CROP YIELD PREDICTION USING ENSEMBLE MODEL

Hemant kumar singh, Sailesh Krishnan
hemant.singh5@mail.dcu.ie, sailesh.krishnan2@mail.dcu.ie

I. Introduction

Crop yield prediction involves estimating the expected amount of crop production from a particular piece of land. It is crucial for farmers, policymakers, and the world's population. There are various methods for predicting crop yield worldwide, including manual field surveys and sophisticated machine-learning algorithms. Combining machine learning models with other technologies can provide a practical and economical way to predict and analyze crop yield production, which is essential for sustainable development.

Traditionally, crop yield prediction has relied on statistical models considering a limited set of variables such as weather conditions, soil characteristics, and crop type. However, with the emergence of machine learning models, structured and unstructured data from various sources can be integrated to make more precise and timely predictions. This literature review aims to provide an overview of recent research on crop yield prediction using machine learning models, explicitly emphasizing integrating structured and unstructured data and incorporating predictions from other sources, such as simulation models, as input for machine learning models.

II. Structured and Unstructured Data in Prediction Models

In addition to traditional approaches based on field observations and fixed datasets, ML experts are exploring to combine structured and unstructured data sources to better predict the amount of yield. Structured data refers to data organized in a predefined format, such as spreadsheets, databases, and tables. However, in many cases, relevant data may also be available in unstructured formats such as text, images, and audio recordings - which could be leveraged to gain additional insights and factors related to various environmental changes and the impact this will have on crop yield. Other sectors have already explored integrating structured and unstructured data for yield predictions. For instance, in a study by [1], machine learning models were used to predict bloodstream infections among children by integrating structured and unstructured data from electronic medical records. Similarly, a study by [2] used a machine learning model to predict sepsis early and improve diagnosis by integrating structured and unstructured data. The study found that using both data types improved the prediction models' accuracy. [4] employed machine learning models to detect probable dementia cases in undiagnosed patients by integrating structured and unstructured data from electronic medical records. [3] developed a machine-learning model for predicting the risk of 30-day hospital readmissions among heart failure patients through this combination. [6] used structured and unstructured data to identify patients' needs for services that address the social determinants of health. [5] used machine learning models to predict the demand volatility of pharmaceutical products during disruption through news sentiment analysis.

In predictive analytics, [7] used a deep learning-based approach to analyze structured and unstructured data to make predictions.In the context of public security, a hybrid methodology was used by [8] to analyze structured and unstructured data to support decision-making. Additional applications include predicting the risk of derailment in freight trains [16] and classifying types of surgical procedures [9].

Empirical data based on the abovementioned studies suggest that integrating structured and unstructured data in machine learning models can lead to more accurate and timely predictions. Similarly, in crop yield prediction, integrating these two types of data can improve accuracy, which is why their use is gaining momentum, with several studies demonstrating their value. For instance, [17] utilized both data types to develop a machine-learning model that provides accurate crop yield predictions. Similarly, [18] used both structured and unstructured data to develop a deep-learning model for crop yield forecasting. This approach allows for more comprehensive and accurate predictions. It can capture information about the environment, soil, climate, crop types, and other factors, such as farmers' experience, which are difficult to capture using traditional methods.

III. Ensemble Model Approach

Besides integrating structured and unstructured data sources, machine learning models can benefit from integrating external sources like simulation models' predictions. Simulation models are computer programs that use mathematical equations to predict the behaviour of a system. Simulation models can predict crop yields, such as soil moisture, pest infestations, and plant growth rates. These predictions can be used as inputs to machine learning models, which can then be trained to predict crop yields.

Several sectors are already leveraging the use of external predictions, such as simulation models, in their hybrid ML approaches. For instance, [10] presented an application for pre-hospital emergency service, which combined a hybrid simulation model and a machine-learning approach for prediction of the likelihood of cardiac arrest in patients. The simulation model predicted factors such as patient demographics, medical history, and vital signs, while the machine learning model used this information to predict the likelihood

of cardiac arrest. The integrated approach improved the accuracy of predictions compared to using either approach alone. [14] developed the concept of physics-informed ML models for various scientific problems, such as drug discovery, safety engineering, and atmospheric science.Physics-informed machine learning involves incorporating physical laws and principles into machine learning models, which can improve the accuracy of predictions and make the models more interpretable.

Similarly, in micro-structure estimation, a simulation-assisted machine learning approach was used to predict the properties of materials [11]. The simulation model predicted the material's microstructure, while the machine learning model predicted the properties of the material based on this microstructure. The integrated approach is used to accurately estimate the micro-structure of diffuse white matter in brain tumours.

In elderly discharge planning, a hybrid simulation and machine learning approach was used to predict the likelihood of readmission [12]. The simulation model predicted factors such as patient demographics, medical history, and comorbidities, while the machine learning model used this information to predict the likelihood of readmission. The integrated approach improved the accuracy of predictions compared to using either approach alone.

Machine learning and multiscale modelling were integrated into the biological, biomedical, and behavioural sciences to improve predictions [13]. Multiscale modelling involves predicting the behaviour of a system at multiple scales, from the molecular to the cellular to the organismal level. By integrating machine learning with multiscale modelling, researchers can predict the behaviour of complex systems more accurately.

In the context of crop yield prediction, few recent studies used a hybrid machine learning and simulation approach to improve crop yield predictions.[15] and [16] proposed a model combining machine learning and crop modelling to improve yield predictions in the US Corn Belt, which made use of Agricultural Production Systems Simulator (APSIM) along with Machine learning models to generate predictions for crop yield. The model was tested by comparing its predictions with an internal simulation model and machine learning models individually and was shown to have a high degree of accuracy in yield estimates.

These studies demonstrate the potential benefits of integrating predictions from multiple sources to improve the accuracy and timeliness of crop yield predictions. External simulations and models capture the dynamic nature of complex heterogeneous systems with a high degree of accuracy, thereby providing a comprehensive approximation of the underlying dynamic behaviour. Moreover, they prevent ML models from relying solely on local data and providing a global representation of the system, allowing them to generalize better.By incorporating information from simulation mod- els, machine learning models can make more accurate and informed predictions, which can help farmers make better decisions and improve their crop yields.

IV. Conclusion

Researchers have widely explored ML models for crop yield prediction and successfully developed models based on traditional approaches, which showed considerable performance. However, based on various studies like [17][18][19], utilising structured and unstructured data sources improve predictions. People have recently explored the potential of leveraging external predictions as input in ML models. The work of [15], coupled with the findings of [16] on the use of simulations and ML models, demonstrate the promise of Ensemble model approaches and their potential to accompany the development of accurate yield prediction models.[16] observed that the ensemble model boosted ML performance up to 27% and achieved 8%-9% better corn yield predictions. An extension of these works [16][17] could include remote sensing data or irrigation information in the prediction model and investigate the level of importance each data source can exhibit.

References

[1] Tabaie A, Orenstein EW, Kandaswamy S, Kamaleswaran R. Integrating structured and unstructured data for timely prediction of bloodstream infection among children. Pediatr Res. 2022 Jul 19. doi: 10.1038/s41390-022-02116-6. Epub ahead of print. PMID: 35854085

[2] Goh KH, Wang L, Yeow AYK, Poh H, Li K, Yeow JJL, Tan GYH. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun. 2021 Jan 29;12(1):711. doi: 10.1038/s41467-021-20910-4. PMID: 33514699; PMCID: PMC7846756

[3] Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, Hisamitsu T, Kojima G, Felsted J, Kakarmath S, Kvedar J, Jethwani K. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Med Inform Decis Mak. 2018 Jun 22;18(1):44. doi: 10.1186/s12911-018-0620-z. PMID: 29929496; PMCID: PMC6013959.

[4] Shao Y, Zeng QT, Chen KK, Shutes-David A, Thielke SM, Tsuang DW. Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. BMC Med Inform Decis Mak. 2019 Jul 9;19(1):128. doi: 10.1186/s12911-019-0846-4. PMID: 31288818; PMCID: PMC6617952.

[5] Angie Nguyen, Robert Pellerin, Samir Lamouri Béranger Lekens (2022) Managing demand volatility of pharmaceutical products in times of disruption through news sentiment analysis, International Journal of Production Research, DOI: 10.1080/00207543.2022.2070044

[6] Vest JR, Grannis SJ, Haut DP, Halverson PK, Menachemi N. Using structured and unstructured data to identify patients' need for services that address the social determinants of health. Int J Med Inform. 2017 Nov;107:101-106. doi: 10.1016/j.ijmedinf.2017.09.008. Epub 2017 Sep 20. PMID: 29029685.

[7] Dey, Lipika Meisheri, Hardik Verma, Ishan (2017). Predictive Analytics with Structured and Unstructured data - A Deep Learning based Approach. IEEE Intelligent Informatics Bulletin. 18.

[8] Turet, Jean and Seixas Costa, Ana Paula Cabral, Hybrid Methodology for Analysis of Structured and Unstructured Data to Support Decision-Making in Public Security. Available at SSRN: https://ssrn.com/abstract=4046471 or http://dx.doi.org/10.2139/ssrn.4046471

[9] Khaleghi T, Murat A, Arslanturk S. A tree based approach for multi-class classification of surgical procedures using structured and unstructured data. BMC Med Inform Decis Mak. 2021 Nov 23;21(1):328. doi: 10.1186/s12911-021-01665-w. PMID: 34814905; PMCID: PMC8612004.

[10] David Olave-Rojas, Stefan Nickel, Modeling a pre-hospital emergency medical service using hybrid simulation and a machine learning approach, Simulation Modelling Practice and Theory,Volume 109,2021,102302,ISSN 1569-190X.

[11] Rafael-Patino, J. et al. (2020). DWI Simulation-Assisted Machine Learning Models for Microstructure Estimation. In: Bonet-Carne, E., Hutter, J., Palombo, M., Pizzolato, M., Sepehrband, F., Zhang, F. (eds) Computational Diffusion MRI. Mathematics and Visualization. Springer, Cham. https://doi.org/10.1007/978-3-030-52893-5-11

[12] Mahmoud Elbattah and Owen Molloy. 2016. Coupling Simulation with Machine Learning: A Hybrid Approach for Elderly Discharge Planning. In Proceedings of the 2016 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation (SIGSIM-PADS '16). Association for Computing Machinery, New York, NY, USA, 47–56. https://doi.org/10.1145/2901378.2901381

[13] Alber M, Buganza Tepole A, Cannon WR, De S, Dura-Bernal S, Garikipati K, Karniadakis G, Lytton WW, Perdikaris P, Petzold L, Kuhl E. Integrating machine learning and multiscale modeling-perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. NPJ Digit Med. 2019 Nov 25;2:115. doi: 10.1038/s41746-019-0193-y. PMID: 31799423; PMCID: PMC6877584.

[14] Karniadakis, G.E., Kevrekidis, I.G., Lu, L. et al. Physics-informed machine learning. Nat Rev Phys 3, 422–440 (2021). https://doi.org/10.1038/s42254-021-00314-5

[15] S. S. Sajid, I. Huber, S. Archontoulis and G. Hu, "Integrating Crop Simulation and Machine Learning Models to Improve Crop Yield Prediction," 2022 17th Annual System of Systems Engineering Conference (SOSE), Rochester, NY, USA, 2022, pp. 120-125, doi: 10.1109/SOSE55472.2022.9812678.

[16] Shahhosseini, Mohsen, et al. "Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt." Scientific reports 11.1 (2021): 1-15.

[17] Muruganantham, Priyanga, et al. "A systematic literature review on crop yield prediction with deep learning and remote sensing." Remote Sensing 14.9 (2022): 1990.

[18] Gavahi, Keyhan, Peyman Abbaszadeh, and Hamid Moradkhani. "DeepYield: A combined convolutional neural network with long short-term memory for crop yield forecasting." Expert Systems with Applications 184 (2021): 115511.

[19] Rashid, Mamunur, et al. "A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction." IEEE Access 9 (2021): 63406-63439.