt, J., M. R. G. Marques, S. Botti, and M. A. L. Marques. 2019. Recent advances and applications of machine learning in solid-state materials science. *NPJ Comput. Mater*. 5(1):1–36.

Selvin, S., R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon, and K. P. Soman. 2017. Stock price prediction using LSTM, RNN and CNN-sliding window model. *In:* 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), September 13–16, 2017, Udupi, India; IEEE Computer Society's Conference Publishing Services (CPS), Piscataway, New Jersey. pp. 1643–1647.

Shmulsky, R., F. J. N. França, J. T. Ratcliff, B. Farber, C. A. Senalik, R. J. Ross, and R. D. Seale. 2021. Compression properties of small clear southern yellow pine specimens tested across five decades. *Forest Prod. J*. 71 (3):240–245.

Shobana, G. and K. Umamaheswari. 2021. Forecasting by machine learning techniques and econometrics: A review. *In:* 2021 6th International Conference on Inventive Computation Technologies (ICICT), April 2–4, 2021, Pune, India; IEEE Computer Society's Conference Publishing Services (CPS), Piscataway, New Jersey. pp. 1010–1016.

Shrestha, A. and A. Mahmood. 2019. Review of deep learning algorithms and architectures. *IEEE Access* 7:53040–53065.

Singh, S. N. and A. Mohapatra. 2021. Data driven day-ahead electrical load forecasting through repeated wavelet transform assisted SVM model. *Appl. Soft Computing* 107730.

Song, N. 2006. Structural and forecasting softwood lumber models with a time series approach. Louisiana State University and Agricultural & Mechanical College.

Su, H., E. Zio, J. Zhang, M. Xu, X. Li, and Z. Zhang. 2019a. A hybrid hourly natural gas demand forecasting method based on the integration of wavelet transform and enhanced Deep-RNN model. *Energy* 178:585–597.

Su, M., Z. Zhang, Y. Zhu, D. Zha, and W. Wen. 2019b. Data driven natural gas spot price prediction models using machine learning methods. *Energies* 12(9):1680.

Tai, A. M. Y., A. Albuquerque, N. E. Carmona, M. Subramanieapillai, D. S. Cha, M. Sheko, Y. Lee, R. Mansur, and R. S. McIntyre. 2019. Machine learning and big data: Implications for disease modeling and therapeutic discovery in psychiatry. *Artif. Intelligence Med*. 99:101704.

Tao, H., A. O. Al-Sulttani, Salih Ameen, A. M. Z. H. Ali, N. Al-Ansari, S. Q. Salih, and R. R. Mostafa. 2020. Training and testing data division influence on hybrid machine learning model process: Application of river flow forecasting. *Complexity* 2020:8844367.

Verly Lopes, D. J., G. d. S. Bobadilha, and A. Peres Vieira Bedette. 2021. Analysis of lumber prices time series using long short-term memory artificial neural networks. *Forests* 12(4):428.

Wang, W., C. Men, and W. Lu. 2008. Online prediction model based on support vector machine. *Neurocomputing* 71(4–6):550–558.

Wuest, T., D. Weimer, C. Irgens, and K.-D. Thoben. 2016. Machine learning in manufacturing: advantages, challenges, and applications. *Prod. Manufacturing Res*. 4(1):23–45.

Xie, D. and S. Zhang. 2021. Machine learning model for sales forecasting by using XGBoost. *In:* 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), January 15–17, 2021, Guangzhou, China; IEEE Computer Society's Conference Publishing Services (CPS), Piscataway, New Jersey. pp. 480–483.

Yoon, J. 2021. Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. *Comput. Econ*. 57(1):247–265.

Zhao, X., B. Yu, Y. Liu, Z. Chen, Q. Li, C. Wang, and J. Wu. 2019. Estimation of poverty using random forest regression with multi-source data: A case study in Bangladesh. *Remote Sensing* 11(4):375.

Zhou, X., X. Zhu, Z. Dong, and W. Guo. 2016. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *Crop J*. 4(3):212–219.