Advances in Machine Learning–Based Cloud Detection from IASI Observations

UNIVERSITÀ DEGLI STUDI FIRENZE

Istituto per le Applicazioni del Calcolo

MAIL:
chiara.zugarini@fi.iac.cnr.it
chiara.zugarini@unifi.it

Scan for Preprint Version

5th ECMWF-ESA Machine Learning Workshop

ECMWF esa

Authors: *Chiara Zugarini [a,b], Cristina Sgattoni [c], Mathias Chung [d], Francesco Pio De Cosmo [a,b], Luca Sgheri [b]*

a. Department of Information Engineering - University of Florence (IT)
b. Institute for Applied Mathematics (IAC) – National Research Council (CNR) – Florence (IT)
c. Institute of BioEconomy (IBE) – National Research Council (CNR) – Florence (IT))
d. Department of Mathematics and Computer Science – Emory University – Atlanta (GA, USA)

Machine Learning technique [1,2,3]

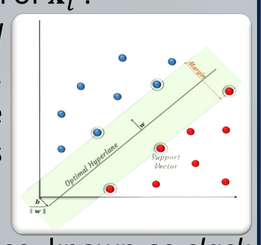
SVM (Support Vector Machine)

The *SVM method* is based on the principles of optimization theory, convex analysis, and linear algebra. This class of algorithms is also used in the case of non-separable data. In this case, the technique, called the *kernel method*, exploits a function, known as a *kernel function*, which transforms the input data into a higher-dimensional *feature space*.

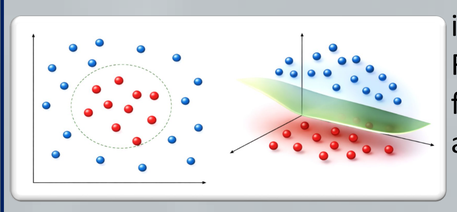
Let:

- $\mathbf{x}_i$ for $i = 1, \dots, l$ is a set of training points, $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_l\}$ represents the measurements (*IASI spectra*).
- y_i for $i = 1, \dots, l$ with $y_i = 1$ (clear) or $y_i = -1$ (cloudy); $\mathbf{y} = \{y_1, \dots, y_l\}$ represents the "truth" values associated with $\mathbf{x}$ and y_i identifies the classification of $\mathbf{x}_i$.

Separable data: there exists an *optimal hyperplane* such that the two data classes, distinguished by their truth values, are included in two different half-spaces separated by the hyperplane.



Non-separable data: by adding new variables, known as *slack variables*, with additional constraints, and using the *kernel method* to map the data into feature space, it is possible to obtain an optimal hyperplane, similar to the case of separable data.

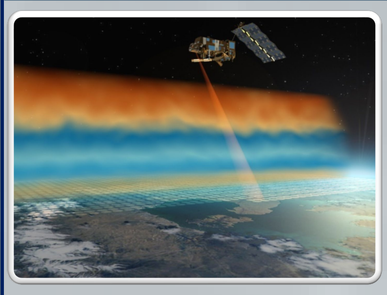


Kernel function

in our case, the Kernel Function is a polynomial function K, of degree q, applied to $\mathbf{X}$

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^q.$$

IASI instrument [4]



The *Infrared Atmospheric Sounding Interferometer* (IASI) is an instrument onboard the Metop series of satellites.
These measurements are crucial for weather forecasting and climate changing prediction.

The *IASI L1c product* contains infrared radiance spectra at $0.25\ \mathrm{cm}^{-1}$ sampling. The product has, for each sounded pixel, 8461 spectral samples covering the range between $645\ \mathrm{cm}^{-1}$ and $2760\ \mathrm{cm}^{-1}$ wavenumbers.

Variables extracted from L1c IASI product

- '`measurement_date`' – scan date;
- '`lat`' – latitude of scan;
- '`lon`' – longitude of scan;
- '`gs_1c_spect`' – radiance spectra;
  (It is necessary to use the scaling factors to reconstruct spectra in the range of given wavenumbers)
- '`geum_avhrr_1b_cloud_fraction`' – cloud fraction.

ERA5 reanalysis database [5]

The *ERA5* reanalysis database combines model data with observations from different instrument into a consistent dataset. Data has been regridded to a regular latitude-longitude grid with spatial resolution of *0.25° x 0.25°* .

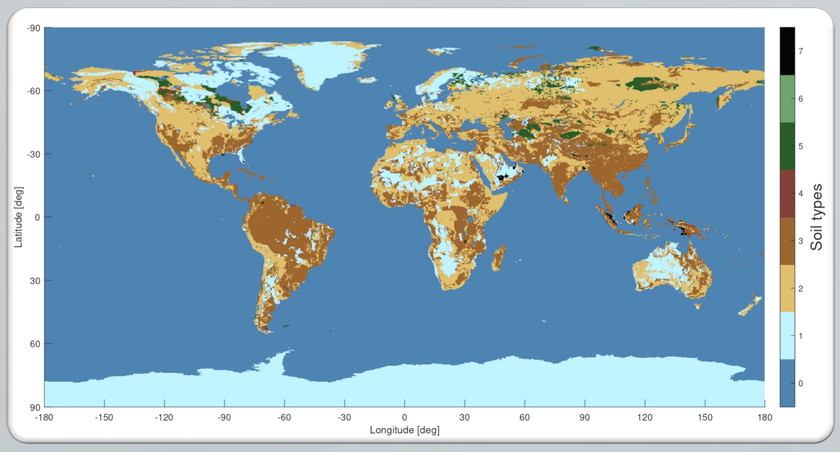
Variables extracted from ERA5 database:

- '`latitude`' – latitude of ERA5 grid point;
- '`longitude`' – longitude of ERA5 grid point;
- '`soil type`' – the texture (or classification) of surface.

The eight *Surface Types* are:

0 – Water	*4* - Fine
1 - Coarse	*5* - Very fine
2 - Medium	*6* - Organic
3 - Medium fine	*7* - Tropical organic

Surface Types from ERA5 database



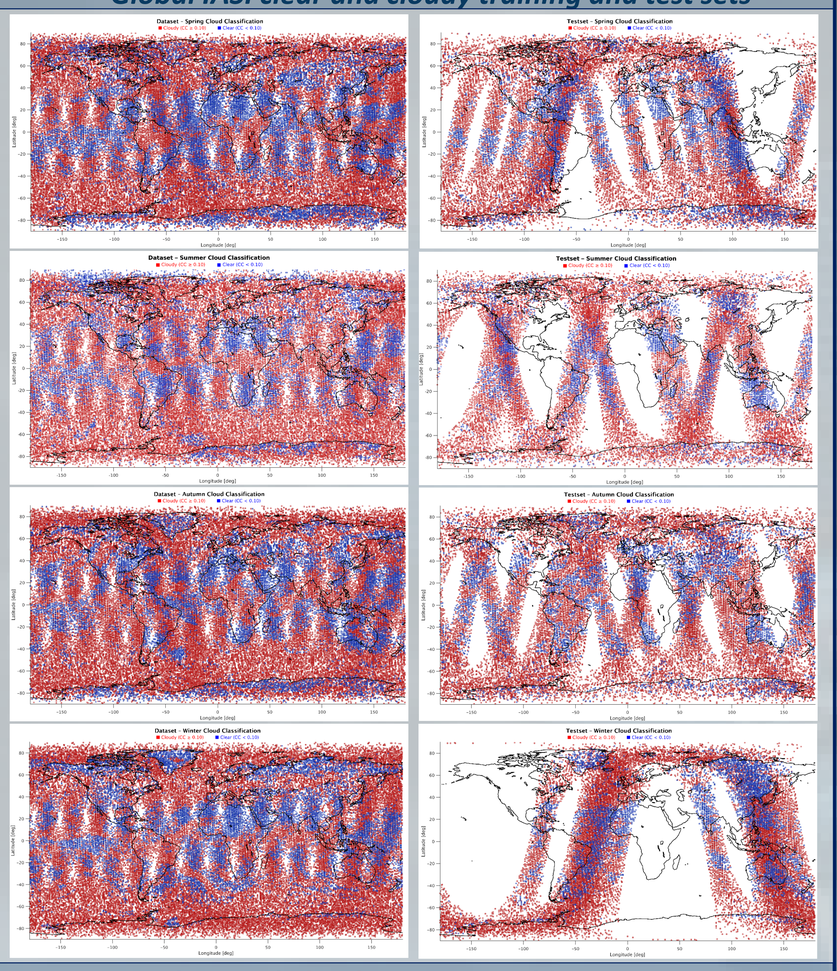
IASI data

Training Sets — *IASI Orbits*	*Test Sets* — *IASI Orbits*
Spring: N. 16 Orbits March/April 2023	*Spring:* N. 8 Orbits March 2020 April 2021
Summer: N. 16 Orbits June/July 2023	*Summer:* N. 8 Orbits June 2020 July 2021
Autumn: N. 16 Orbits Sept./October 2023	*Autumn:* N. 8 Orbits September 2020 October 2021
Winter: N. 16 Orbits Dec. 2023/Jan. 2024	*Winter:* N. 8 Orbits December 2020 January 2021

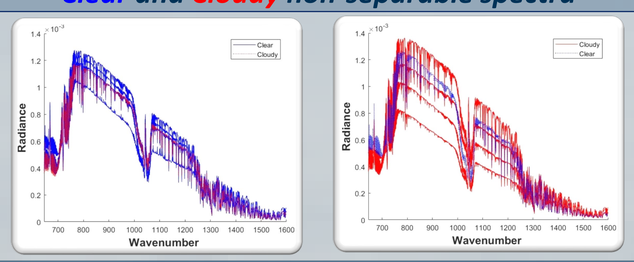
Selected IASI measurements for training set and test set

- For each IASI pixel, a Surface Type is associated using the nearest ERA5 grid point;
- IASI pixels with *cloud fraction<10%* coverage were selected as clear cases;
- IASI pixels with *cloud fraction≥10%* coverage were selected as cloudy cases;
- in the spectral range of IASI $[645; 2760]\ \mathrm{cm}^{-1}$, *different custom range or channels* was selected.

Global IASI clear and cloudy training and test sets



Clear and Cloudy non-separable spectra



SVM applied to IASI measurements

To create a method that works for any geolocation:

1. *Different training sets were created for each surface type and compared with the test sets using the SVM technique.*
2. *The training sets for each surface type were further subdivided in two different ways to reduce the dimensionality.*

SVM Classification Results

Total Training and Test Sets

SURFACE TYPE	N. training sets	N. test sets	SURFACE TYPE	N. training sets	N. test sets
0	~ 650000	~ 300000	*4*	~ 35000	~ 15000
1	~ 130000	~ 66000	*5*	~ 600	~ 250
2	~ 100000	~ 50000	*6*	~ 8000	~ 4000
3	~ 35000	~ 20000	*7*	~ 600	~ 350

High Test Results

Global Dataset + PCA channel selection

↓

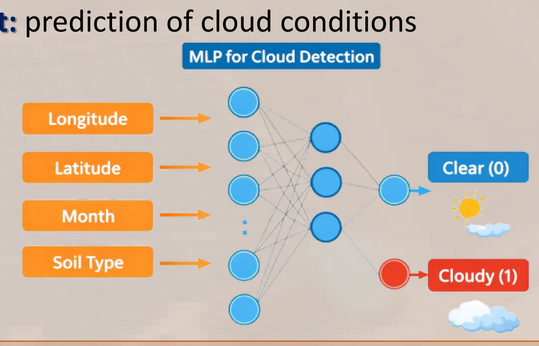
Accuracy > 88%

PCA used for computational efficiency and extraction of the most informative data features

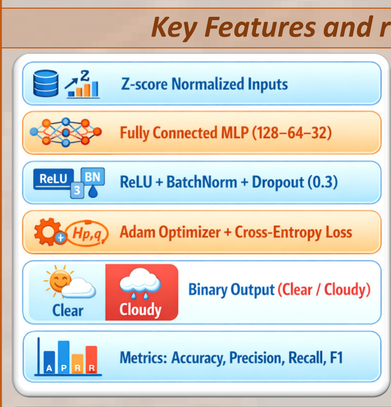
Advanced Mathematical Approach: MLP

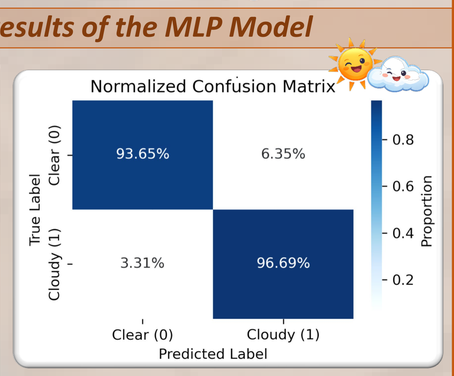
The **M**ulti-**L**ayer **P**erceptron (MLP) cloud detection approach:

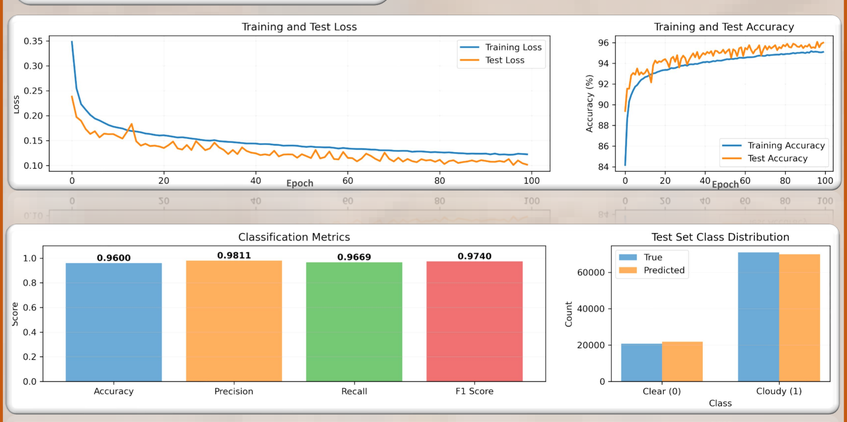
- **Input features:** geospatial data, temporal, and surface type information
- **Feature processing:** inputs are fed into a fully connected neural network
- **Hidden layers:** the MLP learns non-linear relationships among input variables through successive transformations
- **Decision function:** the network maps the extracted features into a binary classification space
- **Output:** prediction of cloud conditions



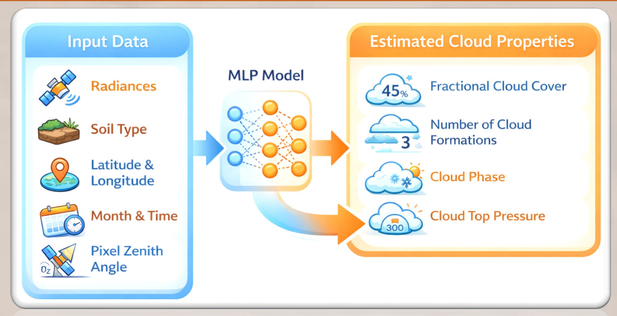
Key Features and results of the MLP Model







Next Steps: Cloud Property Estimation



CONCLUSIONS

*The proposed machine learning framework demonstrates that both SVM and MLP approaches are effective for cloud detection from IASI observations. While **SVM** provides a solid baseline for classification, the **MLP** model shows enhanced capability in capturing complex non-linear relationships between geophysical and radiative inputs. The results highlight robust and generalizable performance across different surface types and global conditions. This work establishes a reliable foundation for future developments toward advanced cloud property retrieval.*

REFERENCES

[1] V. N. Vapnik and A. Ya. Chervonenkis. *A class of algorithms for pattern recognition learning, volume 25. Avtomatika i Telemekhanika, 1964.*
[2] Boser, Bernhard & Guyon, Isabelle & Vapnik, Vladimir. *(1996). A Training Algorithm for Optimal Margin Classifier. Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory. 5. DOI: 10.1145/130385.130401.*
[3] Matlab. *Train support vector machine (svm) classifier for one-class and binary classification.* https://it.mathworks.com/help/stats/fitcsvm.html.
[4] EUMETSAT. *Iasi level 1: Product guide,* https://user.eumetsat.int/s3/eup-strapi-media/pdf_iasi_pg_487c765315.pdf , *2020. Last accessed: 2024-06-27*
[5] Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N. *(2023): ERA5 monthly averaged data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.f17050d7 (Accessed on 27-JUN-2024)*
[6] Whitburn, S., Clarisse, L., Crapeau, M., August, T., Hultberg, T., Coheur, P. F., & Clerbaux, C. *(2022). A CO_2-independent cloud mask from Infrared Atmospheric Sounding Interferometer (IASI) radiances for climate applications. Atmospheric Measurement Techniques, 15(22), 6653–6668.*
[7] Palchetti, L., Buehler, S. A., Camy-Peyret, C., Cortesi, U., Di Natale, G., Dinelli, B. M., ... & Ridolfi, M. (2020). *FORUM: Unique far-infrared satellite observations to better understand how Earth radiates energy to space. Bulletin of the American Meteorological Society, 101(12), E2030–E2046.* https://doi.org/10.1175/BAMS-D-19-0322.1