Future of Food Pricing: Predicting PPI Trends using Machine Learning and Deep Learning Models

Shawana Maliha
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
shawana.maliha@g.bracu.ac.bd

Tasnim Zaman
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
tasnim.zaman3@g.bracu.ac.bd

Zarin Tasnim
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
zarin.tasnim22@g.bracu.ac.bd

***Abstract*—Commodity price hikes remain one of the significant problems in the general people's life. Our study helps to predict the food prices for future years using the historical data among 45 countries in Asia from the Food and Agriculture Organization (FAO). To predict food prices, in this study, we used various machine learning and deep learning models such as Random Forest, Decision Tree, Linear Regression, LightGBM, XGBoost and long-short-term memory(LSTM). We used the R^2, MSE, RMSE, and MAE values to compare between them to evaluate the models. Among them, XGBoost gives the highest R^2 score (0.9991) and predicts the price of 2024 most effectively. Then XGBoost and LSTM were applied to predict the price of 2025. This study suggests that machine learning models, especially gradient boosting, can predict the values more accurately than deep learning models.**

***Index Terms*—Commodity price, Machine Learning, Feature importance, Linear Regression, Decision Tree, Random Forest, LightGBM, XGBoost, LSTM.**

I. Introduction

Commodity price is a key part to measure the economic stability of a country, particularly among Asian countries. A certain amount of people depend on commodities for their livelihood. Sudden increase in the price of a commodity can affect the people with low income. In 2022, approximately 2.4 billion people or 29.6% of the global population, were moderately or severely food insecure [1]. Bangladesh still faces unpredictable price hikes which makes it difficult to afford the essential foods. Hence, an accurate forecasting of price is very crucial for a country for their economic development, stability and market risk management.

To solve this issue, predicting food prices can help the consumers to prepare better and take necessary steps to avoid unexpected situations. In this study, across 47 counties in Asia, we forecast commodity prices using the historical data that was obtained from the Food and Agriculture Organization (FAO). The data contains prices from the year 1991 to 2024. We applied several well known machine learning and deep learning models to predict the future prices and analyze which models give accurate results. Models including Random Forest, Decision Tree, Linear Regression, LightGBM, XGBoost, and Long Short-Term Memory (LSTM) network were estimated based on the known prices of commodities. These approaches can identify the incoming risk in a sudden price hike by forecasting the price of the year 2025.

II. Literature Review

In our research, we tried to focus on the economic loss faced by the Asia subcontinent due to food price uncertainty. Every year, people in this area face financial disability, which creates economic unrest. To address this problem, there has been significant research work in this field.

In a related paper, the work by Sapakova et al.(2023) [2] used a dataset from the World Food Programme, which predicted food prices in Kazakhstan. They applied machine learning models like Decision Tree, Random Forest, and Gradient Boosting regression models. The dataset consists of prices of various food items like rice, beans ,fish and sugar across different places. After pre-processing, the categorical values were converted into numerical values. The models were evaluated using metriceslike RMSE, MSE, MAE and R^2. Among all the ML models, random forest outperformed all other models by achieving an R^2 value of 0.99 which indicates high predictive accuracy.

In another paper by Arya et al.(2025) [3] used nine machine learning models and a deep learning model named LSTM to forecast rice production. They used 62 years of data across 192 countries. Among the models,Linear Regression R^2 = 0.9819 outperformed LSTM R^2 = 0.9819. Their study portrayed Linear Regression as appropriate for temporal learning, whereas LSTM showed results to be more robust.

A different approach has been taken by Desai et al.(2023) [4] where Facebook Prophet was used to predict wheat crop yields in India, using historical data from 1997-2022 across four districts. To evaluate error, they used MAPE with a result of 10.03 and an RMSE of 0.39, which indicates strong predictive performance. Their study also showed that compared to traditional ML techniques, PROPHET showed excellence in forecasting prices, especially when the given data has an information shortage.

In addition, Azkaenza et al. [5] shared a different perspective by targeting food inflation trends across Indonesia using advanced ML models specially XGBoost and LSTM using a dataset from (PHIPS). Traditional models like SARIMA could not perform well compared to XGBOOST and LSTM. This model handled sudden price fluctuations . features like rice price index and exponential moving averages improved the model's responsiveness.

Kupferschmidt et al. [6] conducted a deep study to forecast food prices in CANADA by using ML models. Using Canada's food price report(CFPR), they went along with a hybrid approach and used indices like Food manufacturing price index and Food Commodity Price Index. Here they used LLMs to evaluate their performance in predicting food price fluctuations. In another relevant study, Chitikela et al. [7] tried to forecast Indian food grain production where they used both traditional time series methods and machine learning models.They comparaded between 4 models named ARIMA, Interrupted ARIMA, artificial neural network(ANN) and interrupted ANN. Among all, the interrupted ann model outperformed others by reducing error up to 99.06%.

In a comparative study, Mencuilini et al. [8] tried to forecast wholesale food prices using models like ARIMA, Facebook Prophet and deep learning methods like LSTM along with CNN. While researching, they found that ARIMA and LSTM models showed comparative forecasting accuracy whereas CNN-LSTM models achieved superior performance.

Diwane et al.(2024) [9] conducted research by evaluating the performance of Arima Sarima and the facebook prophet to forecast the demand of the food industry. To evaluate them, they used the model's accuracy method MAPE where Facebook prophet achieved 6.5% outperforming Arima and Sarima as they scored MAPE value as 8.2% and 7.9% respectively. Here prophet outperformed for its ability to automatically detect changes which made prophet a robust tool for forecasting demand in the food industry.

Lastly, Tami et al. [10] explored an LSTM based model to predict the prices of daily used food commodities including bread, meat,oil etc.The model achieved a strong predictive performance which scored MAPE of 3.04%, RMSE of 0.14 and R^2 of 98%. This model outperformed traditional ML models with perfect accuracy in producing reliable commodity price forecasts.

The findings indicate that machine learning can be a powerful tool for predicting food prices across Asia, even when working with incomplete data. Among the models tested, XGBoost delivered the most accurate results, making it especially useful for real-world forecasting. By comparing multiple methods from simple regressors to deep learning, we identified what works best in large, messy datasets like those from FAO.

This work holds uniqueness as it did not focus on one country or product. Instead, it looked across a subcontinent consisting of 45 countries,offering insights that are more regionally relevant. The result gave us a comparison of food expenses in different countries, which is also helpful to understand and show the economic distribution. Our goal was to support better planning and decision-making using data. We hope this research can help guide future efforts in building more stable and food-secure systems.

III. Methodology

A. Dataset Description

The dataset used in this study was sourced from the Food and Agriculture Organization (FAO) which is publicly available for analysis. It contains commodity price data collected from various Asian countries between 1991 and 2024. The dataset includes a wide range of food items and related economic indicators.

In the initial phase of working with our datasets, we discovered a wide amount of null numerical inputs across different time periods, particularly in the year 2024. The null values were more than 2000 for all the years before the year 2001. Since we wanted to keep all the years in count to get a better trained model, we didn't want to drop our columns necessary for future prediction. So, the initial dataset was filtered to focus on the "Producer Price Index (2014–2016 = 100)" as the primary variable of interest to ensure consistency and maintain the focus on the relevant data. After filtering, the dataset was reduced to 3,876 rows and 37 columns from 28,422 rows and 45 columns. It became evidently suitable with almost less than 10 null values for all last 10 years.

"Producer Price Index (2014–2016 = 100)" reflects the average change in prices received by producers for their goods, normalized to the 2014–2016 period, and is crucial for analyzing price trends across different Asian markets. The dataset also contains various characteristics such as Area, Item, Year, and Value for each country and food item. It has 45 unique Areas representing different Asian countries or regions and 216 unique Items representing various food products analyzed.

B. Dataset Preprocessing

The dataset underwent a detailed preprocessing to ensure quality and compatibility with machine learning models. The raw data, which was rich in information, contained multiple missing values, redundant features, and noise that needed to be addressed before proceeding to model training and evaluation.

A significant number of columns were metadata fields, such as codes for Area, Item, Element, and Months, which did not affect directly to the analysis were removed. We initially had 45 features, and it was reduced to 37 features after removing redundant ones. To handle missing values, we filled them using

row-wise interpolation across years. This ensured that all data points remained numerically valid and consistent for training without introducing significant bias.

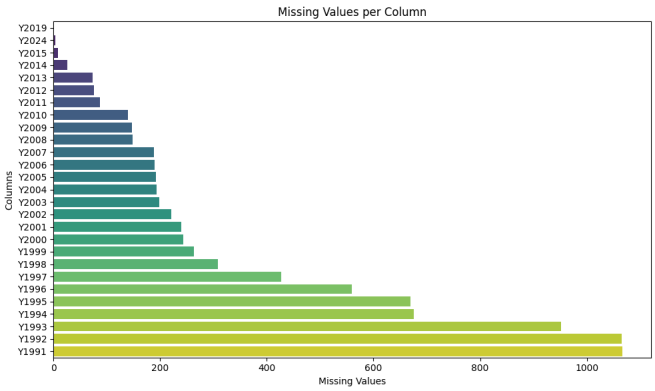

Fig. 1. **Distribution of missing values across different years**

Furthermore, categorical data such as 'area' and 'item' were encoded into a numerical format using label encoding techniques. After preprocessing, the final dataset was prepared with only relevant and meaningful features. The cleaned and structured dataset allowed effective training of machine learning and deep learning models to predict food prices in Asian countries.

C. Learning Phase

During the learning phase of the Producer Price Index for food items across Asia, we started off with the comprehensive dataset of 3876 rows and 37 column values and prepared for testing with multiple machine learning and deep learning models. We wanted to find out the best suited model to forecast for the year 2025.

a) Random forest (RF): It is one of the robust used models to handle nonlinear types of data and has a good resistance to overfitting. Since our dataset contains a mixture of categorical and numerical features, we chose this non-parametric model handling for all the countries in Asia and its division of different food items along it. The random forest starts by building multiple decision trees which it bags later on to train each tree on random subsets allocation and chooses for the best fitted prediction that it considers through traversing on trees.

b) Decision Tree Regressor (DT): The decision tree is a basic yet powerful tool that learns from recursively splitting the data on various feature sets. It has better flexibility than many other models and is better interpretable to demonstrate the price drivers to the policymakers due to the transparency it can provide. Out of different ways to select the best feature for each node, Information gain and Gini impurity approach are chosen by the majority.

c) Linear Regression (LR): It builds a linear communication lineup between the dataset feature characteristics with the target feature. We chose linear regression to get the idea of how much linearity in the price scale differences is impacting from years longing. This can also be considered to be a standard scale in price based on which plans for balancing the prices can be made easily. It works by minimizing the sum of squared errors.Prediction along with precision is optimized by comparing and analyzing using the efficient methodology [11].

d) LightGBM (LGB): This Gradient Boosting method is a special type of machine learning model which is utilized to work faster with better accuracy. It has its own style of building and training decision trees and making it learn the patterns in the dataset. The speeding up is ensured with using histogram-based splitting mechanism. The model automatically detects the categorical features and encodes accordingly as well as supports missing data making the learning process proceed in a smoother way. We chose this in our data training expecting higher accuracy with fewer trees with less overfit results.

e) XGBoost (XGB): It has even a higher tendency to reduce overfitting for it does experimentation in L1-L2 regularization of introducing penalties and forms a strong predictor combining some existing weak learners. Different tabular types of data give a precise result with this extreme gradient boosting [12]. It spreads the data train segment in parallel computation due to which a big chunk of data in the dataset can work fine in a minimal time. The cache memory is collected and the similarity points and results obtained are measured which is a faster approach than collecting from the main memory system which makes the boosting system very catchy and unique to use in our big sized datasets.

f) Long Short-Term Memory (LSTM): It is a form of Recurrent Neural Network (RNN) which works upon time series data elements by remembering the long span of time over years. This algorithm would fit the most on our data set since it stacked LSTM layers traversing dense layers which follows a procedural sequence getting habituated with its pattern inputs gradually and can achieve perfection in learning our 34-year price sequences.

Learning about the different functionalities of a number of models, these six models were decided to train and visualize their performance capabilities individually.

IV. Result and Discussion

In this paper, we will discuss the result of the machine learning models that we applied and the challenges we encountered. After filtering our dataset, to remove the extensive null values, we got a more reliable dataset to work with. With this improved dataset, we were able to train and evaluate our models effectively. The following subsection explains which features had more influence and how well each of our models performed.

A. Feature Importance

While researching, we felt the need to understand which historical years contribute the most, predicting values for 2025. To understand the relative impact of features like historical year-wise values, we applied three widely used methods of feature importance analysis, which gave us a clear indication of the most impactful feature. The methods that we used to analyze are: XGBOOST's built-in frequency(weight) based importance, Gain-based tree importance and SHAP value-based global explanations. We performed the methods on the XGBOOST model since it gave us the highest accuracy while predicting the price.

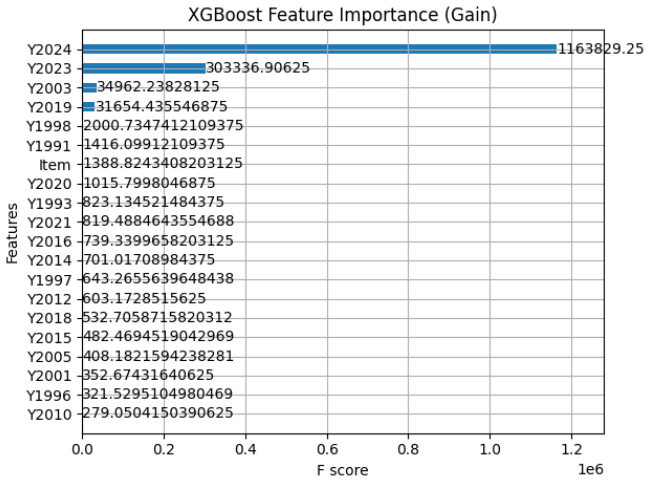

Fig. 2. **XGBOOST feature importance ranked by gain**

At first, we used the `feature_importances` attribute from our trained XGBOOST regressor model. This method calculates how frequently each feature is used in the decision tree splits. The resulting horizontal bar chart shows that years like (2024,2023) appear frequently in decision splits, which hold a significant amount of importance in the model's decision-making process.

We applied another method that uses Gain-Based Importance. This method calculates the gain(improvement in model performance) when a particular feature is used in a split. It prioritizes features that add the most predictive value and make it suitable for model optimization. We can observe from beneath Fig. that this method highlighted the year 2023 and 2024 as the most influential features in reducing model error.

To analyze both global and individual-level feature contributions and to dig even deeper, we used SHAP(Shapley Additive explanations). This method considers both how much a year matters, along with why a prediction is higher or lower for each country and food item. The summary plot displayed that: higher values in recent years (2022-2024) contributed positively to the 2025 food cost predictions, while older years like (2000-2005) often had a neutral effect. It also helped us to identify how each row was affected by its past values, which is not possible in gain or frequency-based methods.

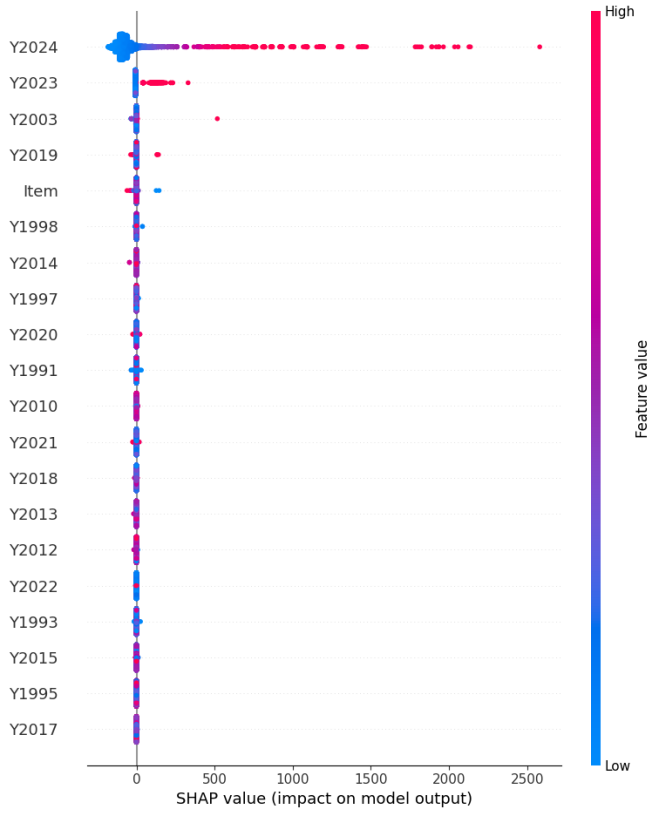

Fig. 3. **SHAP summary plot for XGBOOST feature contributions**

B. Performance Metrics

Going ahead with the training process with different ML models and time series analysis, we discovered precise results with minimum faults for some models giving an unexpected outcome. Our goal was to identify which model could work most efficiently with the historical dataset that we have fed into the training. We completed our evaluation with model performance metrics MSE, RMSE, MAE, and R^2 score which describe improvement factors of various models.

TABLE I
PERFORMANCE COMPARISON OF MACHINE LEARNING AND DEEP LEARNING MODELS

Model	MSE	RMSE	MAE	R^2 Score
Random Forest	2316.64	48.13	16.14	0.8858
Decision Tree	4162.34	64.52	23.90	0.7948
Linear Regression	2245.02	47.38	14.03	0.8893
LightGBM	7479.20	86.48	21.59	0.6313
XGBoost	72.45	8.51	2.12	0.9991
LSTM	0.0021	0.1688	0.0258	0.8627

Following up to the score, a diversion and mismatch is seen according to the type score that is calculated. While MSE is used in training loss calculation, here it is observed the squared error rose very high for the Decision Tree which

is 4162.34 following are Random forest, Linear Regression, LightGBM, XGBOOST which are 2316.14, 2245.02, 7279.2, 72.45 whereas LSTM having the least MSE score 0.0021 showing a big difference which arose due to higher errors being penalized more. With this, it is quite clear the Decision Tree had a very poor result for not being able to deal with the complex dataset having noisy outliners, whereas LSTM easily succeeded in optimizing weights with backpropagating. RMSE and MAE almost had a similar ranking where LSTM stood out with (RMSE=0.1688, MAE=0.0258) and Decision Tree and LightGBM couldn't due to their high error scores.

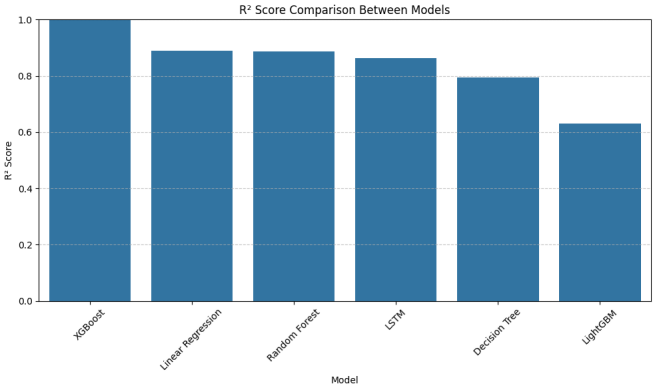

Fig. 4. **R² score comparison between the models**

Above all calculations, R² Score value is given the greater importance since it does overall model fit calculation considering all the factors together. In terms of R² score, clearly, XGBoost had the most accurate percentage of about 99. 91%, which gave almost the correct values to predict the price values of the test data. Following that other regressor models Linear Regression, Random Forest and time series model LSTM had done an outstanding performance in learning and utilizing the patterns of the data tuning well.

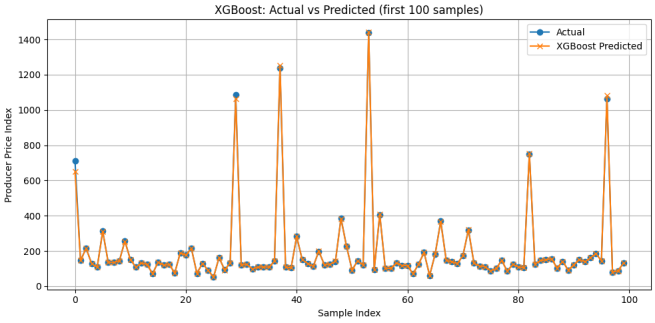

Fig. 5. **XGBoost: Actual vs. Predicted: (First 100 samples)**

From the results, it is evident that the most scalable implementation in our data learning process was done by XGBoost, being able to capture non-linear relationships and learn feature interactions in a shorter time and LSTM outperform significantly across the accuracy parameter. Therefore, we decided to give preference to the XGBoost or LSTM based approach to predict the most accurate price value for the present year 2025 after choosing it as the most probable prediction set of values with the right feedback model from all.

V. Conclusion

In this study, we aimed to forecast food prices among Asian countries using machine learning and deep learning techniques. Despite all the limitations and challenges that we faced, our model gave a robust forecasting that could be valuable for the real life applications. Our model can help to provide early warnings and help stabilize food supply chains across vulnerable regions. As food pricing directly affects millions of lives, our model can help make timely predictions that will make people ensure better planning and reduce food insecurity. In the future , we can work with a more consistent dataset that will help us to stabilize the market and save people from sudden price shocks.

References

[1] FAO, IFAD, UNICEF, WFP, and WHO, *The State of Food Security and Nutrition in the World 2023: Urbanization, agrifood systems transformation and healthy diets across the rural–urban continuum*. Rome: FAO, 2023. [Online]. Available: https://doi.org/10.4060/cc3017en

[2] S. Sapakova, A. Sapakov, N. Madinesh, A. Almisreb, and M. Dauletbek, "Using Machine Learning to Predict Food Prices in Kazakhstan," in *DTESI (Workshops, Short Papers)*, 2023. [Online]. Available: https://ceur-ws.org/Vol-3680/S3Paper9.pdf

[3] S. Arya and N. A. R. Anju, "Prediction of international rice production using long short-term memory and machine learning models," *Int. J. Inf. Commun. Technol.*, vol. 2252, no. 8776, pp. 8776, [Online]. Available: https://www.researchgate.net/publication/387321409_Prediction_of_international_rice_production_using_long-short-_term_memory_and_machine_learning_models

[4] M. Desai and A. Shingala, "Time Series Prediction of Wheat Crop based on FB Prophet Forecast Framework," in *ITM Web of Conferences*, vol. 53, p. 02014, 2023. [Online]. Available: https://www.itm-conferences.org/articles/itmconf/abs/2023/03/itmconf_icdsia2023_02014/itmconf_icdsia2023_02014.html

[5] M. Azkaenza, "Employing Machine Learning Techniques to Forecast Food Inflation in Indonesia: The Role of Food Price Indices," *CBS Research Portal*, May 2023. [Online]. Available: https://research-api.cbs.dk/ws/portalfiles/portal/98733666/1584520_Employing_Machine_Learning_Techniques_to_Forecast_Food_Inflation_in_Indonesia_The_Role_of_Food_Price_Indices.pdf

[6] K. L. Kupferschmidt et al., "Food for thought: How can machine learning help better predict and understand changes in food prices?," *arXiv preprint*

arXiv:2412.06472, 2024. [Online]. Available: https://arxiv.org/abs/2412.06472

[7] G. Chitikela, S. Rathod, and S. Vijayakumar, "Change point-driven interrupted time series and machine learning models for forecasting Indian food grain production," *Discov. Food*, vol. 5, p. 68, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s44187-025-00350-5

[8] L. Menculini et al., "Comparing prophet and deep learning to ARIMA in forecasting wholesale food prices," *Forecasting*, vol. 3, no. 3, pp. 644–662, 2021. [Online]. Available: https://www.mdpi.com/2571-9394/3/3/40

[9] R. B. Diwane, K. S. Oza, A. M. Chougule, and P. G. Naik, "Food Industry Analytics and Forecasting: Utilising Time Series Models for Improved Decision-Making," [Online]. Available: https://www.researchgate.net/publication/388406748_Food_Industry_Analytics_and_Forecasting_Utilising_Time_Series_Models_for_Improved_Decision-Making

[10] M. Tami and A. Y. Owda, "Efficient commodity price forecasting using a long short-term memory model," *Int. J. Artif. Intell.*, vol. 2252, no. 8938, 2024. [Online]. Available: https://pdfs.semanticscholar.org/fabe/35eee1e3cba552a68b8930c0fb3e8f62bb8c.pdf

[11] D. Maulud and A. M. Abdulazeez, "A review on linear regression comprehensive in machine learning," *J. Appl. Sci. Technol. Trends*, vol. 1, no. 2, pp. 140–147, 2020. [Online]. Available: https://jastt.org/index.php/jasttpath/article/view/57

[12] O. M'hamdi et al., "A comparative analysis of XGBoost and neural network models for predicting some tomato fruit quality traits from environmental and meteorological data," *Plants*, vol. 13, no. 5, p. 746, 2024. [Online]. Available: https://www.mdpi.com/2223-7747/13/5/746