Ameliorating Farmer Suicides by Predicting Crop Price Trends using a Deep Learning Approach

Maryam Tabar [1] and **Chi-Yang Hsu** [2] and **Hangzhi Guo** [3] and **Amirreza Bagherzadehkhorasani** [4] and **Amulya Yadav** [5]

Abstract. Farmer suicides have reached a concerning level in India recently. This issue mainly stems from farmers' inability to sell their products at the desired profit level, which is caused by price fluctuation in the agriculture market. To help the farmers with this issue, this paper proposes a deep learning algorithm, PECAD-CLS, which can predict the future crop price trends (Increase, Decrease, Stable) based on the historical patterns of crop price and volume. Even though previous studies have attempted to tackle market price trend prediction via Machine Learning (ML) algorithms, they do not model the spatio-temporal dependence of future prices on past data explicitly. Hence, they do not have a desirable performance on the spatio-temporal datasets. To address this deficiency, our proposed method makes two main contributions. First, we collect real-world daily price and volume of different crops over a period of 11 years and then impute it to deal with missing values. Second, we modify a state-of-the-art model, called PECAD, to predict the future produce price trends. Our experiment results show that PECAD-CLS improves state-of-the-art baseline models by $\sim$5% in terms of F1 (in the best case scenario). In addition, the PECAD architecture performs even better in direct crop price prediction; it outperforms baselines by $\sim$24% in terms of *coefficient of variation*.

1 Introduction

One of the complicated issues that Indian farmers deal with is the volatility of the crop price especially in harvesting season. As a result, they may not only lose their saving but also end up not being able to pay back their loans. Accumulation of these problems has led to a large number of farmers to commit suicide. According to official reports, over 12,000 farmer suicides have occurred annually since 2013 [10]. More importantly, this trend is exasperating every day; e.g., over 600 farmer suicides have been reported in the Indian state of Maharashtra from January to March 2019 [6].

Farmer suicides stem from several socio-economic factors such as crop failures, irrigation issues, the distress of repaying loans, etc. Failing to pay back the agricultural loans is one of the most important problems that mostly results from price fluctuations in the agricultural market [8]. Given the importance of the issue, we have decided to address that using predictive analytics.

The research question that we are going to address is as follows: *Can we build an accurate predictive model to foresee the future crop price trends in different markets based on the historical patterns of crop price and volume?* This model can help farmers make intelligent decisions on selling their crops using the future crop price trends in various markets.

To address this question, several real-world conditions need to be taken into account. First, there are a large number of missing values in datasets. Second, there is a long-term temporal dependency between the future crop price and past pricing patterns. Third, there is a spatial dependency between crop prices in different markets; the crop prices in nearby markets may be closer to each other compared to those in distant markets. Therefore, it is crucial to design a model capable of dealing with these challenges (i.e., data sparsity and spatio-temporal dependency between crop prices).

There have been several studies on predicting future crop price trends using various ML approaches. However, they do not model the spatio-temporal dependence of future prices on past data explicitly. As a result of this deficiency, they do not achieve a desirable performance on the real-world spatio-temporal datasets.

To fill this gap, we propose a deep neural network architecture, called PECAD-CLS (**P**rice **E**stimation for **C**rops using the **A**pplication of **D**eep Learning for **CL**a**S**sification), to predict future price trends based on the historical pricing and volume patterns. This model makes two major contributions: First, it collects real-world data on crop prices and volumes at 1,352 Indian agricultural markets over a period of 11 years from Agmarknet.gov.in[6] and imputes missing values with a state-of-the-art algorithm. Second, it adapts a state-of-the-art model, PECAD [4], to predict crop price trends.

To evaluate the proposed model, we consider two classification tasks. In the first one, we only predict the upward and downward directions of price changes. However, the price movement doesn't always show a clear upward or downward trend in the market. Therefore, in the second task, in addition to those two classes, another class representing price stability is considered. According to the experiment results, PECAD-CLS improves the F1 of several state-of-the-art baseline models by $\sim$5%, which results from explicitly modeling the spatio-temporal dependencies during neural architecture design. Besides, the PECAD architecture performs even better when predicts crop prices directly: it outperforms baselines by $\sim$24%.

Related work. Several studies have attempted to tackle farmer distress via ML algorithms. You et al. [11] proposed a deep neural network for predicting crop yield using remote sensing data. However, this model needs satellite images of fields which can be hard to access in low-resource environments. In [2, 7], the authors took advantage of classical ML approaches to predict crop price trends. How-

[1] Pennsylvania State University, USA, email: mfg5544@psu.edu
[2] Pennsylvania State University, USA, email: cxh5437@psu.edu
[3] Wenzhou Kean University, China, email: guoha@kean.edu
[4] Pennsylvania State University, USA, email: abb6024@psu.edu
[5] Pennsylvania State University, USA, email: amulya@psu.edu

[6] http://agmarknet.gov.in/

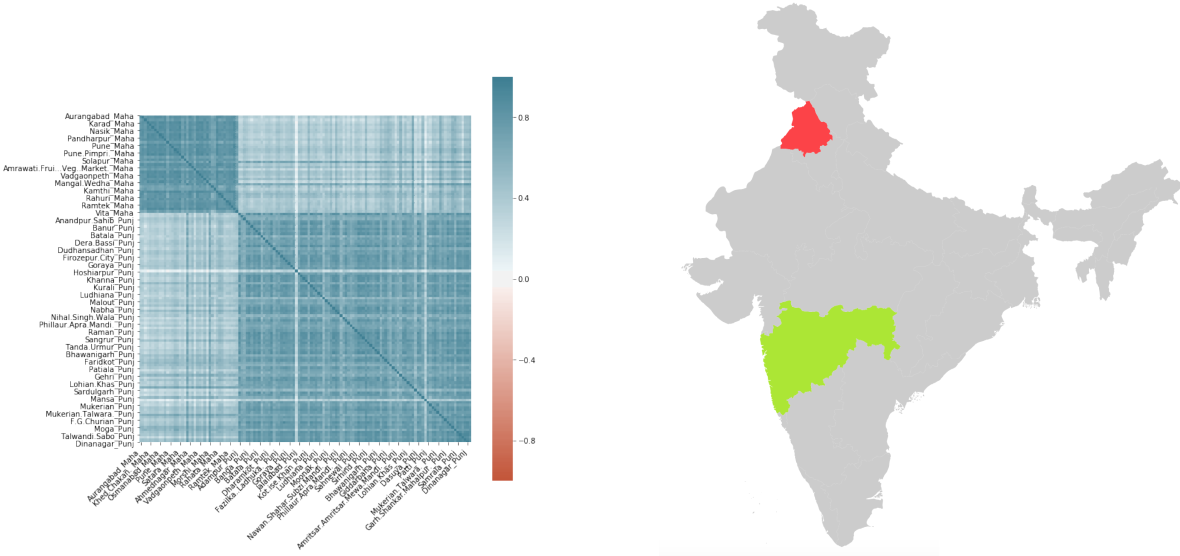

Figure 1: Correlation between Green Chilli prices at different markets across the Indian states of Maharashtra (green) and Punjab (red).

ever, they did not benefit from the spatio-temporal dependency between future price and past pricing patterns. Recently, Guo et al. [4] proposed a deep learning method for price prediction based on the spatio-temporal dataset.

2 Real-World Dataset

Data Collection. To conduct this study, two datasets are collected. The first one is raw data on the price and volume of three different crops (Brinjal, Tomato, and Green Chili) at 1352 agricultural markets across India over the period of 2008 through 2018. This data is fetched from the Agmarknet.gov.in[6] website, which is administrated by the government of India.

Furthermore, to capture the effect of physical proximity on crop prices, the geographical locations of markets are needed. For this purpose, we collect latitude and longitude of the locations of markets using Google Maps API. Each market and crop is assigned a unique ID, which is converted to a one-hot vector representation.

Data Preparation. One of the challenges that we ran into is the large number of missing values. These missing values are caused by several different factors, e.g., on certain days, markets may have been closed or no transactions have been recorded. To deal with this issue, we apply a state-of-the-art data imputation method named SoftImpute [5] to replace the missing values with an estimated value. In addition, we conduct time quantization to deal with the vanishing (or exploding) gradients issue, which may happen during learning long-term temporal dependencies. In fact, we consider a time window of w days (we test with w values of 4, 6, and 9) as a single time step and average the crop prices and volumes to obtain those values for a specific time step.

To develop a price movement prediction model, we consider two groups of classification. In the first group, we divide the dependent variable into two categories, “Increase” and “Decrease”, to monitor the upward and downward movements of the price. However, sometimes, price movement does not show a clear upward or downward trend in the market. Therefore, in the second type of classification, we added a third category, “Stable”, which keeps track of the relatively stable price patterns. For this purpose, we round the price values to the nearest 10 (e.g., if the price is 123, it will be rounded to 120). We denote the new price $\tilde{P}^c_{m,t}$. Therefore, the label of a datapoint for the crop c and t^{th} time-step at market m ($y^c_{m,t}$) is defined as follows. This helps in the case of minor increase or decrease which may cause malfunction to our model on more obvious increases or decreases.

$$\mathrm{y}^c_{m,t} = \begin{cases} \text{Increase} & \text{if } \tilde{P}^c_{m,t} > \tilde{P}^c_{m,t-1} \\ \text{Stable} & \text{if } \tilde{P}^c_{m,t} = \tilde{P}^c_{m,t-1} \\ \text{Decrease} & \text{Otherwise.} \end{cases}$$

Data Characteristics. The final dataset consists of ∼40000 labeled data points. An input feature vector consists of the following elements: vector embeddings of the specific crop c and market m, longitude and latitude of the location of the market m, and price and volume of crop c at market m over the past n time steps (the feature space includes the pricing and volume data for previous 360 days).

Furthermore, prices at nearby markets are highly likely to be correlated. We generate a heatmap indicating the correlation between Green Chilli prices at a subset of markets (i.e., the spatial dependency between prices). As shown in Figure 1, the prices at those markets located in the same state, e.g., Maharashtra, are highly correlated. However, there is a slight correlation between the price at the markets across two relatively distant states, Maharashtra and Punjab.

3 Deep Learning Algorithm

PECAD-CLS adapts a state-of-the-art algorithm called PECAD [4] for classification. It takes advantage of two best practices, a wide and deep network [3] and Temporal Convolutional Network (TCN) [1] (shown in Figure 2). The features that make PECAD stand out compared to similar algorithms are capturing long-term historical patterns (by using TCN) and dealing with highly sparse inputs in an efficient manner (by using wide and deep architecture). However, unlike the original wide and deep network which uses cross-product transformation features as the features to train the linear model, PECAD uses a novel combination of two separate TCNs for memorizing the sequence of price and volume, respectively.

To predict the price trends, we add a Softmax layer on top of the last layer of the network and employ a categorical Cross-Entropy loss function. In addition, we consider pricing and volume data from 2008 to 2016 as the training set. The remaining data (price and volume data from 2017 to 2018) is considered as the test set.

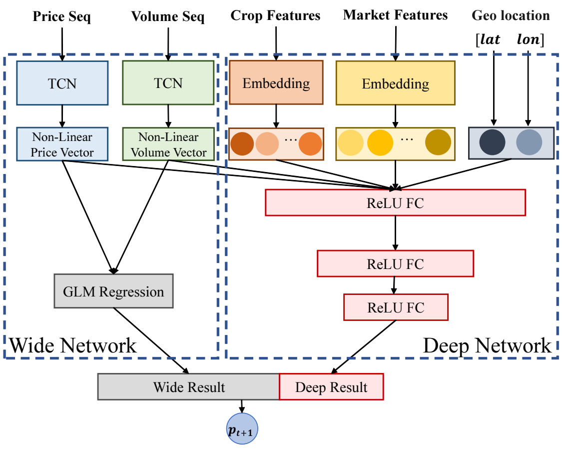

Figure 2: Architecture of the PECAD model [4].

4 Experimental Evaluation

For each window size (4, 6, and 9 days), a separate model is trained for 150 epochs. Then, the average performance of the model over 20 runs is reported. Later, we compare the original PECAD (regression) and PECAD-CLS (classification) with some baselines and analyze the findings in the Discussion section.

Baselines. To evaluate the effectiveness of our model, we compare it with Random Forest (RF) and four deep learning models capable of capturing spatio-temporal dependencies. The first model is the standard TCN model (*TCN*) [1]. The second one is the standard wide and deep networks (*Standard Wide & Deep*) [3] in which cross-product transformation features are used as the input of the wide network, instead of using TCN transformation. The third model is (*Attention-LSTM*) [9] which utilizes attention mechanism to memorize long sequential input. The last baseline (PECAD-CLS - Single-TCN) is a variation of the PECAD-CLS in which a single TCN is used for capturing historical pricing and volume patterns.

Regression. As the performance measure, "*coefficient of variation*" is used to compare different ML models. *Coefficient of variation* is the root mean squared error (RMSE) divided by the mean produce price. Table 1 shows the performance of different models in predicting prices of three crops (Brinjal, Tomato, and Green Chilli) across three different time window sizes ($w = 4, 6$ and 9 days) [4]. According to the results, PECAD achieves $\sim$20% lesser *coefficient of variation* as compared to RF. Further, PECAD significantly outperforms other deep learning models by reducing *coefficient of variation* by $\sim$25% as compared to the average case performance of the other deep learning models. In particular, PECAD outperforms *Standard Wide & Deep* [3] and *PECAD-Single TCN* by achieving $\sim$13% and $\sim$13.5% lesser coefficient of variation, respectively [4].

Classification. We use *accuracy* and weighted *F1* to assess the predictive performance of different algorithms. For the two-class classification, the achieved *accuracy* and *F1* are shown in Table 2 and Table 3, respectively. According to the results, on average, PECAD-CLS improves the accuracy and F1 of RF by 0.67% and 1.1%, respectively. In addition, on average, PECAD-CLS achieves 1.1% higher accuracy and 2% higher F1 as compared to the other deep learning methods.

For the three-class classification, the achieved *accuracy* and *F1* are shown in Table 4 and Table 5, respectively. The results show that PECAD-CLS improves the *accuracy* and *F1* of RF by 1.6% and 6.5%, respectively. Also, on average, it achieves 1.9% higher accuracy and 5.2% higher F1 as compared to the other deep learning methods.

Discussion. In the three-class classification, quality measures have shown considerable improvement in terms of F1 compared to the two-class classification. This improvement is largely due to considering the third class for small price changes. In the two-class classification, short price rise or fall will result in the trend change. However, this small fluctuation may not be quite effective in the long-term trend. The three-class classification considers a "Stable" class for such small price fluctuations because their effect on the general trend of the market over the long-run is negligible.

As demonstrated before, PECAD-CLS performs better when it comes to memorizing sequential data and at the same time, it can generalize the crop and market features. The other baseline models fail in either of these tasks. In particular, PECAD-CLS outperforms *Standard Wide & Deep* about 3.8% in F1 (in the three-class version), which shows the benefits of using TCN for modeling sequential data.

5 Conclusion & Future Research Pathway

In this paper, we build a deep neural network model that can predict price trends based on past price and volume patterns, and geo-location information. Previous ML algorithms did not model the spatio-temporal property explicitly. However, our method tries to model this relationship by using the Wide and Deep architecture along with the TCN model. As a result, F1 improved by $\sim$5% compared to the state-of-the-art baseline models.

For the future research projects in this area, adding the weather data in the model is recommended. The weather forecasting data affects the supply level and as a result, it could improve the prediction model. The other factor which would improve the prediction accuracy is the demand for each crop in different markets. According to the law of demand and supply, if demand for a product goes up, it will positively affect the price. Therefore, studying the impact of demand would create a more accurate model.

REFERENCES

[1] Shaojie Bai, J Zico Kolter, and Vladlen Koltun, 'An empirical evaluation of generic convolutional and recurrent networks for sequence modeling', *arXiv preprint arXiv:1803.01271*, (2018).

[2] George H Chen, Kendall Nowocin, and Niraj Marathe, 'Toward reducing crop spoilage and increasing small farmer profits in india: a simultaneous hardware and software solution', in *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*, (2017).

[3] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al., 'Wide & deep learning for recommender systems', in *Proceedings of the 1st workshop on deep learning for recommender systems*, pp. 7–10. ACM, (2016).

[4] Hangzhi Guo, Alexander Woodruff, and Amulya Yadav, 'Improving lives of indebted farmers using deep learning: Predicting agricultural produce prices using convolutional neural networks', in *The Thirty-Second Annual Conference on Innovative Applications of Artificial Intelligence*, (2020).

	4 Days/90 Cells			6 Days/60 Cells			9 Days/40 Cells		
	Brinjal	**Tomato**	**Chilli**	**Brinjal**	**Tomato**	**Chilli**	**Brinjal**	**Tomato**	**Chilli**
RF	21.12	22.88	19.45	23.47	38.48	21.60	24.50	44.30	23.54
Attention-LSTM	19.88	20.52	17.49	21.98	24.36	18.44	21.00	31.94	21.04
TCN	20.59	19.87	17.36	54.42	33.25	27.69	27.59	98.02	81.83
Standard Wide & Deep	23.63	24.47	19.07	24.34	28.22	18.67	27.36	34.29	21.10
PECAD - Single TCN	21.90	23.50	17.77	29.43	30.86	20.21	26.26	33.65	20.46
PECAD	19.64	21.62	17.07	21.14	24.20	17.65	21.75	28.46	19.31

Table 1: Coefficient of Variation of different ML models with varying time window sizes [4].

	4 Days/90 Cells			6 Days/60 Cells			9 Days/40 Cells		
	Brinjal	**Tomato**	**Chilli**	**Brinjal**	**Tomato**	**Chilli**	**Brinjal**	**Tomato**	**Chilli**
RF	0.56	0.59	0.54	0.58	0.60	0.56	0.61	0.63	0.58
Attention-LSTM	0.56	0.57	0.55	0.58	0.58	0.58	0.60	0.60	0.59
TCN	0.55	0.58	0.55	0.58	0.59	0.50	0.61	0.62	0.60
Standard Wide & Deep	0.55	0.57	0.55	0.58	0.58	0.58	0.61	0.61	0.59
PECAD-CLS - Single TCN	0.56	0.57	0.55	0.58	0.59	0.56	0.60	0.62	0.59
PECAD-CLS	0.57	0.58	0.55	0.59	0.60	0.58	0.61	0.63	0.60

Table 2: Accuracy of different ML models with various window sizes in the two-class classification task.

	4 Days/90 Cells			6 Days/60 Cells			9 Days/40 Cells		
	Brinjal	**Tomato**	**Chilli**	**Brinjal**	**Tomato**	**Chilli**	**Brinjal**	**Tomato**	**Chilli**
RF	0.55	0.59	0.54	0.58	0.60	0.56	0.60	0.63	0.58
Attention-LSTM	0.56	0.57	0.55	0.57	0.58	0.58	0.60	0.60	0.59
TCN	0.56	0.58	0.54	0.58	0.59	0.33	0.61	0.62	0.59
Standard Wide & Deep	0.54	0.57	0.54	0.58	0.58	0.57	0.61	0.61	0.59
PECAD-CLS - Single TCN	0.56	0.57	0.55	0.58	0.59	0.56	0.60	0.62	0.59
PECAD-CLS	0.57	0.58	0.56	0.59	0.61	0.58	0.61	0.63	0.60

Table 3: F1 of different ML models with various window sizes in the two-class classification task.

	4 Days/90 Cells			6 Days/60 Cells			9 Days/40 Cells		
	Brinjal	**Tomato**	**Chilli**	**Brinjal**	**Tomato**	**Chilli**	**Brinjal**	**Tomato**	**Chilli**
RF	0.50	0.56	0.53	0.53	0.59	0.55	0.56	0.62	0.57
Attention-LSTM	0.53	0.54	0.53	0.56	0.57	0.56	0.59	0.60	0.59
TCN	0.53	0.49	0.48	0.56	0.50	0.49	0.59	0.61	0.58
Standard Wide & Deep	0.53	0.47	0.53	0.55	0.55	0.56	0.59	0.62	0.58
PECAD-CLS - Single TCN	0.54	0.55	0.53	0.56	0.58	0.55	0.59	0.61	0.57
PECAD-CLS	0.53	0.57	0.54	0.56	0.59	0.56	0.59	0.62	0.60

Table 4: Accuracy of different ML models with various window sizes in the three-class classification task.

	4 Days/90 Cells			6 Days/60 Cells			9 Days/40 Cells		
	Brinjal	**Tomato**	**Chilli**	**Brinjal**	**Tomato**	**Chilli**	**Brinjal**	**Tomato**	**Chilli**
RF	0.39	0.53	0.5	0.46	0.57	0.54	0.51	0.61	0.56
Attention-LSTM	0.51	0.53	0.52	0.54	0.56	0.55	0.58	0.59	0.58
TCN	0.52	0.32	0.31	0.55	0.33	0.32	0.58	0.6	0.58
Standard Wide & Deep	0.39	0.53	0.50	0.46	0.57	0.54	0.51	0.61	0.56
PECAD-CLS - Single TCN	0.55	0.56	0.54	0.57	0.59	0.56	0.60	0.62	0.58
PECAD-CLS	0.55	0.58	0.55	0.58	0.60	0.57	0.60	0.63	0.60

Table 5: F1 of different ML models with various window sizes in the three-class classification task.

[5] Trevor Hastie, Rahul Mazumder, Jason D Lee, and Reza Zadeh, 'Matrix completion and low-rank svd via fast alternating least squares', *The Journal of Machine Learning Research*, **16**(1), 3367–3402, (2015).

[6] INDIA TODAY. Over 12,000 maharashtra farmers committed suicide in 3 years: Government. `https://www.indiatoday.in`, 2019.

[7] Wei Ma, Kendall Nowocin, Niraj Marathe, and George H Chen, 'An interpretable produce price forecasting system for small and marginal farmers in india using collaborative filtering and adaptive nearest neighbors', in *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development*, p. 6. ACM, (2019).

[8] Arvind Panagariya, *India: The emerging giant*, Oxford University Press, 2008.

[9] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, 'Sequence to sequence learning with neural networks', in *Advances in neural information processing systems*, pp. 3104–3112, (2014).

[10] THE TIMES OF INDIA. Over 12,000 farmer suicides per year, centre tells supreme court. `https://timesofindia.indiatimes.com/`, 2017.

[11] Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon, 'Deep gaussian process for crop yield prediction based on remote sensing data', in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 4559–4565. AAAI Press, (2017).