T-FORS

WP2: LSTID ML forecasting model

C. Cesaroni on behalf of WP2 participants

INGV

Claudio.cesaroni@ingv.it

1st External Expert Advisory Board Meeting

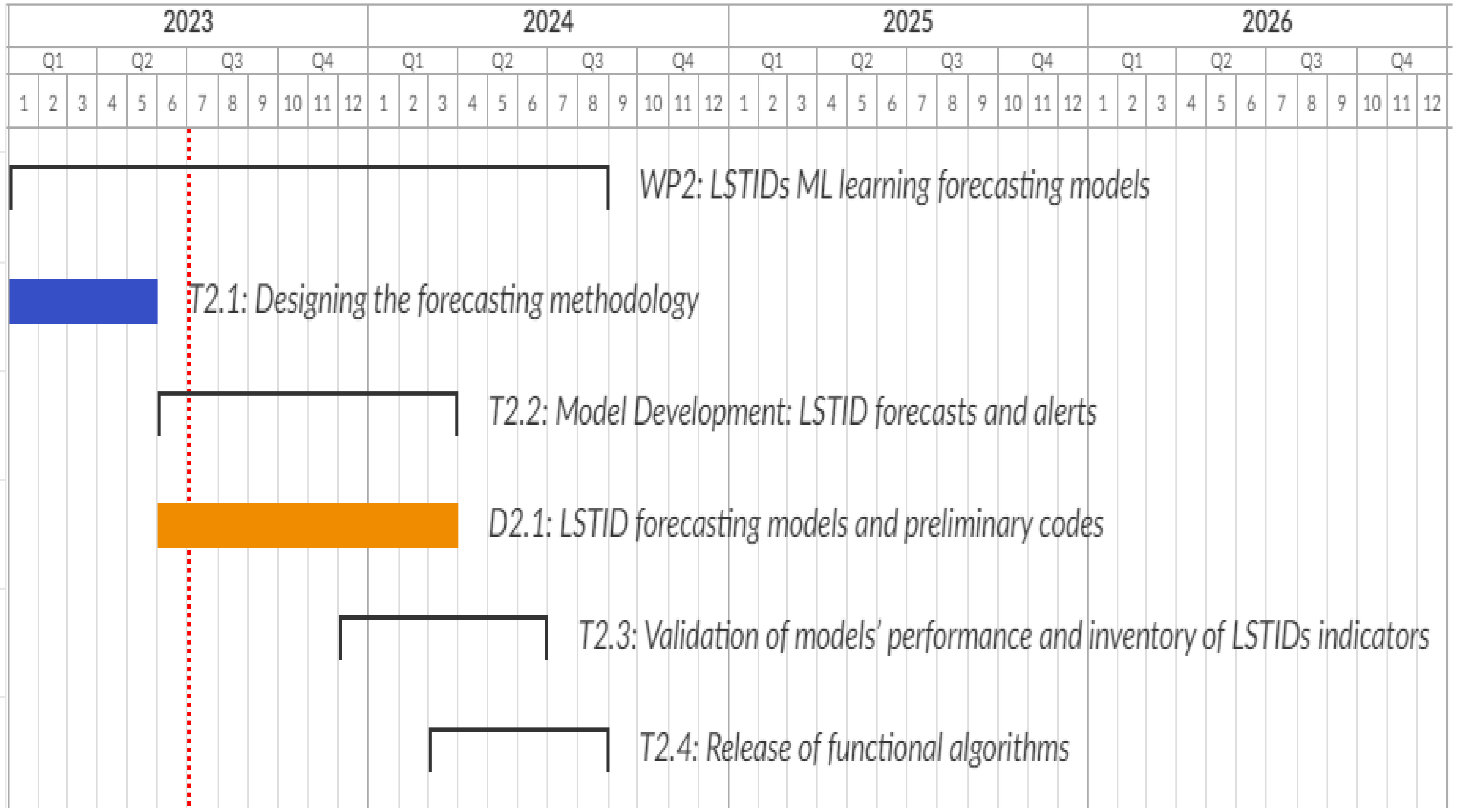
Thursday 6 July 2023

Outline

- **The data:**
  - HFI-EU index
  - TIDs catalog
- **Possible approaches:**
  - HFI-EU index: LSTM-regression (NOA)
  - HFI-EU index: KNN-regression / FNN-classification (EBRO)
  - TIDs catalog: classification (INGV)
- **Future work**

GANTT and Milestones

SUBJECT	START DATE	FINISH DATE
WP2: LSTIDs ML learning forecasting models	01-01-2023	31-08-2024
T2.1: Designing the forecasting methodology	01-01-2023	31-05-2023
T2.2: Model Development: LSTID forecasts and alerts	01-06-2023	31-03-2024
D2.1: LSTID forecasting models and preliminary codes	01-06-2023	31-03-2024
T2.3: Validation of models' performance and inventory of LSTIDs indica...	01-12-2023	30-06-2024
T2.4: Release of functional algorithms	01-03-2024	31-08-2024



FIRST MILESTONE completed

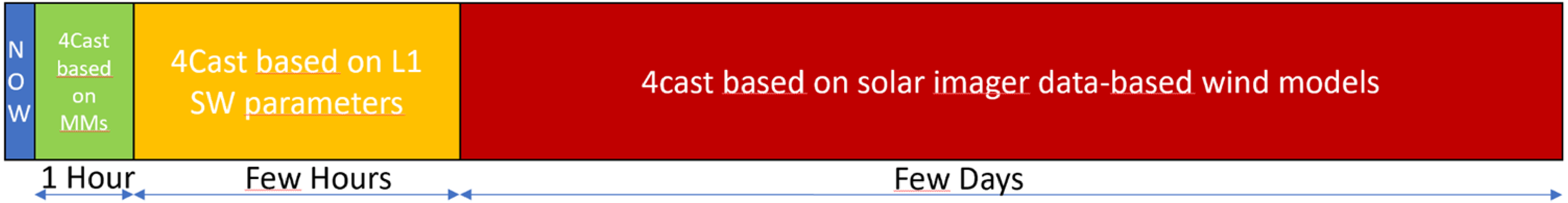
MS3	Definition of the LSTID forecasting models – design of ML learning experiments	WP2	INGV	A report will be available in the project wiki.	5	31-05-2023	Achieved

NEXT MILESTONE: due to 31/12/23 first release of forecasting codes

MS3 - Report on the ML approach

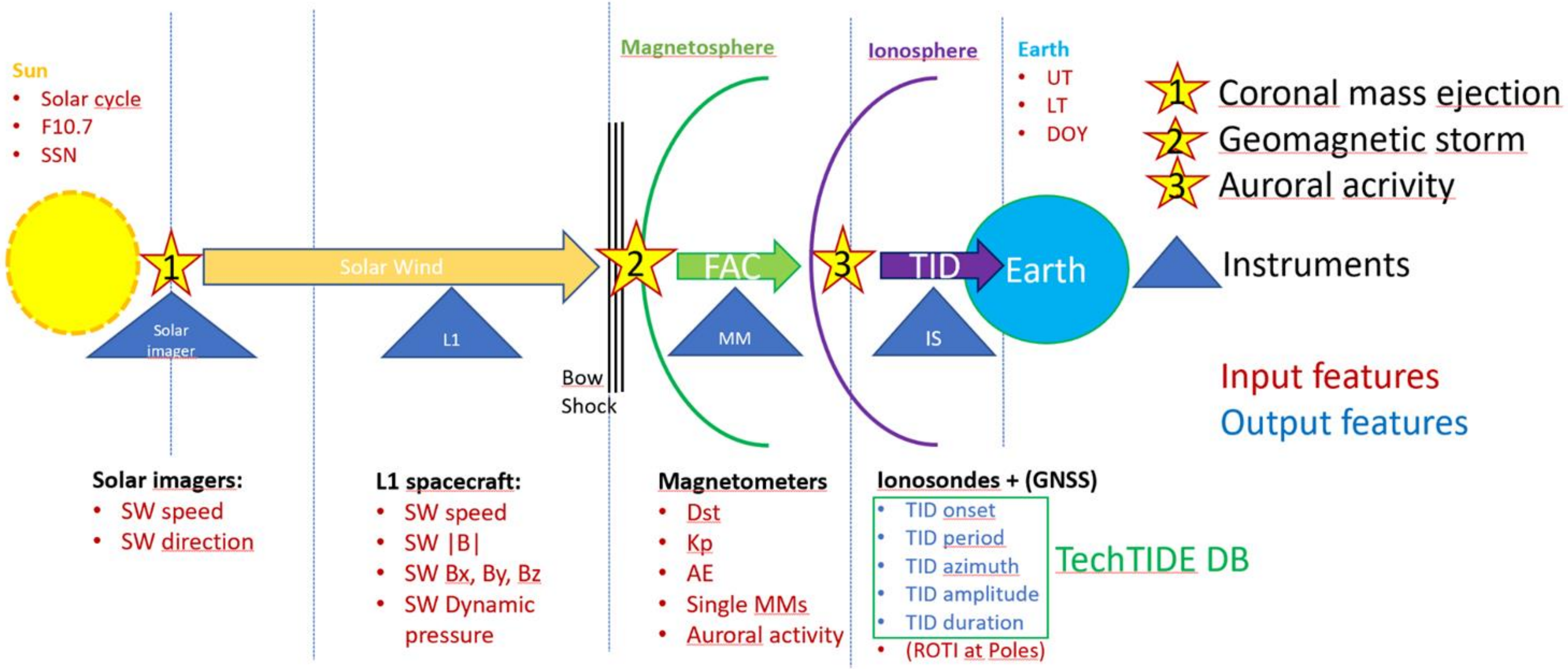
The report presents the strategy for the development of the Machine Learning algorithm dedicated to forecasting LSTIDs over the European sector:

- It describes the objectives of the Machine Learning Modelling for LSTIDS.
- Presents the approach of the modelling, providing insights on input data, model features, datasets and labels
- Provides the conceptual workflow of the three foreseen families of modelling (ST-HA, MT-MA and LT-LA) and some early implementation.
- It presents the foreseen validation strategy.



ML approach - The physical problem

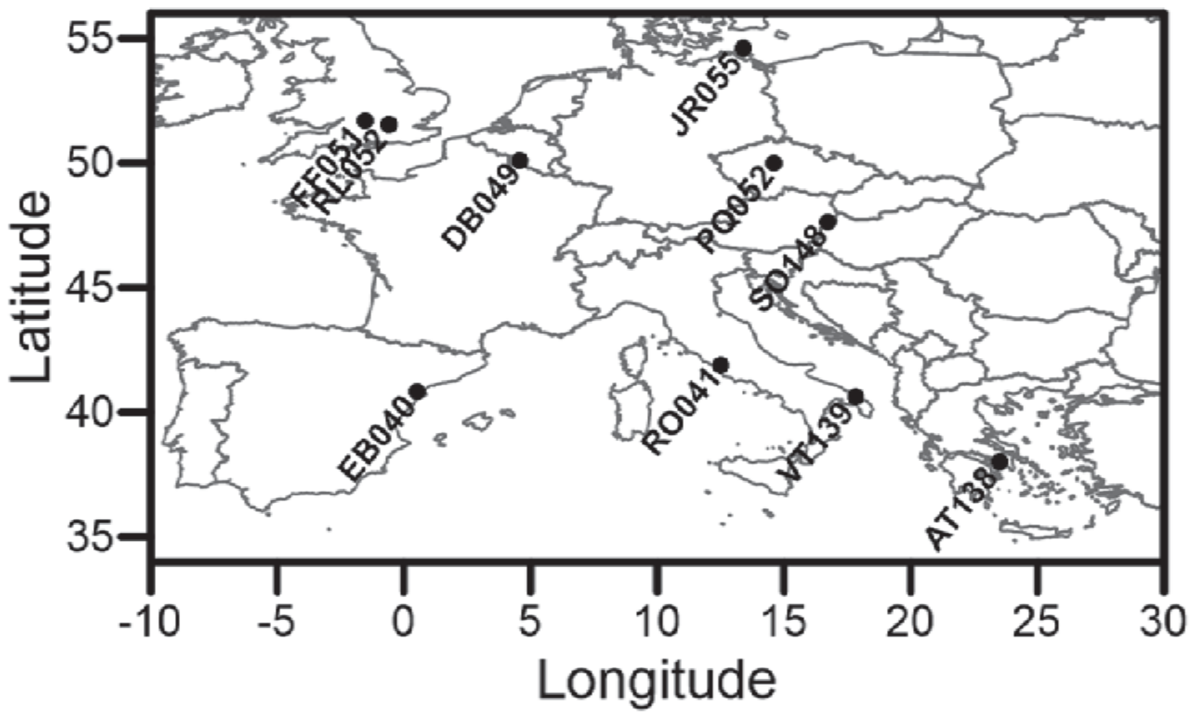
As anticipated, three families of models are necessary to cover the complex chain of events\interactions causing LSTIDs.



Input features to include are still being investigated through ML experiments

The Detection method: HF-Interferometry

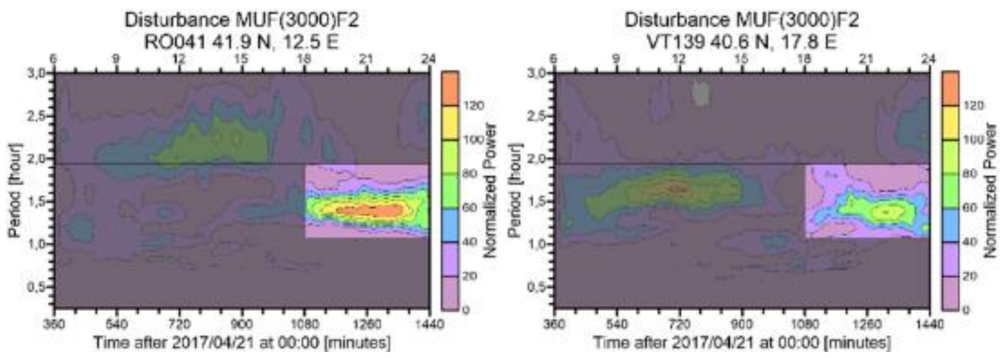
INPUT



- Characteristics from VI Ionospheric sounding (**MUF(3000)F2**).
- Network of DPS4D with stations working **synchronized**.
- GIRO DIDBase Fast Chars database http://giro.uml.edu/didbase/scaled.php

- Detection of TID-like variation

Detect coherent TID-like variations by spectral analysis.



- TIDs contribution to data variability.

Application of the Parseval's relation

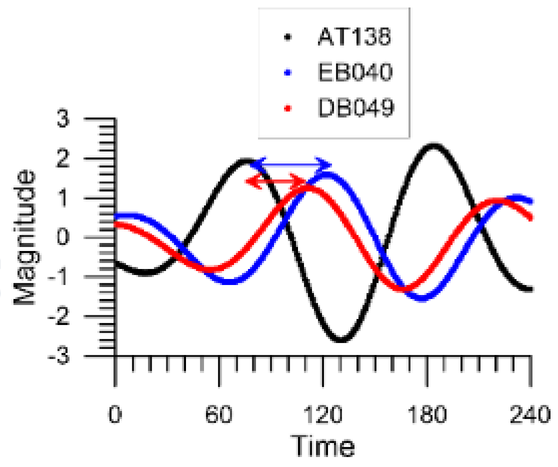
$$\sum_{n=-\infty}^{\infty}|x[n]|^2=\frac{1}{2\pi}\int_{-\pi}^{\pi}|X(\omega)|^2 d\omega \sim \sum_{T=T_S}^{T=T_E}A(\omega)^2$$

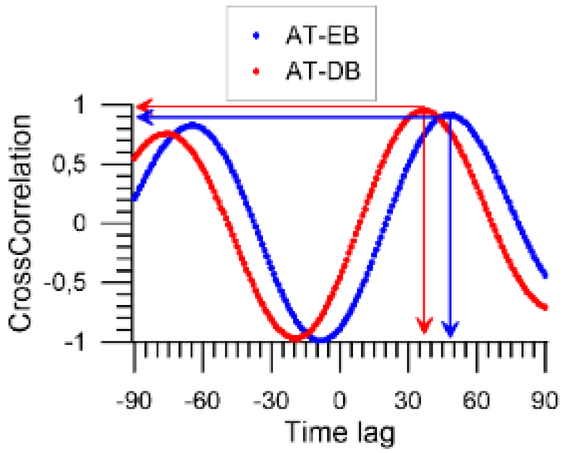
$$SEC(\%)=\frac{\sum_{T=T_{TID_S}}^{T=T_{TID_E}}A(T)^2}{\sum_{T=T_S}^{T=T_E}A(T)^2}$$

- Estimation of the velocity and azimuth of the TID

Estimate time delays for different sites by cross-correlation, ΔTM_i. Estimate velocity of disturbance $\vec{v}$ assuming planar propagation.

$$\Delta TM_i-\vec{s}\cdot\Delta\vec{r}_i=0\ ;$$





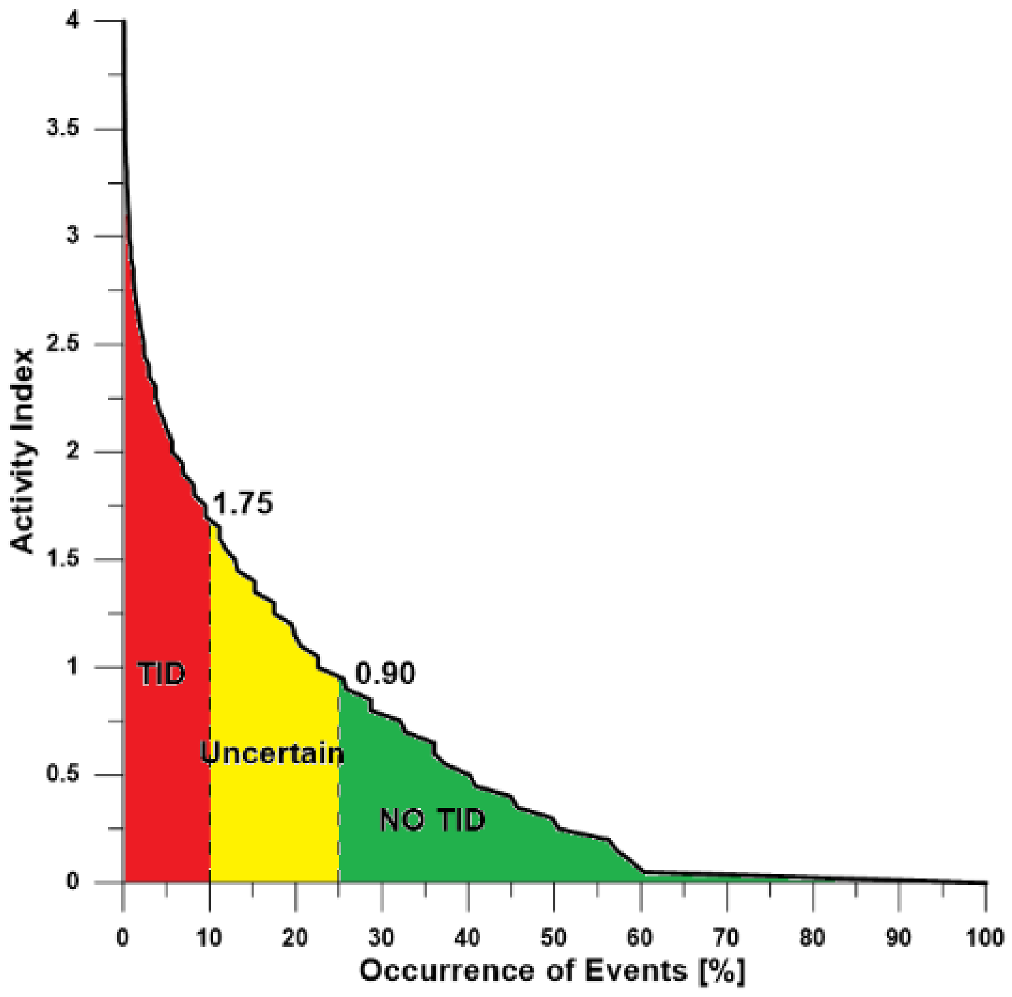
Products associated

- **HF-EU index**

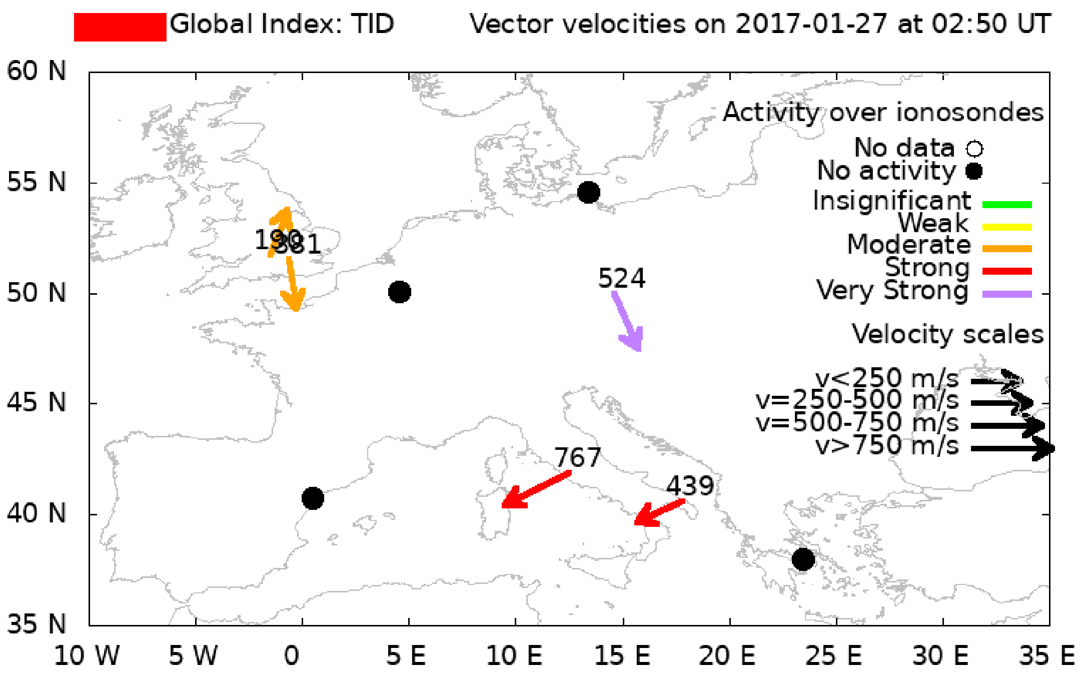
```
EU 201701271330 TrL AV= 3.00 Area= 66.00% ActivityIndex= 1.98 TID
SA 201701271330 TrL AV= 0.00 Area=  0.00% ActivityIndex= 0.00 NO TID
```

OE_HFI_YYYYMMDDHHmm_COND.log files every 5 minutes

- One index for the whole network.
  - Is the product of the average of intensity of the TID (related to the spectral contribution) multiplied by the area affected (number of stations).
  - The thresholds have been established by statistics
    - 0 means no data
    - 0.1 means nothing detected
- Only data from April 2019 is available in the TechTIDE portal. OE will provide data from January 2014.



Problems of the method



Sporadic E layer, Es

- We cannot see what is happening in F layer.
  - Affects specially on summertime at central hours of the day.

Lack of data

- Technical problems in some stations
- Connectivity problems with GIRO DIDBase
  - The TechTIDE portal storage the real-time data. To fix connectivity problems, time to time a reanalysis is carried out. But it is not storage in the TechTIDE portal

Uncertainty in the azimuth determination at the edge of the network

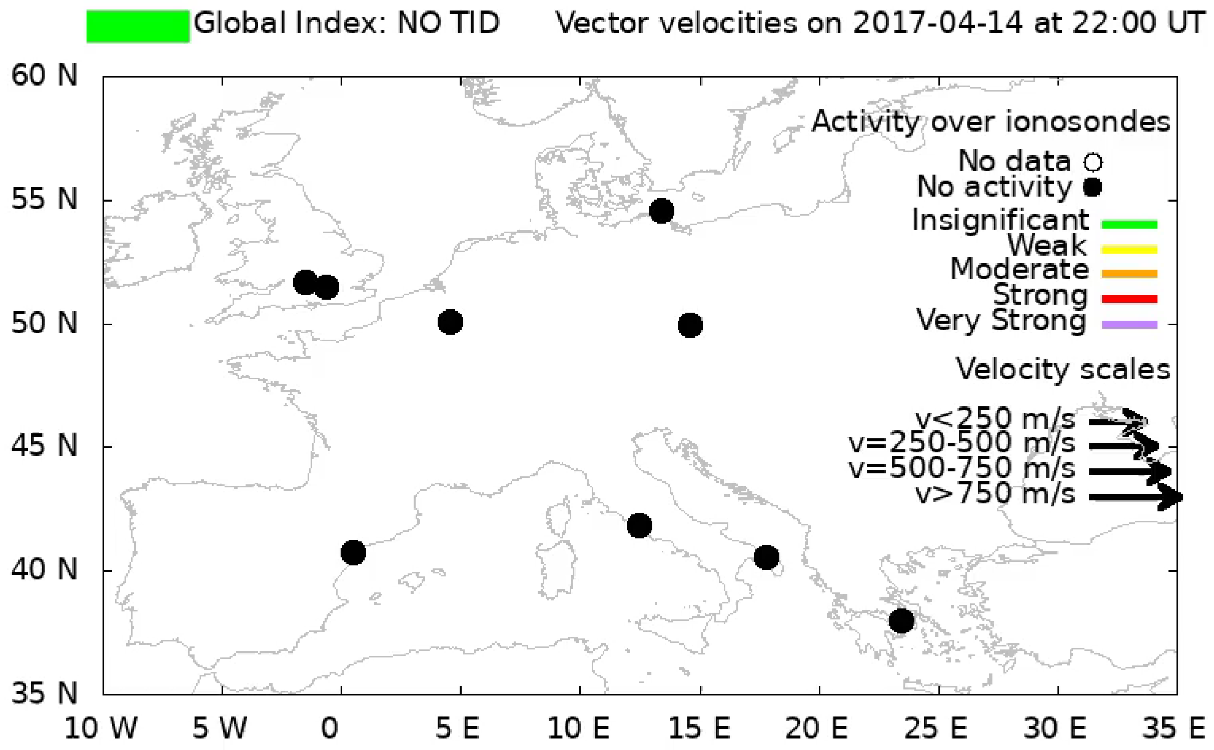
- The methodology to find the azimuth has an intrinsic uncertainty of 360º for stations located at the edge of the network (not usual but sometimes happens).

Intrinsic delay (Need to adopt a criteria for time detection)

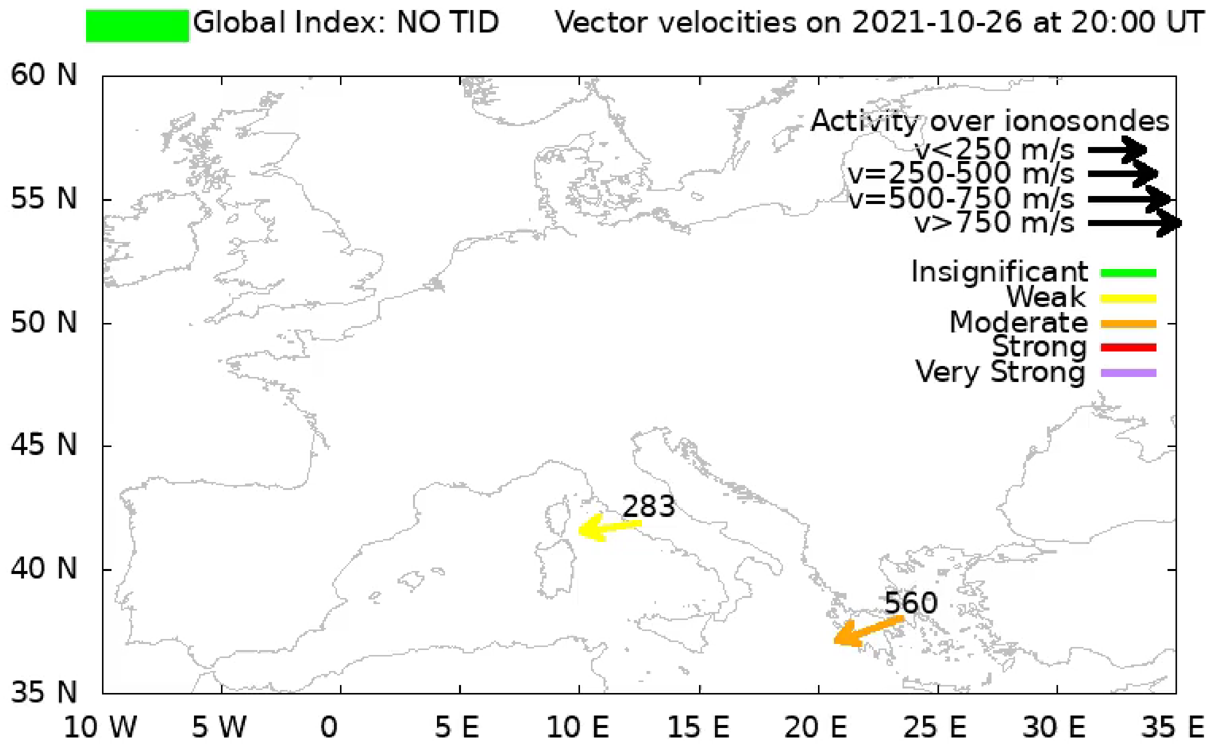
- The detection time refers to the last download of the data. Then the method looks for periodicities in the previous 6 hours.
- As we look for periodicities in the input data, we need a full period to detect it.
- The method considers a detection if there is a coherent periodicity in as minimum 4 stations. Then, a propagation time is needed to affect 4 station, it will depend on the azimuth of the perturbation and the velocity.
- Impact on the distribution of the time of detection

HF-INT: Catalogue of events

- **Visual inspection to determine LSTIDs events**
  - ➢ Looking for coherent velocity propagation
  - ➢ 760 TIDs events detected and recorded above Europe between FEB 2014 to DEC 2022
- **Determination of onset time and duration**
  - ➢ Approximative (see slide about delay)
- **Average of the main characteristics of the TID for all stations and during the whole event.**



Included in the catalogue



Not included in the catalogue

EHF-EU index

- **Pros**
  - ➢ Automatic determination of the index clear criteria.
- **Cons**
  - ➢ Not all events with large index (above 1.75) are LSTIDs. Presence of the solar terminator effects and situations with a perturbation but with a non-coherent velocity. *Idea for mitigation strategy: keep only events with continuous large index for at least 60-75 minutes.*
  - ➢ No spatial information (you must be back to the raw data)
  - ➢ Although, you can determine an onset time automatically, you must keep in mind the delay problem of the method.

CATALOG

- **Pros**
  - ➢ We are sure that all events in the Catalogue are LSTIDs.
  - ➢ One file per year. Easy to work with.
- **Cons**
  - ➢ Not all TIDs are in the Catalogue, maybe not detected, no data, etc.
  - ➢ No spatial information (you must be back to the raw data)
  - ➢ Created by human inspection, probably biased.
  - ➢ Difficulties to determine the starting time and the duration of the event.

HFI-EU index: LSTM-regression

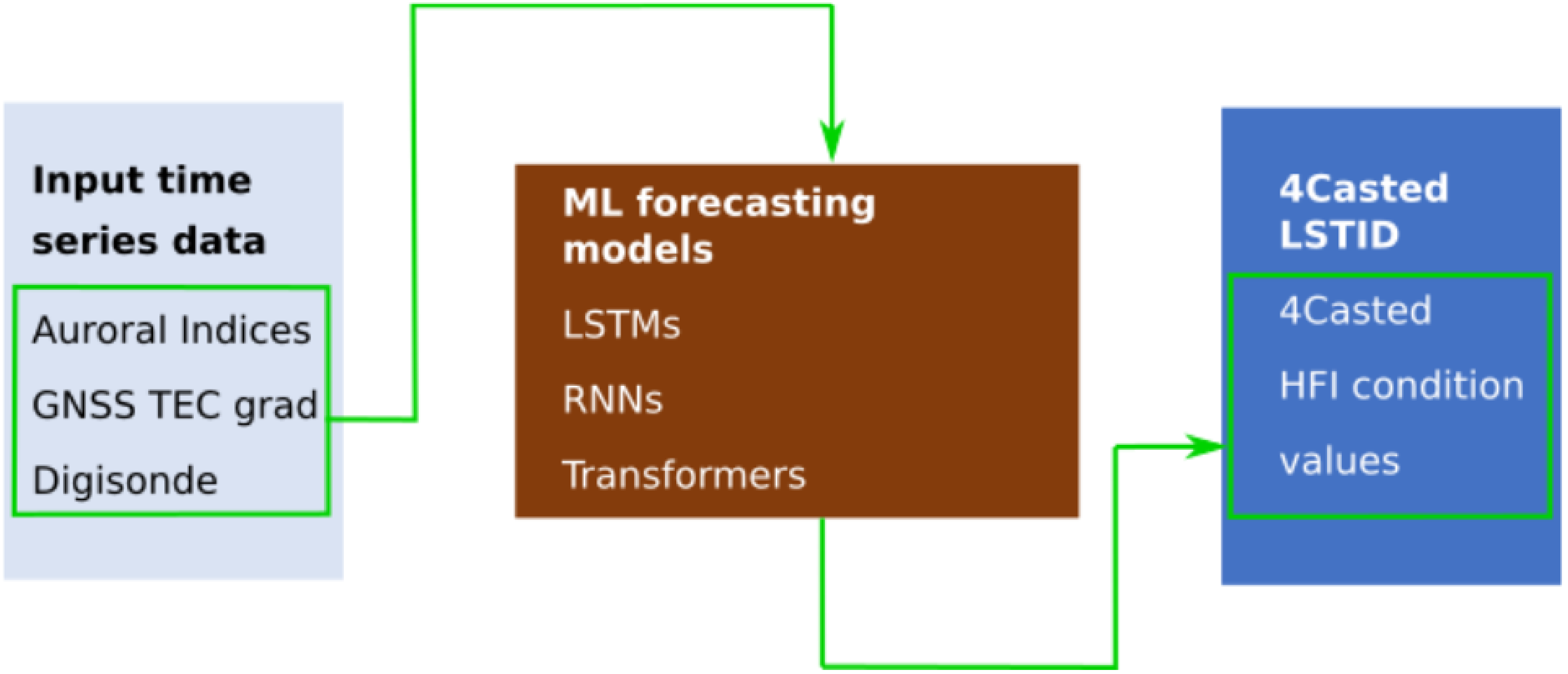
We formulate a forecasting problem using Machine Learning and Deep Neural Networks

Input data

- Auroral Indices: IL, IU and IE IMAGE electrojet indicators
- GNSS TEC gradient over Europe, as provided by DLR.
- Digisonde observations: HFI activity index derived from the Ebro Observatory

Output data

- Forecasted HFI condition values

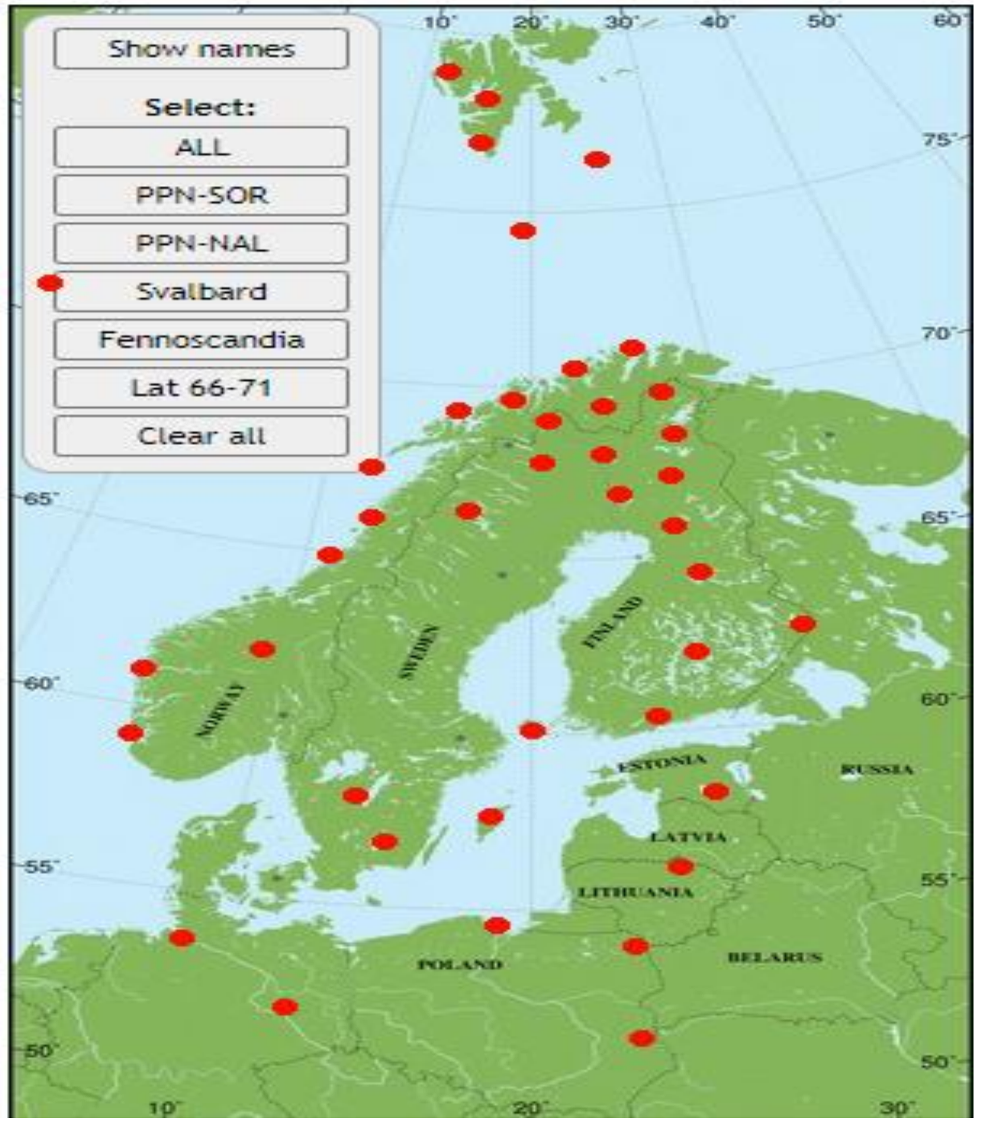


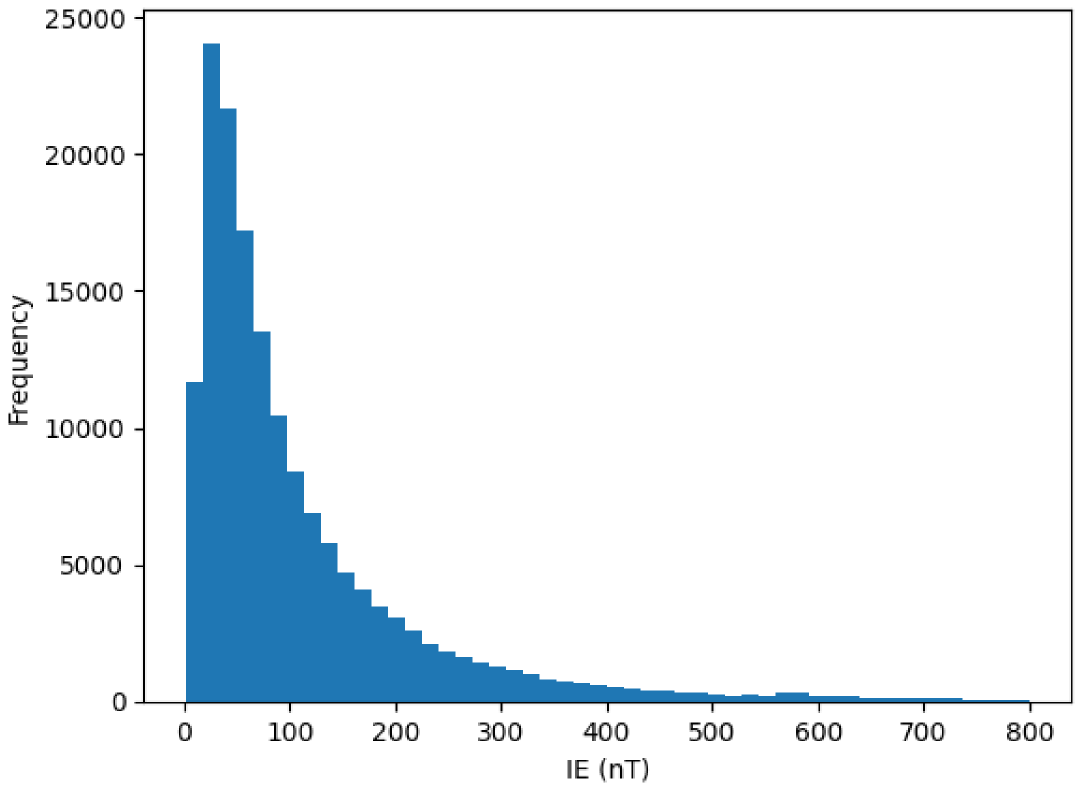
Electrojet indicator (IE) from IMAGE

IMAGE electrojet indicators are simple estimates of the total eastward and westward currents crossing the magnetometer network.

Analogously to the auroral electrojet indices, IE is a measure of the horizontal component variation of the magnetic field.

- ➢ $IL(t) = \min(\{\Delta X(t)\})$,
- ➢ $IU(t) = \max(\{\Delta X(t)\})$, and
- ➢ $IE = IU - IL$





Dataset preparation

Data cleaning

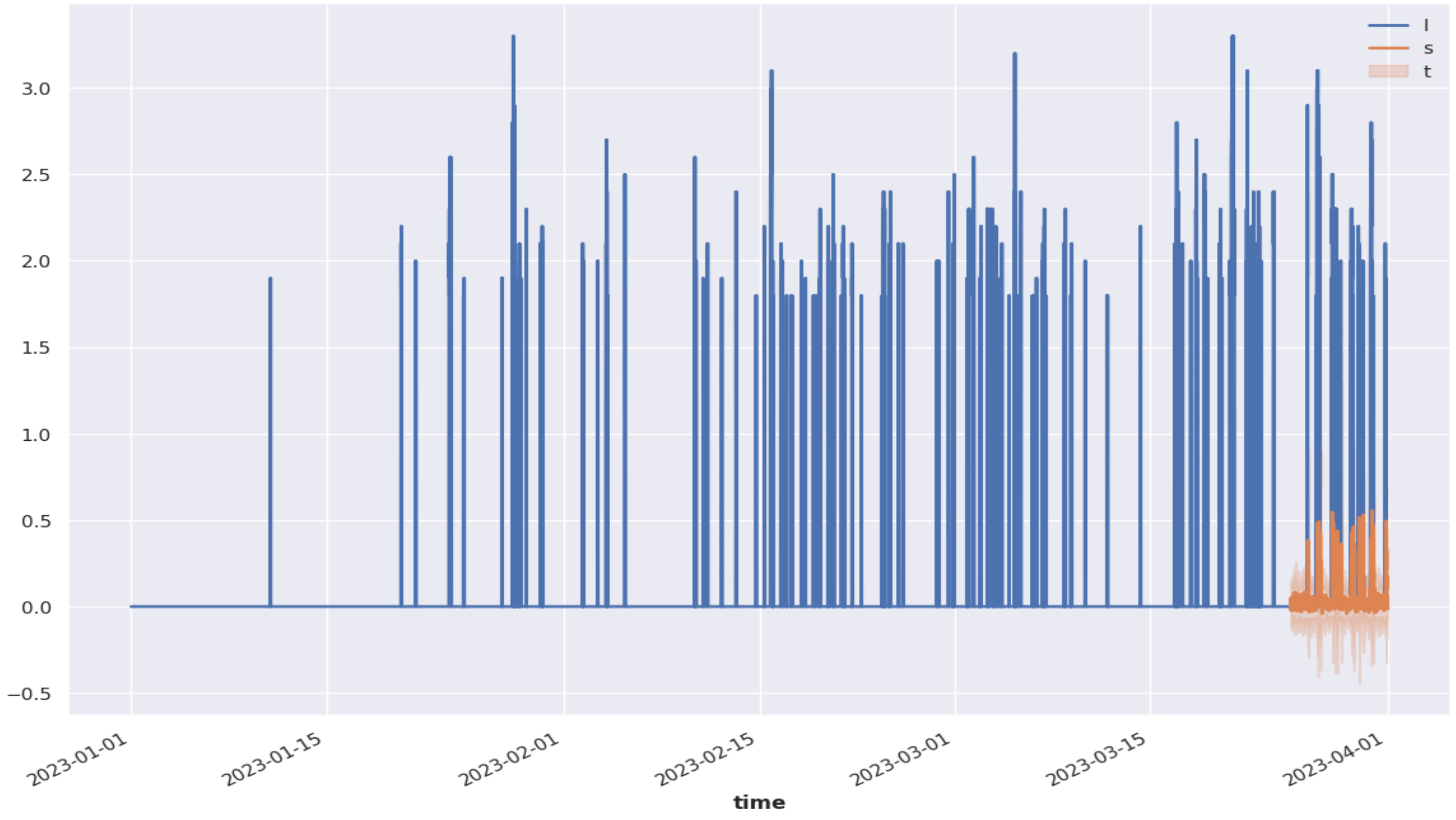
- Input time series are sampled at a time interval of 5min
- Missing measurement values have been imputed
- Data have been scaled accordingly in interval [0, 1].
- We utilize the Darts library, https://github.com/unit8co/darts.

Data splitting

- Split data in training and validation set
- LSTM model is trained on the train data set and its performance is evaluated using the validation set
- We have selected the last six days of our data set as validation set

Results

Example of 1h-forecast output(forecasted values of hfi_cond)



Short term HFI-EU activity index prediction

The **problem** has been treated both as

- a **regression/prediction** problem: Given the previous d HFI-EU activity index measurements **predict** the value s-steps ahead.

- a **classification** problem: Given the previous d HFI-EU activity index measurements, **classify** the situation s-steps ahead as "**disturbance**" or "**no disturbance**".

NOTE 1: In the following only the one-step ahead case is considered.
NOTE 2: No other quantities have been utilized (e.g., GNSS data).

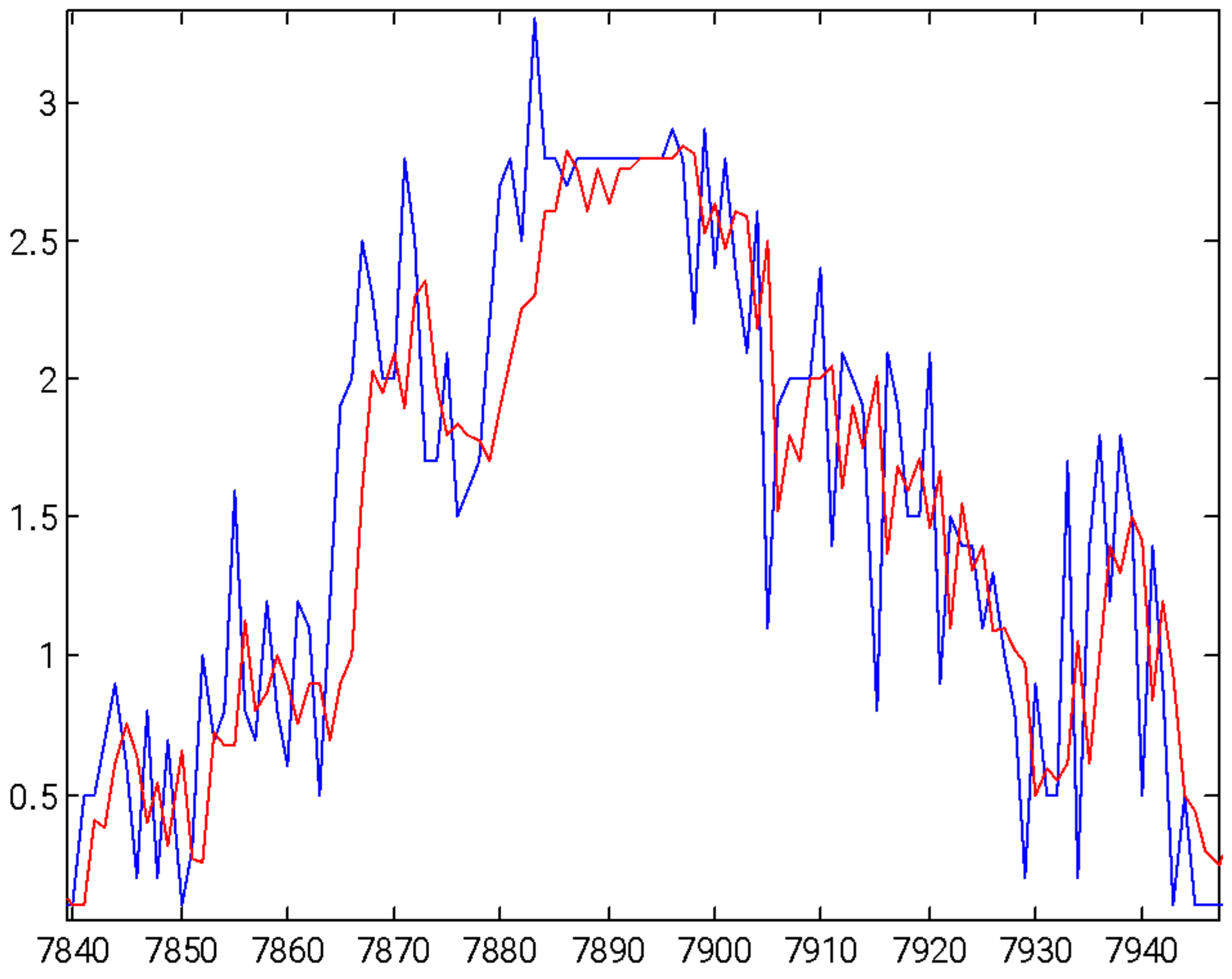
The **data:** Two HFI-EU time series
- $\boldsymbol{D}$: 1/9/2020 - 31/12/2020 (used to create training data where needed)
- $\boldsymbol{D1}$: 1/1/2023 – 31/3/2023 (used to create test data)

NOTES:
- The measurement at a specific time step is the **mean** of the **measurements** taken from ~20 stations distributed over Europe (missing values are ignored).

HFI-EU activity index: Regression

(Classic) **Prediction methods used:**

- **Least squares-based** regression/estimation
- k **nearest neighbor- based** regression/estimation



Performance criteria used:

- the **minimum prediction error**
- the **mean prediction error**
- the **median prediction error**
- the **maximum prediction error**
- the **relative prediction error**

Results

- ➢ k **nearest neighbor**- based method **outperformed** **the least squares**-based method.
- ➢ **Relative prediction error** < 0.5 has been achieved.

HFI-EU activity index: Classification

Rationale in defining classes:
- $\boldsymbol{hfi}$ values $\leq \mathbf{0.1}$ correspond to class 1 (**no disturbance**)
- $\boldsymbol{hfi}$ values $> \mathbf{0.1}$ and $\leq \mathbf{1.7}$ correspond to class 2 (**uncertainty** about a disturbance event)
- $\boldsymbol{hfi}$ values $> \mathbf{1.7}$ correspond to class 3 (**disturbance**)

NOTE: Class 3 is the more interesting class.

Classification algorithms considered**:**
- **k-NN** classifier with **(I)** $k = 15$ and **(II)** $k = 25$.
- **FNN** (Feedforward Neural Network) classifier **(III)** 1-hidden layer (30 nodes), **(IV)** 2-hidden layer ($30 - 10$ nodes), **(V)** 3-hidden layer ($50 - 10 - 5$).

Experiments have been conducted on:
(a) the **original** data set
(b) on a data set **augmented** with additional artificially generated class 3 data (disturbance)

Performance indices used:
- Class j **Recall (**R**):** Percentage of the vectors that stem from class j and are classified correctly by the classifier.
- Class j **Precision (**P**):** Percentage of the vectors that have been classified to class j and they are actually belong to that class.

Results

Results:

The **FNN** classifiers have slightly **better performance** compared to the ***k*-NN** classifiers.

The classifiers applied

- on the **original** data give **lower Recall** R and **higher precision** P.
- on the **augmented** data give **higher Recall** R and **lower precision** P.

In up to 90% of the cases, class 3 has been identified correctly (for FNNs) (**augmented** data set)

For class 3 cases, in up to 96% of them, the classifiers give probability $> 20\%$ for them.

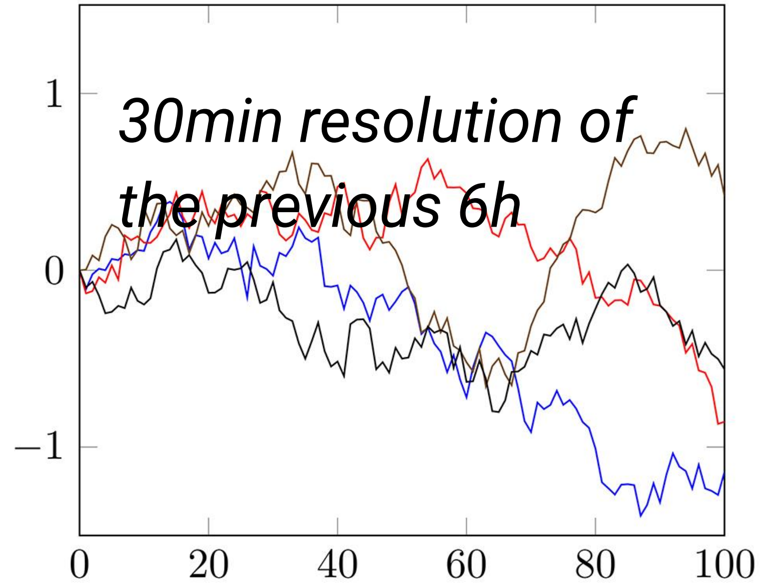
In 75% of the cases where **class 3** is not the dominant one, it still has probability $> 20\%$.

TID Catalog classification (TID next 3h?)

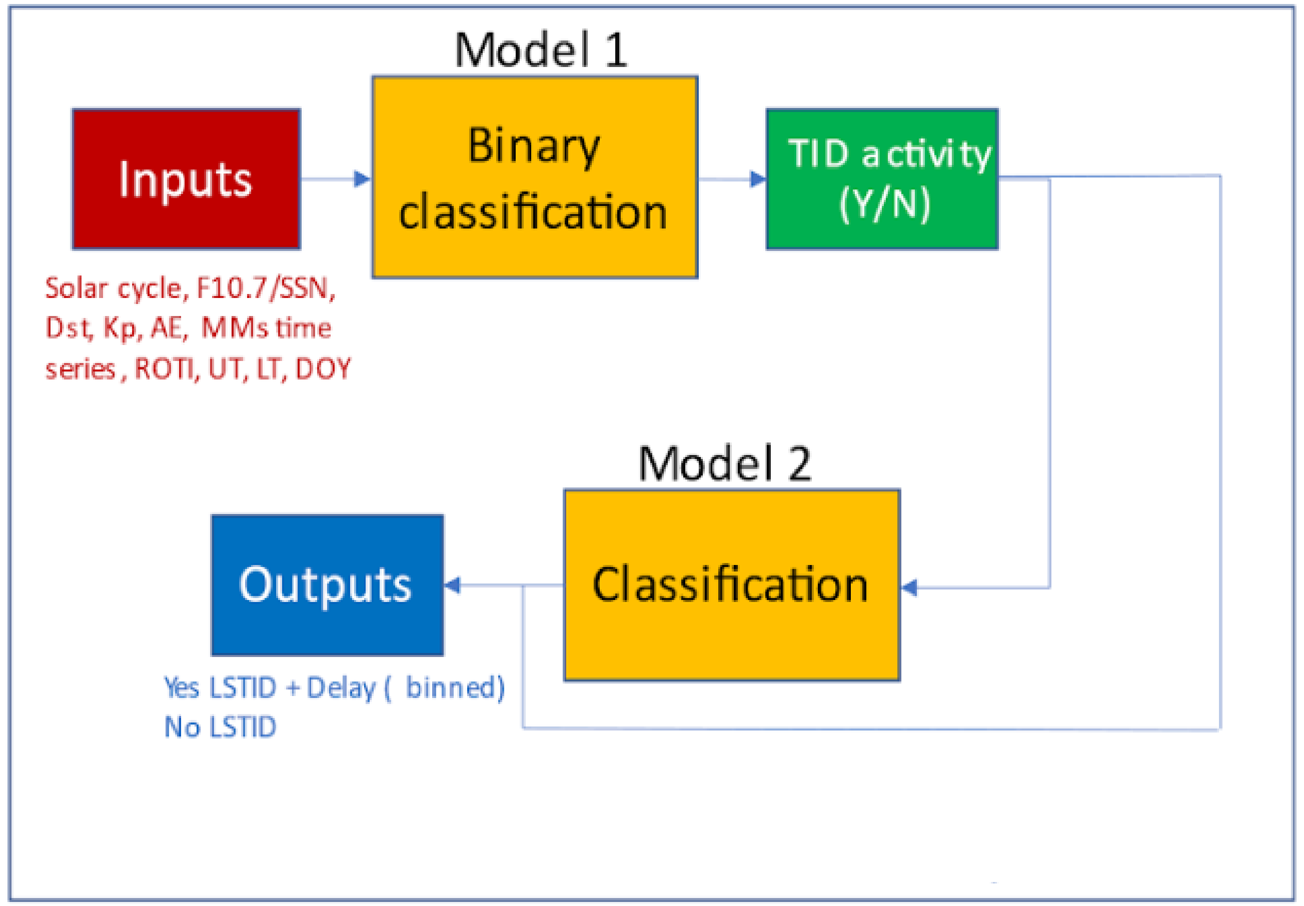
The dataset built is made of 157.777 couples (**X**(T),y(T)) for each T every 30min between FEB 2014 and DEC 2022, where:

$$\mathbf{X}(T) = \begin{bmatrix} X_1(T-6.5h) & X_1(T-6h) & X_1(T-5.5h) & X_1(T-5h) & \dots & X_1(T-1h) & X_1(T-0.5h) \\ X_2(T-6.5h) & X_2(T-6h) & X_2(T-5.5h) & X_2(T-5h) & \dots & X_2(T-1h) & X_2(T-0.5h) \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ X_N(T-6.5h) & X_N(T-6h) & X_N(T-5.5h) & X_N(T-5h) & \dots & X_N(T-1h) & X_N(T-0.5h) \end{bmatrix}$$

$$y(T) = \begin{cases} 1 & \text{TID detected in 3h starting from T,} \\ 0 & \text{else.} \end{cases}$$



→ TID detected in the next 3h from T? Yes/No

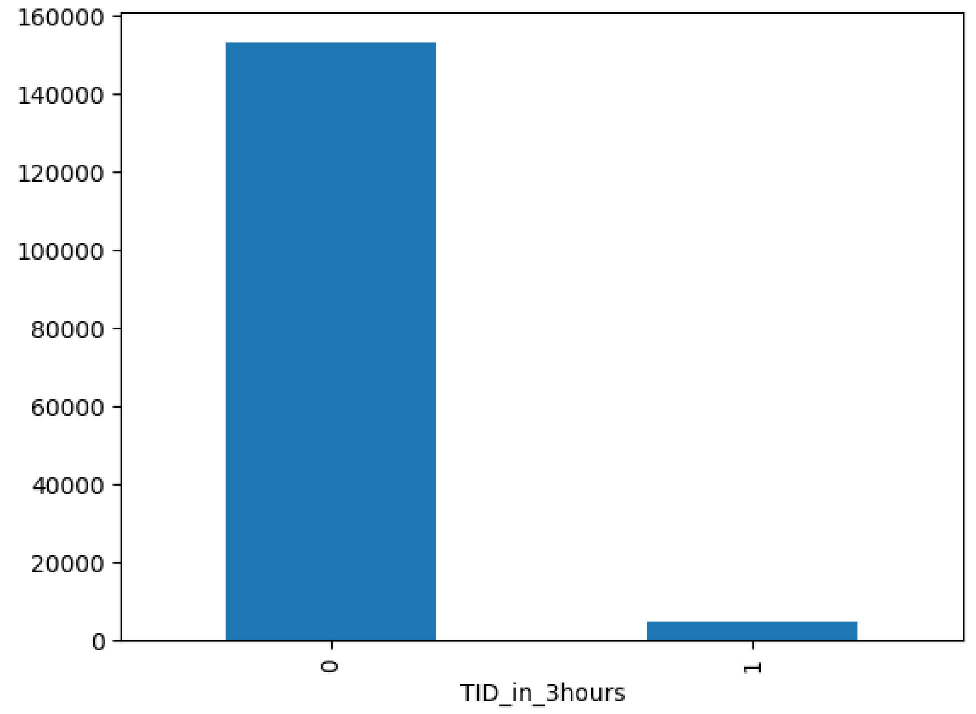


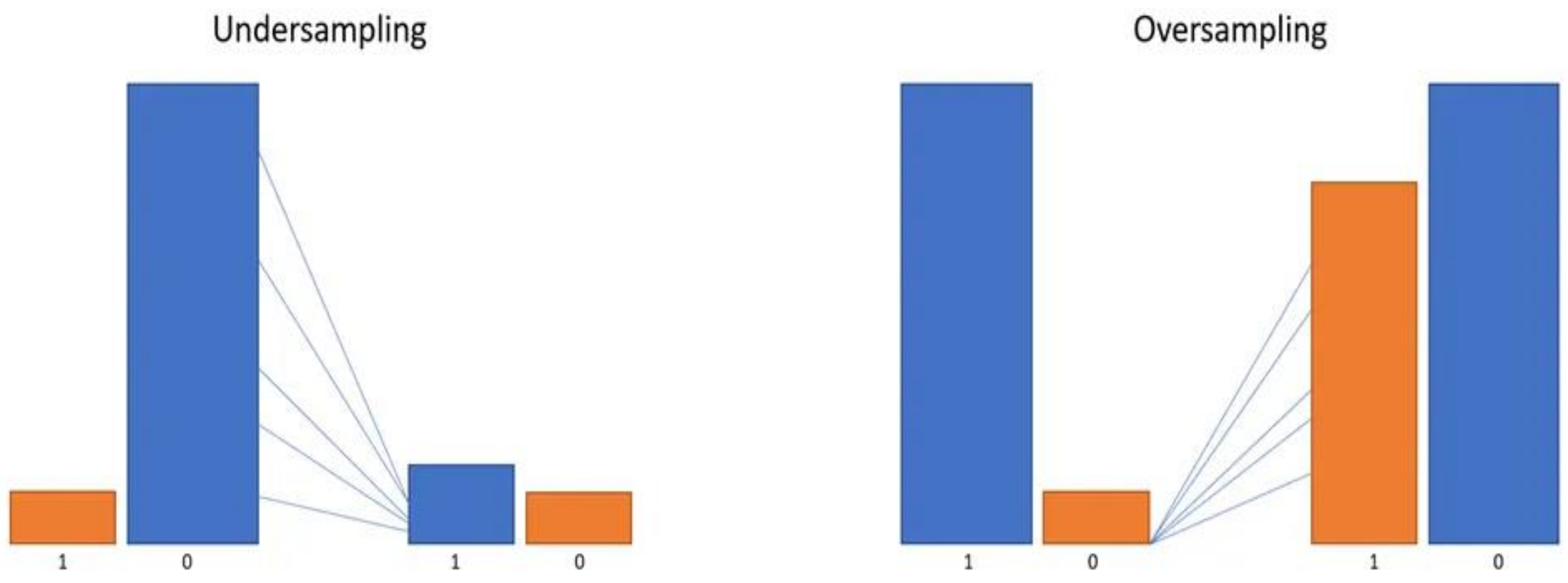
Dataset problems

- The dataset is incomplete: (misdetections related to the technique used to create it)
- The class are severely unbalanced: 3% of Yes and 97% of No
- The input is shifted in space and time with respect to the output

POSSIBLE SOLUTIONS:

- Use different TID datasets (HF-INT raw dataset, GNSS TEC)
- Compare available TID dataset with different techniques on different case events
- Methods to balance the dataset (under/over-sampling)
- Ad hoc machine learning models





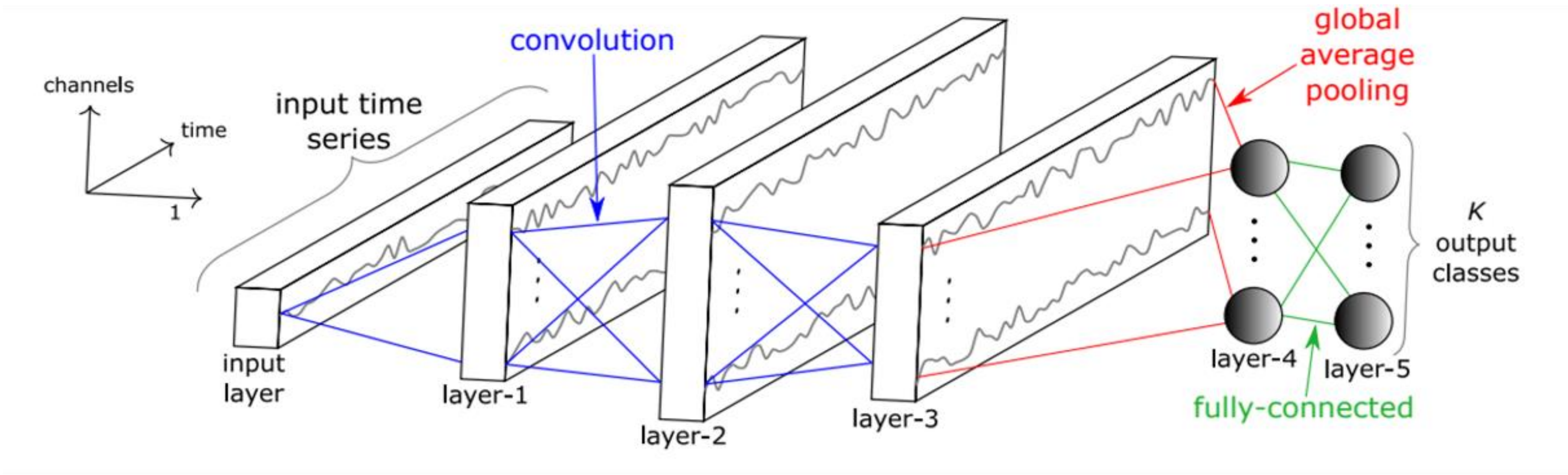
ML algorithms for classification

Classical models:

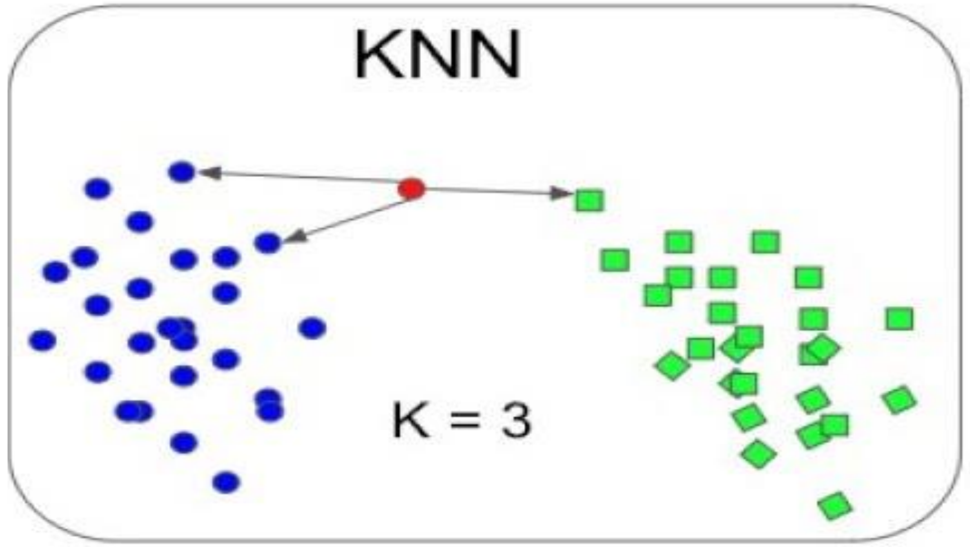
- Logistic Regression (LR)
- Random Forest Classifier (RFC)
- Multilayer perceptron (FFNN)
- Convolutional Neural Network (CNN)

SoA models for time series:

- KNN [with dynamic time warping (DTW)]
- HIVE-COTE (HC)
- Inception time NN (ITNN)



Sample CNN Architecture



Future directions

- **Here just different preliminar results for various approaches**
- **Try different**:
  - ➢ datasets (catalog, EU-HF index, GNSS, ...)
  - ➢ approaches (regression vs classification vs (multi-step) forecasting vs anomaly detection,...)
  - ➢ feature choice (indices, TEC, auroral electrojet,...)
  - ➢ feature engineering
  - ➢ models (standard, ad-hoc)
  - ➢ models configuration and hyperparameters

Thank you for your attention!

Funded by the European Union

The T-FORS project is funded by the European Union (GA-101081835). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.