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

Miro Miranda [*,1,2], *Deepak Pathak* [*,1,2], *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 yield prediction. We address limitations of previous works and introduce a framework that incorporates local neighborhood information using a convolutional neural network approach. Additionally, we address the complexity of sensor fusion, showcasing both early fusion and late fusion frameworks. Notably. This study reports an R^2 of 0.83 for soybean in Argentina. The results are demonstrated on a large yield dataset for Soybean, Wheat, and Rapeseed distributed across multiple countries, including 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 resolution up to $10m$. Such data captures crop-specific features frequently used for crop yield modeling. When combined with additional data sources (ADS) such as weather, soil, and terrain information, a robust foundation for effective yield modeling is established. This, in turn, leverages powerful models, supporting decision-making in the agricultural industry.

Nonetheless, open questions remain in RS-based yield prediction. When it comes to multimodal data, originating from sensors with varying temporal and spatial resolutions, the identification of an appropriate data fusion scheme becomes imperative. In addition, existing research has primarily concentrated on pixel-based yield mapping, often neglecting local neighborhood effects by treating each pixel independently. In contrast, the inclusion of local neighborhood information, provides a more comprehensive understanding of spatial relationships and interactions within the agricultural landscape.

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. Moreover, different data fusion strategies are compared, accounting for varying temporal and spatial resolutions using a modality-specific encoder architecture. We highlight, that the incorporation of neighborhood information improves over a state-of-the-art baseline model. Additionally, it is demonstrated that S2 is impressively suited for crop yield prediction. The inclusion of ADS can additionally contribute to a superior model performance.

2. METHODS

2.1. Data

Yield Data A large yield data set is created, containing yield data over multiple years, countries, and crop types. Field records contain sub-field level data points, collected by combine harvesters and containing geo-reference information of the yield in tons/hectare (t/ha). More precisely, 1061 yield maps are available for major field crops, including Soybean, Wheat, and Rapeseed, distributed over Argentina, Uruguay, and Germany. A detailed description of available ground truth yield data is given in Tab. 1. Such data is frequently plagued by inaccurate measurements, necessitating thorough data cleaning. This process involves eliminating samples

[*]both authors contributed equally to this work

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/).

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**

with zero yield. Additionally, samples beyond three standard deviations are filtered out. Following this, yield maps are transformed into a 10m resolution by utilizing S2 data as the reference.

Data Modalities In all experiments, S2 L2A multispectral time series data is employed, encompassing all 12 spectral bands. Bands available in low resolution are upsampled to achieve a spatial resolution of $10m$. S2 is collected from seeding to harvesting. Further, ADS are acquired for each sample. More specifically, weather, soil, digital elevation map (DEM), and the coordinates as latitude (lat) and longitude (lon), respectively. Coordinates are projected into three dimensions as $[\cos{(lat)} \cdot \cos{(lon)}, \cos{(lat)} \cdot \sin{(lon)}, \sin{(lon)}]$. A detailed description of data modalities is given in Tab. 2. From DEM, we further derive slope, aspect, curvature, and topographic wetness index (TWI).

Table 2: Selected modalities complementing Sentinel-2

Data Source	Product	Unit	Source
Weather	Precipitation	m	ECMWF[1]
	Max Temperature	K	
	Min Temperature	K	
	Average Temperature	K	
Soil	Soil Organic Carbon	dg/kg	SoilGrids[2]
	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[3]

Data Preprocessing

Two different data preprocessing are employed, encompassing a monthly sampling with Early Fusion (EF), and a Late Fusion (LF) with the source data.

Early Fusion For EF, 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, ADS are further included, by first upsampling modalities to $10m$ 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.

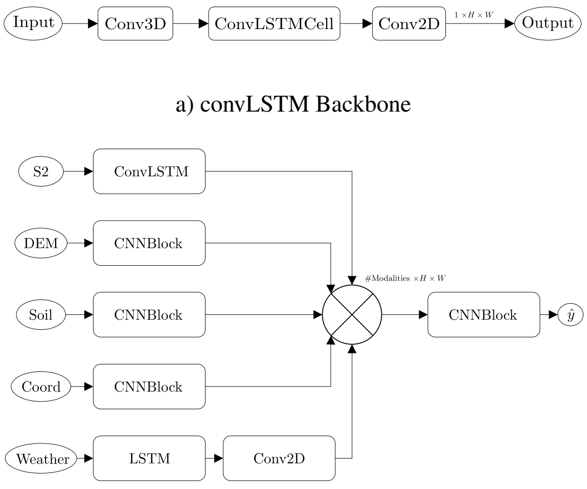
Late Fusion For late fusion, each modality is handled independently. We distinguish between temporal features and static features. For temporal modalities, including S2 and weather, the complete time series is used. We further use padding, with the padding value being -1. Static features, including soil,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

2.2. Model Architecture & Training

We formulate a pixel-based prediction that leverages local neighborhood information. Specifically, a window is employed to extract the neighborhood for each sample, where the center 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 $X \in R^{N \times B \times T \times 3 \times 3}$, and for the target, $Y \in R^{N}$. In this representation, N refers to the total number of samples, B the number of bands, T the number of time steps, respectively. For each sample, we span a window of size 3×3. In the case of late fusion, X is a set of $\{X_{S2}$, X_{DEM}, X_{Soil}, X_{Coord}, $X_{Weather}\}$, where each modality is handled individually. We establish two scenarios for both model architecture and training oriented on the data fusion strategy.

Early Fusion In the EF approach, the model architecture incorporates a convolutional Long short-term memory (convLSTM) [5] backbone architecture, depicted in Fig. 1 (a). The input, is passed to a conv3-D block to expand the number of channels, incorporating batch normalization and LeakyReLU activation. Further, the output is passed into a convLSTM cell with 64 hidden units, following a conv-2D block with a single output channel, reflecting the predicted yield value. rectified linear unit (ReLU) activation is used. From the output of size $1 \times 3 \times 3$, the center pixel is extracted as the final prediction.

Late Fusion For LF, we employ independent modality encoders. We differentiate between spatio-temporal (S2), temporal (weather), and spatial features (soil, DEM, coordinates). The architecture of the late-fusion model is illustrated in Fig.1(b). The foundational elements of this architecture include the convLSTM unit depicted in Fig.1(a), an Long short-term memory (LSTM)[6] network as a stack of 2 LSTM layers, and a CNN block comprising two 2D-convolution layers with batch normalization and ReLU activation after each layer. Spatio-temporal features are directed to a convLSTM block, while spatial features are channeled to the CNN block. For temporal features, we utilize an LSTM block for representation extraction, followed by a 2D-convolution layer with padding. Each modality-specific encoder extracts features in the form of a window with a single channel represented as $1 \times 3 \times 3$. Finally, we concatenate the extracted window representations along the channel dimension. Subsequently, the input is passed into the CNN block to obtain predictions, considering the center pixel of the output window.



a) convLSTM Backbone

b) Late Fusion

Fig. 1: a) convLSTM backbone, to process spatio-temporal data. b) Late fusion architecture. Each modality is separately encoded. Following, modalities are concatenated and subsequently inserted into a CNN Block.

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. Additionally, as part of the data augmentation strategy, random rotation is applied to the input window, temporal dropout is employed for temporal features.

To assess the impact of local neighborhood information, we present results of a *baseline* model. More specifically, a pixel-based model founded on LSTM [6]. We utilize the same architecture as described in [4], including only S2 as input. To further assess the impact of fusion strategies and multimodal input, we report results of a convLSTM model (Fig. 1(a)) with only S2 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. This includes a mix of S2 and ADS. Further, results of the convLSTM and baseline model are presented, incorporating only S2 data. For each model, the incorporated modalities are specified, and the best scores are highlighted. Note that all proposed architectures improve over the baseline model, highlighting the potential of including neighborhood information. For wheat, the EF approach performs best, with an R^2 of 0.71 on the field level. This signifies an impressive improvement of 16 percentage points (p.p) over the baseline, and 2 p.p over the convLSTM. Reviewing the other dataset, it becomes apparent, that the difference in performance is marginal. Surprisingly, it is noteworthy that the convLSTM trained on S2 alone achieves a remarkable performance.

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

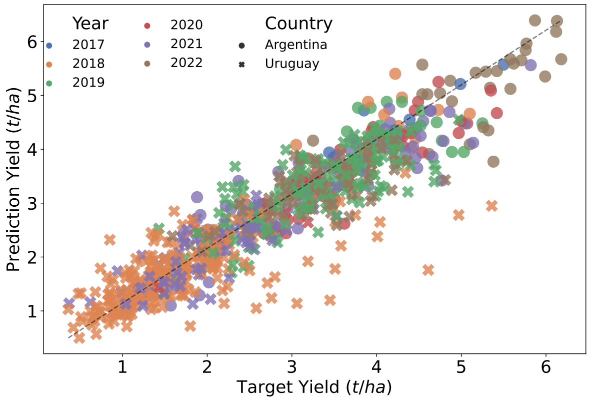
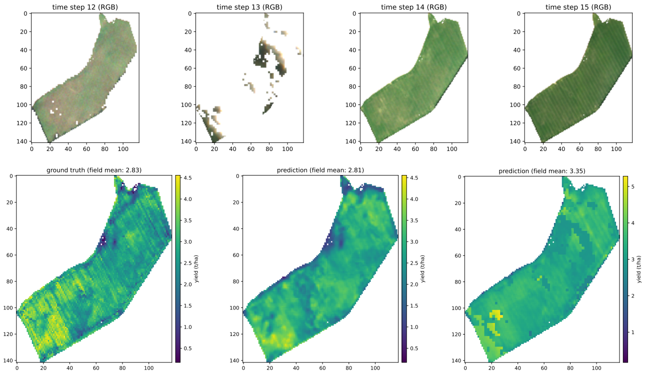


Fig. 2: Scatter plot of field predictions and ground truth yield data for Soybean in Uruguay and Argentina. Colors are used to differentiate between harvesting years.

In terms of quantitative assessment, all spatial models demonstrate an enhancement over the baseline. Particularly, when dealing with fields affected by cloud cover, the inclusion of spatial information proves beneficial for the overall performance. Fig. 3 showcases a rapeseed field in Germany, where the top section presents a segment of S2 time series, highlighting cloud corruption in the second image. Below this, the ground truth yield map is displayed, followed by the prediction from the spatial model (LF), and the baseline prediction. Notably, the baseline exhibits difficulties with cloud-corrupted pixels, while the spatial approach yields more realistic and improved results.

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
Soybean (Argentina)	Early Fusion	0.38	0.49	0.11	0.78	0.03	0.66	0.88	0.25	0.64	**0.01**
	Late Fusion	**0.31**	**0.42**	**0.08**	**0.83**	-0.06	**0.62**	**0.85**	**0.22**	**0.67**	-0.05
	convLSTM	0.34	0.45	0.1	0.82	-0.4	0.66	0.89	0.24	0.63	-0.04
	Baseline	0.4	0.53	0.11	0.74	**-0.1**	0.69	0.92	0.25	0.61	-0.06
Soybean (Uruguay)	Early Fusion	0.37	**0.51**	0.19	**0.77**	**0.02**	0.8	**1.22**	0.92	0.41	**0.02**
	Late Fusion	0.35	**0.51**	0.18	**0.77**	-0.04	0.8	**1.22**	**0.9**	**0.4**	-0.04
	convLSTM	**0.34**	0.52	**0.17**	**0.77**	**-0.02**	0.79	**1.22**	**0.9**	**0.4**	**-0.02**
	Baseline	0.36	0.53	0.19	0.75	-0.06	**0.78**	**1.22**	0.96	0.41	-0.05
Rapeseed (Germany)	Early Fusion	0.44	**0.58**	**0.13**	0.81	-0.12	**0.88**	**1.18**	**0.34**	**0.5**	-0.08
	Late Fusion	0.44	0.6	**0.13**	0.8	-0.08	0.9	1.23	0.35	0.46	-0.09
	convLSTM	**0.42**	**0.58**	**0.13**	**0.82**	**-0.01**	**0.88**	1.2	0.36	0.49	**0.02**
	Baseline	0.61	0.77	0.17	0.67	-0.17	0.98	1.31	0.38	0.38	-0.11
Wheat (Germany)	Early Fusion	**0.78**	**1.03**	**0.09**	**0.71**	-0.03	1.69	**2.32**	0.28	**0.38**	**-0.02**
	Late Fusion	0.8	1.14	**0.09**	0.64	-0.14	**1.67**	**2.32**	**0.27**	**0.38**	-0.17
	convLSTM	**0.78**	1.06	**0.09**	0.69	**0.02**	1.72	2.36	0.29	0.36	0.04
	Baseline	0.91	1.27	0.1	0.55	-0.15	1.73	2.38	0.29	0.35	-0.19

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



Ground Truth   Late Fusion   Baseline

Fig. 3: Example Rapeseed field in Germany, harvested in 2020. On top, part of the S2 time series, displayed in RGB. Below, the ground truth yield map, depicted on the right, next to it are the pixel-wise prediction from the late fusion model, followed by the baseline predictions.

4. CONCLUSION

The findings of this research underscore the high potential of remote sensing-based yield prediction. It is evident that predicting yields at subfield resolution is particularly advantageous when leveraging input from globally diverse sensors, with Sentinel-2 emerging as a predominant contributor. Furthermore, we emphasize the advantages of incorporating neighborhood information at various levels. The optimization of fusion strategies additionally enhances the quality of results. However, it is crucial to note that the selection of suitable data sources and their effective combination remains an open area of research, demanding thorough exploration.

5. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[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] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

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