Machine Learning Based Egg Supply Forecasting for Sorting and Grading Institutes

Avraham Cohen* and Shay Horovitz

Business & Computer Science Schools, College of Management Academic Studies, Rishon LeZion, Israel;
Email: horovitz@colman.ac.il (S.H.)
*Correspondence: avicohen466@gmail.com (A.C.)

***Abstract*—Animal agricultural productivity is highly influenced by the environment with many variables that affect it which makes forecasting a challenging task in this industry. We propose a new and practical approach forecasting egg-supply by utilizing Machine Learning (ML) algorithms, powered by limited data, which is regularly collected in this industry. The proposed approach does not require additional organizational resources for the purpose of collecting information but examines the behavior of farmers as expressed in the supply of eggs to the Egg Sorting Institute (ESI). We propose several possible models and present forecasts for egg supply with 90% accuracy, thus allowing both effective and economical operational decisions to be made.**

***Keywords*—egg, poultry, forecasting, machine learning**

I. Introduction

The egg industry has central planning and restrictive production quotas for supply. During the years 2010 to 2020, the production of edible eggs increased from 1.8 billion to 2.3 billion eggs, and the demands are steadily growing. Moreover, the requirement must be completed by importing eggs from abroad [1]. As can be studied from Fig. 1, egg demand changes over time and companies need to design an efficient supply chain that strikes a balance between customer demand and farmers' egg production [2].

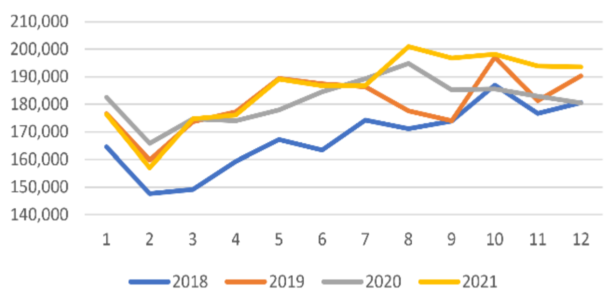

Figure 1. Monthly demand for eggs in Israel in years 2018–2021.

Agricultural supply chains differ in many aspects from its industrial counterpart, where consistent supply chain management at all levels of value creation is a common approach. Implementation in agricultural processes requires a rethinking of the supply chain concept since as it is affected by heuristic processes (decisions in different situations of uncertainty) and stochastic environmental conditions (depending on accidental factors) [3].

In the poultry industry, production problems cause an economic loss that can be reduced and even prevented by a timely action [4]. Predicting the egg production curve is a complex task, because the egg production curve (from week 20 to week 72) is characterized by nonlinear probabilities and the percentage of laying in the egg production process is unknown [5].

The main difficulty in mathematical or statistical models is due to the varying shapes and nature of the 3 phases of the egg production cycle. Fig. 2 depicts the average percentage curves of most US commercial egg-laying white-shelled layers. Factors contributing to curve variations include genetic potential, feed and feeding formulation, ambient temperature, light pattern, and myriad other factors. Such approaches are too complex, as in the case of a compartmentalization model or that require too many variables to determine production, as in the case of a stochastic model. Such models are impractical and do not correspond to the type of variables and production data in most commercial situations [6].

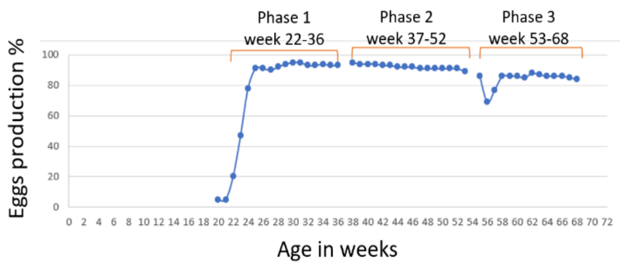

Figure 2. Phases of egg production curves of white-shelled laying hen.

The goal of Machine Learning algorithms is to optimize task performance through examples and past experience. Based on training data, in the learning phase, the machine learns to perform the task from experience. In this data-driven methodology, the more data used, the better machine learning performs. Once the learning performance reaches a sufficient accuracy level, it ends. After that, the model developed in the coaching process can be used for predictive purposes [7].

Manuscript received February 10, 2023; revised April 16, 2023; accepted July 19, 2023; published October 24, 2023.

II. Related Work

There are several studies that use ML to make predictions about egg production: A study conducted in Turkey examined the applications of fuzzy logic to predict egg production data. Egg production records were obtained from commercial poultry farms. In order to predict the quantity of eggs, the fuzzy logic model used several parameters such as the position of the cage (upper, middle or lower), age at sexual maturity, body weight at sexual maturity, body weight at mature age [5]. A study done in Brazil examined the application of an ANN model for predicting egg production performance in commercial laying breeder flocks. For this purpose, 26 characteristics of egg production were selected as input variables for the model [8]. From a study conducted in Japan with the aim of predicting the cumulative egg production of Japanese quail using linear regression, partial linear regression, and multivariate adaptive regression algorithms, included variables such as age at sexual maturity, weight at sexual maturity, average weight of the first ten eggs [9].

The above mentioned studies share the use of a large number of variables collected from each farmer to create an egg production forecast, and reached limited performance results.

III. Data Preprocessing

A. Data Collection

This study is based on data collected from the ERP system records of the Eggs Sorting Institute (ESI) "MIN HATEVA BEEROTAIM LTD" throughout the year 2020. The study collected data from 280 farmers who supplied eggs to the ESI along the year. The registration of the quantities of eggs sorted at the ESI is done automatically according to the owners of egg supply quotas as appears in the records of the "Chicken Coop Council". The farmers who supply eggs to the ESI are divided into two types, the first are farmers with quotas for growing laying eggs and they are the ones who grow the laying hens themselves. The second are farmers who grow laying hens, but they are not the owners of the quotas, rather, rent the growing quotas from the owners of the growing quotas, and grow all the ingots in concentrated chicken coops. The data on which the study is based on was collected by the ESI for regular reporting purposes and a required by the needs of the ESI accounting management. Data from farmers whose quotas were rented and raised together in concentrated chicken coops were inaccurate in associating the eggs with the quota holders. Due to this, it was decided to remove farmers who rented out their quotas from the database on which the model will operate. These farmers were manually screened in accordance with the information provided by the ESI.

B. Data Presentation

A significant feature that affects the behavior of the data is the way of transporting the eggs to the ESI. The eggs are transported in designated carts containing 144 egg molds, each mold containing 30 eggs, a total of 4320 eggs in each cart. Regularly, full egg carts are being transported to ESI, except in rare cases where a partial cart is sent due to the completion of laying or a large break of eggs in the cart (collapse of a shelf in the cart). That is why egg counts in deliveries are in multiples of 4320. Fig. 3 shows the distribution of the number of eggs shipped by farmers, in most shipments farmers send between 4320–8640 eggs, which is equivalent to 1–2 carts.

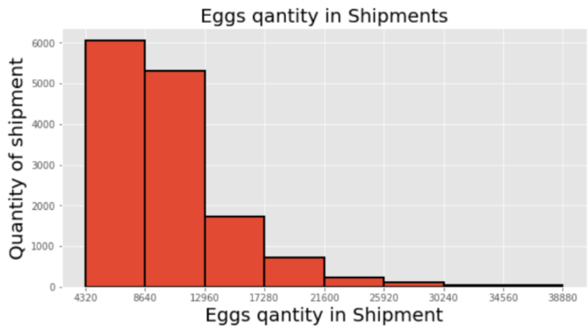

Figure 3. Distribution of the number of eggs sent by farmers.

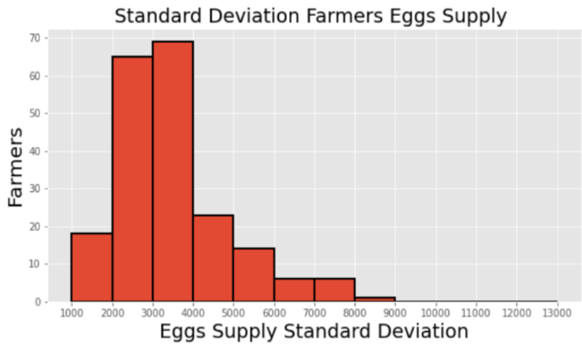

Figure 4. The distribution of standard deviation of eggs supply by farmers.

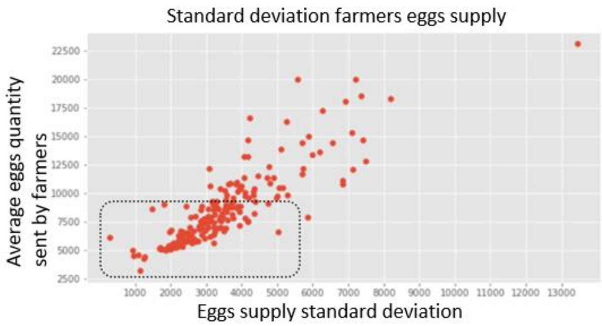

Figure 5. The average amount of eggs sent by a farmer in relation to his standard deviation.

In addition, the standard deviation between all shipments of each farmer was examined. Fig. 4 shows the farmers standard deviations distribution, for most farmers there is a fluctuation of 2000–4000 eggs between shipments, that is, the volatility of at least one cart.

From Figs. 3 and 4, it can be concluded that for most farmers the difference between deliveries can vary by 100% or even more. Fig. 5 shows the farmer standard deviation between deliveries relative to his average supply of eggs; It can be noticed (in the black rectangle) that most farmers who send 1–2 carts are in a standard deviation of 2000–4000 eggs. It means that most of farmers who send an average of 1–2 carts have 100% or even more difference between egg shipments.

The frequency of egg collection from farmers is planned for fixed dates. However, the pick-up dates can vary depending on the different needs of the farmers and the sorting institute. These changes can be caused due to lack of eggs at the institute, lack or excess space in the collection truck, and changing needs of the farmers. However, in any case, egg collection is no later than a week from the last collection. Fig. 6 shows the distribution of the difference in days between consecutive shipments. Notice that most often the day difference between shipments is 3–4 days. Differences over 8 days are usually the result of new growing cycles in farms.

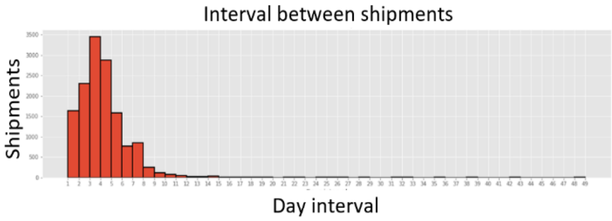

Figure 6. Distribution of the day difference between shipments.

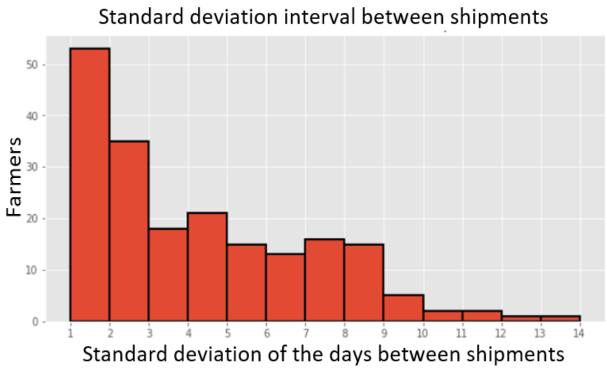

Figure 7. Variance in egg collection dates.

Fig. 7 shows the standard deviation in egg collection times for each farmer. For most farmers, the deviation is one day. Deviations over 9 days reflect differences between new growing cycles in the same farm.

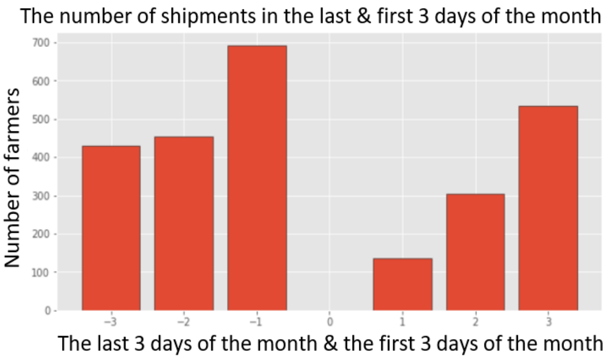

Figure 8. Distribution of shipments in the last three and first days of the month.

Fig. 8 shows the behavior of egg deliveries in the last three days and the first three days of the month. The number of egg senders increased in the last three days of the month compared to the first three days of the month. This anomaly is basically attributed to the date of the end of the month according to which the payment to the farmer is calculated and therefore great efforts are made by farmers to send as many eggs as possible.

C. Data Manipulation

The data collected from the sorting institute contain information of one set of data on the amount of egg supply and the dates of supply of each farmer. Data from one or more series of repeated observations on a single subject, wherefor the most part, each observation depends on its past observations and is especially suitable for models of time series. The ability to demonstrate accurate, and reliable relationships between variables that present sequential correlation is one of the main advantages of models [10]. Accordingly, Time series model models were selected as suitable models for testing the predictability of this data.

Raw data sets may contain inaccuracies, missing data, incorrect encoding, and other problems affecting data analysis. One of the biggest challenges is discovering and fixing dirty data. Failure to carry out this move can lead to inaccurate analysis and unforeseen conclusions [11]. Since the data on which the study is based were not collected for the purposes of the study, but for the purposes of reporting to the chicken coop council and the needs of the accounting management of the ESI, the need arose to adapt these data to the way machine learning is learned. To allow the model data depth for learning, the model sifted through data sets smaller than three months. Fig. 9 shows the distribution of farmers by the number of months of supply.

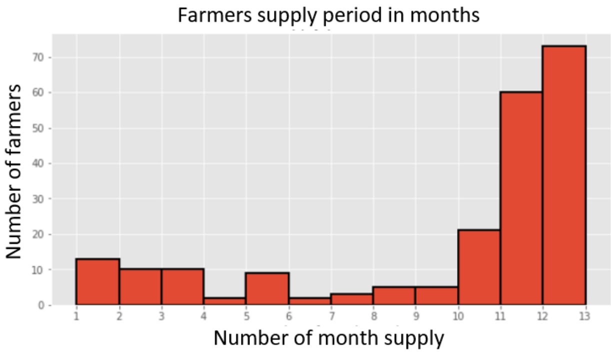

Figure 9. Distribution of farmers by the number of months of supply.

The examination of the data revealed several abnormal behaviors that do not represent typical egg-supplying behavior. For each of them, appropriate actions were performed to integrate them correctly into the data series for the model. Sometimes there are situations in which eggs are supplied two days in a row because of lack of space in the transport truck. The effect of this behavior on the data creates a situation of few eggs in one supply and immediately followed by a large supply of eggs. To cope with such situations, the shipments were consolidated to the date of the last shipment, thus bridging unusual supply behavior that could mislead the model, since the number of eggs was not supplied that day due to lack of space in the truck and not due to lack of eggs at the farmer's side. To maintain the veracity of the data, the shipments were consolidated to the last of the two shipment dates because it accurately expresses the number of eggs that would have been shipped had the truck not collected them on the first date.

The egg deliveries are carried out in the carts and only full carts are sent, but sometimes a malfunction occurs and some of the eggs in the cart break at the transport stage due to a technical malfunction (collapse of the egg shelf in the cart) or this is the last shipment of a farmer at the end of the pack, and he sends all the remaining quantity. These situations are unusual and do not reflect the normal behavior of the farmer. In these cases, a rounding of the number of eggs is performed into a whole cart or an empty cart to create stability in the series Data.

Fig. 10 shows, for example, the behavior of egg shipments sent by a farmer as collected by the Institute and after performing the preliminary actions described earlier. The data history for that farmer contains 85 shipments that sometimes included 8640 or 4320 eggs per shipment. Fig. 11 shows the distribution of the difference in days between the shipments of the same farmer. The distribution shows a range of two to seven days of supply difference between shipments.

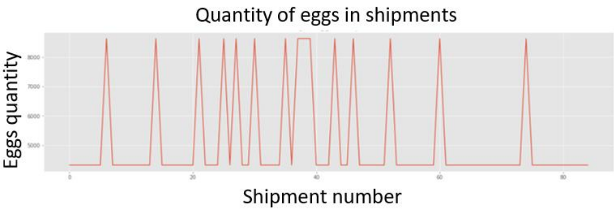

Figure 10. The number of eggs in each farmer's shipment.

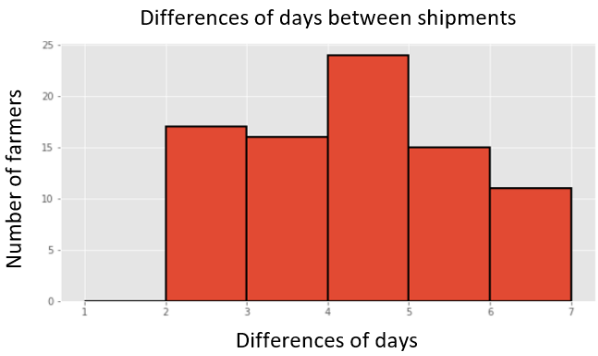

Figure 11. The distribution of differences between the days of supply of a farmer, for example.

Many models of time series assume that the observations follow a fixed time interval; However, real-world data may not necessarily meet this convention [10]. Since egg supply data at the Institute indicate different points in time (the variance of time differences between shipments is described in Fig. 7), it is necessary to create a data series at fixed intervals. Since the purpose of the study is to be able to predict the number of eggs in the next supply of a farmer according to daily accuracy, The series that best suits the needs of the study is a series in daily intervals. In the transition to a series of daily time differences, it is necessary to change the view from the number of eggs in the shipment to the number of eggs that theoretically was collected each day. The change in the point of view is necessary to estimate the number of eggs that can be shipped at any requested time if given the last pick-up date and the next pick-up date. Summarizing the entire number of eggs theoretically collected between these two dates will provide the number of eggs to be shipped. Two main steps were required to create such a series. One is to create a list of all the daily dates that are in the period that the farmer supplied eggs to ESI, where the number of days between two consecutive shipments will be calculated as the time window for collecting eggs for the last shipment of them. The second is to fill these dates with a theoretical number of eggs collected where the sum of this quantity between the delivery dates (shipping time window) will be equal to the number of eggs in the original shipment. This calculation is done by calculating the average of eggs for the difference of days between shipments. The number of eggs in each shipment was divided as an average the entire number of days preceding the shipment including the day of delivery itself. The result of this calculation is shown in Fig. 12, which shows the number of eggs given in Fig. 10, but in a daily distribution over a sequence of 330 days. The figure shows the time windows (the time intervals between shipments) and the average amount of eggs for that period. Organizing the data in this way created a step-like behavior of the data between the time windows. That is, sharp transitions between a state of static within the time window and a sharp decrease or increase in the transition to the next time window.

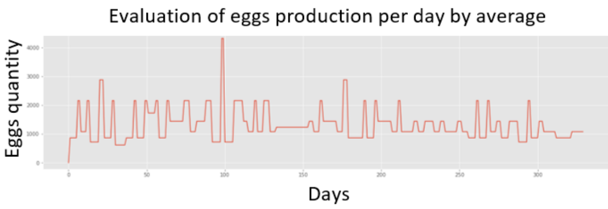

Figure 12. The number of eggs collected by the farmer each day according to daily average between deliveries.

In the data collected from the ESI, there are farmers whose egg supply sequence is interrupted followed by continued supply. That is, a period of 14–30 days when no eggs are supplied to ESI at all. An interruption in the supply sequence can be due to two main reasons, one is the dropout of a laying flock which lasts about ten days to two weeks (an explanation of the dropout period appears at the beginning of the article), the second is the end of the breeding cycle of an existing flock and the start of breeding of a new flock, usually after two weeks to a month. To maintain the continuity of a stable data series that characterizes the normal behavior of the data, these periods of time (with no supply) were removed from the data series of farmers.

As described earlier, time series models operate on time series in which each observation depends on the previous observations, but in organizing the data as averages within the time window of delivery, the principle of dependence of one variable on another is absent. To solve this problem, another change was made in the arrangement of the data, which is the smoothing of the data values close to the limits of the time window. That is, instead of all the data in the time window being equal, the data values of the number of eggs collection found in each of the time window boundaries were changed towards the values appearing in the adjacent time window on each side. This change, of course, was made without affecting the original number of eggs collected in that window, that is, without affecting the number of eggs sent to the ESI. For this purpose, the reduced/added number of eggs at each end of the window was moved to the center of the window. For example, if the time window is 4 days and it is necessary to reduce the number of eggs by 20% at each of the endpoints of the time window of the delivery. the result will be a transition from a division of day1 = 25%, day2 = 25%, day3 = 25%, day4=25%, of the total quantity of eggs in the original shipment, to a division of day1 = 20%, day2 = 30%, day3 = 30%, day4 = 20% of the total quantity of the shipment for the 4 days. Fig. 13 shows the result of this change, it is possible to notice a wavy behavior that expresses the dependence of each observation on the previous one.

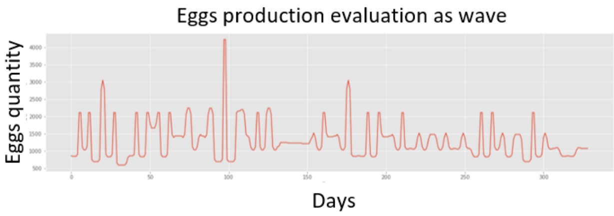

Figure 13. Wave behavior of the number of eggs collected by the farmer each day according to daily average between deliveries.

IV. Results

For examining the research hypothesis, three different models ARIMA, PROPHET, LSTM (30 LSTM units in the hidden layer, output layer of single numerical value, batch size of 7 shipments) were chosen to make predictions according to time series. ARIMA, Auto-Regressive Integrated Moving Average, is an automatic time series technique that calculates short-term future forecasts from time series analysis of historical data [12]. PROPHET, a modular regression model for time series predictions with high accuracy by using simple interpretable parameters that consider the effect of custom seasonality and holidays [13]. LSTM, Long Short Term Memory, model based on an artificial neural network to produce the internal representation of time series data [14].

The models were required to predict the number of future eggs supply to ESI on each of the days of December 2020. The prediction was made for those farmers who supplied eggs to the ESI on each of the days. The verification of the prediction results is done by comparing to eggs supplies that were made to the ESI at that day. Fig. 14 shows the prediction results values for the three models for each of the days of December 2020. This figure shows that the behavior of the shipments is very jumpy, and all three models were able to follow these changes. However, the results of the ARIMA model behaved the closest to the actual results.

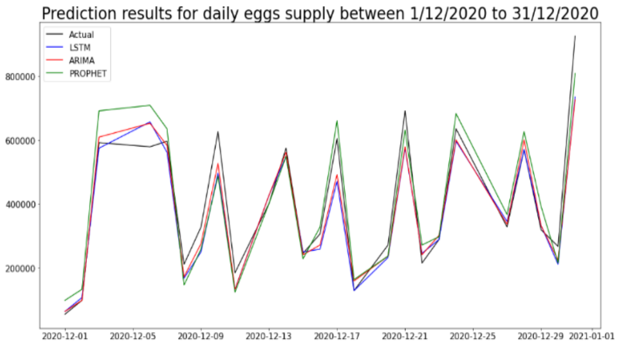

Figure 14. Presentation of the egg supply forecast each day throughout the month of December 2020.

To gain a better understanding of the prediction results, the performance of the models was examined in the R2, MAPE, MSE, MAE indices and their results are shown in Table I.

TABLE I. Comparison of Model Performance According to R2, MSE, MAE, MAPE

	LSTM	ARIMA	PROPHET
R2	0.902	0.917	0.909
MAPE	0.117	0.11	0.181
MSE	4557358080	3865356288	4218448896
MAE	47520	43545	55123

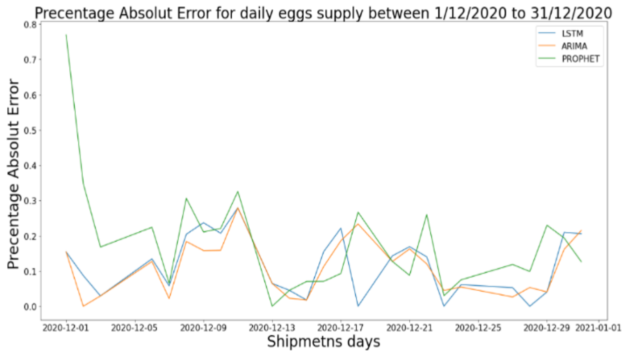

Figure 15. Comparison of the percentage of absolute prediction error on each day for the 3 models – LSTM, ARIMA, PROPHET.

To evaluate the predictive ability of the three models, the coefficient of determination R2 was used. The three models showed prediction results greater than 0.9. It means that past data of egg supply to the ESI (predictor variables), manage to explain 90% of the variance in the prediction results of future egg supply. MAPE is an important measure, because the variable's units are scaled to percentage units, which makes it easier to understand from a business point of view. ARIMA and LSTM get less than 12% MAPE for the forecast results. This means that according to this measure, the prediction average accuracy is 88% for LSTM, 89% for ARIMA, and 82% for PROPHET. But to examine the performance of the models the MAPE measure is not enough, the absolute errors must be examined every day and all so their standard deviation. Fig. 15 was created for this purpose, it shows the absolute percentage error for each day. All so, the standard deviation for absolute percentage error was calculated for each model, LSTM STD = 8.3%, ARIMA STD = 7.6%, and PROPHET STD = 15.3%. ARIMA shows the lowest forecast absolute percentage error line and the most stable standard deviation.

To complete an investigation, it remains to examine the distribution of error percentages for each model to rule out an unbalanced distribution that could indicate model underfitting. Figs. 16–18 shows the distribution of the percentage of errors. In all models, the errors are distributed almost equally between the positive and negative part of the graph. In addition, these graphs indicate a normal distribution around 0 of the error percentage. This means that the models are not underfitting.

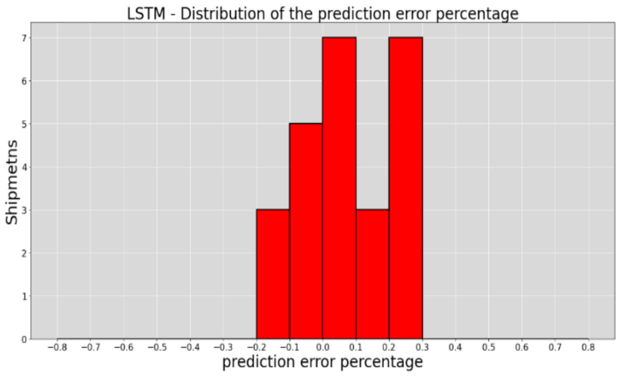

Figure 16. LSTM-Distribution of the prediction error percentage for daily eggs supply between 1/12/2020 to 31/12/2020.

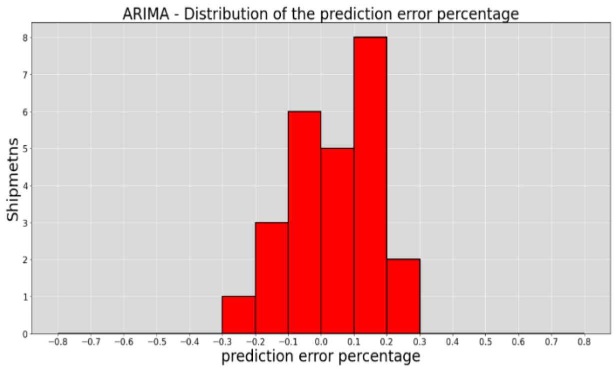

Figure 17. ARIMA-Distribution of the prediction error percentage for daily eggs supply between 1/12/2020 to 31/12/2020.

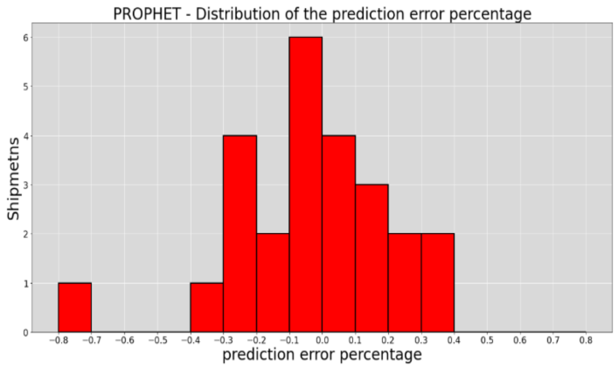

Figure 18. PROPHET-Distribution of the prediction error percentage for daily eggs supply between 1/12/2020 to 31/12/2020.

MSE and MAE are mainly used to compare performance between the models. From all measures, it can be seen that the ARIMA model produced the best results with the lowest prediction error.

V. Conclusion

In this study, the ability to predict egg supply to ESI was tested by using ML and ANN models operating on time series. Three models, LSTM, ARIMA, and PROPHET were chosen to predict the future daily supply of eggs to the ESI for an entire month (December 2020). The models were fed with historical supply data of 213 farmers who supplied eggs to the ESI in the 6–12 months preceding the forecast date. The data included the farmer's code, the number of eggs in each delivery and delivery dates only. This data was collected by the ERP system of ESI.

All three models showed prediction results of R2 greater than 0.9, the ARIMA model presented the highest level of prediction R2 = 0.917. In addition to this, ARIMA presented the lowest error in MAPE measure, MAPE = 0.11 and the lowest prediction error in MSE and MAE measures. Predicting egg supply with an accuracy level of 90% by R2 or 88%–89% by MAPE, is excellent for making more efficient (and economical) operational decisions of an ESI. This study shows that it is possible to combine ML and ANN models for the purpose of predicting daily agricultural supplies based on historical time series data in a relatively simple way that does not require many resources and without the need to collect additional data.

Future research based on historical data organized as time series can be extended to other branches of animal agriculture. Decisions in uncertainty is one of the main issues of the agricultural sector, accurate and easy-to-implement forecasting tools can greatly improve the management of agricultural production. Research analysis shows that weather variables such as rainfall and temperature significantly affect the productive output of animals [15]. We wish to thank the Research Authority Fund of the College of Management Academic Studies, Rishon Le'Zion, Israel and the Schools of Business Administration & Data Science, for funding & supporting the study leading to this publication.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Avi Cohen wrote the paper; Shay Horowitz is a supervisor to this paper; all authors had approved the final version.

References

[1] Knesset, Research and Information Center, Budget Control Department. (2022). Description and analysis of the broiler and egg. [Online]. Available: https://fs.knesset.gov.il/globaldocs/MMM/a04e595b-122f-ec11-8140-00155d0401c3/2_a04e595b-122f-ec11-8140-00155d0401c3_11_19556.pdf

[2] K. Sethanan, T. Jamrus, and P. Boonthavee, "Solving production scheduling of egg production in hen farms with capacity of slaughterhouse by heuristic algorithms," *International Journal of Engineering and Technology*, vol. 5, no. 1, p. 50, 2013.
[3] A. T. Braun, E. Colangelo, and T. Steckel, "Farming in the era of Industrie 4.0," *Procedia Cirp*, vol. 72, pp. 979–984, 2018.
[4] I. R. Morales, D. R. Cebrián, E. F. Blanco, *et al.*, "Early warning in egg production curves from commercial hens: A SVM approach," *Computers and Electronics in Agriculture*, vol. 121, pp. 169–179, 2016.
[5] T. G. Omomule, O. O. Ajayi, and A. O. Orogun, "Fuzzy prediction and pattern analysis of poultry egg production," *Computers and Electronics in Agriculture*, vol. 171, 105301, 2020.
[6] H. A. Ahmad, "Egg production forecasting: Determining efficient modeling approaches," *Journal of Applied Poultry Research*, vol. 20, pp. 463–473, 2011.
[7] L. Benos, A. C. Tagarakis, G. Dolias, *et al.*, "Machine learning in agriculture: A comprehensive updated review," *Sensors*, vol. 21, 3758, 2021.
[8] E. B. Oliveira, L. G. B. D. Almeida, D. T. D. Rocha, *et al.*, "Artificial neural networks to predict egg-production traits in commercial laying breeder hens," *Brazilian Journal of Poultry Science*, vol. 24, 2022.
[9] Y. Altay, "Prediction of cumulative egg production in Japanese quails by using linear regression, linear piecewise regression and MARS algorithm," *Large Animal Review*, vol. 28, pp. 93–99, 2022.
[10] S. M. Lavrenz, E. I. Vlahogianni, K. Gkritza, *et al.*, "Time series modeling in traffic safety research," *Accident Analysis & Prevention*, vol. 117, pp. 368–380, 2018.
[11] M. Z. H. Jesmeen, J. Hossen, S. Sayeed, *et al.*, "A survey on cleaning dirty data using machine learning paradigm for big data analytics," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 10, pp. 1234–1243, 2018.
[12] H. Alabdulrazzaq, M. N. Alenezi, Y. Rawajfih, *et al.*, "On the accuracy of ARIMA based prediction of COVID-19 spread," *Results in Physics*, vol. 27, 104509, 2021.
[13] Z. Z. Oo and P. H. Y. U. Sabai, "Time series prediction based on Facebook prophet: A case study, temperature forecasting in Myintkyina," *International Journal of Applied Mathematics Electronics and Computers*, vol. 8, pp. 263–267, 2020.
[14] I. E. Livieris, E. Pintelas, and P. Pintelas, "A CNN-LSTM model for gold price time-series forecasting," *Neural Computing and Applications*, vol. 32, pp. 17351–17360, 2020.
[15] S. Kumar, B. M. K. Raju, C. A. Ramarao, *et al.*, "Sensitivity of livestock production to climatic variability under Indian drylands and future perspective," *Current Agriculture Research Journal*, vol. 3, pp. 142–149, 2015.