Deep Learning Based Approach for Fresh Produce Market Price Prediction

Lobna Nassar
Electrical and Computer Engineering Department
University of Waterloo
Ontario, Canada
lnassar@uwaterloo.ca

Ifeanyi Emmanuel Okwuchi
Systems Design Engineering Department
University of Waterloo
Ontario, Canada
ifeanyi.okwuchi@uwaterloo.ca

Muhammad Saad
Electrical and Computer Engineering Department
University of Waterloo
Ontario, Canada
m32saad@uwaterloo.ca

Fakhri Karray
Electrical and Computer Engineering Department
University of Waterloo
Ontario, Canada
karray@uwaterloo.ca

Kumaraswamy Ponnambalam
Systems Design Engineering Department
University of Waterloo
Ontario, Canada
ponnu@uwaterloo.ca

***Abstract*— Building highly precise prediction models for Fresh Produce (FP) market price is crucial to protect retailers from overpriced FP. In this paper we are comparing the price prediction models performance of deep learning (DL) models with statistical as well as standard machine learning (ML) models. Five types of FP are considered in performance testing. It is found that the conventional ML models outperform the statistical models such as ARIMA. On the other hand, the winning model among the conventional ML models (the Gradient Boosting model) proves to be less performant as compared with the simple or compound DL models. Moreover, the simple DL models, such as the Long Short-Term Memory (LSTM), are outperformed by the compound one, the Convolutional Long Short-Term Memory Recurrent Neural Network (CNN-LSTM), whose performance improves by adding attention. The model is capable of precisely predicting FP prices for up to three weeks ahead.**

Keywords—Price Prediction, Deep Learning, Neural Networks, Attention, Fresh Produce.

I. Introduction

Fruits, flowers and vegetables are referred to as Agri-fresh produce, food products, Agri-food, short shelf-life food products, perishables, deteriorating products [1, 2]. Only fruits and vegetables are considered in this paper and are referred to as fresh produce or FP. The fresh produce has high contribution to the total revenues of many countries which can reach nearly 40% of the total revenues as the case in Australia [3]. It is worth mentioning that FP supply chain management (FSCM) is more complex as compared to ordinary SCMs due to the nature of the produce and high fluctuations in its prices [2]. FP ordering depends mostly on the demand for FP since it is expected to meet that demand. Adequately timed orders with fair prices bring financial benefits to the buyers in our case the distribution centers (DC) and at the same time minimize waste. For example, U.S. estimates an annual loss of 5.9 and 6.1 billion pounds for fresh fruit and vegetables, respectively [4]. The estimation of the order quantity and procurement price currently rely on the experience of the procurer who utilizes history of immediate past purchases of the same produce which is a bit simplistic and may not be financially optimal for the buyer. By providing better price prediction using automated AI machine learning (ML) and deep learning (DL) based approaches, error-prone manual activities by DC procurers should become minimal, and people will be supplied with less expensive fresh foods more reliably.

Prediction models can be univariate and multivariate models. The univariate models include the time series forecasting models [5-7] that use the previous prices to predict the future one such as the auto-regressive integrated moving average (ARIMA) model [8]. Time series modeling has also been widely used in finance for forecasting and simulating uncertain variables of interest in stock market. In case of prediction of FP, most of the available prediction models have focused on yield prediction and not on price modeling or forecasting.

Additionally, most of the current tools and models used for price forecasting for perishable commodity markets are underperforming because they do not utilize the existing capacity in the elicitation of complex patterns and rules inherent in big data sources of price transactions and the influential factors to provide valuable information and knowledge on underlying complex processes affecting FP prices; something that has become possible by the advent of state-of-the-art AI and ML and DL techniques.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

The goal of the DC procurer is to predict the *ask price* that the supplier should accept via the bilateral transactions. This predicted price should protect against overpriced agricultural commodities, minimize the *bid ask* spread which should benefit both the distribution centers (procurers) and the end customers. Therefore, the main objective of this work is developing FP price prediction service by designing DL based forecasting tools for improved and efficient procurement offer price for FP via bilateral transactions of the DC procurers and suppliers. Validation, and comparison of the proposed models against more conventional models like time series and ML models is important for performance evaluation.

The Mean Absolute Percentage Error (MAPE) is used to compare the ARIMA statistical model performance against three conventional ML models. It is found that all conventional models outperform the statistical one. The best performing ML model, the gradient boosting (GB), is chosen to be compared against the simple and compound deployed DL models with and without attention [9]. Multiple evaluation models are utilized in performance assessment such as the mean absolute error (MAE), Root Mean Square Error (RMSE), R2 measure [10] along with an aggregated measure that summarizes the results of those measures to decide the best performing model. Based on this aggregated measure, it is found that the compound DL model, ATT-CNN-LSTM, is the best performing in price prediction compared to the ML and simple DL models especially after adding the attention.

The rest of this paper is structured as follows: In Section 2, the details on the datasets sources and data preprocessing are provided. Section 3, includes the two main types of the FP price prediction models including the conventional (Statistical and ML) models as well as the (simple and compound) DL models. In Section 4, details of the applied performance measures are presented. Section 5 has the conducted experiments and discussion of the results. The paper is concluded with Section 6 that highlights the main findings and future work.

II. PRICES DATA SETS

Two datasets are used; the first is a dataset for four types of FP extracted from a website providing daily crop prices for Taiwan markets used in [11]. The second is a dataset for Strawberry daily transactions provided by a confidential DC in Cambridge, Ontario, Canada for their FP daily transactions over the past seven years.

A. *Taiwan Dataset – (Cabbage, Cauliflower, Bok Choi, Watermelon)*

First, the Taiwan dataset is used to test performance of deployed models. The set includes daily prices for four sets of FP: Cabbage, Bok Choi, Watermelon and Cauliflower. The dataset for each of these FP contains daily records in the period from January 4th, 2011 to July 26th, 2015. The sizes of these datasets are 34986, 27896, 17589, 30547 for Cabbage, Bok Choi, Watermelon and Cauliflower respectively (6997, 5579, 3518, 6109 data items are reserved as test set for each while the rest of the records are used for training).

B. *The DC Dataset - (Interpolated Strawberry Dataset)*

Second, the DC dataset is a dataset provided by a distribution center in Cambridge, Ontario, Canada. For the DC dataset the study focused on the strawberry FP as an initial stage. The purchase prices for strawberry are extracted from the DC dataset. To have one purchase price per day despite of having more than one daily supplier for strawberry, purchases made from more than one supplier on a given date are averaged. The average price paid is considered as the price of that day. Rows with missing values are filled with interpolated prices using a linear interpolation function.

III. UTILIZED PREDICTION MODELS FOR FP

Two DL models are deployed in this work, one has simple structure while the other is compound. They are both compared to more condensational models; three ML models and one statistical model. Brief description of the deployed models is provided in this section.

A. *Conventional Models*

1) Statistical Models (ARIMA): ARIMA model is an acronym that stands for Autoregressive Integrated Moving Average which is a class of statistical models for analyzing and forecasting time series data. It explicitly caters to a suite of standard structures in time series data. It is a generalization of the simpler Autoregressive Moving Average (ARMA) with the notion of integration.

2) Machine Learning (ML) models: Three standard ML models are considered. By standard ML models is meant the traditional tools of machine learning approaches (which do not include recent tools of deep learning).

a) Support Vector Regression (SVR): This regression algorithm is based on the same principles as the Support Vector Machines (SVM). It is a discriminative algorithm which seeks to maximize the boundary between two different classes. A new point is determined by assigning it to the side it is closest to. For the regression case, a margin of tolerance is chosen which then approximates the original SVM [12]. SVR is slow, memory intensive and problematic with very large datasets.

b) Gradient Boosting (GB): Gradient Boosting Machine (GB) is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak decision trees [13]. XGBoost is an implementation of GB which stands for eXtreme Gradient Boosting which is an optimized distributed gradient boosting framework designed to be highly efficient, flexible for most kinds of problems while being robust to overfitting. It implements parallel tree boosting in fast and accurate way [14]. The major drawback of gradient boosting is that like other traditional machine learning algorithms, its performance doesn't improve as much with increase in data.

c) Artificial Neural Networks (ANN): This model is based on the biological brain neurons, they have artificial net of neurons they receive external or internal input from other neurons after the input is weighted and added the result is then transformed by a transfer function (Sigmoid, hyperbolic, tangent functions or a step) into the output [15,16,17,18].

The work is supported by Grant's from NSERC CRD Funding and Loblaws Research Chair in Artificial Intelligence.

B. Deep Learning Models

1) Simple models.

a) Long Short-Term Memory (LSTM): LSTM was introduced by Sepp Hochreiter and Jürgen Schmidhuber back in 1997 [19]. Their model suggested an architecture of a recurrent network which included cells, input gates, and output gates. They also introduced Constant Error Carousel (CEC) units to tackle the issue of vanishing and exploding gradients. Later in 1999, a 'Forget Gate' (also called 'Keep Gate') was introduced in the LSTM architecture which allowed the LSTM to reset its own state [20]. The basic idea behind an LSTM is that the 'cell' (the memory part of LSTM unit) is responsible for maintaining dependencies between all elements in the input sequence.

The 'input gate' has the control over all new values that flow into the cell, the 'forget gate' controls the extent to which a value stays within a cell and the output gate controls the magnitude of the value in the cell that is used to calculate the activation output of the LSTM unit [21].

The activation function of the LSTM gates is often a logistic function [22]. In this paper the LSTM chosen structure is as follows: one hidden layer, one dropout layer of 0.2, single activation layer of relu, , 100 nodes, 150 epochs and the optimizer is Adam.

b) Convolutional Neural Network (CNN): A CNN is a deep learning algorithm which is a regularized version of multilayer perceptron (MLP). It typically consists of convolution, pooling and normalization layers. They are widely used for computer vision and image recognition but have also been used with text, time series and sequence data [23]. CNNs reduce dimensionality by implementing weight sharing. Unlike MLPs, they also have the ability to take into account, the spatial relationships that exist in the data [24]. CNNs have some drawbacks though by being difficult to tune, requiring very large datasets and being unable to extract temporal features.

2) Compoound models.

a) CNN-LSTM: After considering CNNs inadequacy in extracting temporal features and the failure of LSTMs to decipher spatial features, stacks of CNNs and LSTMs have been used so as to take advantage of their joint strengths [24, 25]. The structure of the deployed CNN-LSTM can be seen in Fig. 1. Stacked CNNs and LSTMs have been used for sentiment classification, inventory time series analysis, etc. Interestingly, only very little applications of this composition have been proposed for the FP price prediction. In this work, the CNN-LSTM is deployed for the first in modeling such problem.

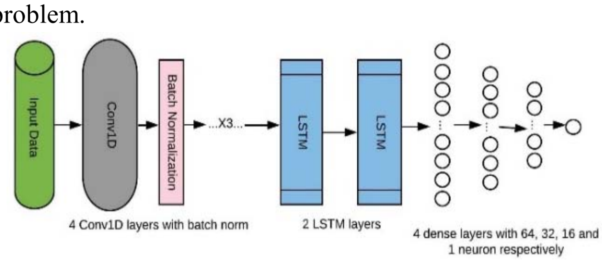

Fig. 1. CNN-LSTM architecture.

b) Attention-CNN-LSTM: A model similar to the CNN-LSTM mentioned above is proposed in this section with added self-attention after the LSTM layers. This attention layer uses additive attention with a sigmoid activation function, see Fig 2.

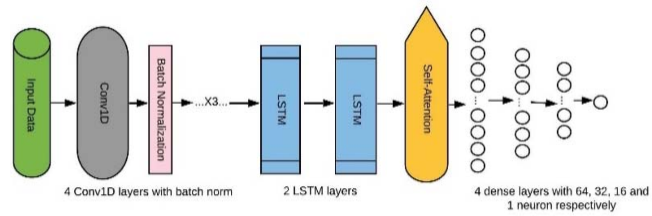

Fig. 2. CNN-LSTM with Attention.

The idea behind adding self-attention is to help models give more importance on relevant parts of the input data. It proved to be very useful in machine reading, abstractive summarization, and image description generation. An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors [9]. Self-Attention is a special kind of attention where the query, key and value are all the same. Attention is applied to the unit of each sequence and the unit of all sequences. We implement additive attention. Additive attention computes the compatibility function using a feed-forward network with a single hidden layer [9]. The attention is computed as follows:

$$h_{t,t'} = \tanh(x_t^T W_t + x_{t'}^T W_x + b_t) \quad (1)$$

$$e_{t,t'} = \sigma(W_a h_{t,t'} + b_a) \quad (2)$$

$$a_t = \mathrm{softmax}(e_t) \quad (3)$$

$$l_t = \textstyle\sum_{t'} a_{t,t'} x_{t'} \quad (4)$$

Where σ is the element wise sigmoid function, Wt and Wx are weight matrices corresponding to x_t^T and $x_{t'}^T$, Wa is the weight matrix corresponding to their non-linear combination and b_t, b_a are bias vectors [9]. Equation (4) shows how the attention l_t is calculated. To calculate this attention, the probability distribution a_t in (3) of the compatibility score $e_{t,t'}$ in (2) should be found first. This compatibility score is computed based on $h_{t,t'}$, the hidden representation of x_t and $x_{t'}$ computed as in (1).

It has been noticed that using attention improves the performance of deep learning models [25-28]. This is apparent in text sentiment analysis [25], vegetable yield prediction [27], healthcare question-answering, etc. but it is not applied till now to fresh produce price prediction [28].

c) Performance Measures: To choose the most appropriate forecasting model, performance evaluation measures should be utilized. The most frequently used performance measures in literature are used for measuring the forecasting accuracy level of the tested models as in [29,30, 31]; Mean Absolute Percentage Error (MAPE) [32], Mean absolute error (MAE) [33], Root Mean Square Error (RMSE) [34].

The lower the value of these measures the better. The three measures are utilized in this paper in addition to the R2 measure [10]. The R2 is a fraction that has no units corresponding to the correlation between the actual prices and predicted ones. The higher the value of R2 the better the model.

An aggregated measure is then applied to help in finding one value that summarizes the previously discussed measures. The validity of the measure stems from the fact that the two error measures, MAE and RMSE, have the same unit and both measure prediction errors or how erroneous the model is performing so they are both negatively-oriented scores; hence it is decided to aggregate them by taking their average error. In order to incorporate the R2 measure with the calculated average, the meaning of R2 should be considered. Since R2 is basically a no unit fraction with acceptable values between 0.0 and 1.0 proportionate to the goodness of the model, i.e. a positively-oriented score, it is decided to calculate 1 – R2 to represent how bad the model is performing and convert R2 to a negatively-oriented measure to match the MAE and RMSE measures. The resulting value is a fraction of 1 that should be multiplied by the previously calculated average error of MAE and RMSE to get an aggregated measure representing how erroneous the model is. Therefore, the resulting measure is negatively-oriented as well; i.e. the lower the aggregated measure value the better. The winning model is the one with the least aggregated value.

IV. EXPERIMENTS AND RESULT ANALYSIS

Periods of forecasting from 1 up to 20 steps ahead are experimented for Cabbage, Cauliflower, Bok Choi, Watermelon and Strawberry, see Fig. 3. As in [11], a sliding window of ten previous prices is used to forecast future ones up to 20 steps (nearly three weeks) ahead. Therefore, a lag of 10 is adopted, and the steps ahead are iteratively tested from 1 to 20.

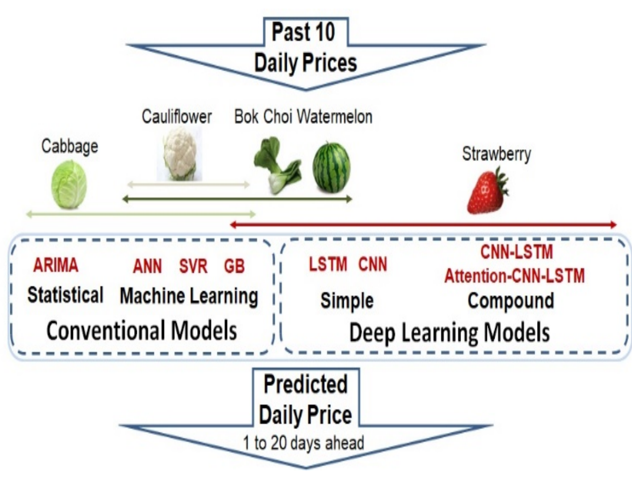

Fig. 3. ML & DL prediction models used to predict 1 to 20 days ahead prices of 5 types of FP using their previous 10 days prices.

The test sets sizes are 20 percent of the original data items. Each of these test sets are reserved as a test set for each crop while 80 percent of the data is used for training.

Performance evaluation is achieved using both the mean absolute percentage error (MAPE) as well as the mean absolute error (MAE); the lower the errors the better the performance. Experiments are conducted first using the Taiwan dataset for Cabbage, Cauliflower, Bok Choi, and Watermelon; then the interpolated version of the DC dataset is used to confirm the findings. The following subsections summarize obtained results:

A. Statistical vs Conventional Models (Taiwan - Cabbage)

The four deployed models are preconfigured before the experiment as follows:

1) ARIMA: In this experiment, the lag parameter for the ARIMA model is chosen to be 10 as in [11] and the prediction steps ahead go from 1 to 20 steps ahead which allows for nearly three weeks ahead prediction.

2) SVM: For the deployment of the support vector machine (SVM) or the support vector regressor, a grid search cross validation is done in Sci-kit Learn to select the best parameters. The grid consists of the following: C = [1, 5, 10], Gamma = [Scale, Auto] with 5-fold cross validation. This grid search tries all combinations of the different parameters, trains the model 5 times in every case based on the 5-fold cross validation and picks the combination that gives the best average cross validation score.

3) GB: For the GB model a grid search cross validation is done in Sci-kit Learn to select the best parameters [35, 36]. The grid consists of the following: Loss = [Huber, Ls], n_estimators = [30, 50, 100, 150], max_depth = [2, 3, 4] and min_samples_split = [2, 3, 5]. A 5-fold cross validation is used as well. This grid search tries all combinations of the different parameters, trains the model 5 times in every case based on the 5-fold cross validation and picks the combination that gives the best average cross validation score.

4) ANN: The ANN model used in this section is trained using the Levenberg-Marquardt algorithm to adjust the ANN weights [11,37,38]. The weights in the weight vector are adapted using repeated training until either the maximal number of iterations is reached, or the error is minimized to the goal. The model is initially examined using a Strawberry dataset obtained from the average monthly Strawberry prices received by growers in the US from Jan 1985 to Dec 2012 [39] before applying it on the Taiwan dataset.

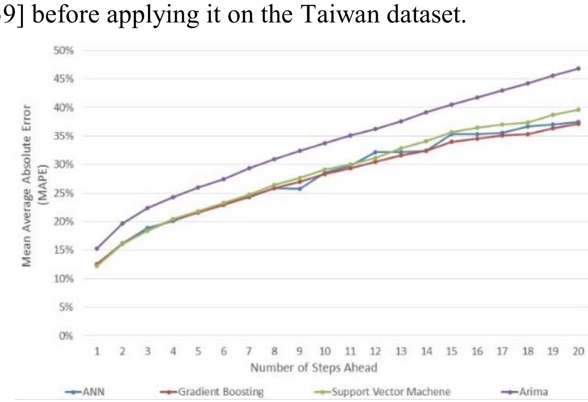

Fig. 4. Comparison of ARIMA & Conventional ML models in forecasting Cabbage prices for 1 to 20 days ahead, (Taiwan dataset).

The configured models (ARIMA, SVR, GB, and ANN) are trained and utilized in predicting the future prices of the Cabbage dataset. The results are compared based on the MAPE measure as in [11]. The ARIMA model is found to have the highest mean absolute percentage error (MAPE) over all the 20 steps ahead compared to the conventional ML models hence the lowest performance as evident in Fig. 4. Therefore, it is decided to exclude it in the remaining set of experiments.

B. Finding the Best Conventional Model (Taiwan - Cauliflower)

The wining ML models (ANN, GB, and SVR) are deployed in the Cauliflower dataset price prediction from 1 to 20 steps ahead using a lag of 10 previous prices. The results are compared based on the MAPE measure. As illustrated in Fig. 5, it is found that the Gradient Boosting model outperforms all other conventional models by having the least prediction error over all the 20 steps ahead which reflects the highest prediction accuracy.

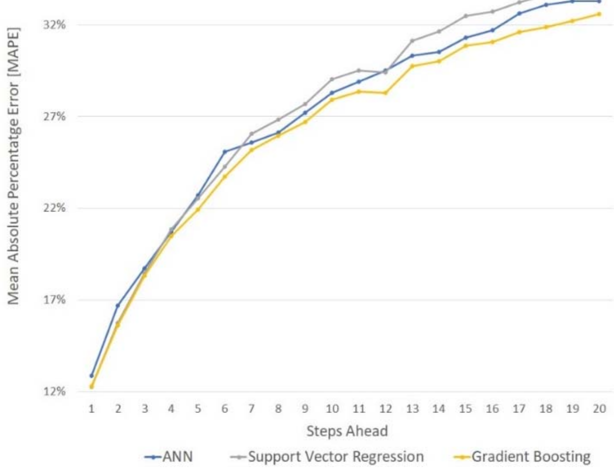

Fig. 5. Finding the best conventional model (GB) in predicting *Cauliflower* prices for 1 to 20 days ahead using the most recent 10 daily prices. (Taiwan dataset).

C. Conventional Models vs DL LSTM (Taiwan - Watermelon & Bok Choi)

The three conventional models (SVR, ANN, and GB) prediction performances are compared with that of the Simple LSTM DL model in this experiment. All models are used to predict the prices of two FP that are used in testing, the Watermelon and Bok Choi datasets. According to the MAPE measure, it is found that the simple deep learning model (LSTM) is performing better than the other conventional models in each of the twenty steps ahead as illustrated in Fig. 6.

These results can be interpreted in light of the fact that the LSTM has an inherent structure to deal with the time series data while XGBoost works on regression model. Due to having time series datasets for the Watermelon and Bok Choi, LSTM performs better as it is able to gauge the pattern taking into account the variability of time. Gradient Boosting on the other hand utilizes the decision tree regression and hence it is expected to be outperformed by LSTM.

Furthermore, regression is a method that deals with linear dependencies, while Neural Networks can deal with nonlinearities. Since the datasets have some nonlinear dependencies, Neural Networks should perform better than regression models therefore the LSTM DL model manages to have the lowest prediction error and highest accuracy compared even to the winning conventional model; the Gradient Boosting.

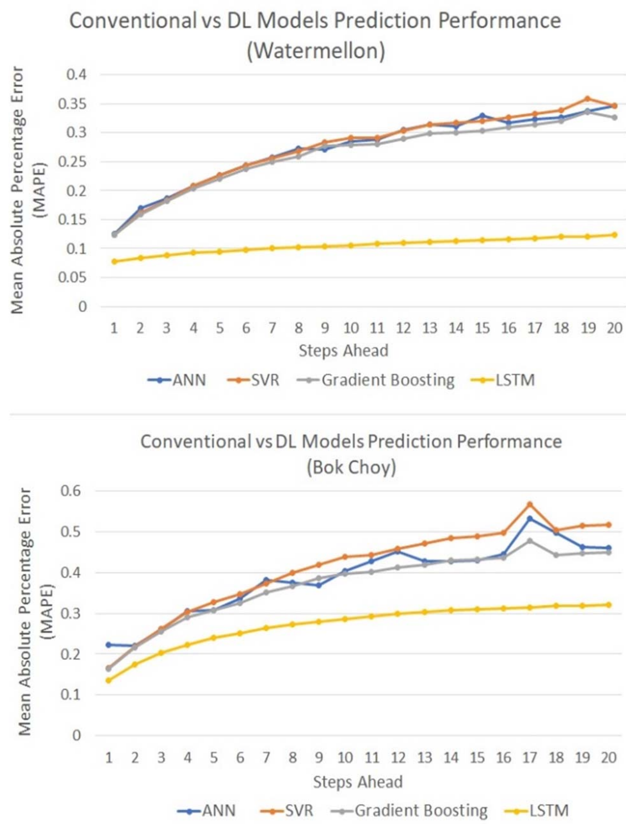

Fig. 6. LSTM performance vs conventional ML models in forecasting prices of 2 FP 1-20 days ahead using the 10 recent prices, (Taiwan dataset).

D. Best Conventional (GB/XGB) vs DL models (Taiwan DS - DC Strawberry)

RNN is known to perform better with complete time series, hence the experiment is using a more complete time series version of the DC strawberry dataset filled with interpolated prices. The previously winning GB model is then compared to the simple DL LSTM model and the compound DL CNN-LSTM model with and without attention.

It is found that the ATT-CNN-LSTM outperforms the winning standard ML GB model as well as the simple DL model, especially after adding attention, in forecasting the prices of strawberry for all the forecasting time steps greater than one. This is based on the aggregated measure results illustrated in Fig. 7. In addition, the simple DL LSTM model has lower MSE and MAE forecasting errors for all steps compared to those of the ML GB model as depicted in Fig. 7.

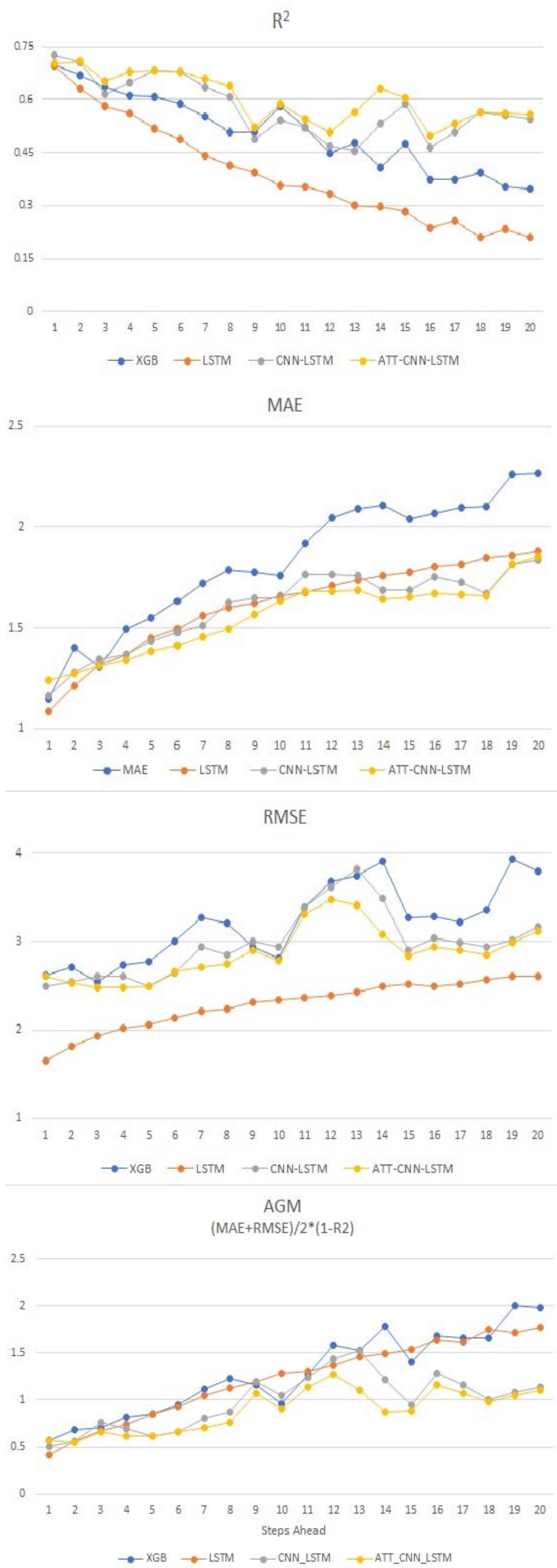

Fig. 7. Performance of the winning XGB Conventional model vs that of 3 DL models utilized in predicting Starawberry prices for 1 to 20 days ahead using the most recent 10 daily prices. (DC interpolated strawberry dataset).

To have one value summarizing the performance of each model across the twenty steps ahead, the average of all the 20 steps ahead AGM values (AAGM) is calculated for each of the tested models, the resulting values are listed as in Table I. As evident, the ATT-CNN-LSTM model still persists the best performance with the lowest AAGM value. The contribution of adding the CNN DL model to the simple LSTM model as well as adding attention to the resulting compound DL model is summarized in Table I. The AAGM reduction in percentage due to adding CNN then attention is reported to quantify the contributed performance enhancement after each modification to the model structure. This reduction which represents the performance enhancement is calculated as follows:

$$Improvement\% = \frac{AAGM_{before} - AAGM_{After}}{AAGM_{before}} \text{ x } 100 \quad (3)$$

Where $AAGM_{before}$ & $AAGM_{after}$ are the prediction model *AAGM* before the modification in architecture and after. *Improvement%* is the prediction percentatage improvement after the modification due to AAGM reduction; -ve values indicate AAGM increase hence deteriaration in performance.

Table 1 shows how models' performance improve after each addition, adding the CNN to LSTM improves performance by 19%. The resulting CNN-LSTM performance has then enhanced by additional 10% after adding attention.

TABLE I. AAGM REDUCTION IN PERCENTAGE DUE TO ADDING THE CNN TO THE SIMPLE LSTM DL MODEL THEN ATTENTION TO THE RESULTING COMPOUND DL MODEL

Metrices	XGB	LSTM	CNN-LSTM	Attention-CNN-LSTM
AAGM	1.28	1.22	0.99	0.89
Modification			+ CNN	+ Attention
AAGM Improvement			19%	10%

V. CONCLUSION AND FUTURE WORK

From experiments results it can be concluded that the ARIMA model has the highest mean absolute percentage error (MAPE) hence the lowest performance compared to the conventional ML models. In addition, among the conventional ML algorithms, the Gradient Boosting (GB) is the best due to having the least MAPE error. Finally, the performance of LSTM simple DL model is higher than all of the tested conventional ML models for two FP (Watermelon and Bok Choy). This is due to having less markets for these two FP which leaves us with data that is closer in nature to time series.

It is also found that the best performing model according to the aggregated measure is the compound DL model, ATT-CNN-LSTM, which outperforms the ML and simple DL models in accuracy of price prediction especially after adding the attention.

For future work, more compound DL models with attention should be compared to the winning ATT-CNN-LSTM. Varying lags other than 10 days and longer future steps ahead greater than 20 should be tried. More complete price time series should be created using advanced interpolation techniques to improve the performance of the LSTM as it is a recurrent Neural Network that should perform better with complete time series.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

ACKNOWLEDGMENT

We appreciate the contribution of Prarabdha Agrawal, who has assisted with the SVR. We acknowledge the generous contribution of Loblaws which funded the Research Chair in AI at the University of Waterloo. Warm thanks are addressed to managers and officials of Loblaws for their support and help during this project.

REFERENCES

[1] A. M. Ticlavilca, D. M. Feuz, and M. McKee, "Forecasting agricultural commodity prices using multivariate Bayesian machine learning regression," Proc. NCCC-134 Conferencen on Applied Commodity Price Analysis, Forecasting, and Market Risk Management. St. Louis, MO, April, 2010.

[2] M. Shukla and S. Jharkharia, "Agri‐fresh produce supply chain management: a state‐of‐the‐art literature review," IJOPM, vol. 33 (2), pp.114-158, 2013.

[3] J. Bacon, N. Khanna, and V. Naithani, "High velocity supply chains for perishables: A Fresher Proposition," Cognizant 20-20 Insights. Aug, 2015. Available online: https://www.cognizant.com/whitepapers/high-velocity-supply-chains-for-perishables-a-fresher-proposition-codex1325.pdf (Accessed on 7 Jan 2020).

[4] J. C. Buzby, J. T. Bentley, B. Padera, C. Ammon, and J. Campuzano, "Estimated Fresh Produce Shrink and Food Loss in U.S. Supermarkets," Agriculture 2015, 5, pp. 626-648, 2015.

[5] V. Ramasubramaniyam, "Forecasting Techniques in Agriculture, " J Ind Soc Agril Statist, IASRI, New Delhi, 2015.

[6] A. Pankratz, "Forecasting with univariate Box-Jenkins models: concepts and cases," Wiley Series in Probability & Statistics. New York: John Wiley & Sins. ISBN: 978-0-471-09023-6, 1983.

[7] P. J. Brockwell and R. A. Davis, "Introduction to time series and forecasting," Second Edition. Springer Texts in Statistics International Publishing, 2016. DOI: 10.1007/978-3-319-29854-2.

[8] J. Contreras, R. Espinola, F. J. Nogales, and A. J. Conejo, "ARIMA models to predict next-day electricity prices," IEEE Trans. Power Syst., vol. 18, pp. 1014–1020.

[9] A. Vaswani et al., "Attention is all you need," In 31st Conference on Adv Neural Inf Process Syst, pp.6000–6010, 2017.

[10] R. Fonticella, "The usefulness of the R2 Statistic," Associate of the Casualty Actuarial Society (ACAS) Forum. Winter 1998 ed.

[11] Y. Peng, C. Hsu and P. Huang, "Developing crop price forecasting service using open data from Taiwan markets," Technologies and Applications of Artificial Intelligence (TAAI), Taiwan, pp. 172-175, 2015.

[12] V. Vapnik, "The nature of statistical learning theory," Springer: New York, NY, USA, pp. 768, 1995.

[13] J. H. Friedman, "Greedy function approximation: a Gradient Boosting machine," The Ann. Stat., Vol. 29 (5) , pp. 1189-1232, Oct. 2001.

[14] J. Chen et. al "Social media popularity prediction based on visual-textual features with XGBoost," ACM MM '19: the 27th ACM Int. Conf. on MM, pp. 2692–2696, Oct. 2019.

[15] F. Karray and C.W. De Silva, "Soft computing and intelligent systems Design: theory, tools and applications," Addison Wesley, New York, 2004.

[16] P. Doganis, A. Alexandridis, P. Patrinos, and H. Sarimveis, "Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing," Journal of Food Engineering, vol. 75 (2), pp.196-204, 2006.

[17] L.R. Medsker and L. C. Jain, "Recurrent Neural Networks: design and applications," 1st ed. Boca Raton, FL, USA: CRC Press, Inc., pp. 41, 1999.

[18] B. M. Sonali and W.Priyanka, "Research paper on basic of Artificial Neural Network," International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), vol. 2(1), pp. 96-100, 2014.

[19] S. Hochreiter and J. Schmidhuber, "Long Short-term Memory," Neural Computation, vol. 9 (8), pp. 1735-1780, 1997.

[20] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," NIPS'08: 21st International Conference on NIPS, pp. 545–552, Dec. 2008.

[21] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech), USA, pp. 338–342, 2014.

[22] F. A. Gers, J. Schmidhuber and F. Cummins, "Learning to Forget: Continual Prediction with LSTM," Neural Comput., vol. 2(10), pp. 2451-2471, 2000.

[23] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,". CoRR, abs/1803.01271, 2018.

[24] L. Zheng et. al, "A fault prediction of equipment based on CNN-LSTM network," IEEE ICEI, Nanjing, China, pp. 537-541. Aug. 2019.

[25] Z. Yangsen, Z. Jia, J. Yuru, H. Gaijuan and C. Ruoyu, "A text sentiment classification modeling method based on coordinated CNN-LSTM-Attention model," CHINESE J. ELECTRON., vol. 28 (1), pp. 120-126, Jan 2019.

[26] G. Zheng, S. Mukherjee, X. L. Dong, and F. Li, "Opentag: Open attribute value extraction from product profiles," CoRR, vol. abs/1806.01264, 2018.

[27] S. De Alwis, Y. Zhang, M. Na and G. Li. "Duo attention with deep learning on tomato yield prediction and factor interpretation," 16 th PRICAI: Trends in Artificial Intelligence. Cuvu, Yanuca Island, Fiji, E1, proceedings, part III, vol. 11672, pp. 704-715, Aug. 2019.

[28] R. Cai et al., "An CNN-LSTM attention approach to understanding user query intent from online health communities," IEEE ICDMW, New Orleans, LA, pp. 430-437, 2017.

[29] N. K. Ahmed, A. F. Atiya, N. El Gayar, and H. El-Shishiny. "An empirical comparison of machine learning models for time series forecasting," Econom. Rev., vol. 29 (5-6), pp. 594-621, 2010.

[30] M. Shukla and S. Jharkharia. "Agri-fresh produce supply chain management: a state-of-the-art literature review," IJOPM, vol. 33(2), pp. 114-158, 2013.

[31] R. J. Hyndman, and A. B. Koehler, "Another look at measures of forecast accuracy," Int. J. Forecast., vol. 22 (4), pp. 679–688, 2006.

[32] J. S. Armstrong, F. Collopy, "Error measures for generalizing about forecasting methods: Empirical comparisons," Int. J. Forecast., vol. 8 (1), pp. 69 – 80, 1992.

[33] A. Botchkarev, "Performance metrics (Error Measures) in machine learning regression, forecasting and prognostics: properties and typology," IJIKM, vol. 14, pp. 45–79, 2019.

[34] T. Chai and R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," GMD, vol.7(3), pp. 1247–1250, 2004.

[35] R. Lawrence, A. Bunn, S. Powell, and M. Zambon, "Classification of remotely sensed imagery using stochastic gradient boosting, as a refinement of classification tree analysis," Remote Sens. Environ., vol. 90, pp. 331–336, 2004.

[36] J. H. Koo, D. K. Seo, and Y. D. Eo, "Variable importance assessment for colorization of grayscale aerial images based on random forest, adaptive boosting and stochastic gradient boosting," Disaster Adv. 2018, vol. 11, pp. 22–29, 2018.

[37] K. Levenberg, "A method for the solution of certain problems in least squares," Q. APPL. MATH., vol. 2, pp.164-168, 1944.

[38] D. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," SIAM Journal, vol. 11, pp. 431-441, 1963.

[39] USDA Economic Research Service, "Fruit and tree nut yearbook: dataset (89022)," Oct. 2017. [Available Online]: http://usda.mannlib.cornell.edu.