energies

MDPI

Article

Forecasting Brazilian Ethanol Spot Prices Using LSTM

Gustavo Carvalho Santos [1,*,†], Flavio Barboza [2,*,†], Antônio Cláudio Paschoarelli Veiga [1,*,†] and Mateus Ferreira Silva [3,*,†]

[1] Electrical Engineering School, Federal University of Uberlândia, Uberlândia 38408-100, Brazil
[2] School of Business and Management, Federal University of Uberlândia, Uberlândia 38408-100, Brazil
[3] School of Accounting, Federal University of Uberlândia, Uberlândia 38408-100, Brazil
* Correspondence: gustavocavsantos@gmail.com (G.C.S.); flmbarboza@ufu.br (F.B.); acpveiga@ufu.br (A.C.P.V.); mateusferreira2@ufu.br (M.F.S.); Tel.: +55-34-3230-9472 (F.B.)
† These authors contributed equally to this work.

Abstract: Ethanol is one of the most used fuels in Brazil, which is the second-largest producer of this biofuel in the world. The uncertainty of price direction in the future increases the risk for agents operating in this market and can affect a dependent price chain, such as food and gasoline. This paper uses the architecture of recurrent neural networks—Long short-term memory (LSTM)—to predict Brazilian ethanol spot prices for three horizon-times (12, 6 and 3 months ahead). The proposed model is compared to three benchmark algorithms: Random Forest, SVM Linear and RBF. We evaluate statistical measures such as MSE (Mean Squared Error), MAPE (Mean Absolute Percentage Error), and accuracy to assess the algorithm robustness. Our findings suggest LSTM outperforms the other techniques in regression, considering both MSE and MAPE but SVM Linear is better to identify price trends. Concerning predictions per se, all errors increase during the pandemic period, reinforcing the challenge to identify patterns in crisis scenarios.

Keywords: price prediction; trend prediction; LSTM; SVM; Random Forest; MAPE; MSE; commodity price

Citation: Santos, G.C.; Barboza, F.; Veiga, A.C.P.; Silva, M.F. Forecasting Brazilian Ethanol Spot Prices Using LSTM. *Energies* **2021**, *14*, 7987. https://doi.org/10.3390/en14237987

Academic Editor: Ana-Belén Gil-González

Received: 27 October 2021
Accepted: 22 November 2021
Published: 30 November 2021

1. Introduction

Ethanol has become an interesting alternative to fossil fuels in the world. In Brazil, this biofuel is widely used, and Brazilian production is the second largest in the world, behind only the United States. The importance of this biofuel in Brazil is because the country has succeeded in replacing oil with ethanol in 20% of the automotive fuel and thus 80% of Brazilian cars can carry several mixtures of gasoline and ethanol. This substitution took place due to the fact the country was severely affected by the 1973 oil crisis, in which the local government invested in an ambitious program "Proalcool" for the production of ethanol as a substitute for gas [1].

The main form used in the global industry for the production of ethanol is alcoholic fermentation through a microbiological process [2]. In Brazil, the main raw material for this process is sugarcane, where sugars are transformed into ethanol, energy, cell biomass, CO_2 and other secondary products by yeast cells [3].

In terms of economic impact, the international market of ethanol produced in Brazil is expressive and widespread. According to government data [4], the local producers export to more than 60 countries and imports to almost 20, and the country is among the top 3 exporters of this commodity [5]. Moreover, the Energy Information Administration (EIA, [6]) reports the importance of the Brazil-US relationship, as well as highlights political aspects that influence transactions in these countries.

Some studies revealed ethanol prices can affect many products, mainly food and gasoline [7]. Besides, it is not difficult to find researchers that find contrary outlines. For instance, David et al. [8] evaluates the price transmission via cointegration among ethanol

and other commodities, especially coffee and Carpio [5] noticed the oil prices have affected ethanol prices when analysing more than 20 years. In addition, the volatility of spot prices encourages agents for hedging against market risk in the Brazilian and American markets [9].

The ethanol price is peculiar as long as it is not standardized for a global trade like oil, corn or coffee [5]. Although the spot price of ethanol in Brazil is given by the market, there are timid financial devices on the local stock exchange, which leads many producers and consumers to assume market risk. However, it has growth potential as renewable energy but also due to efficiency improvements in its production [10] and enormous current and future world necessity of energy [11]. Thus, as Brazil is one of the great markets [5,12], the behavior of local price is an element to be observed around the world.

Another issue involved here is the complexity of this task. Predicting price is one of the greatest challenges in the financial environment. There are obvious reasons (the future and dynamic nature) and methodologies employed for that. On the one hand, Statistical modelling needs to simplify the phenomena, when attempting to see linear structures. On the other hand, newer techniques arise from artificial intelligence development and demonstrated interesting outcomes [11,13–15] when outperformed statistical ones in many cases, notably for complex backgrounds.

Based on that, our study intends to use artificial neural networks with LSTM architecture to forecast the spot prices of the Brazilian market for this fuel based on sugar cane. To the best of our knowledge, this is the first study dedicated to forecast ethanol price. The results obtained in this paper demonstrate that it is possible to forecast ethanol prices in Brazilian sight with a degree of correctness of direction between 68 and 80% for the periods of 63, 125 and 252 working days (which is equivalent to 3, 6 and 12 months). Also, for validation of the algorithm, 3 models are compared to LSTM: Random Forest (RF), and Support Vector Machine (SVM)—with Linear and RBF kernels, since these techniques have shown satisfactory performance in the financial market contexts [13–15]. The model revealed interesting prediction power of the Brazilian biofuel price with a small error in periods of low volatility but poor performance occurred during the pandemic caused by COVID-19 due to the sharp drop of the commodity prices in this period.

This study contributes to the literature while it is the first study that examines machine learning models for forecasting Brazilian Ethanol, especially LSTM networks. The implemented algorithms can help practitioners to improve their performances, as well as enable the application of advanced strategies for hedging portfolios, as well as speculating ones.

The paper is organised in the following parts: In Section 2, we review related studies. Section 3 describes the proposed LSTM model and our methodology. The empirical results are presented in Section 4. The last part of the paper (Section 5) presents the concluding remarks and some recommendations.

2. Related Work

Several studies have been dedicated to applying statistical techniques in order to understand the behaviour of ethanol prices and establish dependency relationships in different markets. David et al. [9] used several tools such as Autoregressive Integrated Moving Average, Autoregressive Fractional Integrated Moving Average, Detrended Fluctuation Analyzes and Hurst and Lyapunov exponents to investigate the mechanism of ethanol prices in Brazil in the period from 2010 to 2015. According to the author, results demonstrate that the price of biofuel is antipersistent.

Bouri et al. [16] stated that the generalized autoregressive conditional heteroscedasticity (GARCH) models can incorporate structural breaks and improve the prediction of the volatility of the ethanol market in the United States. They also noted that the influence of good and bad news is properly assessed under such breaks.

It is also possible to find in the literature numerous papers that study the relationship between ethanol prices with other commodities: Carpio [5] relates the long-term and

short-term effects of oil prices on ethanol, gasoline, and sugar price predictions. The author concludes that ethanol is sensitive to short- and long-term changes in the oil. David et al. [8] state that in general, ethanol has a lower predictability horizon than other commodities. Pokrivčák and Rajčaniová [17] also find a relationship in oil and ethanol prices. Bastianin et al. [18] suggest evidence that ethanol can be predicted by returns on corn.

However, studies involving artificial intelligence and the forecasting of ethanol prices are still scarce, despite the large number of works related to machine learning applied to commodity time series. In particular, Bildirici et al. [19] tested a hybrid model (GARCH + LSTM) to analyze the volatility of oil prices, including the effects of the COVID-19 pandemic. Their findings bring to light the contribution of LSTM, especially because of the complexity usually prevails in such data.

Dealing regression algorithms, Ding and Zhang [20] examined the effects of oil, copper, gold, corn, and cattle among them in terms of correlations. More specifically, the authors applied the cointegration method and found a link between oil and copper, and pieces of evidences connected with governments' impact in the other commodities markets.

Kulkarni and Haidar [21] developed an ANN model-based to forecast crude oil price trends. One interesting comment in this paper emphasizes the problematic use of econometric models can deliver "misleading outputs" due to robust assumptions required to them. In terms of results, they reached an impressive rate of 78% for predictions of oil price one day ahead.

In another use of neural networks applied to commodity prediction task, Alameer et al. [14] adopted an LSTM architecture to forecast coal price movements in Australia. Based on a large dataset (about 30 years) with monthly observations, the main findings are: LSTM is better than SVM and MLP when comparing RMSEs; and, there are correlations with other commodities, such as oil, natural gas, copper, gold, silver and iron. Still using LSTM, Liu et al. [22] combined the variational mode decomposition method and LSTM to construct a forecasting model for non-ferrous metals prices. They achieved remarkable performance close to 95% of correctly price trends for Zinc, Copper and Aluminum by working with the 30th last prices to predict the next day as inputs.

Other studies have brought relevant progress to the literature in this field. For example, Herrera et al. [11] compares neural networks and autoregressive integrated moving average (ARIMA) in forecasting Cattle and Wheat prices. Hu et al. [23] implemented a hybrid deep learning approach by integrating LSTM networks with the generalized autoregressive conditional heteroskedasticity (GARCH) model for copper price volatility prediction. Zhou et al. [24] uses a hybrid classification framework to forecast the price trend of bulk commodities over upcoming days, results show an f-score of up to 82%; Ouyang et al. [25] uses long- and short-term time series network for agricultural commodity futures prices prediction.

The papers cited demonstrate several techniques used for analyzing and forecasting commodities, in addition to studying the correlations of different assets with each other and their effects on the world and local economy. Thus, observing the papers developed on the topic of commodity price prediction using artificial intelligence, it is possible to verify a predominance of neural network algorithms, especially the implementation of the LSTM architecture [14,19,22,23].

3. Methodology

3.1. Data

The Center for Advanced Studies on Applied Economic (CEPEA) is an economic research department at Luiz de Queiroz School of Agriculture (ESALQ) from the University of São Paulo (USP) that gathers and provides data from economic, financial, social and environmental aspects of about 30 agribusiness supply chains [26]. The time series analyzed in this research holds daily prices of hydrous ethanol, collected from CEPEA/ESALQ/USP database, which covers the period from 25 January 2010 to 11 December 2020. This time interval includes all data available for ethanol prices up to the conclusion of this research.

We chronologically separated the data in the proportion of 80% for training the neural network and 20% to validate the model. Figure 1 illustrates the prices for the period specified above.

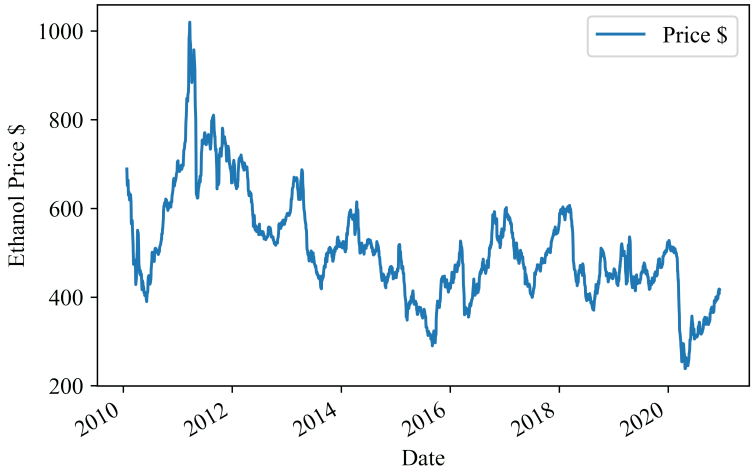

Figure 1. Brazilian Ethanol Spot Prices in U$ Dollars over the period between 2010 and 2020.

Data Pre-Processing

The inputs used in the proposed model vary according to the forecast horizon used. We use a rolling window containing the last 5× days for each model. For example, let's assume we want to predict the price d days ahead. Then, we use data (Close Price of the ethanol in the day d) from d, $2d$, $3d$, $4d$, and $5d$ days before as inputs to the LSTM. In this paper, we apply 3 horizons in business days which are close to 3 and 6 months and 1 year of a calendar time. Table 1 shows the rolling windows used as inputs. These horizons are based on the required time for producing sugar cane, the main input of ethanol. One year covers the whole production [27] and shortest ones get partial perspective and can give the best point to hedge for anyone (buyers and sellers)[28].

Table 1. Steps used as features. C_t means Close Price at time t.

Forecast Horizon (Days)	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6
63	C_t	C_{t-63}	C_{t-126}	C_{t-189}	C_{t-252}	C_{t-315}
126	C_t	C_{t-126}	C_{t-252}	C_{t-378}	C_{t-504}	C_{t-630}
252	C_t	C_{t-252}	C_{t-504}	C_{t-756}	C_{t-1008}	C_{t-1260}

In order to increase the efficiency of the predictors [23], the features were normalized through a StandardScaler algorithm provided by the Scikit-learn package. Basically, this scaler transforms the data into a normal distribution. It's important to note that the parameters of the distribution are given by the training sample and reused for transforming the test sample.

The price Y to be forecast on the Nth day is determined by looking at d days ahead to the current price C. Equation (1) shows this process:

$$Y_N = C_{N+d}. \quad (1)$$

3.2. LSTM Networks

A neural network is a data processing system that is based on the structure of brain neurons. Thus, it consists of a large number of simple processing and highly interconnected elements in an architecture [29]. There are several types of network architectures, this paper uses a model with dense and LSTM layers. This last type of architecture is widely used for

learning sequences (time series, word processing and others) and is very sensitive when choosing hyperparameters [30]. According to Breuel [31] the performance of the LSTM slightly depends on the learning rate and the choice of non-linear recurrent activation functions (tanh and sigmoid) make the network perform better. Based on that, we chose sigmoid as the activation function.

Proposed by Hochreiter and Schmidhuber [32] as a solution for vanishing gradient problem and improved by Gers et al. [33] by introducing a forget gate into the cell, LSTM is a type of recurrent neural networks architecture. As in Yu et al. [34] based on Figure 2, the LSTM cell can be mathematically described as:

$$f_t = \sigma(W_f h_{t-1} + W_f x_t + b_f) \quad (2)$$

$$i_t = \sigma(W_i h_{t-1} + W_i x_t + b_i) \quad (3)$$

$$\tilde{c}_t = \sigma(W_c h_{t-1} + W_c x_t + b_c) \quad (4)$$

$$c_t = f_t \,.\, c_{t-1} + i_t \,.\, \tilde{c}_t \quad (5)$$

$$o_t = \sigma(W_o h_{t-1} + W_o x_t + b_o) \quad (6)$$

$$h_t = o_t \,.\, tanh(c_t) \quad (7)$$

where f_t represents the forget gate, which allows the LSTM to reset its state [35]. When the f_t value is 1, it keeps that information, while the value 0 means that it deletes all that information. The input, the recurrent information, and the output of the cell at time t is portrayed by x_t, h_t and y_t respectively. The biases represented by b, i_t and o_t represent the input and output gates at time t. The cell state is symbolized by c_t and W_i, W_c, W_o and W_f are the weights. The operator '$\cdot$' expresses the pointwise multiplication of two vectors.

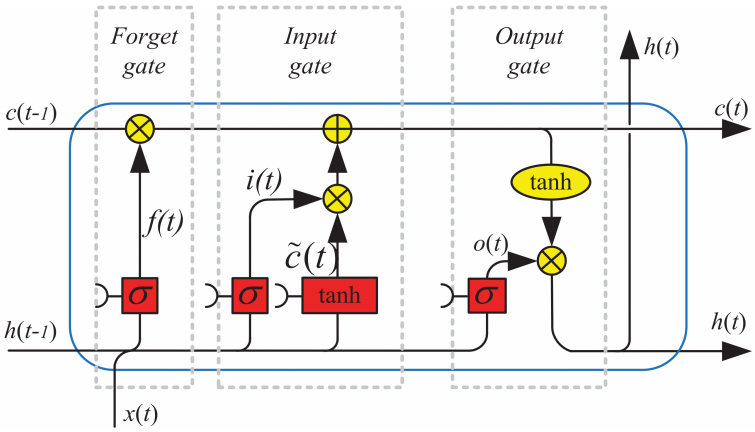

Figure 2. Architecture of LSTM with a forget gate. Reproduced from [34].

LSTM networks have a wide range of applications. It is possible to find in the literature several works in different areas that use this type of recurrent network to build machine learning models. Due to the ability to learn data sequences, numerous papers use this architecture for language processing and text classification [36–41], financial predictions [23,42–45], and other problems involving time series [19,46–50].

3.3. Proposed Model and Benchmarks

The model implemented in this work run in Python 3.8.6 (64-bit) with Jupyter Notebook. The hardware setup includes an Intel Core i5-4310u CPU 2.0GHz, 8GB RAM. The neural network is built using Tensorflow with Keras version 2.3.1 as interface.

The model's architecture includes four hidden layers with 64 units interspersed by a dropout of 20%, the first three are of the LSTM type and the last is dense. The purpose of the dropout layers is to randomly drop units from the neural network during training

to avoid overfitting [51]. This architecture is also known as *vanilla LSTM* and has been applied in similar contexts [35,52].

Figure 3 illustrates an example of the implemented neural network architecture:

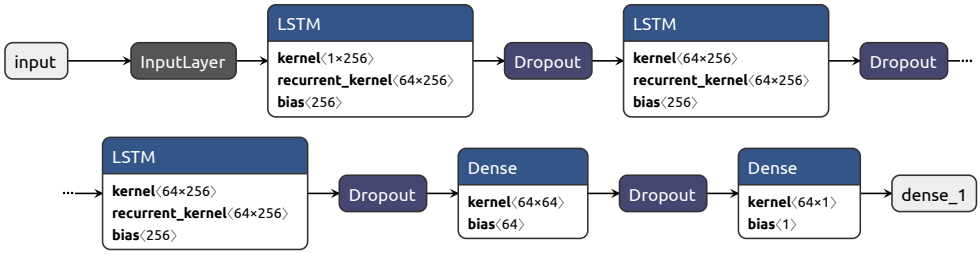

Figure 3. Neural network architecture.

In attempt to compare performances, we evaluate other 3 models: Random Forest (RF), Support Vector Machine (SVM) with two kernels: Linear (SVML) and Radial Basis Function (SVMR), which are considered as suggested techniques for this kind of problem [14,15]. Both are considered machine learning techniques [14,53]. RF is based on a collection of decision trees, in which classifies each instance by majority vote while SVM builds a hyperplane in attempt to optimize the division between classes.

We run all models in Python either and supported to scikit-learn libraries. In particular, we used parameters cost C = 1 and $\epsilon = 0.2$ to SVM. According to Carrasco et al. [54], the parameter C is responsible for the regularization, focusing in to avoid large coefficients and then contributing to lower misclassification rates, and epsilon is the width of the region (also called tube) centered in the hyperplane. This procedure tends to prevent overfitting. Other parameters remained as default settings. In the case of RF, all features keep as standard.

To check the error of the predictions made, the root mean squared error (RMSE) and the mean absolute percentage error (MAPE) were used. They can be defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y_t - \hat{y}_t)^2} \tag{8}$$

$$\text{MAPE} = \frac{1}{n}\sum\frac{|y_t - \hat{y}_t|}{y_t} \tag{9}$$

where $\hat{y}_t$ is the forecast price and y_t the actual value, both in the time t, and n represents the number of forecast observations in the sample.

Another way to evaluate the predictions is by observing the ability of the models to adjust the change in price direction. We used the accuracy, precision and recall measures to do it. As in Wang et al. [55], these measures can be defined as:

$$\text{precision} = \frac{\text{correct predictions as } x}{\text{total predictions as } x} \tag{10}$$

$$\text{recall} = \frac{\text{correct predictions as } x}{\text{number of actual } x} \tag{11}$$

$$\text{accuracy} = \frac{\text{correct predictions}}{\text{total of predictions}} \tag{12}$$

where x can be the Upward or Downward trend.

4. Results and Discussions

First, we evaluate the model performance with the concern of detecting any bias in there. Second, we present the outputs and discuss them. Lastly, the visual analysis complements our investigation.

4.1. Learning Curves

Learning curves are a way to assess the ability of a deep learning model to generalize the realized information in the training phase. The curves of training and validation errors are plotted in Figures 4–6, representing short, medium and long-term horizons, respectively. In these cases, the curves can be observed through the number of epochs, which allows detecting possible overfitting of the model. If the curves of the training and validation errors decay together in a uniform trail, this issue can be discarded [56].

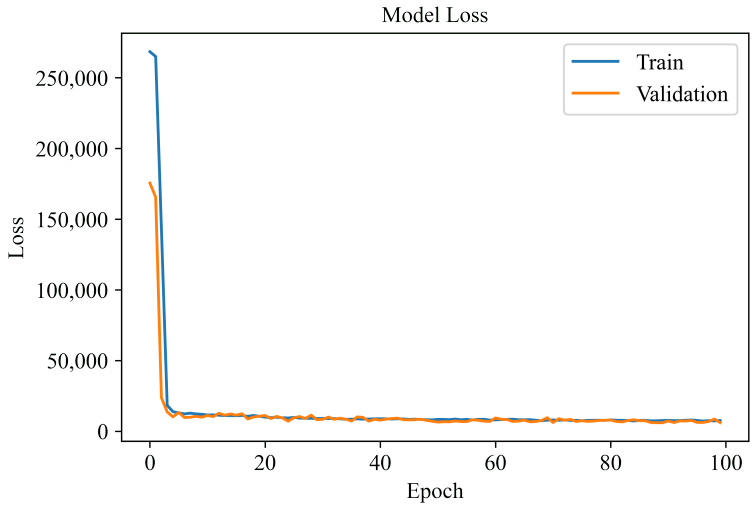

Figure 4. Loss for 63 days model.

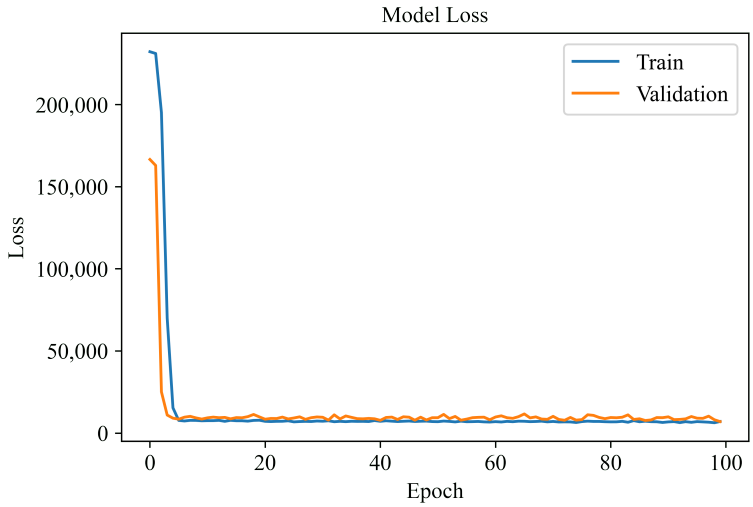

Figure 5. Loss for 126 days model.

Visually, the 63-day model obtained the best learning curves, and it is possible to observe that both error curves converged to a small value. In other words, this model is robust and sufficiently reliable. The same convergence can be viewed for the other models (126 and 252 days). However, it can be affirmed that the higher the forecast interval, the slower the error curves convergence.

The convergence of the learning curves of the three models presented coincide with their respective results in the validation: The 63-day model had the best convergence and also the smallest errors (MAPE and RMSE) in the predictions. On the other hand, the 252-days model had the lowest convergence and consequently the highest errors. These results can be verified in the next subsection.

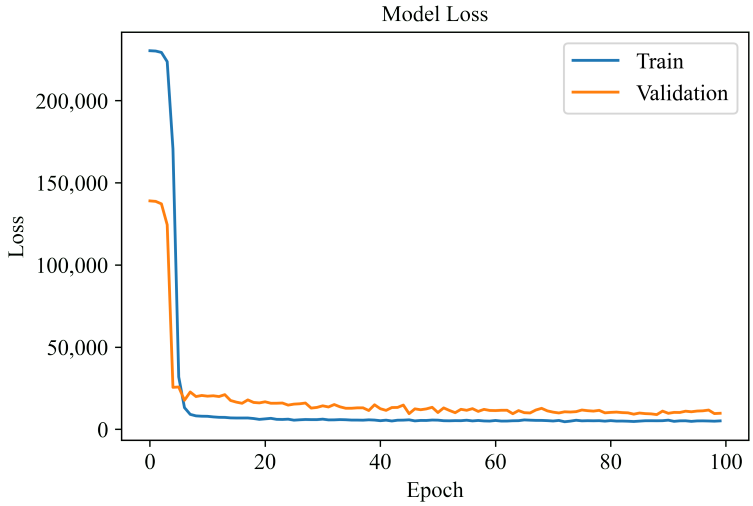

Figure 6. Loss for 252 days model.

4.2. Forecasting Results

We trained and validated the models of Long-Short Term Memory (LSTM), Random Forest (RF), Support Vector Machine with linear (SVML) and Radial Basis Function (SVMR) kernels, by using a chronological split of the full sample. Table 2 illustrates the error measures (MAPE and RMSE) obtained in the validation process for each forecast horizon.

Table 2. Error rates in each model for validation sample.

	63 Days Ahead		126 Days Ahead		252 Days Ahead	
Model	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**	**RMSE**
LSTM	17.23	78.53	19.91	83.38	26.15	98.69
RF	21.49	94.48	22.28	95.78	32.12	127.26
SVML	17.24	86.12	22.58	97.55	26.58	98.58
SVMR	20.65	92.62	23.32	98.00	33.72	120.22

It can be stated that the LSTM achieved lower errors (MAPE and RMSE) for all forecast horizons, except for the RMSE of the 252-day model. Similar to Alameer et al. [14], we can see the higher the average error increases the higher horizon-time. However, the results of our work show that the accuracy levels improve by extending the horizon of prediction (in the case of LSTM and SVML). These values show that the result is unwanted, but it is worth noting that such measures are commonly indicated for regression outlines. For example, an error rate of about 17% (on average, considering predictions in 63 days ahead) can be considered poorly contributory information and needs other measures to better understand the quality of the model. To do so, Table 3 provides details in terms of precision, recall and accuracy (hit ratios) to 63, 126, and 252 days, respectively, in this analysis.

For 63-day forecasts, Table 3 shows that accuracy of 72% has been achieved, which can be considered a good result, and comparable to Zhou et al. [24]. On the one hand, the Upward movements obtained a better accuracy (90%), that is, of all predictions that pointed high, 9 out of 10 were correct. On the other hand, the recall of the Downward movements was higher (88%). In other words, out of every 100 occurrences of a fall in the asset value (actual values), the model presented 88 correct predictions. This result is can be considered good, but it is possible to improve. Hence, the optimized cutoff values can be found if any.

Table 4 shows that the LSTM has greater accuracy when sorting Uptrends for 126-horizon forecasts. This can be interpreted as bullish movements being clearer and less noisy signals. In addition, the total accuracy of the model was higher than the previous ones. On the other hand, for the 252-day forecasts, Table 5, although the total accuracy was

80%, the forecaster demonstrated to have low accuracy when classifying uptrend ethanol prices. This information casts doubt on the generality of LSTM to perform better in longer periods. Nevertheless, the only unbalanced sample was the latter (only 37 uptrend events against 200 downtrends). When looking specifically at this medium/long-term forecast, it is necessary to consider the analysis period, which includes moments not experienced by the training sample, like the pandemic caused by COVID-19. This effect probably justifies its poor performance in understanding the upwards and collaborates with the results of Bildirici et al. [19] when analyzing the impact of the pandemic on the prices of a commodity.

Table 3. Hit ratios of predictions to 63-days horizon.

Model–Trend	Precision	Recall	Support	Accuracy
LSTM–Downtrend	0.59	0.88	173	72%
LSTM–Uptrend	0.90	0.63	291	
RF–Downtrend	0.80	0.65	173	81%
RF–Uptrend	0.81	0.90	291	
SVML–Downtrend	0.68	0.60	173	74%
SVML–Uptrend	0.78	0.86	291	
SVMR–Downtrend	0.74	0.84	173	83%
SVMR–Uptrend	0.90	0.82	291	

Table 4. Hit ratios of predictions to 126-days horizon.

Model–Trend	Precision	Recall	Support	Accuracy
LSTM–Downtrend	0.70	0.93	201	74%
LSTM–Uptrend	0.86	0.52	187	
RF–Downtrend	0.95	0.71	201	83%
RF–Uptrend	0.76	0.96	187	
SVML–Downtrend	0.84	0.78	201	81%
SVML–Uptrend	0.78	0.83	187	
SVMR–Downtrend	0.83	0.84	201	83%
SVMR–Uptrend	0.83	0.81	187	

Table 5. Hit ratios of predictions to 252-days horizon.

Model–Trend	Precision	Recall	Support	Accuracy
LSTM–Downtrend	0.88	0.88	200	80%
LSTM–Uptrend	0.35	0.35	37	
RF–Downtrend	0.87	0.40	200	44%
RF–Uptrend	0.17	0.68	37	
SVML–Downtrend	0.94	0.89	200	86%
SVML–Uptrend	0.53	0.70	37	
SVMR–Downtrend	0.97	0.50	200	57%
SVMR–Uptrend	0.25	0.92	37	

In general, these results show to be different from the work of Kulkarni and Haidar [21], in which the accuracy of the forecasts decays with the enlargement of the forecast horizon. However, it is important to highlight that in the cited article the forecast periods are only 1,

2 and 3 days. On the contrary, Bouri et al. [16] remind that the volatility (risk) perceived in periods is lower, which has greater meaning and more corresponding to this study.

Comparing LSTM outlines with benchmarks, its error rates are better, except for RMSE in the longer horizon (SVML is slightly better). Based on that, we can consider LSTM as the best predictor on a relative and regression basis. If we focus the analysis in the classification report, the findings are essentially unlike. LSTM presented just one single better indicator (recall in 63-day predictions, only bearing trends). Surprisingly, SVM achieved interesting performance in all horizons. For 63-day predictions, SVMR was the most accurate with the higher precision for uptrends. Curiously, the accuracy persisted to the mid-term horizon but decrease in the long-term while SVML was the best and reported two higher ratios (precision for uptrends and recall for downtrends). The RF precision and recall deserve to be emphasized. Except in the 252-day horizon, RF presented competitive ratios and more accurate than LSTM.

4.3. Visualising the Predictions

In the work of Bildirici et al. [19], it is possible to verify the impact of the COVID-19 pandemic on oil prices, in which the values suffered a great fall, affecting the entire international market. We can say the same for ethanol, as the big drop in prices also occurred. In addition, the author's analysis that prices will return to their highest values—potentially causing inflation—may also be valid for ethanol.

Another important point is the impact of the pandemic on forecast errors. As it is an adverse and unpredictable event, the error rates (MAPE and RMSE) were greater in this period compared to moments prior to the COVID-19 crisis.

Figures 7–9 illustrate the predictions of the algorithms versus the price verified in the analyzed period.

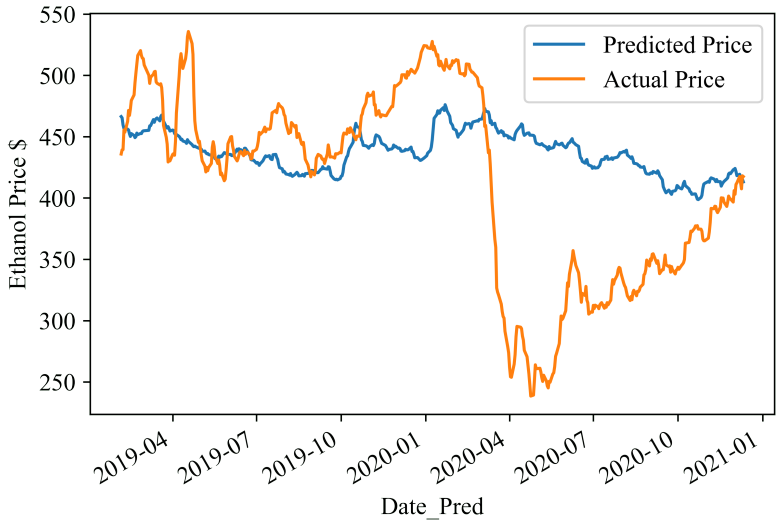

Figure 7. Predictions for 63 days ahead.

Furthermore, we can see that 63-days forecasts are relatively close to the actual price before the most intense period of the pandemic (more specifically between March and September of 2020) and, after that, recover similar performance. Thus, regarding regression outlines, there are evidences that incremental volatility could negatively interfere in the predictions. Besides, it is hard to conclude for the price directions, as we found accuracy higher than 80% for all the best predictors. The precision achieves values over 70%, except for uptrend in 252-days, in which the test set is heavily unbalanced (only 37 upwards against 200 downwards. Since the interval of time for predicting is relatively short (less than a year), these findings would be useful and sufficiently interesting for people who need or want to trade ethanol, whatever is the interest: hedge or speculation positions.

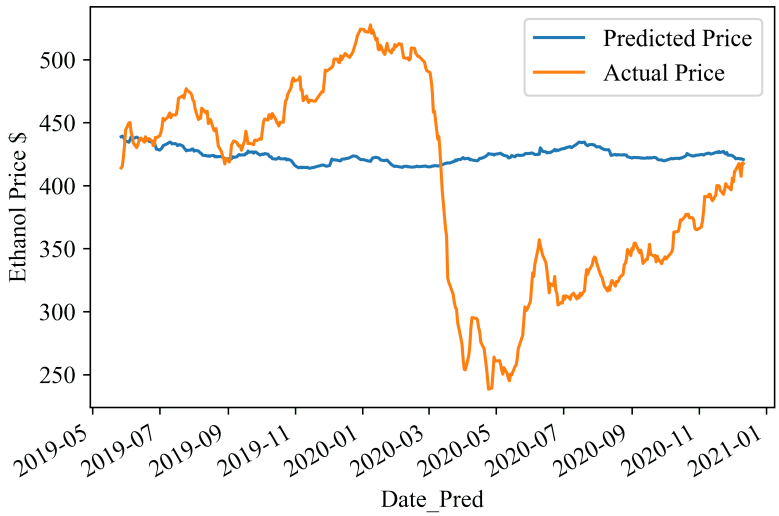

Figure 8. Predictions for 126 days ahead.

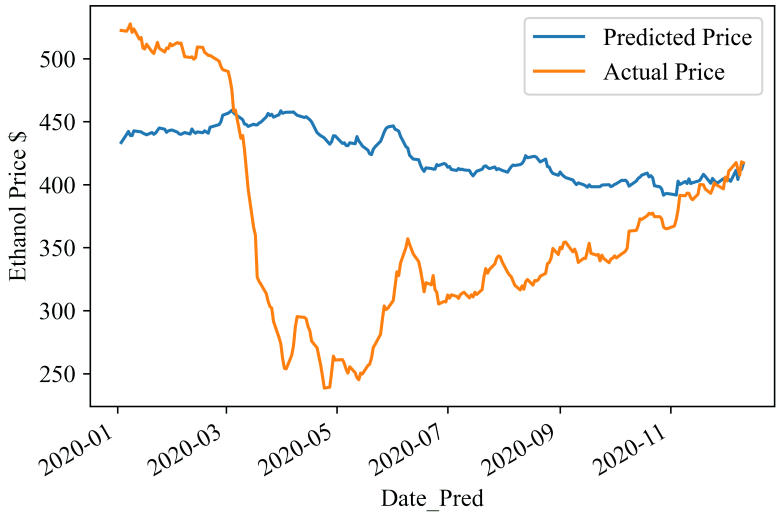

Figure 9. Predictions for 252 days ahead.

5. Conclusions

This paper presents a Brazilian spot ethanol price prediction model using artificial neural networks with the LSTM architecture and compared it to SVM and RF forecasts. The algorithms provide outlines for periods of 63, 126 and 252 business days. The results evaluated in this work show that it was possible to predict prices with a reasonable degree of accuracy in market directions for all horizons used.

Tests to verify overfitting were performed using learning curves, and the models converged in a satisfactory way, demonstrating a good fit of the neural network. Benchmark results show that LSTM produced the smallest regression errors (MAPE and RMSE). However, regarding the correctness of the direction in the predictions, other algorithms had better accuracy for specific horizons.

SVML proved to be the best algorithm for detecting trends achieving good results for all forecast windows used. Still, LSTM also managed to achieve satisfactory results for all forecasts, unlike RF and SVMR which had poor results for the 252-day horizon.

It was possible to observe in the LSTM outcomes an increase in the accuracy of the algorithms in longer forecast horizons, 72%, 74%, and 80% for 63, 126, and 252-day horizons, respectively. However, the mean absolute percentage error (MAPE) of the forecasts increased: 17.2%, 19.9% and 26.1% for 3, 6 and 12 months respectively. The same

was found in the RMSE outputs. Furthermore, it is important to note that the COVID-19 pandemic caused an unexpected drop in prices, increasing model errors.

The high degree of correctness of models in the direction of prices can be useful in the development of new hedging strategies for market agents. In addition, it can help producers and cooperatives to protect their capital through planning that takes into account these forecasts.

This work contributes by demonstrating that LSTM networks are able to perform efficiently when predicting ethanol prices, a biofuel widely used in Brazil and worldwide, which has the capacity to replace fossil energy sources.

Nevertheless, this paper has limitations: (i) despite of the satisfactory results, we built models based on pure techniques, (ii) our research takes into account one single commodity with prices traded in one country but these data can be exclusively found in Brazil (Top 3 producer in the world) and there is no global market with standardized price for ethanol, (iii) potential effects (e.g., macroeconomic indicators) are not considered. However, the proposed (and best) model requires only historical values, and (iv) comparison with previous results was impracticable since data, performance measures and horizons did not exist in the literature, what shows the pioneering of this study.

For future work, we can add hybrid models that mix different network architectures and machine learning algorithms, such as Empirical Mode Decomposition. Thus, new features can be implemented, such as endogenous variables (technical indicators) and exogenous variables (exchange rate, inflation and prices of other commodities).

Author Contributions: Conceptualization, G.C.S., F.B. and A.C.P.V.; methodology, G.C.S. and F.B.; software, G.C.S. and M.F.S.; validation, F.B. and A.C.P.V.; formal analysis, A.C.P.V.; investigation, G.C.S.; resources, A.C.P.V.; data curation, M.F.S.; writing—original draft preparation, G.C.S.; writing—review and editing, F.B. and A.C.P.V.; visualization, G.C.S. and M.F.S.; supervision, A.C.P.V.; project administration, F.B.; funding acquisition, G.C.S. and F.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Council for Scientific and Technological Development (CNPq) grant number 438314/2018-2. The APC was funded by Sapiens Agro.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://www.cepea.esalq.usp.br/en/indicator/ethanol.aspx (accessed on 11 November 2021). The predictions generated by the LSTM models can be accessed in the following repository: https://github.com/gustavocavsantos/Ethanol-Price-Predictions (accessed on 11 November 2021).

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

References

1. Goldemberg, J. The ethanol program in Brazil. *Environ. Res. Lett.* **2006**, *1*, 014008. [CrossRef]
2. Silva, G.P.D.; Araújo, E.F.D.; Silva, D.O.; Guimarães, W.V. Ethanolic fermentation of sucrose, sugarcane juice and molasses by Escherichia coli strain KO11 and Klebsiella oxytoca strain P2. *Braz. J. Microbiol.* **2005**, *36*, 395–404. [CrossRef]
3. Lopes, M.L.; de Lima Paulillo, S.C.; Godoy, A.; Cherubin, R.A.; Lorenzi, M.S.; Giometti, F.H.C.; Bernardino, C.D.; de Amorim Neto, H.B.; de Amorim, H.V. Ethanol production in Brazil: A bridge between science and industry. *Braz. J. Microbiol.* **2016**, *47*, 64–76. [CrossRef] [PubMed]
4. Ministry of Agriculture, Fisheries and Supply—Ethanol Archives. Available online: https://www.gov.br/agricultura/pt-br/assuntos/sustentabilidade/agroenergia/arquivos-etanol-comercio-exterior-brasileiro/ (accessed on 11 November 2021).
5. Carpio, L.G.T. The effects of oil price volatility on ethanol, gasoline, and sugar price forecasts. *Energy* **2019**, *181*, 1012–1022. [CrossRef]
6. EIA—Today in Energy. Available online: https://www.eia.gov/todayinenergy/detail.php?id=47956 (accessed on 11 October 2021).
7. Hira, A.; de Oliveira, L.G. No substitute for oil? How Brazil developed its ethanol industry. *Energy Policy* **2009**, *37*, 2450–2456. [CrossRef]
8. David, S.A.; Inácio, C.; Tenreiro Machado, J.A. Quantifying the predictability and efficiency of the cointegrated ethanol and agricultural commodities price series. *Appl. Sci.* **2019**, *9*, 5303. [CrossRef]

9. David, S.; Quintino, D.; Inacio, C.; Machado, J. Fractional dynamic behavior in ethanol prices series. *J. Comput. Appl. Math.* **2018**, *339*, 85–93. [CrossRef]
10. Tapia Carpio, L.G.; Simone de Souza, F. Competition between second-generation ethanol and bioelectricity using the residual biomass of sugarcane: Effects of uncertainty on the production mix. *Molecules* **2019**, *24*, 369. [CrossRef] [PubMed]
11. Herrera, G.P.; Constantino, M.; Tabak, B.M.; Pistori, H.; Su, J.J.; Naranpanawa, A. Long-term forecast of energy commodities price using machine learning. *Energy* **2019**, *179*, 214–221. [CrossRef]
12. de Araujo, F.H.A.; Bejan, L.; Rosso, O.A.; Stosic, T. Permutation entropy and statistical complexity analysis of Brazilian agricultural commodities. *Entropy* **2019**, *21*, 1220. [CrossRef]
13. Barboza, F.; Kimura, H.; Altman, E. Machine learning models and bankruptcy prediction. *Expert Syst. Appl.* **2017**, *83*, 405–417. [CrossRef]
14. Alameer, Z.; Fathalla, A.; Li, K.; Ye, H.; Jianhua, Z. Multistep-ahead forecasting of coal prices using a hybrid deep learning model. *Resour. Policy* **2020**, *65*, 101588. [CrossRef]
15. Sun, W.; Zhang, J. Carbon Price Prediction Based on Ensemble Empirical Mode Decomposition and Extreme Learning Machine Optimized by Improved Bat Algorithm Considering Energy Price Factors. *Energies* **2020**, *13*, 3471. [CrossRef]
16. Bouri, E.; Dutta, A.; Saeed, T. Forecasting ethanol price volatility under structural breaks. *Biofuels Bioprod. Biorefining* **2021**, *15*, 250–256. [CrossRef]
17. Pokrivčák, J.; Rajčaniová, M. Crude oil price variability and its impact on ethanol prices. *Agric. Econ.* **2011**, *57*, 394–403. [CrossRef]
18. Bastianin, A.; Galeotti, M.; Manera, M. Ethanol and field crops: Is there a price connection? *Food Policy* **2016**, *63*, 53 – 61. [CrossRef]
19. Bildirici, M.; Guler Bayazit, N.; Ucan, Y. Analyzing Crude Oil Prices under the Impact of COVID-19 by Using LSTARGARCHLSTM. *Energies* **2020**, *13*, 2980. [CrossRef]
20. Ding, S.; Zhang, Y. Cross market predictions for commodity prices. *Econ. Model.* **2020**, *91*, 455–462. [CrossRef]
21. Kulkarni, S.; Haidar, I. Forecasting model for crude oil price using artificial neural networks and commodity futures prices. *arXiv* **2009**, arXiv:0906.4838.
22. Liu, Y.; Yang, C.; Huang, K.; Gui, W. Non-ferrous metals price forecasting based on variational mode decomposition and LSTM network. *Knowl.-Based Syst.* **2020**, *188*, 105006. [CrossRef]
23. Hu, Y.; Ni, J.; Wen, L. A hybrid deep learning approach by integrating LSTM-ANN networks with GARCH model for copper price volatility prediction. *Phys. A Stat. Mech. Its Appl.* **2020**, *557*, 124907. [CrossRef]
24. Zhou, B.; Zhao, S.; Chen, L.; Li, S.; Wu, Z.; Pan, G. Forecasting Price Trend of Bulk Commodities Leveraging Cross-Domain Open Data Fusion. *ACM Trans. Intell. Syst. Technol.* **2020**, *11*, 1–26. [CrossRef]
25. Ouyang, H.; Wei, X.; Wu, Q. Agricultural commodity futures prices prediction via long- and short-term time series network. *J. Appl. Econ.* **2019**, *22*, 468–483. [CrossRef]
26. CEPEA—Center for Advanced Studies on Applied Economics. Available online: https://www.cepea.esalq.usp.br/en/cepea-1.aspx (accessed on 13 January 2021).
27. Ariyawansha, T.; Abeyrathna, D.; Kulasekara, B.; Pottawela, D.; Kodithuwakku, D.; Ariyawansha, S.; Sewwandi, N.; Bandara, W.; Ahamed, T.; Noguchi, R. A novel approach to minimize energy requirements and maximize biomass utilization of the sugarcane harvesting system in Sri Lanka. *Energies* **2020**, *13*, 1497. [CrossRef]
28. Franken, J.R.; Parcell, J.L. Cash Ethanol Cross-Hedging Opportunities. *J. Agric. Appl. Econ.* **2003**, *35*, 510–516. [CrossRef]
29. Uhrig, R.E. Introduction to artificial neural networks. In Proceedings of IECON'95-21st Annual Conference on IEEE Industrial Electronics, Orlando, FL, USA, 6–10 November 1995; Volume 1, pp. 33–37.
30. Nakisa, B.; Rastgoo, M.N.; Rakotonirainy, A.; Maire, F.; Chandran, V. Long Short Term Memory Hyperparameter Optimization for a Neural Network Based Emotion Recognition Framework. *IEEE Access* **2018**, *6*, 49325–49338. [CrossRef]
31. Breuel, T.M. Benchmarking of LSTM Networks. *arXiv* **2015**, arXiv:1508.02774.
32. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
33. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to Forget: Continual Prediction with LSTM. *Neural Comput.* **2000**, *12*, 2451–2471. [CrossRef]
34. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef]
35. Greff, K.; Srivastava, R.K.; Koutník, J.; Steunebrink, B.R.; Schmidhuber, J. LSTM: A Search Space Odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *28*, 2222–2232. [CrossRef] [PubMed]
36. Graves, A.; Fernández, S.; Schmidhuber, J. Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. In *Artificial Neural Networks: Formal Models and Their Applications—ICANN 2005*; Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2005; pp. 799–804.
37. Habibi, M.; Weber, L.; Neves, M.; Wiegandt, D.L.; Leser, U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **2017**, *33*, i37–i48. [CrossRef] [PubMed]
38. Zhou, C.; Sun, C.; Liu, Z.; Lau, F. A C-LSTM neural network for text classification. *arXiv* **2015**, arXiv:1511.08630.
39. Zhou, P.; Qi, Z.; Zheng, S.; Xu, J.; Bao, H.; Xu, B. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. *arXiv* **2016**, arXiv:1611.06639.

40. Liu, G.; Guo, J. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing* **2019**, *337*, 325–338. [CrossRef]
41. Sachan, D.S.; Zaheer, M.; Salakhutdinov, R. Revisiting lstm networks for semi-supervised text classification via mixed objective function. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6940–6948.
42. Bao, W.; Yue, J.; Rao, Y. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLoS ONE* **2017**, *12*, e0180944. [CrossRef]
43. Siami-Namini, S.; Tavakoli, N.; Namin, A.S. A comparison of ARIMA and LSTM in forecasting time series. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018; pp. 1394–1401.
44. Moghar, A.; Hamiche, M. Stock market prediction using LSTM recurrent neural network. *Procedia Comput. Sci.* **2020**, *170*, 1168–1173. [CrossRef]
45. Tong, G.; Yin, Z. Adaptive Trading System of Assets for International Cooperation in Agricultural Finance Based on Neural Network. *Comput. Econ.* **2021**, 1–20. [CrossRef]
46. Karim, F.; Majumdar, S.; Darabi, H.; Chen, S. LSTM Fully Convolutional Networks for Time Series Classification. *IEEE Access* **2018**, *6*, 1662–1669. [CrossRef]
47. Mahasseni, B.; Lam, M.; Todorovic, S. Unsupervised video summarization with adversarial lstm networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 202–211.
48. Sønderby, S.K.; Sønderby, C.K.; Nielsen, H.; Winther, O. Convolutional LSTM networks for subcellular localization of proteins. In *International Conference on Algorithms for Computational Biology*; Springer: Cham, Switzerland, 2015; pp. 68–80.
49. Trinh, H.D.; Giupponi, L.; Dini, P. Mobile traffic prediction from raw data using LSTM networks. In Proceedings of the 2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Bologna, Italy, 9–12 September 2018; pp. 1827–1832.
50. Ycart, A.; Benetos, E. A Study on LSTM Networks for Polyphonic Music Sequence Modelling. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 23–28 October 2017.
51. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
52. Gers, F.A.; Schmidhuber, E. LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Trans. Neural Netw.* **2001**, *12*, 1333–1340. [CrossRef]
53. Herrera, G.P.; Constantino, M.; Tabak, B.M.; Pistori, H.; Su, J.J.; Naranpanawa, A. Data on forecasting energy prices using machine learning. *Data Brief* **2019**, *25*, 104122. [CrossRef] [PubMed]
54. Carrasco, M.; López, J.; Maldonado, S. Epsilon-nonparallel support vector regression. *Appl. Intell.* **2019**, *49*, 4223–4236. [CrossRef]
55. Wang, X.; Zhou, T.; Wang, X.; Fang, Y. Harshness-aware sentiment mining framework for product review. *Expert Syst. Appl.* **2022**, *187*, 115887. [CrossRef]
56. Zhang, P.; Yin, Z.Y.; Zheng, Y.; Gao, F.P. A LSTM surrogate modelling approach for caisson foundations. *Ocean Eng.* **2020**, *204*, 107263. [CrossRef]