

DeepAg: Deep Learning Approach for Measuring the Effects of Outlier Events on Agricultural Production and Policy

Sai Gurrapu

Department of Computer Science (CS) Department of Electrical and Computer Engineering (ECE)
Virginia Tech
Blacksburg, VA
saig@vt.edu

Feras A. Batarseh, IEEE Senior Member

Virginia Tech
Arlington, VA, USA
batarseh@vt.edu

Pei Wang

Department of CS
Virginia Tech
Blacksburg, VA, USA
pwang1@vt.edu

MD Nazmul Kabir Sikder

Department of ECE
Virginia Tech
Arlington, VA, USA
nazmulkabir@vt.edu

Nitish Gorentala

Department of CS
Virginia Tech
Blacksburg, VA, USA
nitishg@vt.edu

Munisamy Gopinath

Department of Agricultural and Applied Economics
University of Georgia
Athens, GA, USA
m.gopinath@uga.edu

Abstract—Quantitative metrics that measure the global economy's equilibrium have strong and interdependent relationships with the agricultural supply chain and international trade flows. Sudden shocks in these processes caused by outlier events such as trade wars, pandemics, or weather can have complex effects on the global economy. In this paper, we propose a novel framework, namely: DeepAg, that employs econometrics and measures the effects of outlier events detection using Deep Learning (DL) to determine relationships between commonplace financial indices (such as the DowJones), and the production values of agricultural commodities (such as Cheese and Milk). We employed a technique called Long Short-Term Memory (LSTM) networks successfully to predict commodity production with high accuracy and also present five popular models (regression and boosting) as baselines to measure the effects of outlier events. The results indicate that DeepAg with outliers' considerations (using Isolation Forests) outperforms baseline models, as well as the same model *without* outliers detection. Outlier events make a considerable impact when predicting commodity production with respect to financial indices. Moreover, we present the implications of DeepAg on public policy, provide insights for policymakers and farmers, and for operational decisions in the agricultural ecosystem. Data are collected, models developed, and the results are recorded and presented.

Index Terms—Deep Learning, Precision Agriculture, Outlier Events, Public Policy

I. INTRODUCTION AND MOTIVATION

According to the United Nations, the world's population is expected to increase by two billion persons in the next 30 years, from 7.7 billion (current) to 9.7 billion in 2050 and could peak at nearly 11 billion around 2100. To feed this growing population, a similar increase in food production must also be achieved [1]. Several challenges exist in agriculture -with declining productivity of resources such as

land, the environmental footprint of production practices, and the ensuing need for sustainability—limiting human abilities to scale up production to meet the global demand. The integration of technology into the agricultural ecosystem is considered to be an important pathway for not only providing adequate nutrition to the world but also ensuring sustainability of the resources for the benefit of future generations [2]. For instance, precision farming where producers optimize their complex multivariate farming practices by continuously monitoring, measuring and analyzing several variables such as weather, soil, and crop type, enables precise targeting and care for each specific agricultural commodity at scale that was not possible in the 20th century [3]. However, with agriculture highly susceptible to outlier events, e.g., floods, drought, and trade wars, predicting the future with an accounting for possible outlier events remains a major challenge [4]. Big data analytics in precision farming demonstrated the importance of recognizing and extracting insights and trends from historical agricultural data to better guide commodity production decisions and policy making based on context [5]. As the amount of data generated in the agricultural ecosystem continues to increase, Machine Learning (ML) is being used to provide accurate predictive insights and guidance on operational decisions with real-time data [6].

ML allows the machine to learn from the available data without being explicitly programmed, thus revealing more insights than what is normally possible through traditional data analytics. DL extends classical ML by adding more complexity into the models with a large learning capacity as it has a strong advantage in feature learning. This enables DL models to be flexible and highly adaptable for a wide

range of complex tasks. That notion allows it to excel at classification and prediction problems in many domains [1]. The application of DL in agriculture is relatively recent and can be a promising technique considering the impact and potential it has demonstrated in other domains [1]. Most studies and applications of precision farming today are localized to the farm environment without much consideration of the impact of external variables, specifically outlier events [7]. The hypothesis tested in our work is twofold: (1) Whether DL will perform better than traditional statistical (such as regression) techniques for predicting multivariate relationships in farming? (2) Outlier events, such as Economic, Financial, Weather, or Political events influence precision farming practices - regardless if they are positive or negative- and therefore need to be accounted for as an important part of the life-cycle process, can a combination of isolation forests and LSTM accomplish that with high accuracy? During the COVID-19 pandemic for instance, many farmers and producers were struggling with the forecasts provided to them using traditional econometrics because the models used to create such predictions don't account for outlier events. Accordingly, in this paper, we propose an approach called DeepAg that employs DL to measure the effect of outlier events on agricultural production and employ it to predict the future production patterns in the presence of outliers. For this purpose, data on five commonplace financial indices and agricultural commodities production, along with that on outliers' events over the past two decades are assembled on a monthly scale. A combination of isolation forests and LSTM model is employed to train and test for classification. The input features include the commodities production (data collected from United States Department of Agriculture - USDA) and their respective highest *causation* and *correlation* with market's financial indices. We present two forecasts for each commodity's production, one with the presence of an outlier event, and one without. To demonstrate the effectiveness of our proposed approach and the importance of outliers' detection, we also present several baseline models using standard ML techniques and compare that with an LSTM model that doesn't consider outliers. The results indicate that the LSTM model with isolation forests (i.e. outlier detection) outperforms all others and can be used to determine the relationship between commodity production and financial indices with a high accuracy, in the presence of outliers.

The rest of this paper is organized as follows: The Related Work section presents relevant literature on the developments of ML and DL for precision farming; the Experiment section describes the datasets used for the experiment and overall experimental setup; the Methodology section introduces our approach and outlines its evaluation; the Results section discusses the outcomes of our experiment; and finally, the Conclusions section evaluates our approach and provides insights on precision farming and the importance of outlier events detection for farmers and public policy makers.

II. RELATED WORK

As agricultural ecosystems adopt technology to improve their farming practices, the data collected in the background is increasingly valuable. In [2], a comprehensive review was conducted on ML applications for agricultural production systems. They demonstrated examples of certain precision farming practices such as crop and soil management, disease detection, livestock management, and water usage amongst others that can be improved using ML. A Support Vector Machines (SVM) based methodology was presented by [8] for the early detection of problems in egg production. The experiment forecasts egg production for up to three days and sends an alert if the production curve displays any anomalies. The results demonstrate that a poultry management system with production forecasting would prove to be useful to assist producers with preventative measures before a problem occurs. Another approach using SVM was shown in [9] to predict the weight trajectory of livestock given the past evolution of the herd. Additionally, using advanced hardware sensing techniques and artificial neural networks, [10] demonstrated an architecture to predict wheat yield production with a high accuracy of 91%.

Agricultural data has also been shown to be useful outside of the farm environment. For example, [4] and [11] employed deep-learning ML techniques (unsupervised and supervised) to predict trade patterns of seven major agricultural commodities and indicated that unsupervised ML approaches with neural networks provides better prediction fits over the long term. A method in [12] was proposed to measure causal scenarios in trade during outlier events using network-based models, specifically Graph Neural Networks (GNNs) were used to predict outliers effectively and to provide relevant domain explainability. In [13], Association Rules (AR) analysis was employed to identify imports and exports associations (*if a then b*) with the trade flows and used Ensemble Machine Learning (EML) methods for agricultural trade predictions. In [5], the use of ML for econometric practices was presented and demonstrates the challenges of such simulation models and shortcomings when used for quantitative economic analysis. A fast unsupervised algorithm called Isolation Forest was proposed by [14] for detecting anomalies in continuous data (of all domains). Accordingly, no method is found that could be applied to the production of all commodities considering multiple forms of outliers, our study aims to address that gap.

III. EXPERIMENTAL WORK

This section describes the datasets used for the experiment, the indices, outlier detection methods, and the overall pipeline.

A. Financial Indices Dataset

A collection of financial indices data are obtained from Yahoo Finance (<https://finance.yahoo.com/>). Daily market close data are collected for years 2000-2019, the following indices are considered: Gold, Crude Oil, Dow Jones Industrial Average (DOW), S&P 500, and Volatility Index (VIX). Significance of the five financial indices is as follows:

- *Gold*: Price of one ounce of gold [15].
- *Crude Oil*: Price of a barrel of benchmark crude oil [15].
- *DOW*: Index of thirty most prominent companies listed on US stock exchanges [16].
- *S&P 500*: Index of 500 of the largest companies listed on US stock exchanges [16].
- *VIX*: Measure of the market's expected volatility expectation based on the S&P 500 index [17].

B. Commodities Dataset

The data on the commodities was acquired from USDA's National Agricultural Statistics Service (NASS) QuickStats database. The commodities dataset was cleaned and segmented based on the commodity type. The list of commodities is as follows: *Beef, Butter, Cheese, Chickens, Ducks, Eggs, Ice Cream, Lamb and Mutton, Milk, Other poultry, Pork, Sherbet, Turkeys, Veal, and Water Ices*. The available frequency of the commodities is the first day of every month, however, the financial indices frequency is for every weekday. This created a conflict when merging the two datasets because not every first day of the month is a weekday. In this case, we supplemented the data by labeling the most recent weekday as the first day of the month.

C. Outlier Events

Outlier detection can be defined as a rare event identification, an observation that differs significantly from the rest of the data points [14]. Anomalies or outlier events in our experiment typically indicate a major disturbances related to the Economy (eg. trade war), Financial (eg. recession), Weather (eg. natural disaster), or Political (eg. government instability). We use anomaly detection algorithm to find data points that show abnormal behavior from the rest of the data points in financial indices. A proper understanding of the outlier events will help us better predict the relationship between financial indices and commodity production. We used an unsupervised algorithm, *Isolation Forest*, for detecting anomalies. This algorithm works best if the outlier points are easily isolable and contamination rates of abnormal points are low. The overall pipeline of DeepAg is presented in Figure 1.

A. Data Preprocessing

All data are pre-processed before being fed into the base-lines or LSTM models. To minimize bias, we employed data transformation techniques to normalize each of the input features using MinMaxScaler as represented in equation 1.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Formula 1 normalizes the values of the financial indices dataset, commodities dataset, and the outlier events dataset into a range of 0-1. To prepare the datasets for anomaly detection, we then used *DoubleRollingAggregate* from the *ATDK Python* library to track the statistical behavior in a time series dataset. The *DoubleRollingAggregate* transformer rolls two sliding windows side-by-side along with a time series, aggregates using statistical mean, and tracks the difference of the aggregated metrics between the two windows. Figure 2 shows the changes (normalization) to the indices after applying *DoubleRollingAggregate* Transformer.

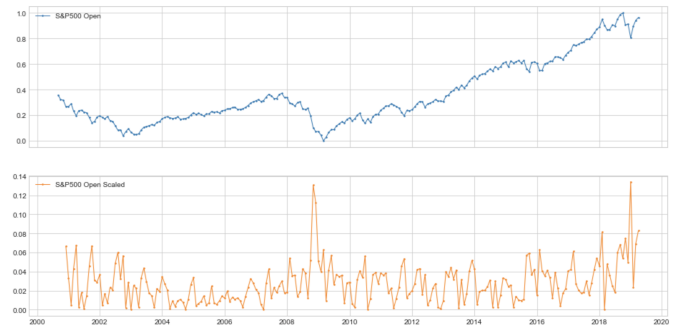


Fig. 2. Original S&P 500 data from 2000-2019 (top) and scaled data (bottom)

B. Outlier Detection

The attribute of Isolation Forest is one that shows the tendency to separate outlier data points from normal data points, as the algorithm randomly selects an attribute and splits values between the minimum and maximum of that attribute. Afterwards, it recursively creates random partitions to find outlier points. Random partitioning creates a short path for outliers. Therefore, a sample is more likely to be an outlier if a forest of random trees collectively produces shorter path lengths for particular samples.

1) *Isolation Forest*: The algorithm takes several hyperparameters as input and among them the most important is *contamination rate*. It decides the percentage of outliers which can be approximately predicted out of the total data points. We determine the contamination rate using a popular statistical measure called Interquartile Range (IQR). IQR describes the middle 50% of the data distribution. Quartiles slice any Gaussian distribution into four equal groups of 25%. Calculating IQR describes the middle half of the data in the distribution set. We are considering these middle data segments as normal data points since outliers are usually at the tails of a Gaussian distribution. The IQR can be represented as:

IV. METHODOLOGY

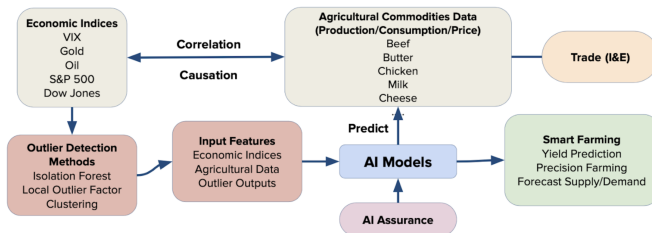


Fig. 1. The DeepAg methodology

$$\text{Interquartile Range} = Q3 - Q1 \quad (2)$$

Where Q1 is the first quartile or 25th percentile and Q3 is the third quartile or 75th percentile. Tables I and II provide the daily and monthly contamination rate of all financial indices using the IQR method. The Isolation Forest algorithm identifies an estimation of the anomaly score for a given instance x using this formula:

$$s(x, m) = 2^{-E(h(x))/c(m)} \quad (3)$$

Where E is the average of h trees and $c(m)$ is the average path length of unsuccessful binary searches.

TABLE I
DAILY IQR CONTAMINATION FOR FINANCIAL INDICES

Daily Data Contamination (Using IQR)	
Financial Index	Contamination Rate
VIX	6.559
Gold	5.382
S&P 500	6.008
DOW	6.125
Crude Oil	3.953

TABLE II
MONTHLY IQR CONTAMINATION FOR FINANCIAL INDICES

Monthly Data Contamination (Using IQR)	
Financial Index	Contamination Rate
VIX	6.250
Gold	2.232
S&P 500	2.232
DOW	2.232
Crude Oil	6.250

2) *Algorithm Accuracy*: After determining the anomalies, we plotted the data as shown by Figure 3. The red markers on the graphs indicate the outlier events found by the isolation forest algorithm. To ensure high accuracy, we manually evaluated the anomalies by cross-checking them with historical events that have occurred during that time. The algorithm proved to be accurate in determining and isolating outlier events as it was able to identify all events such as trade war, COVID-19, hurricanes, and others.

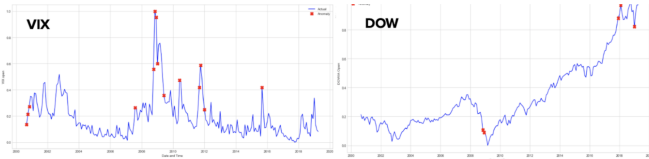


Fig. 3. Graph of anomalies present in VIX and DOW

C. Causation and Correlation

The *DoWhy Python* library is used for causal inference and for measuring causality. Measuring causality enabled us to use the financial index that had the greatest causal impact on each agricultural commodity (for instance, commodity: *Milk* had

a high causation score with Index: *Gold*). While correlation (Figure 4) shows whether two measures are following similar trends in a dataset, causation (Figure 4) can be useful to determine what is truly causing another dataset to change. It's evident from the results that the causation matrix in Figure 5 *DOW* and *meat* commodities generally have a strong causal relationship with green color indicating high causality, and red indicating low. In the Table III of correlation results, VIX indicates the lowest correlation with the commodities compared to the other indices. Moreover, essential commodities such as Butter, Cheese, and Chickens seem to be strongly correlated with important economic indicators such as DOW and S&P 500.

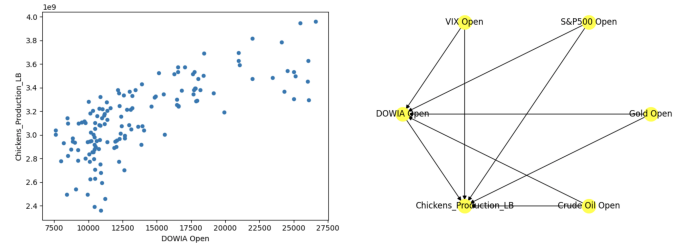


Fig. 4. Chicken production and DOW correlation (Left); chicken production and DOW causation (right)

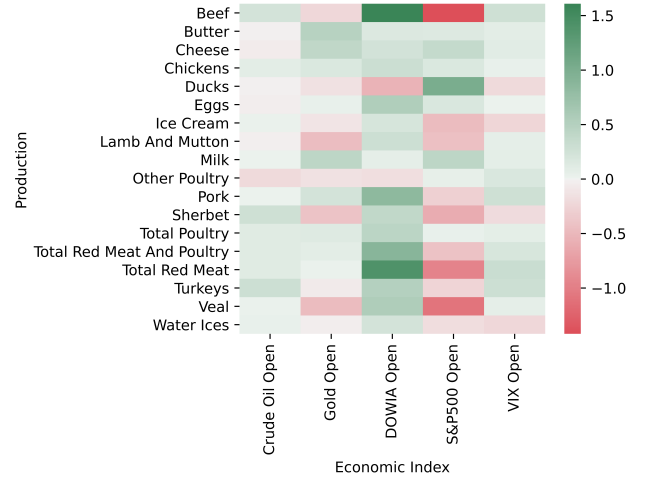


Fig. 5. Causation matrix of commodities and indices

D. Artificial Intelligence Models for DeepAg

1) *Baseline Models*: To ensure that our proposed DeepAg LSTM approach works effectively, and to determine if it outperforms existing approaches, baseline models are created for evaluation purposes. We applied five popular boosting and regression models as baselines: *Linear Regression*, *Linear Regression with Polynomial Features*, *Regression Tree*, *Random Forest Regressor*, and *XGBoost Regressor* to predict commodities' production. The input for the regression model includes the highest causation and correlation financial indices and the

TABLE III
CORRELATION OF COMMODITIES AND INDICES

Commodities	GOLD	DOW	S&P 500	OIL	VIX
Beef	0.096	-0.060	-0.030	-0.043	0.043
Butter	-0.097	0.721	0.571	0.529	0.403
Cheese	-0.276	0.801	0.863	0.825	0.371
Chickens	-0.307	0.676	0.748	0.709	0.389
Ducks	-0.427	0.110	0.556	0.564	-0.088
Eggs	-0.502	0.147	0.869	0.863	-0.406
Ice Cream	-0.054	-0.277	-0.256	-0.230	-0.108
Lamb & Mutton	0.182	-0.771	-0.558	-0.505	-0.490
Milk	-0.261	0.803	0.781	0.744	0.427
Other poultry	0.183	-0.462	-0.322	-0.283	-0.400
Pork	-0.089	0.678	0.701	0.652	0.329
Sherbet	-0.005	-0.418	-0.395	-0.377	-0.064
Turkeys	0.046	0.289	0.205	0.175	0.287
Veal	0.319	-0.791	-0.808	-0.772	-0.414
Water Ices	-0.197	-0.019	0.054	0.070	0.010

historical commodity production data. Every commodity is paired with an index (ones with the *highest* correlation and causation score per commodity).

2) *DeepAg LSTM Approach*: The architecture of Recurrent Neural Networks(RNN) has allowed for more accurate modeling of time-series processes where references to information from the past persist [18]. The Long Short-Term Memory (LSTM) model improves upon the RNN architecture to overcome the vanishing gradient problem and enables sequence learning with high accuracy [19]. The LSTM model is an optimal choice for our experiment considering its state-of-the-art performance in sequence learning and our dataset is heavily time-series based. The process of preserving information and carrying it forward occurs in the LSTM memory cell and it is composed of Input Gate (i_t), Forget Gate (f_t), and Output Gate (o_t). The mathematical representation of an LSTM memory cell can be represented as follows (in formulas 3, 4, and 5):

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \quad (4)$$

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \quad (5)$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \quad (6)$$

Where σ represents sigmoid function, w_x is the weight for the respective gate(x) neurons, h_{t-1} is the output of the previous LSTM block at timestamp $t-1$, x_t is the input at current timestamp, and b_x is the biases for the respective gates(x) [19]. Once the cell state is filtered, it is passed through an activation function, specifically the sigmoid function, and it predicts the output, h_t , of the LSTM unit at timestamp t . The output, h_t , is passed through a softmax layer to obtain the predicted output y_t . This process is represented as follows:

$$\tilde{c}_t = \tanh(w_c[h_{t-1}, x_t] + b_c) \quad (7)$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \quad (8)$$

$$h_t = o_t * \tanh(c^t) \quad (9)$$

Where c_t is the cell state memory at timestamp t and \tilde{c}_t represents candidate for cell state at timestamp t .

The objective of the LSTM model is to better predict the multivariate relationship between the commodity production, and financial indices *with* and *without* considering outlier events. In order to measure the effects of outliers on the prediction, the final LSTM model is developed where one prediction is made with the outlier events as one of the input features (outputted by isolation forest) and another prediction without. In the case that the highest correlation and causation were equal, we performed the same procedure as we did for the baselines of selecting only one of the indices with a preference towards causation scores.

V. RESULTS

A. Baseline Models

Our experiment with the baseline models included default hyperparameters with no feature engineering. The results rank the overall performance of the models as follows: Linear Regression with Polynomial Features, Linear Regression, XG-Boost Regressor, Random Forest Regressor, and Regression Tree. *Linear Regression with Polynomial Features* resulted with the most accurate and best-performing baseline model (as shown in Figure 6).

We use the R^2 score to determine the accuracy of the regression models as calculated by equation 10 and $RMSE$ to evaluate the error as shown in equation 11 given our prediction, \hat{y} , and the actual value, y . The commodity production values typically range in the billions/year depending on the commodity.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (10)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (11)$$

When comparing R^2 scores across models, the linear fit can be seen as unusually low for certain commodities such as Beef, Ice Cream, and Water Ices. This is most likely due to the variance in the data (i.e. variable production due to economic factors). In addition, USDA has more available historical production data for these commodities and this also may have been a factor in affecting the fit.

As shown in Figures 7 and 8 certain commodities such as Beef compared to Chickens had significantly more variance in terms of their range. For instance, Chickens production has a clear upward trend with slight variance over time. In contrast, the Beef production range significantly varies with abrupt changes year to year. Therefore, this is a factor that made a considerable difference in the linear fit R^2 score for the baselines. For the Linear Regression with Polynomial Features model, Beef's R^2 score value was -0.1384 whereas Chickens was 0.5373 which had a much better linear fit for the data supporting our conclusion about data variance.

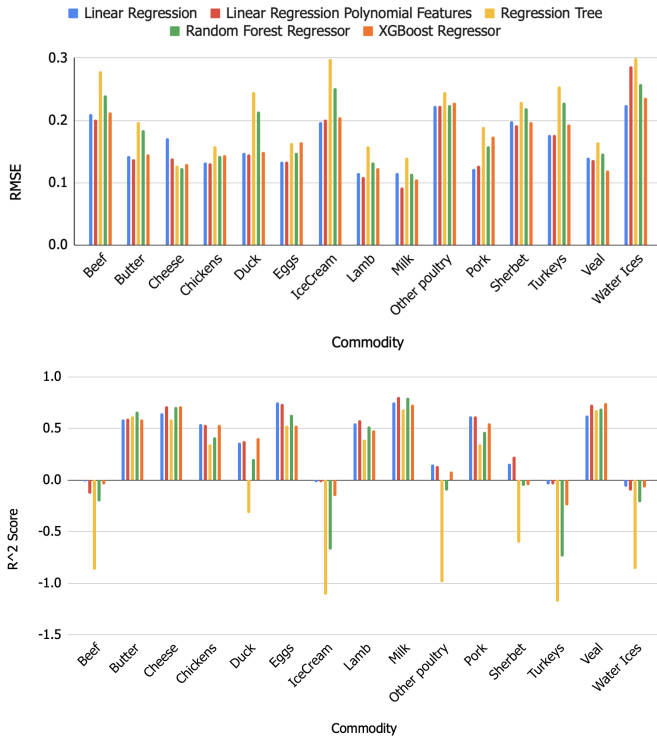


Fig. 6. Baseline models Root Mean Squared Error ($RMSE$) and R^2 scores results

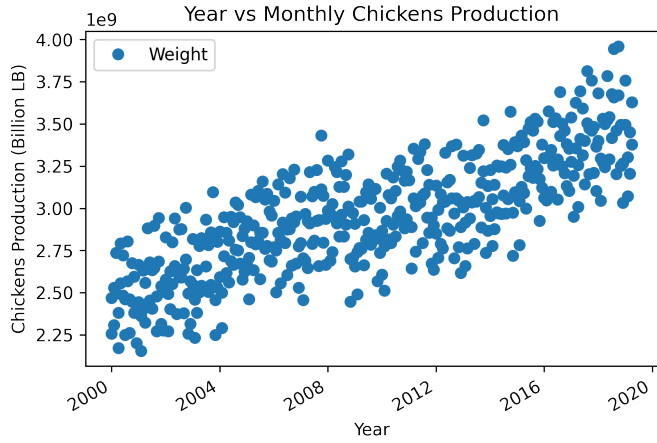


Fig. 7. Historical chickens production

B. DeepAg Outcomes

1) *Forecasting Commodities' Production*: We employed the multi-step multivariate time series forecasting technique to accurately predict the commodity's production in the future. Multivariate time-series forecasting enables us to predict a future dependent variable y , based on more than one independent variable x . Here, our dependent variable y is the commodity production, and independent variables x are the historical production data, highest causation, highest correlation index, and historical outliers. Lastly, multi-step forecasting tunes the model to predict a certain number of time-steps ahead instead

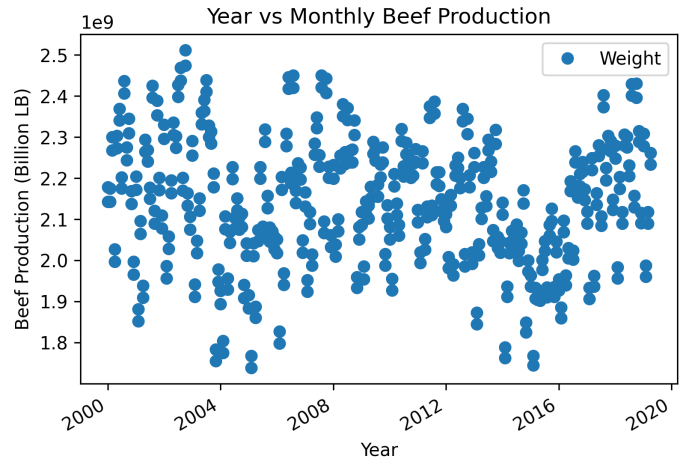


Fig. 8. Historical beef production

of only predicting *one* future value. Our model had a look-back of *60* time-steps from the past to forecast approximately *30* multi-steps into the future. Based on the commodity since every commodity has a different number of data points, our model was able to predict approximately *five* years ahead.

Table IV indicates the results obtained using our DeepAg approach. For each commodity, a prediction is made with outputs from isolation forests as a feature of the model. Each prediction is also paired with an $RMSE$ value to measure error rate and also to evaluate against the baselines. For each commodity, the most relevant indices is noted as (*highest causation*, *highest correlation*) or (*highest causation*) if they are the same. It follows as Beef (DOW), Butter (Gold), Cheese (Gold), Chickens (DOW), Duck (S&P 500), Eggs (DOW), IceCream (S&P 500, DOW), Lamb and mutton (Gold, DOW), Milk (S&P 500, Gold), Other poultry (Oil, VIX), Pork (DOW), Sherbet (S&P 500, DOW), Turkeys (DOW), Veal (DOW), Water Ices (VIX, DOW).

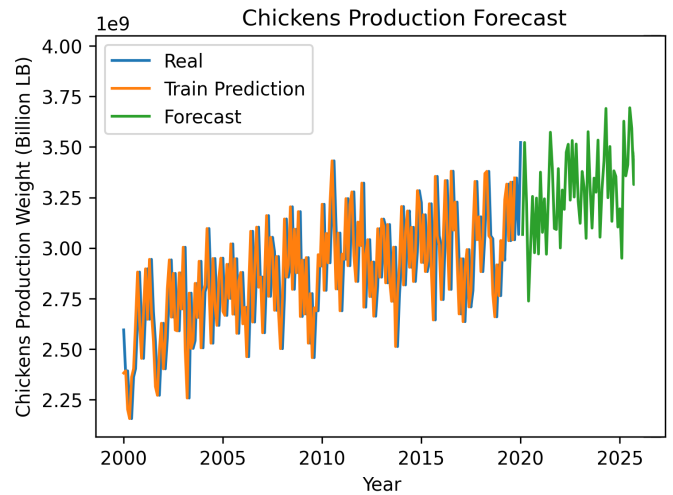


Fig. 9. Chickens production forecast 2020-2025

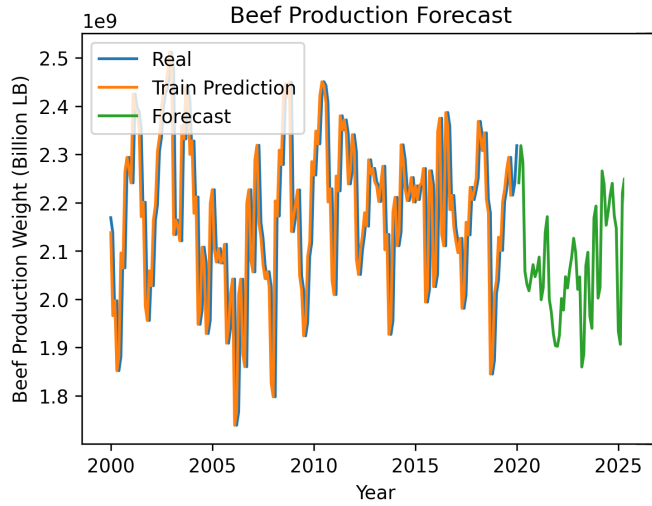


Fig. 10. Beef production forecast 2020-2025

TABLE IV
DEEPAg LSTM MODEL RESULTS

Commodity	RMSE With Outliers	Prediction With Outliers	RMSE W/O Outliers	Prediction W/O Outliers
Beef	0.164	2060199936	0.12	1960800000
Butter	0.001	164524992	0.009	164524992
Cheese	0.004	991345984	0.058	991345984
Chickens	0.258	3531831040	0.197	3302286080
Duck	0.001	6348000	0.001	11891000
Eggs	0.708	8599600128	0.501	8599600128
IceCream	0	12010000	0.004	82232000
Lamb and mutton	0.001	12000000	0.001	10100000
Milk	1	18872999936	1	16984999936
Other poultry	0	142000	0	100000
Pork	0.158	2022099968	0.118	2157299968
Sherbet	0	3541000	0	3541000
Turkeys	0.024	523214016	0.028	458169984
Veal	0.001	6700000	0	6000000
Water Ices	0	4647000	0	4647000

C. The Effect of Outlier Detection

The Figures 9 and 10 illustrate the overall prediction fit and forecast for two of the commodities, Chickens and Beef. It's evident that the model had a close train prediction compared to the real values and was able to accurately fit the data in addition to the forecasting. Early Stopping regularization method is implemented to avoid overfitting during model training. From Table IV, the RMSE values demonstrate the high prediction accuracy of our DeepAg approach compared to the baselines. The lower the RMSE value, the better our approach is able to fit the production data and predict accurately. For instance, the RMSE for Beef production was 0.164, 0.120, 0.201 using DeepAg with outliers, DeepAg without outliers, and with Linear Regression with Polynomial Features baseline model, respectively, indicating that DeepAg with isolation forest outperforms DeepAg without outliers and the baselines. Out of the 15 commodities, four commodities had a lower RMSE with DeepAg (with Isolation Forest);

TABLE V
COMPARISON OF BEST BASELINE MODEL AND DEEPAg

Commodity	Linear with Features RMSE	Regression Polynomial RMSE	DeepAg With Isolation Forest RMSE	DeepAg W/O Isolation Forest RMSE
Beef	0.201	0.164	0.120	
Butter	0.138		0.001	0.009
Cheese	0.139		0.004	0.058
Chickens	0.131	0.258		0.197
Duck	0.146		0.001	0.001
Eggs	0.134	0.708		0.501
IceCream	0.201		0.000	0.004
Lamb and mutton	0.109		0.001	0.001
Milk	0.093	1.000		1.000
Other poultry	0.224		0.000	0.000
Pork	0.127	0.158		0.118
Sherbet	0.192		0.000	0.000
Turkeys	0.177		0.024	0.028
Veal	0.137		0.001	0.000
Water Ices	0.287		0.000	0.000

three commodities had a lower RMSE with DeepAg (without Isolation Forest), and three commodities performed better with Linear Regression with Polynomial Feature - the results are presented in Table V. For the remaining three commodities, baseline RMSE was lower and this may be because those are consistent commodities (Chickens, Eggs, Milk) and tend to have a clear upward trend that are better predicted by linear regression models.

Table V supports that outlier events are useful and contribute to a higher accuracy when predicting commodities' production. Outlier events are the cause of sudden upward or downward spikes in economics, and in our work, we have captured that ambiguity through a more formal process. The use of outliers as one of the input features allows the LSTM model to learn and optimize for these extreme events.

VI. CONCLUSIONS AND DISCUSSIONS

In this manuscript, we proposed a novel approach called DeepAg that employs econometrics and leverages DL to measure the effect of outlier events when predicting the relationship between five commonplace financial indices and agricultural commodities. We compared DeepAg with existing techniques in terms of prediction accuracy. Our findings indicate that DeepAg's use of isolation forests enables it to outperform other techniques and offers a highly accurate prediction of the multivariate relationship in precision farming.

A. Outlier Events

Merging outlier events with production forecasts also reveals more accurate insights. A global supply shock for commodities, weather events, or an important international affair can affect production and can cause sudden spikes. Typically, these outlier events are sudden and cannot be planned for. During these times, producers are left with little insight on how their production will be affected. For instance, the COVID-19 pandemic resulted in a global supply shock for many essential commodities. The event created a large spike in the demand for products such as Beef and Chickens amongst other

commodities and caused the price of these commodities to rise. This caused a shortage of essential commodities and caused the price of these commodities to rise even higher creating a cycle until the production supply of these commodities exceeded market demand. Our approach could potentially help producers plan for such events.

B. Public Policy Insights

Predicting agricultural production is essential for feeding the world in the years ahead. The amount of agricultural production has a direct effect on supply, demand, and trade. While we demonstrate our approach at the aggregate level, they can also be used at micro scales such as at a farm or county. The aggregate analysis can particularly aid in shaping policies, since USDA, the Food and Agriculture Organization (FAO) and many other nations rely on country-specific as well as global forecasts to set policy parameters. Seasonal pattern data can also be used as a source of ground truth to determine trends and anticipate future demand. For example, note the prediction of chicken and beef production to 2025, which accounts for potential outlier events in the future; they are sharply different from existing straight-line forecasts and provide a bounded pathways for policy and other decisions. Our forecasting approach can help producers determine how low the production should drop and help them take preventative measures to continue operating even when production demand is low. During the winter, the producers may reduce the number of workers to save costs, minimize distribution, and reduce production volume - amongst other actions.

C. Future Work

As part of next steps, different splits between training and scoring datasets will be tested to evaluate the best split for the use-case presented in this paper. Additionally, we will evaluate the results for further AI assurance aspects such as over/underfitting, bias detection, and further explainability of outcomes. The scope of our approach can be extended to take specific outlier *types* into consideration. Currently, the outliers considered are a mix of Economic, Financial, Weather, and Political events. Limiting our approach to specific outlier types like weather can offer more precise insights - analysis can be done by *event* and *event type*. For instance, if a hurricane is forecasted, the producer can have more control over their production, and also real-time operational guidance can be provided to policy makers. DeepAg can positively affect agriculture through on-time outcomes and can increase overall farm performance using Deep Learning.

ACKNOWLEDGMENT

The authors would like to acknowledge the Commonwealth Cyber Initiative (CCI), the SmartFarms Initiative, and Center for Advanced Innovation in Agriculture (CAIA) at Virginia Tech. Additionally, a word of thanks goes to Madison J. Williams from the University of Mary Washington for her help in assembling data on outlier events.

REFERENCES

- [1] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Computers and Electronics in Agriculture*, vol. 147, pp. 70–90, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0168169917308803>
- [2] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, 2018. [Online]. Available: <https://www.mdpi.com/1424-8220/18/8/2674>
- [3] R. Gebbers and V. I. Adamchuk, "Precision agriculture and food security," *Science*, vol. 327, no. 5967, pp. 828–831, 2010. [Online]. Available: <https://science.sciencemag.org/content/327/5967/828>
- [4] M. Gopinath, F. A. Batareseh, and J. Beckman, "Machine learning in gravity models: An application to agricultural trade," National Bureau of Economic Research, Working Paper 27151, May 2020. [Online]. Available: <http://www.nber.org/papers/w27151>
- [5] H. Storm, K. Baylis, and T. Heckelee, "Machine learning in agricultural and applied economics," *European Review of Agricultural Economics*, vol. 47, no. 3, pp. 849–892, 08 2019. [Online]. Available: <https://doi.org/10.1093/erae/fbz033>
- [6] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big data in smart farming – a review," *Agricultural Systems*, vol. 153, pp. 69–80, 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0308521X16303754>
- [7] S. Gurrup, N. Sikder, P. Wang, N. Gorentala, M. Williams, and F. Batareseh, "Applications of machine learning for precision agriculture and smart farming," *The International FLAIRS Conference Proceedings*, vol. 34, Apr. 2021. [Online]. Available: <https://journals.flvc.org/FLAIRS/article/view/128497>
- [8] I. R. Morales, D. R. Cebrián, E. F. Blanco, and A. P. Sierra, "Early warning in egg production curves from commercial hens: A svm approach," *Computers and Electronics in Agriculture*, vol. 121, pp. 169–179, 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0168169915003919>
- [9] J. Alonso, A. Villa, and A. Bahamonde, "Improved estimation of bovine weight trajectories using support vector machine classification," *Computers and Electronics in Agriculture*, vol. 110, pp. 36–41, 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0168169914002488>
- [10] X. Pantazi, D. Moshou, T. Alexandridis, R. Whetton, and A. Mouazen, "Wheat yield prediction using machine learning and advanced sensing techniques," *Computers and Electronics in Agriculture*, vol. 121, pp. 57–65, 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0168169915003671>
- [11] M. Gopinath, F. A. Batareseh, J. Beckman, A. Kulkarni, and S. Jeong, "International agricultural trade forecasting using machine learning," *Data amp; Policy*, vol. 3, p. e1, 2021.
- [12] A. Monken, F. Haberkorn, M. Gopinath, L. Freeman, and F. Batareseh, "Graph neural networks for modeling causality in international trade," *The International FLAIRS Conference Proceedings*, vol. 34, Apr. 2021. [Online]. Available: <https://journals.flvc.org/FLAIRS/article/view/128485>
- [13] F. A. Batareseh, M. Gopinath, A. Monken, and Z. Gu, "Public policymaking for international agricultural trade using association rules and ensemble machine learning," *Machine Learning with Applications*, vol. 5, p. 100046, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666827021000232>
- [14] F. T. Liu, K. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Transactions on Knowledge Discovery From Data - TKDD*, vol. 6, pp. 1–39, 03 2012.
- [15] C. Toraman, Başarır, and M. Bayramoglu, "Determination of factors affecting the price of gold: A study of mgarch mode," *Business and Economics Research Journal*, vol. 2, pp. 37–50, 10 2011.
- [16] W. Jiang, "Applications of deep learning in stock market prediction: recent progress," 2020.
- [17] A. Hirs, J. Osterrieder, B. H. Misheva, W. Cao, Y. Fu, H. Sun, and K. W. Wong, "The vix index under scrutiny of machine learning techniques and neural networks," 2021.
- [18] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," 2015.
- [19] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11 1997. [Online]. Available: <https://doi.org/10.1162/neco.1997.9.8.1735>