FLAGSHIP 2.6.3: AI ALGORITHM FOR (SATELLITE) IMAGING RECONSTRUCTION

REPORT FOR WP6 MEETING, 24/10/2023

A. Tricomi[1,2,3], G. Piparo[1], G. A. Anastasi[2]

1. INFN Sezione di Catania
2. Università degli studi di Catania
3. Centro Siciliano di Fisica Nucleare e Struttura della Materia (CSFNSM)

ICSC
Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing

MILESTONES AND KPIS

Milestones

1. M1-M6 (corresponding to MS7): Survey of the State-of-the-Art; tracking of R&D technologies to be used; selection of datasets for use cases (at least one).
   - D1: report on technologies to be used, selection of at least one test dataset.
2. M7-M10 (corresponding to MS8): first experimentation with data sources and algorithms, demonstration on the feasibility of choices
   - D2: report on the experimentation and of technical choices; first code repository available
3. M11-M24 (corresponding to MS10): Implementation of the selected technology(ies); test and validation on selected dataset(s). Proof-of-Concept deployment.
   - D3: Report on the work carried out; release of the developed code on public repository.
   - Intermediate report at MS9

KPIs

KPI ID	Description	Acceptance threshold
KPI2.6.1.1	Publications	2
KPI2.6.1.2	Presentations at conferences	2
KPI2.6.1.3	Publicly available Code repositories	1
KPI2.6.1.4	Use case Test Datasets	1

MILESTONES AND KPIS

Milestones

1. M1-M6 (corresponding to MS7): Survey of the State-of-the-Art; tracking of R&D technologies to be used; selection of datasets for use cases (at least one).
   - D1: report on technologies to be used, selection of at least one test dataset.
2. M7-M10 (corresponding to MS8): first experimentation with data sources and algorithms, demonstration on the feasibility of choices
   - D2: report on the experimentation and of technical choices; first code repository available
3. M11-M24 (corresponding to MS10): Implementation of the selected technology(ies); test and validation on selected dataset(s). Proof-of-Concept deployment.
   - D3: Report on the work carried out; release of the developed code on public repository.
   - Intermediate report at MS9

KPIs

KPI ID	Description	Acceptance threshold
KPI2.6.1.1	Publications	2
KPI2.6.1.2	Presentations at conferences	2
KPI2.6.1.3	Publicly available Code repositories	1
KPI2.6.1.4	Use case Test Datasets	1

Our two main objectives in this first phase of the project: A dataset and an algorithm landscape

MILESTONES AND KPIS

Milestones

1. M1-M6 (corresponding to MS7): Survey of the State-of-the-Art; tracking of R&D technologies to be used; selection of datasets for use cases (at least one).
   - ➢ D1: report on technologies to be used, selection of at least one test dataset.
2. M7-M10 (corresponding to MS8): first experimentation with data sources and algorithms, demonstration on the feasibility of choices
   - ➢ D2: report on the experimentation and of technical choices; first code repository available
3. M11-M24 (corresponding to MS10): Implementation of the selected technology(ies); test and validation on selected dataset(s). Proof-of-Concept deployment.
   - ➢ D3: Report on the work carried out; release of the developed code on public repository.
   - ➢ Intermediate report at MS9

KPIs

KPI ID	Description	Acceptance threshold
KPI2.6.1.1	Publications	2
KPI2.6.1.2	Presentations at conferences	2
KPI2.6.1.3	Publicly available Code repositories	1
KPI2.6.1.4	Use case Test Datasets	1

Our two main objectives in this first phase of the project: A dataset and an algorithm landscape → **But for what purpose?**

POSSIBLE PURPOSES

Regarding the analysis of satellite images applied **to the economic well-being of agricultural firms** (link with Agri@Intesa IG)**,** we have several possible purposes, including:

- **Field segmentation and identification:** Use of ML algorithms to recognise the edges of a field and the crop type.

- **Recognition of the health status of crops:** Understanding plant health, the possible presence of specific diseases, and/or anomalies in extensive crops.

- **Crop yield prediction:** Development of regression techniques to predict crop yield (in tons/acre for example) based on previous harvests.

POSSIBLE PURPOSES

Regarding the analysis of satellite images applied **to the economic well-being of agricultural firms** (link with Agri@Intesa IG)**,** we have several possible purposes, including:

- **Field segmentation and identification:** Use of ML algorithms to recognise the edges of a field and the crop type.

- **Recognition of the health status of crops:** Understanding plant health, the possible presence of specific diseases, and/or anomalies in extensive crops.

- **Crop yield prediction:** Development of regression techniques to predict crop yield (in tons/acre for example) based on previous harvests.

For today, let's focus on this

STATE OF ART OF CROP YIELD PREDICTION WITH ML

- **We found a very interesting paper on the state of the art in this area, which provides a systematic literature review on the subject (up to 2020).**

- **DOI**: https://doi.org/10.1016/j.compag.2020.105709

- “ In this study, we performed a Systematic Literature Review (SLR) to extract and synthesize the **algorithms and features** that have been used in crop yield prediction studies. Based on our search criteria, we retrieved **567 relevant studies** from six electronic databases, of which we have selected **50 studies** for further analysis using inclusion and exclusion criteria.”

- “After this observation based on the analysis **of machine learning-based 50 papers**, we performed an additional search in electronic databases to identify deep learning-based studies**, reached 30 deep learning-based papers**, and extracted the applied deep learning algorithms”

ANALYTICS OF CONSIDERED PAPERS

Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Growth of the sector over the years, which could be higher in the last 2/3

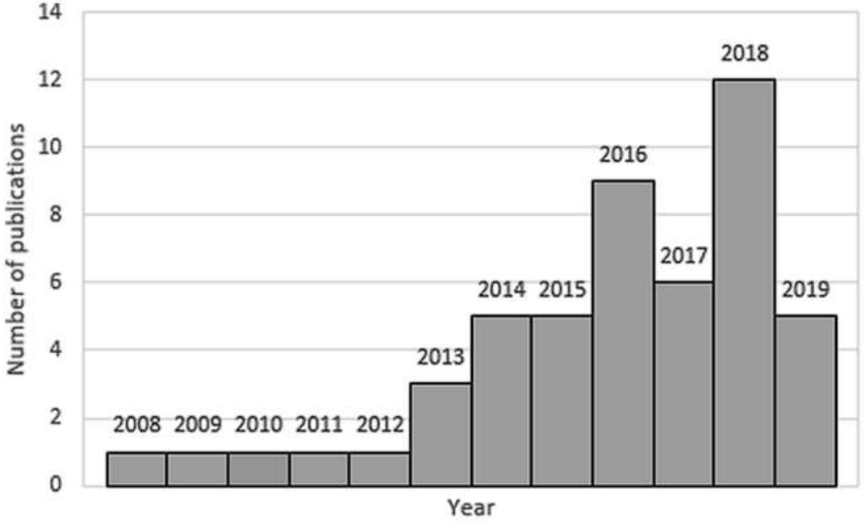


Fig. 4. Distribution of the selected publications per year.

Table 1
Distribution of papers based on the databases.

Database	# of initially retrieved papers	# of papers after exclusion criteria	Percentage of Papers (%)
Science Direct	17	4	8
Scopus	68	11	22
Web of Science	32	0	0
Springer Link	132	10	20
Wiley	20	1	2
Google Scholar	298	24	48
Total	567	50	100

Exclusion criteria 2 – Publication is not written in English
Exclusion criteria 3 – Publication that is a duplicate or already retrieved from another database
Exclusion criteria 4 – Full text of the publication is not available
Exclusion criteria 5 – Publication is a review/survey paper
Exclusion criteria 6 – Publication has been published before 2008

Exclusion criteria 1 - Publication is not related to the agricultural sector and yield prediction combined with machine learning

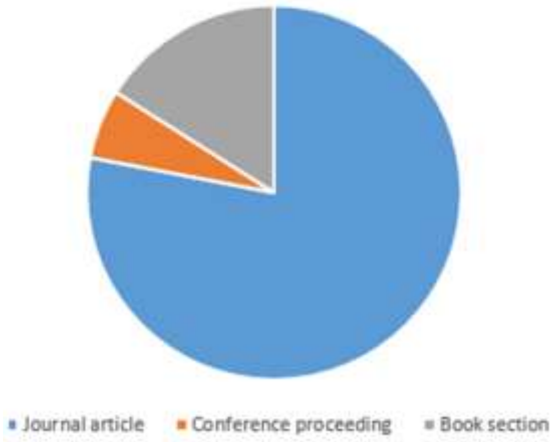


Fig. 5. Distribution of the type of 50 primary publications.

MOST USED FEATURES

Not only remote-sensing information.

Table 3
All features used.

Feature	# of times used
Temperature	24
Soil type	17
Rainfall	17
Crop information	13
Soil maps	12
Humidity	11
pH-value	11
Solar radiation	10
Precipitation	9
Images	8
Area of production	8
Fertilization	7
NDVI	6
Cation exchange capacity	6
Nitrogen	6
Irrigation	5
Potassium	5
Wind speed	5
Zinc	3
Magnesium	3
Shortwave radiation	2
Sulphur	2
Boron	2
Calcium	2
Organic carbon	2
EVI	2
Phosphorus	2
Gamma radiametrics	1
MODIS-EVI	1
Forecasted rainfall	1
Photoperiod	1
Climate	1
Degree-days	1
Time	1
Pressure	1
Leaf area index	1
Manganese	1

Soil, climate and "technical" informations are very important

Table 4
Grouped features.

Group	# of times used
Soil information	54
Solar information	39
Humidity	38
Nutrients	28
Other	24
Crop information	14
Field management	12

ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

MOST USED FEATURES

Not only remote-sensing information.

Table 3
All features used.

Feature	# of times used
Temperature	24
Soil type	17
Rainfall	17
Crop information	13
Soil maps	12
Humidity	11
pH-value	11
Solar radiation	10
Precipitation	9
Images	8
Area of production	8
Fertilization	7
NDVI	6
Cation exchange capacity	6
Nitrogen	6
Irrigation	5
Potassium	5
Wind speed	5
Zinc	3
Magnesium	3
Shortwave radiation	2
Sulphur	2
Boron	2
Calcium	2
Organic carbon	2
EVI	2
Phosphorus	2
Gamma radiametrics	1
MODIS-EVI	1
Forecasted rainfall	1
Photoperiod	1
Climate	1
Degree-days	1
Time	1
Pressure	1
Leaf area index	1
Manganese	1

Soil, climate and "technical" informations are very important

Table 4
Grouped features.

Group	# of times used
Soil information	54
Solar information	39
Humidity	38
Nutrients	28
Other	24
Crop information	14
Field management	12

Not everything can be seen from above!
A good dataset should contain a mixture of the various types of data

MOST USED ARCHITECTURES AND METRICS

Table 5
Most used machine learning algorithms.

Most used machine learning algorithms	# of times used
Neural Networks	27
Linear Regression	14
Random Forest	12
Support Vector Machine	10
Gradient Boosting Tree	4

Table 9
Distribution of deep learning algorithms.

Algorithms used	# of usages	Percentage (%)
CNN	10	30,30
LSTM	7	21,21
DNN	7	21,21
Hybrid	4	12,12
Autoencoder	1	3,03
Multi-Task Learning (MTL)	1	3,03
Deep Recurrent Q-Network (DQN)	1	3,03
3D CNN	1	3,03
Faster R-CNN	1	3,03
Total	33	100

Table 6
All evaluation parameters used.

Key	Evaluation parameter	# of times used
RMSE	Root mean square error	29
R^2	R-squared	19
MAE	Mean absolute error	8
MSE	Mean square error	5
MAPE	Mean absolute percentage error	3
RSAE	Reduced simple average ensemble	3
LCCC	Lin's concordance correlation coefficient	1
MFE	Multi factored evaluation	1
SAE	Simple average ensemble	1
rcv	Reference change values	1
MCC	Matthew's correlation coefficient	1

This gives us an idea of what algorithms and metrics have been used in the field.

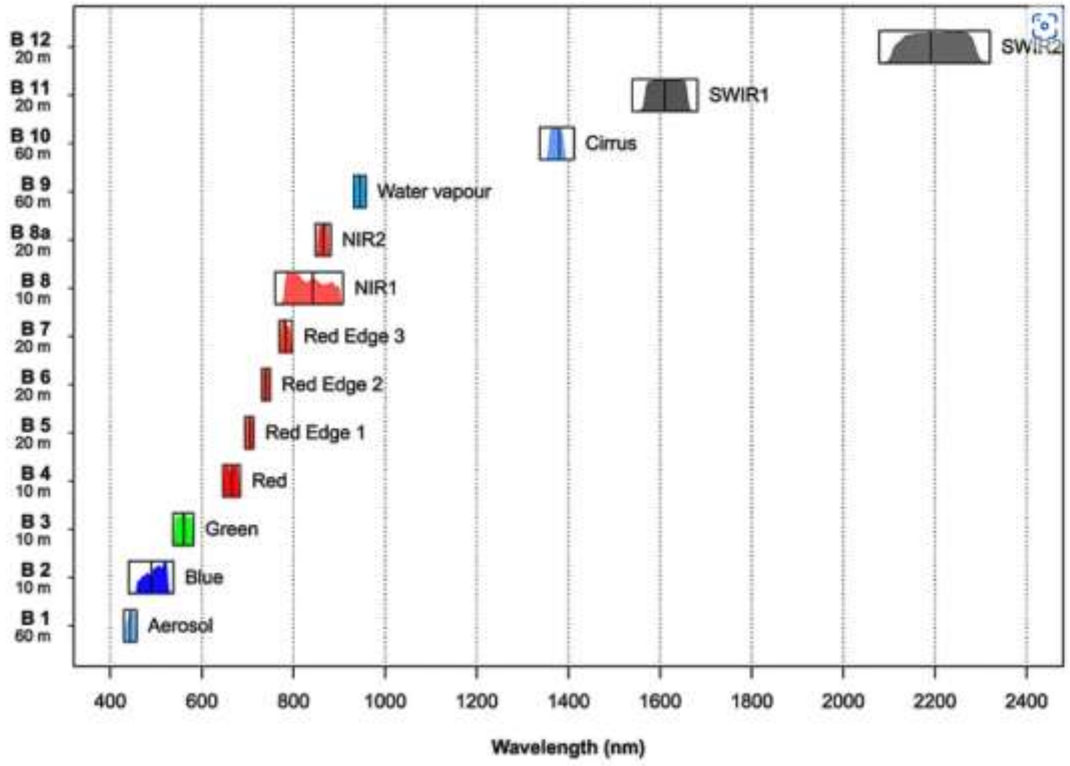
ICSC
Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing

A Practical Exercise: Crop Yield in East Africa

- **We found a nice dataset to perform some tests:** CGIAR Crop Yield Prediction Challenge – Zindi.
- The dataset is about 2552 maize yields in East Africa.
- The dataset is very interesting because it contains features from different domains.

A PRACTICAL EXERCISE: CROP YIELD IN EAST AFRICA

- **We found a nice dataset to perform some tests:** CGIAR Crop Yield Prediction Challenge – Zindi.
- The dataset is about 2552 maize yields in East Africa.
- The dataset is very interesting because it contains features from different domains.
- **Sentinel-2A spectral bands.**



A PRACTICAL EXERCISE: CROP YIELD IN EAST AFRICA

ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

- **We found a nice dataset to perform some tests:** CGIAR Crop Yield Prediction Challenge – Zindi.
- The dataset is about 2552 maize yields in East Africa.
- The dataset is very interesting because it contains features from different domains.

- **Sentinel-2A spectral bands.**
- **TerraClimate informations (TerraClimate - Climatology Lab).**

Variable	Description
Maximum temperature	The highest temperature recorded over a given time period.
Minimum temperature	The lowest temperature recorded over a given time period.
Vapor pressure	The pressure exerted by the water vapor present in the air.
Precipitation accumulation	The total amount of precipitation that has fallen over a specific period of time.
Downward surface shortwave radiation	The amount of solar (shortwave) radiation reaching the Earth's surface, coming from the sun.
Wind-speed	The speed of the wind measured over a specific time period.
Reference evapotranspiration (ASCE Penman-Montieth)	An estimate of the evapotranspiration from a reference surface, such as a green grass cover, under optimum conditions.
Runoff	The portion of precipitation that flows overland to water bodies such as streams, lakes, or oceans, instead of infiltrating the ground.
Actual Evapotranspiration	The actual amount of water that evaporates from the soil and vegetation.
Climate Water Deficit	The difference between potential and actual evapotranspiration, indicating how much water is lacking to meet the demand.
Soil Moisture	The amount of water content held in the soil.
Snow Water Equivalent	The amount of water contained in the snowpack if it were all melted.
Palmer Drought Severity Index	An index measuring the severity of a drought based on recent temperatures and precipitation.
Vapor pressure deficit	The difference between the actual and saturation vapor pressures, indicating potential demand for evaporation.

ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

A PRACTICAL EXERCISE: CROP YIELD IN EAST AFRICA

- **We found a nice dataset to perform some tests:** CGIAR Crop Yield Prediction Challenge – Zindi.
- The dataset is about 2552 maize yields in East Africa.
- The dataset is very interesting because it contains features from different domains.

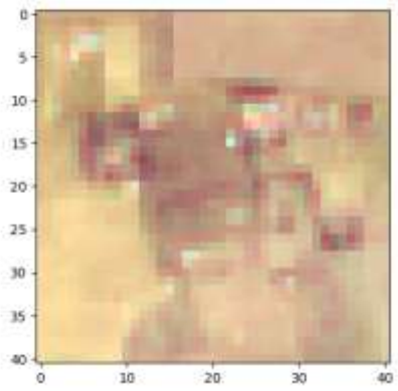
- **Sentinel-2A spectral bands.**
- **TerraClimate informations (TerraClimate - Climatology Lab).**
- **Soil-related Features (SoilGrids — global gridded soil information | ISRIC).**

Variable	Description
soil_bdod_5-15cm_mean	Bulk density of the soil measured between 5-15 cm depth. It's an indicator of soil compaction and porosity.
soil_cec_5-15cm_mean	Cation exchange capacity between 5-15 cm depth. It indicates the soil's ability to retain and supply cations to plant roots.
soil_cfvo_5-15cm_mean	Coarse fragments volume percentage in the soil between 5-15 cm depth.
soil_clay_5-15cm_mean	Percentage of soil particles that are smaller than 0.002 mm in diameter (clay) in the 5-15 cm depth.
soil_nitrogen_5-15cm_mean	Amount of nitrogen content in the soil measured between 5-15 cm depth.
soil_ocd_5-15cm_mean	Organic carbon density in the soil between 5-15 cm depth.
soil_ocs_0-30cm_mean	Organic carbon stock in the soil between 0-30 cm depth.
soil_phh2o_5-15cm_mean	pH of the soil in water solution measured between 5-15 cm depth. Indicates the acidity or alkalinity of the soil.
soil_sand_5-15cm_mean	Percentage of soil particles that are between 0.05 and 2 mm in diameter (sand) in the 5-15 cm depth.
soil_silt_5-15cm_mean	Percentage of soil particles that are between 0.002 and 0.05 mm in diameter (silt) in the 5-15 cm depth.
soil_soc_5-15cm_mean	Soil organic carbon content measured between 5-15 cm depth.

FEATURE ENGINEERING

- From **Sentinel-2A** band information we calculated and used four **vegetation indices,** of which we took median minimum and maximum for each image (1 image/month).
- Only for **CNN approach** we used all pixels (41x41 px).
- For TerraClimate data, we took both the monthly values and the the annual average of all features (**March to October, maize period**).
- We used all the monthly features of the soil.

Index	Description	Calculation
NDVI	Normalized Difference Vegetation Index - A standard measure used to assess whether the observed area contains live vegetation or not. Values range from -1 to 1, where higher values indicate more green vegetation.	NDVI=(NIR+Red)(NIR−Red) where NIR is the near-infrared band and Red is the red band.
GRNDVI	Green Normalized Difference Vegetation Index - Similar to NDVI but uses the green band. It's sensitive to areas with high green vegetation cover.	GRNDVI=(NIR+Green)(NIR−Green) where NIR is the near-infrared band and Green is the green band.
SAVI	Soil Adjusted Vegetation Index - Similar to NDVI but includes a soil adjustment factor to minimize the influence of soil brightness when vegetation is low.	SAVI=(NIR+Red+L)(NIR−Red)×(1+L) where NIR is the near-infrared band, Red is the red band, and L is the soil brightness correction factor (often set to 0.5).
CCCI	Canopy Chlorophyll Content Index - Used to estimate the chlorophyll content in plant canopies. It's calculated using NDVI and the red-edge band.	CCCI=Red−edgeNDVI where NDVI is the Normalized Difference Vegetation Index and Red-edge is the red-edge band.

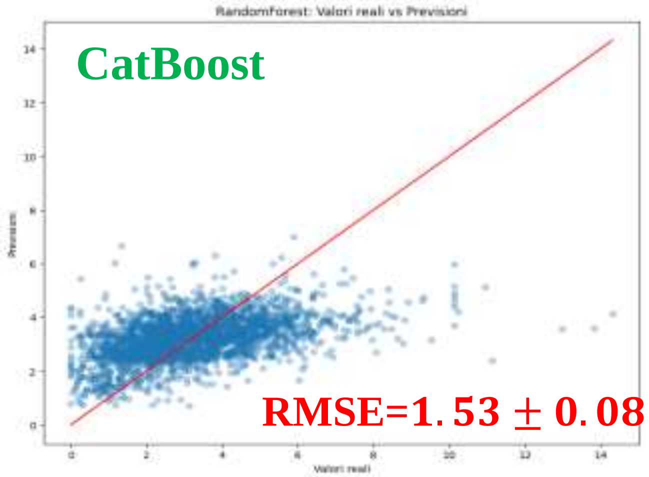


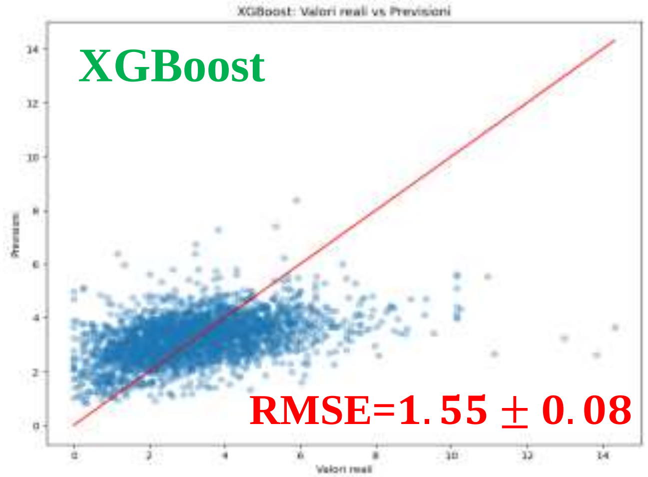
A typical image of a field in rgb

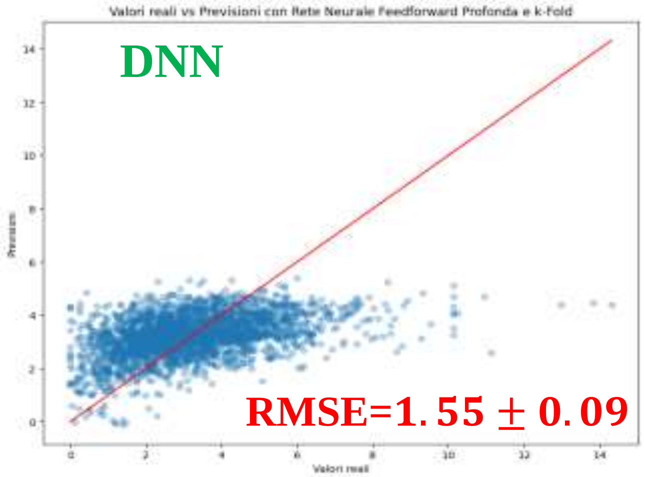
ML AND DL DEVELOPMENT PIPELINE

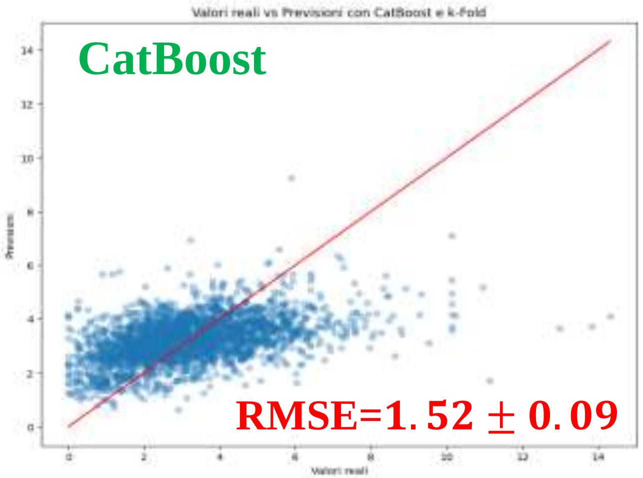
- After preparing the dataset, we experimented with different ML and DL models for **crop yield regression**.
- The following models were tested (rmse metric and k=5-Fold validation):
    - Random Forest
    - XGBoost
    - DNN
    - CatBoost
    - Hybrid CNN (S2A image part)+DNN (climate and soil info) (2 Fold)
    - Long Short Term Memory (LSTM) network (preliminary) (2 Fold)
- Almost all models were fine-tuned with **Optuna (Bayesian optimiser)** with a similar number of iterations (around 100).

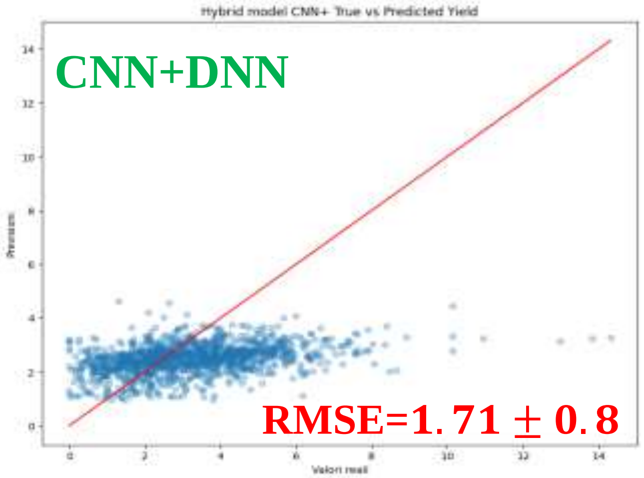
RESULTS

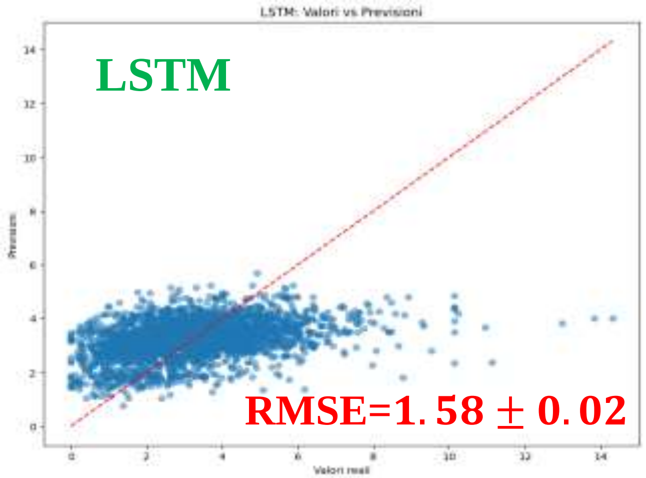

RandomForest: Valori reali vs Previsioni
CatBoost
RMSE=1.53 ± 0.08
Valori reali
Previsioni


XGBoost: Valori reali vs Previsioni
XGBoost
RMSE=1.55 ± 0.08
Valori reali
Previsioni

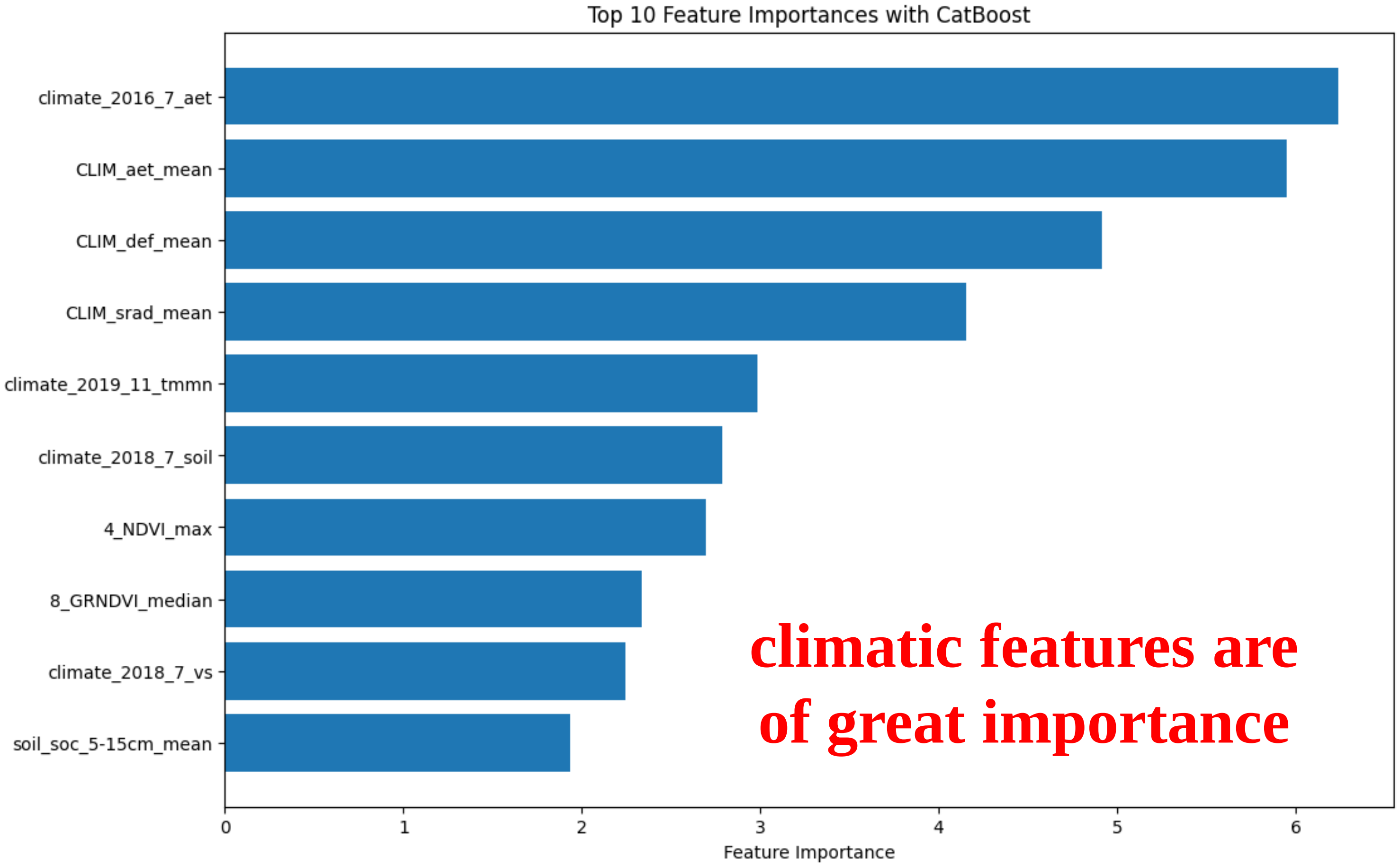

Valori reali vs Previsioni con Rete Neurale Feedforward Profonda e k-Fold
DNN
RMSE=1.55 ± 0.09
Valori reali
Previsioni


Valori reali vs Previsioni con CatBoost e k-Fold
CatBoost
RMSE=1.52 ± 0.09
Valori reali
Previsioni


Hybrid model CNN+ True vs Predicted Yield
CNN+DNN
RMSE=1.71 ± 0.8
Valori reali
Previsioni


LSTM: Valori vs Previsioni
LSTM
RMSE=1.58 ± 0.02
Valori reali
Previsioni

WORD FROM THE WINNER



ICSC
Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing

CONCLUSIONS

- **From the studies and texts in this presentation, we have learned a lot about datasets and methods for crop yield prediction.**

- The datasets must include features from many domains, including satellite, climate and soil data.

- **Feature Engineering** is a key!

- Many ML/DL methods can be used, and **a lot of optimisation to be done**. Making/obtaining the dataset quickly would give time to do a lot of experimentation.