MULTI-MODAL FUSION METHODS WITH LOCAL NEIGHBORHOOD INFORMATION FOR CROP YIELD PREDICTION AT FIELD AND SUBFIELD LEVELS

Miro Miranda [*,1,2], *Deepak Pathak* [*,2], *Marlon Nuske* [2], *and Andreas Dengel* [1,2]

[1]Department of Computer Science, University of Kaiserslautern-Landau (RPTU), Kaiserslautern, Germany
[2]German Research Center for Artificial Intelligence (DFKI), SDS, Kaiserslautern, Germany

ABSTRACT

Yield prediction at both field and subfield level poses a significant challenge, yet it holds paramount importance for decision-making and food security within the agricultural sector. Recent efforts, focused on integrating remote sensing data coupled with machine learning models, thereby creating globally scalable models for various crop types. This study underscores the effectiveness of Sentinel-2 and complementary data sources such as weather, soil, and terrain in enhancing machine learning-based yield prediction. We address the limitations of previous works and introduce a framework that incorporates local neighborhood information using convolutional neural networks and geographical coordinates. Additionally, we address the complexity of sensor fusion, showcasing both input fusion and feature fusion frameworks. We highlight that handling modalities with varying spatial and temporal resolutions requires adequate and advanced fusion mechanisms in crop yield prediction. Notably, this study reports an R^2 of 0.86 for soybean in Argentina using a feature fusion scheme with attention mechanism. The results are demonstrated on a large yield dataset for soybean, wheat, and rapeseed distributed across Argentina, Uruguay, and Germany.

Index Terms— Yield Prediction, Sentinel-2, Multimodal Learning, Neural Networks

1. INTRODUCTION

Yield predictions are pivotal for advancing agricultural productivity and resource efficiency. The integration of Remote Sensing (RS) and Machine Learning (ML) significantly contributed to the recent successes, capitalizing on large datasets and scalable models. Consequently, there is a rising popularity of frameworks showcasing global scalability applicable to diverse crop types. One of the major drivers of such models are RS data sources with global coverage. The Sentinel-2 (S2) mission, as one of the primary examples, offers temporal and multispectral data with a spatial resolution up to $10\,m$. Such data captures crop-specific features, frequently used for crop yield modeling. When combined with additional data modalities (ADM) such as weather, soil, and terrain information, a robust foundation for effective yield modeling is established [1]. This, in turn, leverages powerful models, supporting decision-making in the agricultural industry.

Nonetheless, open questions remain in RS-based yield prediction. When working with multimodal data, originating from sensors with varying temporal and spatial resolutions, the identification of an appropriate data fusion scheme becomes imperative [2]. In addition, existing research has primarily concentrated on pixel-based yield mapping, often neglecting local neighborhood effects by treating each pixel independently [3, 4]. In contrast, the inclusion of local neighborhood information, provides a more comprehensive understanding of spatial relationships and interactions within the agricultural landscape. Only a few studies exist that incorporate spatial information for crop yield prediction, as evidenced in [5].

In this research, a framework for crop yield prediction using multimodal input data is proposed. To explicitly account for local neighborhood effects, a convolutional neural network (CNN)-based architecture is employed. We highlight, that the incorporation of neighborhood information improves over a state-of-the-art baseline model. Moreover, to account for varying temporal and spatial resolutions, different data fusion strategies are compared, namely input and feature fusion. We highlight that, feature fusion based on attention mechanism better captures the non-linear nature of yield formation. Additionally, it is demonstrated that S2 is impressively suited for crop yield prediction.

2. METHODOLOGY

2.1. Data

Yield Data As ground truth, a large yield dataset is utilized, containing yield data from multiple years, countries, and crop types. Field records contain sub-field level data points, collected by combine harvesters and containing geo-referenced information of the yield in tons/hectare (t/ha). Such data is frequently plagued by inaccurate measurements, necessitating thorough data cleaning [6]. This process involves elim-

*equal contribution

inating samples with zero yield values. Additionally, samples beyond three standard deviations are filtered out. Following this, yield maps are rasterized to $10\,m$ spatial resolution by utilizing S2 data as the reference. In total, 1061 yield maps are used for major field crops, including soybean, wheat, and rapeseed, distributed across Argentina, Uruguay, and Germany. A detailed description of the available ground truth yield data is given in Tab. 1.

Table 1: Yield map (fields) per country and crop type for different years.

Country	Years	Rapeseed	Wheat	Soybean	Sum
Germany	2016-2022	111	188	0	299
Uruguay	2018-2022	0	0	572	572
Argentina	2017-2022	0	0	190	190
Sum		111	188	762	**1061**

Data Modalities In all experiments, S2 L2A multispectral time series data is used, encompassing all 12 spectral bands. Bands available in low resolution are upsampled to a spatial resolution of $10\,m$. S2 is collected from seeding to harvesting. Further, ADM are acquired for each field. More specifically, weather, soil, digital elevation map (DEM), and the sample coordinates (coord) as latitude (lat) and longitude (lon). Coordinates are projected into a three-dimensional space as follows: $x = \cos(lat) \cdot \cos(lon)$, $y = \cos(lat) \cdot \sin(lon)$, $z = \sin(lon)$. A detailed description of data modalities is given in Tab. 2. From the DEM, we further derive slope, aspect, and curvature using the RichDEM library [7]. We further derive the topographic wetness index (TWI) [8]

Table 2: Selected modalities complementing Sentinel-2

Modality	Product	Unit	Source
Weather	Precipitation	m	ECMWF [9]
	Max Temperature	K	
	Min Temperature	K	
	Average Temperature	K	
Soil	Soil Organic Carbon	dg/kg	SoilGrids [10]
	Nitrogen	cg/kg	
	Cation Exchange Capacity	mmol(c)/kg	
	Clay	g/kg	
	Silt	g/kg	
	Sand	g/kg	
	pH	pHx10	
	Volumetric fraction of course fragments	cm3/dm3	
Terrain	DEM	m	SRTM [11]

Data Preprocessing

In this research, two different fusion methods are compared. More specifically, Input Fusion (IF) with a monthly sampling, and Feature Fusion (FF) with the complete time series.

Input Fusion: For IF, a 24-month time series is generated by choosing a single S2 image per month, spanning from seeding to harvesting, in accordance with [4]. Time steps falling outside the growing period are masked, utilizing -1 as the masking value. For each time step between seeding and harvesting, ADM are further included, by first upsampling modalities to $10\,m$ resolution before being concatenated, following [4]. Temporal features, including weather data, are aggregated between S2 time steps. In contrast, static features, such as soil, DEM, and coordinates, are replicated across time steps.

Feature Fusion: For FF, each modality is handled independently. We distinguish between temporal features and static features. For temporal modalities, including S2 and weather, the complete time series is used. We further use padding, with the padding value being -1. Static features, including soil, DEM, and coordinates are upsampled to $10\,m$.

2.2. Model Architecture & Training

We formulate a pixel-based prediction that leverages local neighborhood information. Specifically, a window is utilized to extract the neighborhood for each sample, where the center pixel represents the pixel of interest. In the context of labeled training data, where the input is represented as $x \in X$ and the corresponding target as $y \in Y$. The dimensions of the input data are denoted as $X \in R^{N \times B \times T \times 5 \times 5}$, and for the target, $Y \in R^N$. In this representation, N refers to the total number of samples, B the number of bands, and T the number of time steps. For each sample, we span a window of size 5×5. In the case of FF, X is a set of $\{X_{S2}, X_{DEM}, X_{Soil}, X_{Weather}, X_{Coord}\}$, where each modality is handled individually. In the following, we describe the model architecture for IF and FF.

Input Fusion: In the IF approach, we utilize a 3D-CNN (3D-Conv) block with a kernel size of $(1, 5, 5)$ to expand the number of channels to 64, incorporating batch normalization and LeakyReLU activation. Further, the output is passed into a Long short-term memory (LSTM) cell with 2 layers and 64 hidden units. Following, a fully connected network (FC) block is employed, containing a single layer with 64 hidden units, batch normalization, rectified linear unit (ReLU) activation, and dropout. We use a dropout probability of 0.2. Finally, a single linear layer with 64 hidden units and a single output channel is employed, reflecting the predicted yield value.

Feature Fusion: For FF, we utilize independent modality encoders. The core components of this architecture include 3D-Conv, 2D-CNN (2D-Conv), FC, and LSTM blocks, as well as Scaled Dot-Product Attention. Each block incorporates batch normalization and ReLU activation. Spatio-temporal features, such as S2, represented by dimensions $N \times B \times T \times 5 \times 5$, are processed through a 3D-Conv block with a $(1, 5, 5)$ kernel size, followed by a LSTM block with 2 LSTM layers. Spatial features, including DEM, soil, and coordinates with dimensions $N \times B \times 5 \times 5$, are handled by a 2D-Conv block with a $(5, 5)$ kernel size, followed by a fully connected (FC) layer. Temporal features, such as weather data, are processed using a LSTM block for feature extrac-

tion. Each modality-specific encoder outputs features with dimensions $N \times 64$. To achieve modality fusion, we employ scaled dot-product attention [12]. This mechanism enables attention pooling, where a learnable query interacts with each modality's representation through cross-attention, generating attention weights. We apply dropout with a 0.2 probability to the attention weights during pooling. The final fused representation is fed into a linear layer to predict the yield value. The architecture is illustrated in Fig. 1.

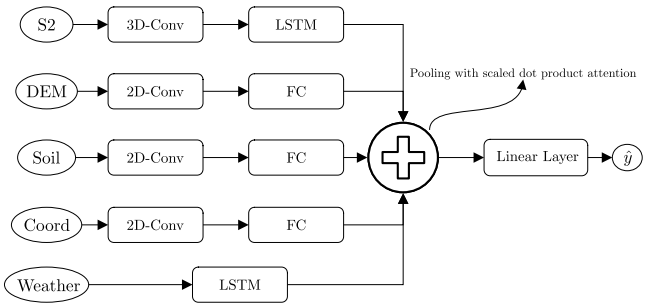

Fig. 1: Feature fusion architecture: Each modality is separately encoded. Subsequently, the modalities are fused using scaled dot-product attention pooling and fed to a linear layer.

Training & Evaluation Training is executed using the ADAM optimizer on MSE loss between prediction of the center pixel and target pixel for a maximum of 50 epochs, with a learning rate of 0.006 and a batch size of 2048. The training incorporates a reduce-on-plateau learning rate scheduler. For regularization, early stopping is applied after 10 consecutive epochs with no improvement on the validation set. Additionally, during training, as part of the data augmentation strategy, random rotation by 90 degrees is applied to the input window, and temporal dropout with a 0.2 probability is employed for temporal features.

To assess the impact of local neighborhood information and data fusion method, we present results of a *baseline* model. More specifically, a pixel-based model based on LSTM [13]. We utilize the same architecture as described in [4], including S2, weather, soil, and DEM as input. We conduct model evaluations on a per-region and per-crop type basis. To achieve this, we employ stratified K-fold (K=10) cross-validation with non-overlapping groups. The stratification is done based on regions, and the grouping is performed by field. In the quantitative evaluation, standard regression metrics are employed and reported as the average over validation folds. The metrics include mean absolute error (MAE), mean absolute percentage error (MAPE), root mean square error (RMSE), the coefficient of determination (R^2), and Bias.

3. RESULTS

We start by presenting quantitative results for each crop type and model. The results are presented in Tab. 3. For each dataset, comprising a unique country and crop type, results are presented for both fusion strategies. Further, results of the baseline model are presented. The best performing model is highlighted. We highlight that the inclusion of spatial information results in an improved or equal performance compared to the baseline. In the case of IF and Germany wheat, an improvement of 3 percentage points (p.p) in R^2 at the field level is presented. Nevertheless, we also observe cases with no improvement such as evidenced in Uruguay, soybean. Considering the FF, we observe consistent improvement over both the baseline model and the IF model across all datasets. In the case of Argentina, soybean we demonstrate an R^2 of 0.86 on the field level, signifying an impressive improvement of 10 p.p over the baseline model and 8 p.p over the IF model. In the case of Germany, wheat an improvement of 15 p.p is reported, highlighting its superior performance in crop yield prediction.

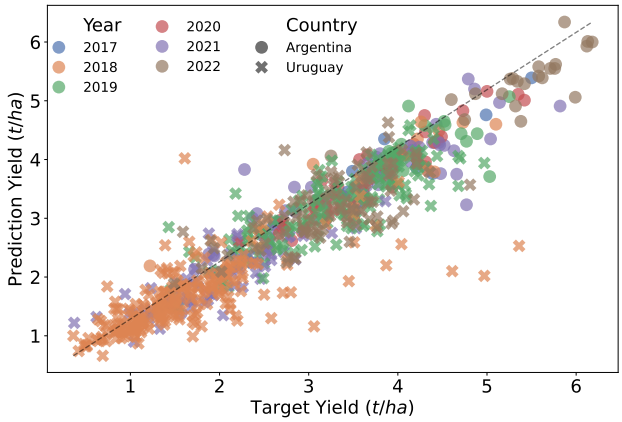

Fig. 2: Scatter plot of field predictions and ground truth yield data for soybean in Uruguay and Argentina. Colors are used to differentiate between harvesting years. Results originate from the FF model.

In Fig. 2, a scatter plot illustrating the relationship between the target yield per field and the corresponding model predictions is presented. The plot showcases results for soybean in Argentina and Uruguay using the FF model. The illustration highlights the differences between countries and years. Notably, we observe low yields in Uruguay in 2018, while Argentina generally exhibits higher yields. Additionally, the figure demonstrates the model's effectiveness in capturing the variability in the data across a broad spectrum of plausible yield values.

We further investigate the attention weights of each data modality for the FF model. Fig. 3 visualizes the average at-

Evaluation		Field					Subfield				
Dataset	Model	MAE t/ha	RMSE t/ha	MAPE %	R2 -	BIAS t/ha	MAE t/ha	RMSE t/ha	MAPE %	R2 -	BIAS t/ha
Argentina, Soybean	Input Fusion	0.37	0.49	0.10	0.78	-0.02	0.67	0.90	0.25	0.62	0
	Feature Fusion	**0.27**	**0.39**	**0.08**	**0.86**	**0.01**	**0.60**	**0.81**	**0.23**	**0.70**	**0.01**
	Baseline	0.40	0.52	0.11	0.76	0	0.66	0.89	0.24	0.63	-0.02
Uruguay, Soybean	Input Fusion	0.37	0.52	0.18	0.76	-0.05	0.81	1.22	**0.91**	0.40	-0.04
	Feature Fusion	**0.32**	**0.46**	**0.17**	**0.81**	**-0.02**	**0.78**	**1.19**	**0.91**	**0.43**	-0.02
	Baseline	0.35	0.51	0.20	0.77	0.01	**0.78**	1.22	1.02	0.42	**0.01**
Germany, Rapeseed	Input Fusion	0.49	0.64	**0.14**	0.77	-0.16	0.90	1.22	0.36	0.46	-0.10
	Feature Fusion	**0.44**	**0.60**	**0.14**	**0.80**	**-0.07**	**0.87**	**1.20**	**0.35**	**0.49**	**-0.04**
	Baseline	0.49	0.65	0.15	0.77	-0.03	0.93	1.23	0.38	0.46	-0.04
Germany, Wheat	Input Fusion	0.80	1.05	0.09	0.70	-0.17	1.67	2.30	0.27	0.39	-0.10
	Feature Fusion	**0.61**	**0.83**	**0.07**	**0.81**	**-0.12**	**1.51**	**2.13**	**0.25**	**0.48**	**-0.13**
	Baseline	0.84	1.11	0.09	0.66	0.16	1.71	2.37	0.29	0.35	0.20

Table 3: Overview of results per crop type, region and model. The best scores are highlighted.

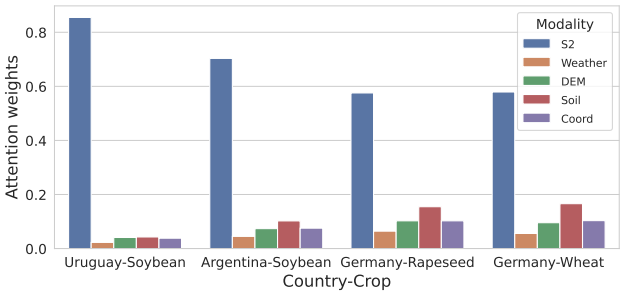

Fig. 3: Bar plot illustrating attention weights derived from scaled dot-product attention for all modalities across different countries and crops.

tention weight for each data modality, by utilizing the scaled dot-product attention. The average attention weight is calculated across all samples in the validation split over all folds in the cross-validation. This illustrates how the model learns to assign different importance to various modalities in terms of attention weights. We notice that attention weights vary across different regions and crops. However, the S2 modality consistently dominates as the main contributor in all cases. In the case of Uruguay, soybean ADM showcase the lowest attention across all samples. In contrast, Germany illustrates higher attentions for ADM with soil exhibiting the highest values.

4. DISCUSSION & CONCLUSION

The findings of this research highlight the potential of remote sensing-based yield prediction using subfield level yield data and machine learning. We highlight Sentinel-2 data as a predominant contributor to the model's performance but being complemented by additional data modalities such as weather, soil and terrain. We address the limitations of previous studies by introducing a method that incorporates local neighborhood information while accounting for the varying temporal and spatial resolutions of the multimodal input data. Our results demonstrate that including neighborhood information enhances model performance in crop yield prediction. Nevertheless, when working with multimodal data, selecting an appropriate data fusion method is crucial. Our study reveals that input fusion suffers under limitations, as it fails to adequately address the different spatial and temporal resolutions of the input. Additionally, such methods require expensive modality selection. To address this issue, we propose using feature fusion with an attention mechanism, which has proven to be a powerful method for crop yield prediction.

5. ACKNOWLEDGMENT

The research results presented are part of a large collaborative project on agricultural yield predictions, which was partly funded through the ESA InCubed Programme (`https://incubed.esa.int/`) as part of the project AI4EO Solution Factory (`https://www.ai4eo-solution-factory.de/`).

6. REFERENCES

[1] T. Van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture*, vol. 177, p. 105709, 2020.

[2] F. Mena, D. Arenas, M. Nuske, and A. Dengel, "Common practices and taxonomy in deep multi-view fusion for remote sensing applications," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[3] P. Helber, B. Bischke, P. Habelitz, C. Sanchez, D. Pathak, M. Miranda, H. Najjar, F. Mena, J. Siddamsetty, D. Arenas *et al.*, "Crop yield prediction: An operational approach to crop yield modeling on field and subfield level with machine learning models," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023, pp. 2763–2766.

[4] D. Pathak, M. Miranda, F. Mena, C. Sanchez, P. Helber, B. Bischke, P. Habelitz, H. Najjar, J. Siddamsetty, D. Arenas, M. Vollmer, M. Charfuelan, M. Nuske, and A. Dengel, "Predicting Crop Yield with Machine Learning: An Extensive Analysis of Input Modalities and Models on a Field and Sub-Field Level," in *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, 2023, pp. 2767–2770.

[5] K. Gavahi, P. Abbaszadeh, and H. Moradkhani, "Deepyield: A combined convolutional neural network with long short-term memory for crop yield forecasting," *Expert Systems with Applications*, vol. 184, p. 115511, 2021.

[6] C. Sanchez, D. Pathak, M. Miranda, M. Charfuelan, P. Helber, M. Nuske, B. Bischke, P. Habelitz, N. Rahman, F. Mena *et al.*, "Influence of data cleaning techniques on sub-field yield predictions," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023, pp. 4852–4855.

[7] R. Barnes, *RichDEM: Terrain Analysis Software*, 2016. [Online]. Available: http://github.com/r-barnes/richdem

[8] M. Kopeckỳ, M. Macek, and J. Wild, "Topographic wetness index calculation guidelines based on measured soil moisture and plant species composition," *Science of the Total Environment*, vol. 757, p. 143785, 2021.

[9] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers *et al.*, "The ERA5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020.

[10] L. Poggio, L. M. De Sousa, N. H. Batjes, G. Heuvelink, B. Kempen, E. Ribeiro, and D. Rossiter, "SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty," *Soil*, vol. 7, no. 1, pp. 217–240, 2021.

[11] T. G. Farr and M. Kobrick, "Shuttle Radar Topography Mission produces a wealth of data," *Eos, Transactions American Geophysical Union*, vol. 81, no. 48, pp. 583–585, 2000.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[13] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.