Marie LATIL

Signal, Image, Communication and Multimedia

2024

Laboratoire Image, Ville, Environnement (LIVE)

3 rue de l'Argonne 67000 Strasbourg

Deep learning for urban tree species classification

---

from 25/03/2024 to 13/09/2024

Under the supervision of:

Laboratory supervisors:

Anne Puissant, Professor, anne.puissant@live-cnrs.unistra.fr

Romain Wenger, Post-doctoral researcher, romain.wenger@live-cnrs.unistra.fr

Germain Forestier, Professor, germain.forestier@uha.fr

Phelma tutor:

Dawood Al Chanti, Associate Professor, dawood.al-chanti@grenoble-inp.fr

Confidentiality: No

Ecole nationale supérieure de physique, électronique, matériaux

Phelma

Bât. Grenoble INP - Minatec

3 Parvis Louis Néel - CS 50257

F-38016 Grenoble Cedex 01

Tél +33 (0)4 56 52 91 00

Fax +33 (0)4 56 52 91 03

http://phelma.grenoble-inp.fr

Acknowledgements

I would like to thank my tutor Professor Anne Puissant for her trust in my capacity to take up this challenge. She welcomed me into this laboratory and this project. This internship would not have been possible without her and her expertise.

I would like to thank Romain Wenger for his constant support both in the technical and thematic parts of this internship. He also trusted me and helped me to better understand urban geomatics. His knowledge-sharing and technical skills were very precious and his incredible daily motivation was very helpful. This internship would not have been possible without him.

I would like to thank Professor Germain Forestier for his share of knowledge to help me in this internship. His advice was very precious and appreciated.

I would also like to thank Antonin Steffanut for our constant knowledge-sharing and crossing work of fields. His thoughts were all interesting and helped me to constantly question and explain correctly my work. He also helped me with the design graphs in this report.

I would like to thank PhD Ali Ismail-Fawaz for his trust in allowing me to share my rewritten code of one model to be publicly available. I would like to thank all the persons from this laboratory who participated in a pleasant working day and shared ideas: PhD Florian Labaude, internship students Baptiste Clemence, Laure Chambaud, Samuel Ladouce-Godier, PhD Armand Pons, post-PhD Guillaume Piasny.

Finally, I would like to the the LIVE and the Faculty of Geography for hosting me. I would like to thank the CNRS for its financial support of this internship.

Contents

Acronyms

A2S	Application for Satellite Survey. 11
AI	Artificial Intelligence. 11
AROSICS	Automated and Robust Open-Source Image Co-Registration Software. 21
BERT	Bidirectional Encoder Representations from Transformers. 13
BIOECO	BIOdiversité et processus ECOlogiques. 11
CNN	Convolutional Neural Networks. 12
CNRS	Centre National de la Recherche Scientifique. 11
DWSC	DepthWise Separable Convolution. 29
DYPA	DYnamique Des PAysages. 11
DYRIM	DYnamique urbaine, Risques et Mobilité. 11
ENGEES	École Nationale du Génie de l'Eau et de l'Environnement de Strasbourg. 11
EPAC	Énergie, Pollution de l'Air et Climat. 11
FCN	Fully Convolutional Networks. 12
FN	False Negative. 33, 34
FP	False Positive. 33, 34
Grad-CAM	Gradient-weighted Class Activation Mapping. 44
HPC	High-performance computing. 11
HYDRO	HYDROsystèmes. 11
ICA	Independent Component Analysis. 16
ICT	Information and communication technology. 11
IMAGE	Image. 11
INEE	INstitut Ecologie et Environnement. 11
INSHS	INstitut Sciences Humaines et Sociales. 11
IPCC	Intergovernmental Panel on Climate Change. 9
IRIMAS	Institut de Recherche en Informatique, Mathématiques, Automatique et Signal. 11
LCZ	Local Climate Zones. 44

LDA	Linear Discriminant Analysis. 16
LIDAR	LIght Detection And Ranging. 9, 44
LITE	Light Inception with boosTing tEchniques. 7, 12, 13, 24, 29–31, 34, 35, 37, 42
LIVE	Laboratoire Image, Ville, Environnement. 7, 11, 59
LSTM	Long Short-Term Memory. 12, 13
MHA	Multi-Head Attention. 7, 26–29
MLP	Multi-Layer Perceptron. 12
OHE	One-Hot Encoding. 16
PCA	Principal Component Analysis. 16
PE	Positional Encoding. 29
PS	PlanetScope. 7–9, 14, 15, 17, 19–21, 30–32, 35, 36, 42, 53, 55, 60
QGIS	Quantum Geographic Information System. 19
ReLU	Rectified Linear Unit. 23
ResNet	Residual Networks. 12
RNN	Recurrent Neural Networks. 12, 26, 29
ROCKET	RandOm Convolutional KErnel Transform. 12
S2	Sentinel-2. 7–9, 14, 17, 19, 21, 22, 30–32, 35, 36, 42, 53, 54, 60
SITS	Satellite Image Time Series. 9, 10, 12
t-SNE	t-distributed Stochastic Neighbor Embedding. 7, 16, 17, 42
TempCNNs	Temporal Convolutional Neural Networks. 12
TN	True Negative. 33, 34
TP	True Positive. 33, 34
TSC	Time Series Classification. 10, 12, 24, 27–29, 42–44
UNISTRA	UNiversité de STRAsbourg. 11

List of Figures

List of Tables

Introduction

Since the pre-industrial area in 1850-1900s, the world average temperature has soared. This has consequences on human society and ecosystems. According to the Intergovernmental Panel on Climate Change (IPCC), +1.5°C compared to the pre-industrial area is the tipping point beyond which the consequences of climate change will become irreversible on human society and ecosystems (Ipcc, 2022, Intergovernmental Panel On Climate Change (Ipcc), 2023). This tipping point is the IPCC projection of global warming by 2030. However, the global average temperature from July 2023 to June 2024 has reached the highest record so far with +1.64°C above the pre-industrial average (Copernicus, 2024). New records are broken each year. Moreover, climate change has impacts on every field, from transport to energy, from agriculture to biodiversity and so on. Nowadays worldwide events are already gigantic, in particular on the world's biggest forests. Wildfires are happening every year in Canada, Australia, South Asia, and even Europe in Greece or Sicilia (Government of Canada, 2024, Clarke et al., 2022, European Comission, 2024). Droughts are spreading rapidly and increasing in frequency, intensity, and duration (Ali et al., 2024, Ullah et al., 2024).

In this context of climate change, urban areas' temperatures are increasing and creating urban heat islands (Oke, 1987). This has consequences on the population's thermal comfort and so their well-being. In addition, 68% of the world's population is expected to live in urban areas by 2050 (UN-DESA, 2018). To face this issue, vegetation is an urgent priority to act on. Indeed, not only are urban greenspaces able to balance the side effects of urbanisation and rising temperatures (Seto et al., 2012), but vegetation including trees is fundamental to protect the biodiversity as ecosystem services provider (Stroud et al., 2022). The impact of vegetation to create more sustainable urban areas is currently being studied (Philipps et al., 2020, European Commission, 2023), and trees are a major part of it. However, an advanced knowledge of trees, their species and their location is necessary. As a matter of fact, climate change also impacts trees, their health and their capacity to survive all the fast changes happening. Trees need to adapt themselves to global warming and changing climates, for instance, more arid or more humid environments (Esperon-Rodriguez et al., 2022).

Thanks to technological development, satellite images are an essential tool for studying many subjects, including trees. Providing images with a regular frequency, it is possible to access the tree's time series. The Satellite Image Time Series (SITS) are a widely used resource for studying trees and particularly urban trees. However, many research works are currently using expensive imagery such as hyper-spectral imagery, aerial or active teledetection like LIght Detection And Ranging (LIDAR) (Karasiak et al., 2017, Kobayashi et al., 2009, Ferreira et al., 2024). These approaches are expensive and so not suitable with an overall aim of generalisation on large geographical spaces. On the contrary, a part of the satellite provided images are free and use high to very high spatial resolution, especially Sentinel-2 and PlanetScope.

Satellite images became a helpful tool for researchers who want to act for a more sustainable future. Whereas classic approaches like machine learning furnish decent results, deep learning is significantly improving the results. Deep learning is often used with high-resolution imagery and high temporal frequency. It is applied in many fields like medicine and they are frequently combined with times series for detection goals in sectors like agriculture, forest management, natural areas and so on (Morid et al., 2023, Paris et al., 2020, Zhang et al., 2023). Recently, SITS combined with deep learning methods have shown more efficiency and promising results than classic machine learning for monitoring urban vegetation (Magalhães et al., 2022).

My research project concerns Time Series Classification (TSC) on urban tree species. The purpose is to classify the 20 most representative urban tree species in Strasbourg. Deep learning is already widespread for classifying urban tree species using SITS (Hartling et al., 2021, Ferreira et al., 2024, Neyns et al., 2024, Wenger et al., 2024). However, most of the studies are classifying a few species, rarely up to 20. This is a new challenge that my research project is taking by testing recent state-of-the-art models and developing new ones. Besides, the aim is also to develop models which can be easily used in other urban areas, using fine-tuning methods. One can then ask:

Can we map urban tree species using freely available high resolution satellite images time series and deep learning? Are these approaches scalable to other cities?

To answer this question, the following three objectives will be covered: (1) apply state-of-the-art deep learning models to SITS for TSC of tree species, (2) develop new approaches to take into account multi-source and multi-temporal satellite data sources; and (3) fine-tune the most efficient model to Nancy to evaluate the generalisation capacities.

1 Laboratory Image, City and Environment (LIVE)

My research work was carried out within the Laboratoire Image, Ville, Environnement (LIVE) - Image, City, Environment Laboratory, a joint and multidisciplinary research unit with around 80 members. The laboratory has 3 administrative supervisors, which are the École Nationale du Génie de l'Eau et de l'Environnement de Strasbourg (ENGEES), the UNiversité de STRAsbourg (UNISTRA) and the Centre National de la Recherche Scientifique (CNRS) (INstitut Ecologie et Environnement (INEE) and INstitut Sciences Humaines et Sociales (INSHS)).

It is composed of 6 research axes, also known as workshops: DYnamique Des PAysages (DYPA), HYDROsystèmes (HYDRO), BIOdiversité et processus ECOlogiques (BIOECO), Énergie, Pollution de l'Air et Climat (EPAC), DYnamique urbaine, Risques et Mobilité (DYRIM) and Image. I was included in the last one with my two supervisors, Anne Puissant and Romain Wenger. Germain Forestier, my last supervisor, is working at the Institut de Recherche en Informatique, Mathématiques, Automatique et Signal (IRIMAS) laboratory at the University of Haute-Alsace in Mulhouse.

My internship was also part of the Application for Satellite Survey (A2S) computing platform (A2S description), which is dedicated to the analysis of massive geospatial datasets and sensor time series (for instance satellite earth observation data such as optical and radar imagery: Sentinel-1 and Sentinel-2) over large territories. The resources of the A2S platform include:

- Computing and storage capacities integrated within the Mésocentre of the University of Strasbourg;
- Dedicated software (modules for managing parallel processing workflows, data ingestion and formatting modules).

A2S operates computation clusters, storage capabilities and dedicated high-level science driven algorithms for detecting changes in in time series of both geospatial datasets and in-situ sensors using advanced Information and communication technology (ICT), Artificial Intelligence (AI) and Cloud/High-performance computing (HPC) technologies.

2 State-of-the-art

2.1 Definitions about time series classification

A (univariate) *time series* X is a set of N real and ordered elements. If all the elements are in a dimension M of reals, it is then a *multivariate time series* (equation 1). A dataset D is composed of an ensemble of pairs of a (multivariate or not) time series with a corresponding label (equation equation 1).

$$\begin{aligned} X &= [X_1, ..., X_N], X_i \in \mathbb{R}^M \\ D &= \{(X_1, Y_1), ..., (X_N, Y_N)\} \end{aligned} \tag{1}$$

The task of Time Series Classification (TSC) consists of learning a classifier on a dataset D. Using the space of possible inputs X, the classifier provides a probability distribution over the classes Y.

2.2 Deep learning evolution up to Time Series Classification

Deep learning first emerged in the 2000s. The perceptron was first invented in 1957 and was able to classify data linearly, into 2 classes. Instead of having one input, one main cell and one output, the Multi-Layer Perceptron (MLP) emerged in 1986 with multiple hidden layers. With the increasing computer capacities over time, the neural networks have gone even deeper with the number of hidden layers, which created more efficient models with the ability to translate, predict text or classify images (LeCun et al., 2015). A lot of deep learning models have been developed since then, for example, Convolutional Neural Networks (CNN), mostly used for image classification (Alzubaidi et al., 2021). Among them, some are useful for time series.

Time series data are widely used. In the medical field, for instance, time series data from exams like electrocardiograms provide information about the patient's health and help to diagnose (Sun et al., 2020, Morid et al., 2023). The first widely used models for time series are Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) (Ismail Fawaz et al., 2019, Faouzi, 2024). The use of deep learning models for SITS has then been developed using the Temporal Convolutional Neural Networks (TempCNNs) model (Pelletier et al., 2019).

Besides, methods based on convolution 1D have emerged or been adapted to time series, such as the RandOm Convolutional KErnel Transform (ROCKET) (Dempster et al., 2020), the Fully Convolutional Networks (FCN), the Residual Networks (ResNet) (Wang et al., 2016), the Convolutional Neural Networks (CNN), but also the InceptionTime (Ismail Fawaz et al., 2019), the Hybrid InceptionTime (Ismail-Fawaz et al., 2022), the Light Inception with boosTing tEchniques (LITE) (Ismail-Fawaz et al., 2023) and many more. The architecture of 11 deep learning models adapted for TSC is presented by Ali Ismail-Fawaz and Maxime Devanne in Irimas.

On the other hand, a very new and different model architecture has emerged recently, the Transformer (Vaswani et al., 2023). This model is using the attention mechanism and was originally made for translation or text prediction. It has been adapted to image classification, object detection (Dai et al., 2021), and even time series very recently (Zerveas et al., 2021).

In addition, the Bidirectional Encoder Representations from Transformers (BERT) model is the precursor of fine-tuning. This model can be fine-tuned in order to answer questions or language inference with high accuracy, without changing the whole architecture, only adding one layer (Devlin et al., 2018).

Finally, a very new innovation has emerged, intending to use multi-modal and multi-temporal data fusion (Audebert et al., 2018, Ienco et al., 2017, Wenger et al., 2022a). This is a very interesting method which proved better results with a sensor fusion with LSTM and an InceptionTime module. There is different ways to create a sensor fusion, (1) the early stage level, (2) the feature level or (3) the decision level. In simple terms, (1) combine the data beforehand to gain information; (2) train the data features in parallel and merge in the classifier layer; or (3) combine the predictions made for each data for the final decision. The most widely used method is the feature level (Dechesne et al., 2017, Audebert et al., 2018), which has been developed for most of the models during this internship.

Using all of these studies, the work in this internship is to develop 4 models using a sensor fusion: InceptionTime, Hybrid InceptionTime, LITE and Transformers.

3 Area of interest and datasets

3.1 Strasbourg and Nancy

The metropolitan area of Strasbourg is the first area of interest of the internship, located in the northeastern France and close to Germany (Figure 1). It has already been studied for classification using the Random Forest machine learning algorithm (Wenger et al., 2024). This previous study chose to classify the 10 most representative species, which represent 40% of the entire dataset furnished by the Strasbourg Euro-metropolis (85,000 trees and about 500 species). Using deep learning models, the challenge is to classify the 20 most representative species, which represent 50% of the dataset.

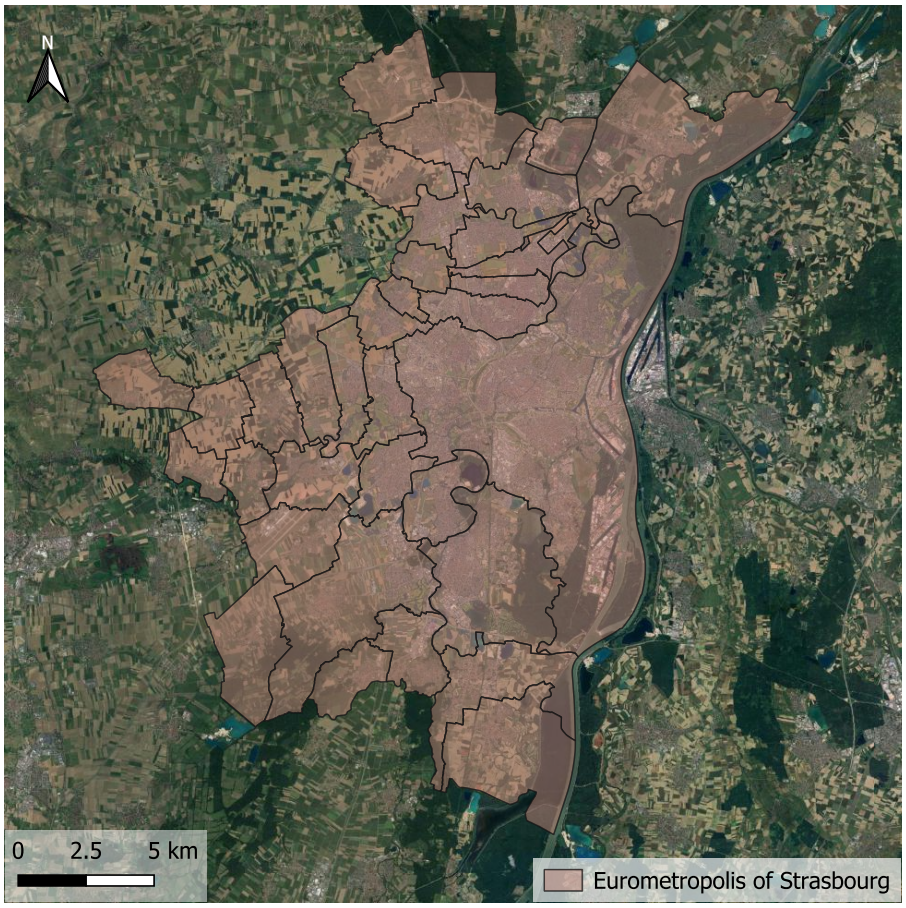


Figure 1 – Area of interest, Strasbourg city

The second area of interest is the metropolitan area of Nancy, located around 150km to the West of Strasbourg. The Métropole du Grand Nancy (Great Nancy Metropolis) also made a catalogue of urban trees used for this study. The challenge is to apply on Nancy the models developed and trained on Strasbourg. The exact same 20 species should then be classified, which is an important criterion for choosing a city close to Strasbourg's climate in order the find the same species. For Nancy, there is one species which does not appear in the inventory (*Styphnolobium japonicum*), and the 19 remaining species represent more than 40% of the dataset (around 26,000 over 60,000 and X species).

3.2 Satellite images

As shown in this study (Wenger et al., 2024), the use of Pleiades images is not significantly improving the accuracy, which is why the current study is only using Sentinel-2 (S2) and PlanetScope (PS) images, free or partly free images. S2 images are part of the European Space Agency's Copernicus program and are freely available within a frequency of 5 days. There are 13 spectral bands from the blue to the short-infrared spectrum and a spatial resolution going from 10 to 60 meters depending on the spectral band. Only the 10 high-resolution spectral bands (10m and 20m) are kept (presented in appendix A Figure 24). The year 2022 is composed of 22 cloud-free images with S2 for Strasbourg. There are 29 cloud-free images in Nancy with S2.

PS imagery is handled by Planet Labs. The images have a 3.125 meters resolution and daily available. At that time of the research, only 4 bands were available (see appendix A Figure 25). There are 53 cloud-free images over Strasbourg, and 37 images for Nancy in 2022.

3.3 Datasets

As explained before, each big city (more than 200.000 inhabitants) has a free tree species inventory. This database contains a great deal of information about every tree, such as the crown size, the height, the coordinates, the species, the plantation date and so on. The Eurométropole of Strasbourg and the Métropole du Grand Nancy provide the dataset for the areas of interest. The respectively 20 and 19 most representative species are presented in appendix A and their distribution in the dataset is shown on Figure 2.

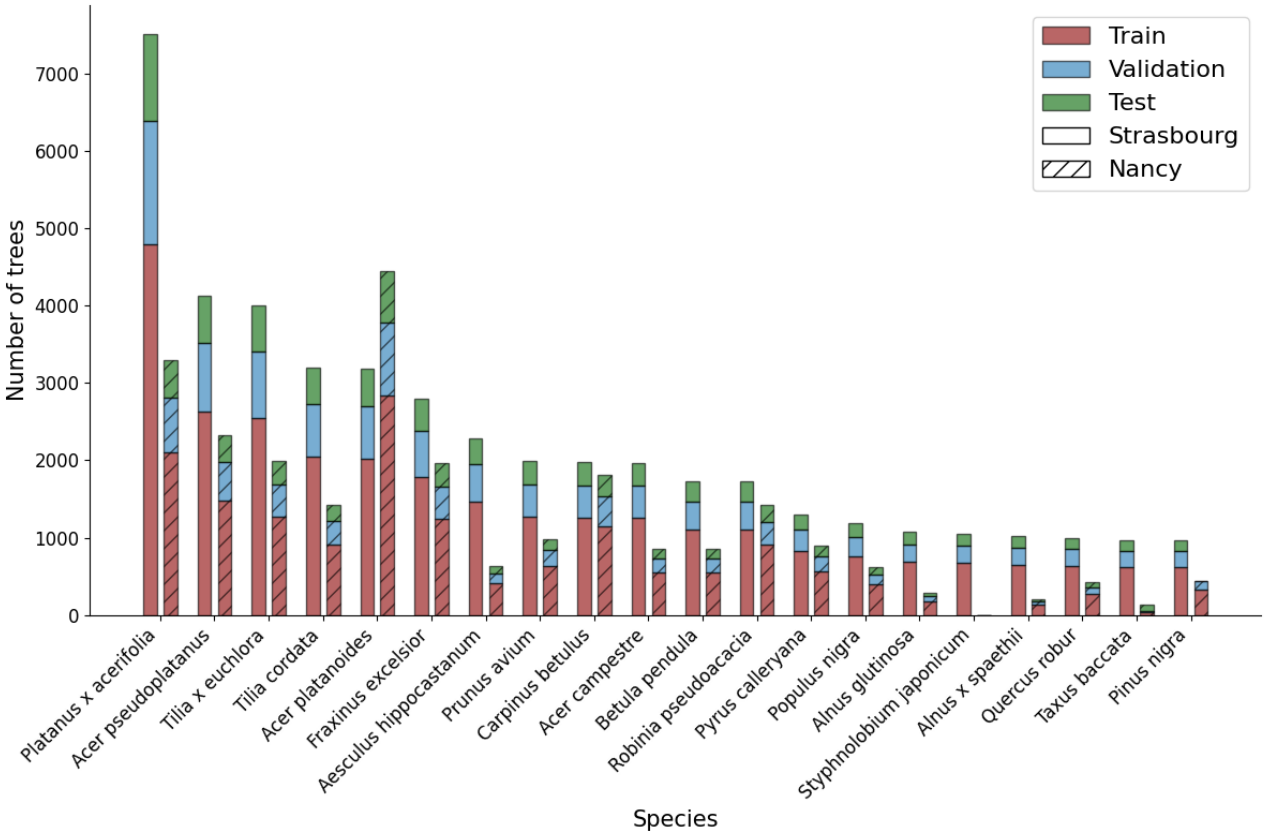


Figure 2 – Species tree distribution in the dataset - Strasbourg and Nancy

4 Preprocessing

4.1 One-hot encoding

Every deep learning model predicts the output label associated with the time series. It usually produces a probability distribution for all the classes and then votes for the best one to predict the final class. However, the class labels are not character strings, they are encoded numbers. Among different methods, a widely used one is the One-Hot Encoding (OHE). It consists of creating a K-dimensional vector ordered (K is the number of classes) for each tree with 0s everywhere except at the class position. Equation 2 shows an example of OHE for 3 species.

$$\begin{aligned} &\text{Species 1: } [1, 0, 0] \\ &\text{Species 2: } [0, 1, 0] \\ &\text{Species 3: } [0, 0, 1] \end{aligned} \tag{2}$$

4.2 t-SNE

When working with huge and complex datasets, a widely used technique from data analysis is dimension reduction. There are different famous methods, such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA) or t-distributed Stochastic Neighbor Embedding (t-SNE). For instance, the PCA method works by finding the directions in which the data varies the most and projects the data into those directions. For different reasons, the t-SNE has been chosen to apply to the data, the main one because it is a better method to take into account non-linear relationships present on time series and complex data, unlike PCA which is mainly used to represent linear relationships. t-SNE has been introduced by (Maaten and Hinton, 2008) with the aim of visualising complex data by reducing its dimensions. As there are many features in time series data, it can be hard to interpret them. The data is simplified into a 2D or 3D map by grouping similar features in order to see patterns and clusters.

With this dataset, it is expected to see 20 species (19 for Nancy) almost distinctly on a 2D or 3D map even with a few overlapping. However, even with different parameters and tries, the result is not visually the same as expected. Figure 3 does not highlight clusters by species. The hypothesis made regarding this observation is that the data is too complex and cannot be represented through a 2D or 3D map.

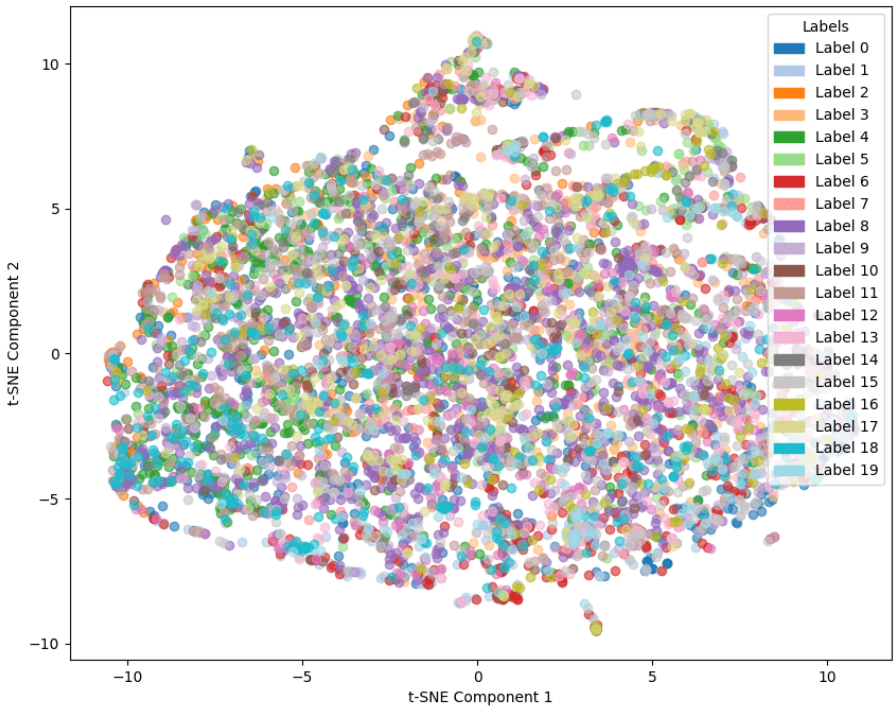


Figure 3 – t-SNE in 2D for S2 data in Strasbourg with 20 species (label 0 to 19)

4.3 Interpolation and smoothing

As mentioned previously, Strasbourg is composed of 22 cloud-free dates from S2 and 53 for PS, which is not a lot in a year. An optional possibility is to create artificially intermediate dates using linear interpolation to reconstruct the complete tree phenology and help the model to better recognise the differences between species. It is especially interesting for one of the models which is used to detect increases, decreases and peaks. As the linear interpolation is jerky, smoothing might help to get more natural and close to real-time series. The Savitzky-Golay smoothing has been used (Chen et al., 2004, Bressant et al., 2024). It is a method of filtering used to reduce noise due to shadows or outliers. A window is sliding through the time series and calculates mean values. That way the outliers are removed while keeping the general trend. The window size is important and depends on the study case. For this investigation, the window size is chosen to be 15 days, a rather stable period for annual vegetation cycles. The interpolation and smoothing have been applied to each band for each tree. An example of the linear interpolation and smoothing on a tree time series is shown in Figure 4. As it was only an experiment idea which did not show better results, no other interpolation type has been tried out.

Figure 4 – Example of linear interpolation and smoothing on a tree time series

4.4 Split and normalisation

Using deep learning models means having 3 sets, the training one, the validation one and the test one. The training set is used to teach the model, all the parameters are evolving with the learning task to reduce the loss. The training set must be a consistent part of the dataset. The validation set is used to evaluate the model right after the training part, during which the parameters are tuned. This set gives information about how well the model is learning, by making small adjustments and also avoiding over-fitting. The validation set does not have to be as large as the training set. When the training is done, the model is then evaluated on the test set. This set must be consistent enough to be sure the actual performances of the model are real but do not need to be as important as the train set. In the literature, the sets are usually split around 0.60, 0.20, and 0.10 respectively for train, validation and test. In this study, the split is **0.64, 0.21, 0.15**.

In order to help the model to learn correctly each tree species, it is important to have every species in every set. Therefore, the species distribution is maintained in the sets. An overview of the species splits is shown in Figure 2. Strasbourg has more trees than Nancy, but the species distribution is very similar.

Finally, normalisation is a key step of every machine or deep learning model. There are different ways to normalise the data, including the classic z-normalisation (subtracting the mean then dividing by the standard deviation for each time series), which is not adapted for vegetation as it leads to a loss of magnitude. Thereby the min-max normalisation has been used in this study (Pelletier et al., 2019). Using the minimum and maximum values among the entire dataset, every value is normalised following equation 3 (subtracting the minimum value then dividing by the range).

$$new\_value = \frac{old\_value - min}{max - min} \tag{3}$$

4.5 Nancy images

As Strasbourg has already been a study area recently, the dataset was already pre-processed and ready to be used. On the contrary, Nancy is a new study area, meaning that the dataset should be prepared using the time series acquisitions from S2 and Planet.

4.5.1 Removing outside trees

First of all, Nancy's tree inventory revealed an important amount of trees outside the satellite acquisitions. All the S2 images are identically calibrated, however, some trees remain outside the acquisition area (Figure 5). In that case, all of these trees should be removed from the dataset, which represent 1364 trees (around 5%). It remains important to have a look at the percentage of removed trees per species, which does not exceed 13%, most likely around 2% (the details are provided in appendix B table 9).

Figure 5 – Trees inventory on S2 images for Nancy

The PS images are usually separated into 2 tiles above Nancy. The first task is to merge them using the Quantum Geographic Information System (QGIS) software. Afterwards, due to specific orbits, black bands appear, cutting the images on the diagonal, not regularly from one date to another. These black bands are sometimes covering trees. It appends on 10 dates. The more dates there are, the more precise time series there will be, however, the more trees

there are, the more data can be trained, both are important to increase the model's accuracy. There is then a compromise to find between removing dates to keep the trees or removing trees to keep the dates.

The idea is to count the number of trees with no data for each date, if it is low in percentage compared to the total number of trees, better to remove the trees, otherwise better to remove the date. In parallel, it is relevant to count the number of dates each tree has no data. A visual representation of the result is presented on Figure 6. Table 1 summarises the dates with no data and highlights the number of no-data trees and their approximate position (by looking at the images, the black bands are usually around the same position, on the left or on the right). In order to work with smaller data, a mask has been used, covering the Nancy metropolis, it is open data available on Data Grand Est.

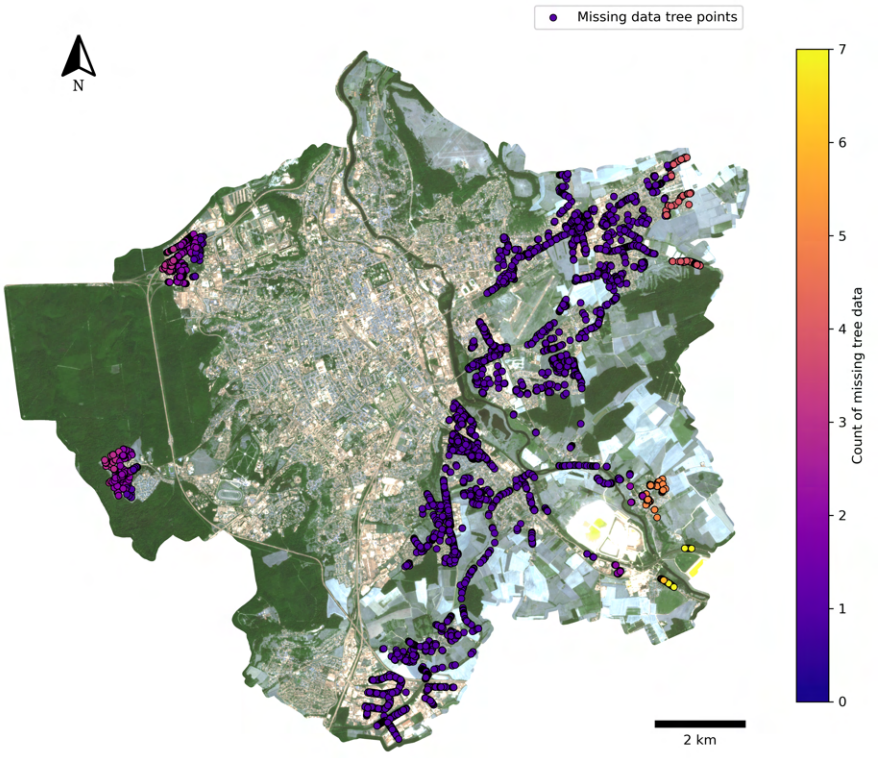


Figure 6 – Frequency of missing tree data in 2022 on PS in Nancy

Table 1 first shows that the 22nd of September should obviously be remove. Except that, there are way more trees with no data from the left than from the right black band. Besides, the left band is involving only 3 dates, whereas the right one is touching 6 dates. For all of these reasons, the dates with a black band on the left have been removed (March 19, May 18 and June 19) to keep the associated trees, and the no data trees on the right band have been removed to keep the dates. The percentage of removed trees per species does not exceed 4% for PS most likely around 1% (details provided in appendix B table 10).

dates (MM/JJ)	number of no data trees	approximate location
02/09	127	right down
03/19	302	left up and down
03/23	66	right down
05/18	1321	left up and down
06/19	1039	left up and down
07/11	40	right down
08/03	236	right up and down
08/22	236	right up and down
08/31	277	right up and down
09/22	6646	right up and down

Table 1 – Dates with the number of no data trees and their approximate position

4.5.2 Coregistration

When studying multiple images taken at different times or by different sensors but all from the same area, there can be a shift between them. Coregistration is an important preprocessing step which is made to perfectly align the pixels from the same geographic location on all the images. It can influence the accuracy or the feature extraction.

The method is the following one. One image (S2 for instance) is chosen as the referenced one, without any cloud and very clear. All the target S2 images are coregistrated on the reference one using COREGIS (Stumpf et al., 2018) coregistration algorithm, already developed and adapted on HPC. Then one PS image is chosen as the referenced one, approximately at the same date as the reference one for S2. The PS reference image is coregistrated on the S2 one. Then all the target PS images are coregistrated on the PS reference one. At the end, all the images are coregistrated both inside each sensor and all together. A summary diagram of the whole method for preparing the dataset is shown in Figure 8.

Coregistration was already done on Strasbourg images between S2 and PS images. Figure 7 is highlighting one noticeable example of a 10m shift at one date, almost equal to 0 after coregistration. This observation was a motivation to apply coregistration on all the images, because the crown is usually lower to the pixel size (10m), and the purest pixel is wanted for precision. For Nancy however, it was done for the S2 images but not for the PS images. Coregistration can be done using the library Automated and Robust Open-Source Image Co-Registration Software (AROSICS) for multisensor satellite data directly on Python (Scheffler et al., 2017). Unfortunately, this software did not work as expected and due to a lack of time in the internship and a visual shift almost undetectable, the decision has been made to keep going without making the coregistration for the PS images.

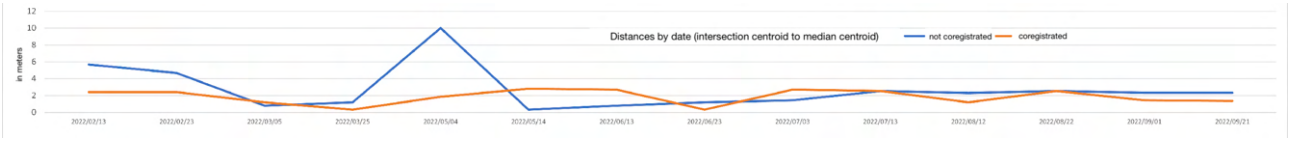


Figure 7 – Shift among time at one point before coregistration (blue) and after coregistration (orange), in meters

4.5.3 Zonal statistics

Based on the satellite images and the tree inventory, the dataset has to be created. There are different methods to do so, but according to (Wenger et al., 2024), buffering the trees is an efficient one, because it does have similar results as with segmentation and it is done easier with lower computational resources needed. Regarding the pixel size, a 1 meter buffer is enough, all the more to generalise the method because the theoretical crowns are not always known for each tree. After buffering all the trees with 1 meter, the zonal statistics can be computed for each tree. In order to deal with roads and trees overlapping on several pixels, the median is computed for each tree, so that the outliers and noise should not be taken into account, and then the most correct tree value is hopefully represented.

The entire dataset is then built for Nancy, including 33 dates for Planet and 29 dates for S2, respectively 4 bands and 10 bands, and 25,020 trees.

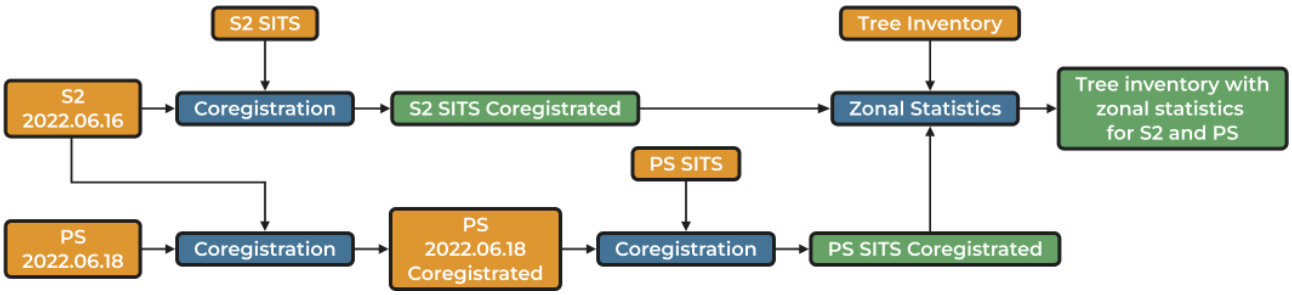


Figure 8 – Processing method for coregistration and zonal statistics computation

5 Methods

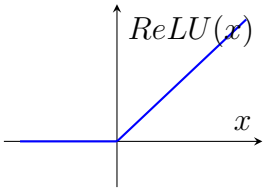
5.1 General information about deep learning models

A lot of the deep learning models use convolution filters in order to capture the data features. For classification tasks, features are the centre of a model decision. A convolution is a way to activate the meaningful regions in the dataset, for instance, the snout for a dog image, or the beginning of flowering for tree phenology in a time series. For time series, it consists of applying a 1D filter along the input. The filter size influences the result. The output of a convolution is shorter than the input unless padding is applied, which consists of adding numbers at the beginning and the end of the input (generally zeros or ones) in order to keep

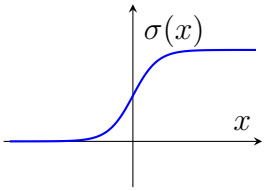
the original dimension after the convolution. Another key function of deep learning models is the activation one. It introduces a non-linearity in the neural networks in order to learn and represent complex relations inside the dataset. The most used functions are Rectified Linear Unit (ReLU) and sigmoïd σ (equation 4). In this study, the activation function always used is ReLU.

$$\text{ReLu}(x) = \max(0, x) \qquad \sigma(x) = \frac{1}{1 + e^{-a}} \tag{4}$$

ReLU Function



Sigmoid Function



Another key practice for the deep learning process is to use K-fold for cross-validation. The aim is for all the data to go through the training and validation sets and to avoid overfitting. To achieve this, the train and validation sets are put together and then divided into K equal folds. There are then n learning splits going on. Each split has a different fold used for the validation part. In the end, all the data has been used either for the training or for the validation step. A visual representation of K-fold cross-validation is shown in Figure 9.

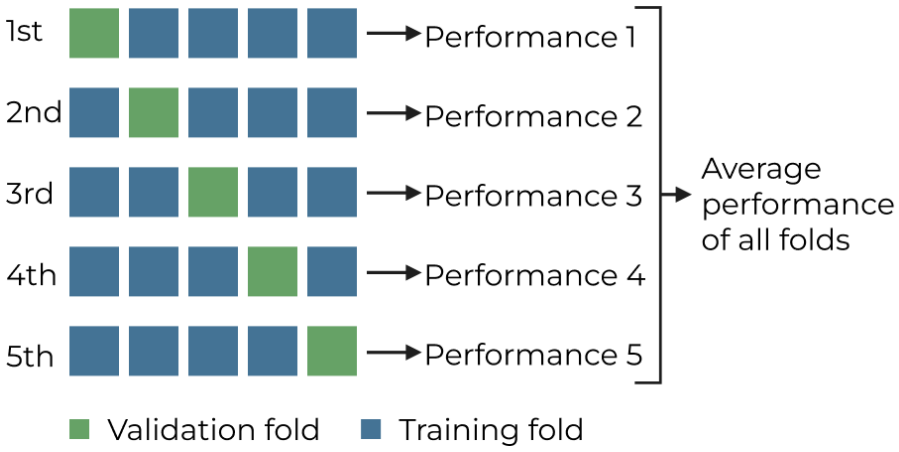


Figure 9 – K-fold process

Finally, each training fold is composed of a certain number of epochs. One epoch is a complete run-through of all the data. The higher the number of epochs, the better the performance will be, until a limit over which overfitting will happen. To compare to a human learning process, if a child has to learn 100 vocabulary words from English to French, every epoch is a complete pass over the 100 words. While repeating the epochs, the child is learning the words and making fewer errors. Methods exist in order to avoid overfitting and stop the learning process after some epochs. For instance, early stopping stops the process if the validation loss is not improving after x epochs; a dropout consists of removing random neurons between each epoch in order to make the model weaker and not overfit during the training.

5.2 Training parameters

All the models have been trained with 5 folds and 30 or 60 epochs per fold. The number of epochs has been fixed empirically by looking at the training and validation loss. For InceptionTime and Hybrid, 30 epochs are enough, while LITE and Transformer need 60 epochs to have a plateau for the validation loss. Besides, the Adam optimizer has been used (Wenger et al., 2022a) with an initial learning rate $\epsilon = 0.001$ and a plateau reduction of patience 5 with a factor of 0.1. It means that if there is no improvement (loss reduction) during 5 epochs, the learning rate is reduced by a factor of 0.1. In addition, a dropout of 0.1 has been used for the Transformer model, meaning that neurons have been set to 0 with a probability $p = 0.1$ and all the other neurons are multiplied by $1/(1-p) = 1/0.9 = 1.1$ to compensate for the weakness. Finally, a different weight has been assigned to each class depending on its frequency. This weight is defined by the inverse of the class frequency (Audebert et al., 2018, Wenger et al., 2022a). This method is forcing the network to pay more attention to the under-represented classes thanks to higher weights.

5.3 InceptionTime

InceptionTime (Fawaz et al., 2020) is the first model which has been tried. The architecture is presented on Figure 10. Even if the Inception model was originally made for image classification, this adaptation for TSC has shown high results. The model is composed of 6 Inception modules, each one computing convolutions in parallel with different filter lengths to capture time series features at multiple scales. According to (Fawaz et al., 2020), the longer the filters are, the better the model's performances will be, with the condition of having enough training data to avoid overfitting. As the dataset is not too large, the classic filter lengths of {10, 20, 40} are used. However, time series are rather short in this study case, which raises a hypothesis on the influence of shorter filter sizes such as {2, 4, 8}. Both sizes are going to be tested and compared. Each Inception module also has a bottleneck convolutional layer to reduce the dimension and computational cost. To keep the same dimension after every convolution layer, a zero-padding is applied.

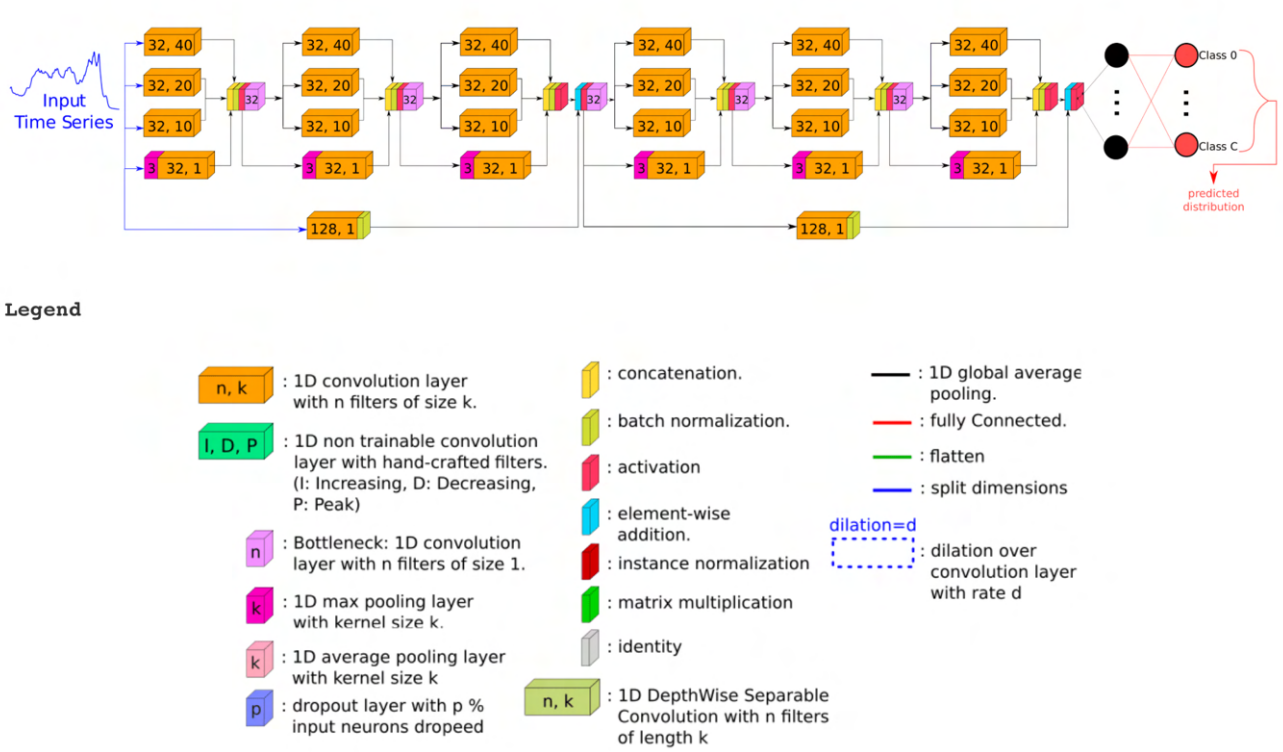


Figure 10 – InceptionTime architecture, taken from Fawaz et al. (2020)

5.4 Hybrid InceptionTime

The Hybrid InceptionTime architecture has been developed by (Ismail-Fawaz et al., 2022) to capture specific patterns in time series. The main idea is to define non-trainable hand-crafted filters to catch 3 specific patterns in times series: increase, decrease and peaks. These patterns are extremely relevant in the study case of tree classification, because tree phenologies have a flowering period in Spring (increase), usually a peak before a leaf fall (decrease). The 3 sets of hand-crafted filters are the ones described in (Ismail-Fawaz et al., 2022), each trend detection has sizes {2, 4, 8, 16, 32, 64} (without size 2 for the peak filter). An example for length 12 is shown in equation 5, respectively increase w_I, decrease w_D, then peak w_P filter.

$$
\begin{aligned}
w_I &= [-1, 1, -1, 1, -1, 1, -1, 1, -1, 1, -1, 1] \\
w_D &= [1, -1, 1, -1, 1, -1, 1, -1, 1, -1, 1, -1] \\
w_P &= [-0.25, -1, -1, -0.25, 0.5, 2, 2, 0.5, -0.25, -1, -1, -0.25]
\end{aligned}
\tag{5}
$$

The global architecture of Hybrid InceptionTime is shown on Figure 11. It is the same as the InceptionTime one (Figure 10), with the hand-crafted filters coming at the beginning of the model, which are not going to be learned during the training set.

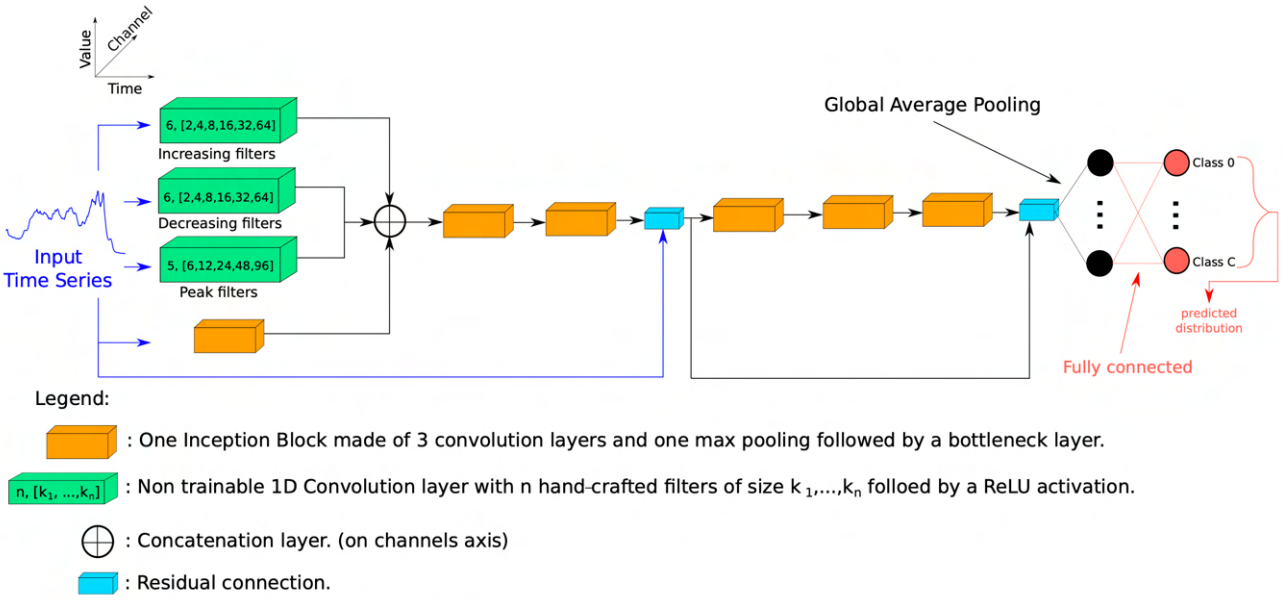
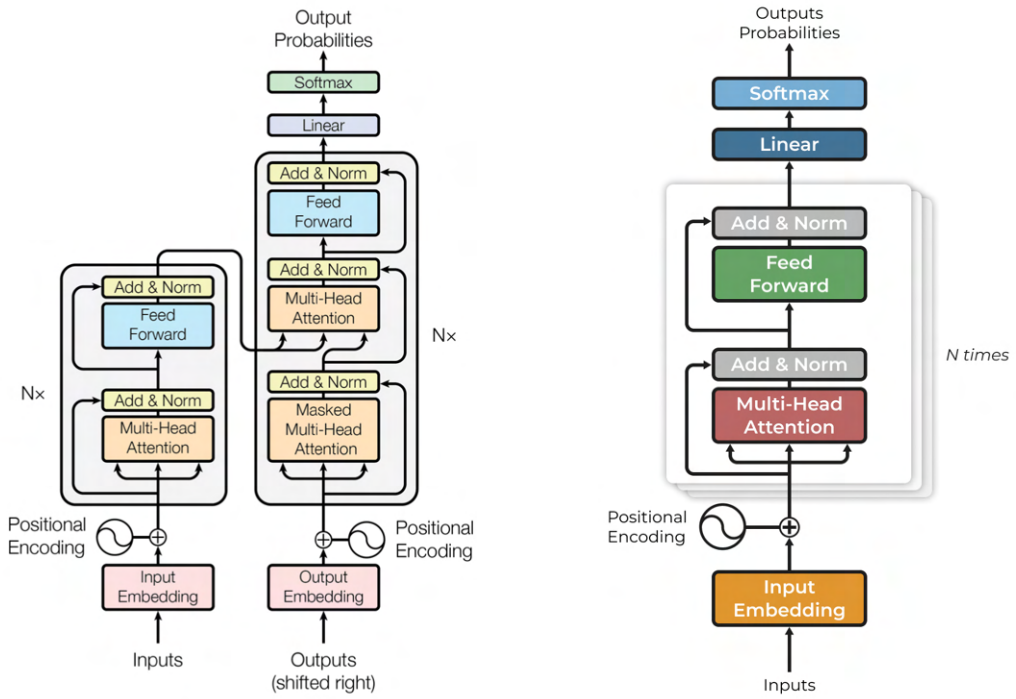


Figure 11 – Hybrid InceptionTime architecture, taken from Ismail-Fawaz et al. (2022)

5.5 Transformers

Transformers have been recently developed in the famous paper *Attention Is All You Need* from Vaswani et al. (2023). The global architecture is shown in Figure 12. The structure is composed of an encoder (left part) and a decoder (right part). This model was originally made for sequence transduction, for example, translation or text prediction. Compared to previous models made for this purpose such as RNN, Transformers have a theoretically infinite memory (very long in reality). The key to this revolutionary neural network is the attention mechanism, especially the Multi-Head Attention (MHA) computation. In general, the encoder maps the input sequence to a continuous representation sequence, while the decoder generates the output, element by element. The memory is made by adding the previous output to the next following input.



(a) Global transformer architecture *(b) Transformer for TSC*

Figure 12 – (a) Transformer global architecture, taken from Vaswani et al. (2023), and (b) Transformer architecture applied for this study

5.5.1 Encoder and decoder stacks

The encoder is composed of N identical layers, in the original paper $N = 6$, but it can be changed to 4 or 8 layers for instance, more layers can improve the learning process. In each layer are 2 sub-layers, a Multi-Head Attention (MHA) and a fully-connected feed-forward one. A residual connection is used in the whole model, and all the sub-layers are followed by a normalisation layer. The layer's outputs are written in equation 6 (Vaswani et al., 2023), where $Sublayer(x)$ is the function inside each sub-layer (MHA or feed-forward), $LayerNorm(x)$ is the normalisation layer, and x is the input data. The general size of the output inside the model depends a lot on the dataset. In the original paper, $dmodel = 512$. However, empirical tests prove that for this study dataset, the model is not learning with this size, but it is with smaller sizes, for instance $dmodel = 64$ or 128. Therefore a hypothesis can be made, the dataset is rather small (22 and 53 dates), and small model sizes are enough for it.

$$Outputlayer = LayerNorm(x + Sublayer(x)) \tag{6}$$

The decoder is similar to the encoder, with N layers and the same sub-layers. The additional sub-layer named the masked MHA, is made to prevent the model from having access to the future because predictions can only be made regarding the past information. For instance, if it tries to finish the sentence *Trees are higher than ...*, it can not have the information after

to predict it. The decoder is very useful for sequence to sequence tasks, like translation or prediction, but in the case of TSC, the encoder is enough to produce a vectorial classification. Thereby, the decoder part has not been used in this model (Zerveas et al., 2021). However, the linear and softmax layers after the encoder are necessary to produce the output probabilities.

5.5.2 Attention mechanism

The attention mechanism requires 3 parameters, the query (matrix Q), the key (matrix K) and the value (matrix V). Here is a simple example to show how they are linked. When searching for videos on the internet, the search bar contains the query (Q) of the user, these words will be mapped against a set of keys (K) like title, description etc. from the database, and finally, the best matching videos (the values V) are showing. The 3 parameters are linked through the Scaled Dot-Product Attention, shown in Figure 13 left and defined in equation 7, where d_k is the dimension of the keys. In the study of time series, d_k is the number of dates.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{7}$$

The Multi-Head Attention shown in Figure 13 right is composed of h attention layers running in parallel. h is the number of heads in the attention. In the original paper, $h = 8$, and Q and V have dimension $dmodel/h$. Finally, a specificity can be added to the model, the *attention size*. This parameter controls the window size of the dates to be taken into account in the attention mechanism. It reduces the attention to a few dates before and after the current date. The hypothesis made is that the values at each date do not depend on remote values in time. For instance, wintertime values will not influence summertime values, but maybe values from a week ago will have a link to current ones. The attention size guides the model to pay attention to a few dates nearby. In this study, 3 attention sizes are tried: 8, 16 and none. The smallest one is to find out if the previous hypothesis is accurate. The second one is to see if the model needs more dates to make the attention mechanism learning. The last one, none, means that all the dates are taken into account in the process, to compare with the small window size.

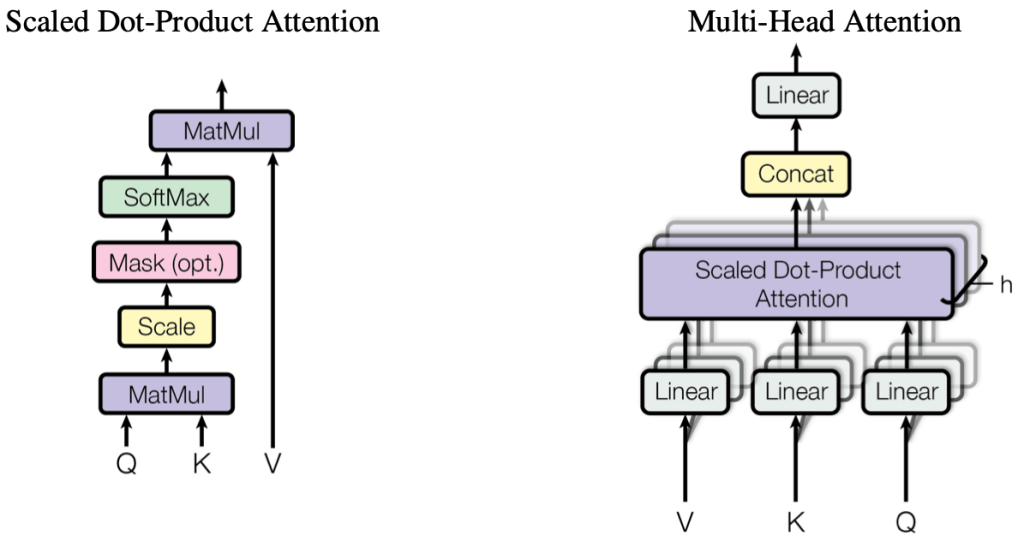


Figure 13 – Scaled Dot-Product Attention (left) and MHA (right), taken from (Vaswani et al., 2023)

5.5.3 Positional Encoding

Unlike RNN models, Transformers process sequences in parallel. There is no recurrence nor convolution, but the date order for time series is fundamental. The Positional Encoding (PE) is adding a positional identifier to each token (words or part of words for text sequences), i.e. each date in this study case. Among different ways to do it, the original paper is using the sine and cosine functions at different frequencies. The main asset is to have an infinite possibility for sequence lengths. Equation 8 is showing the PE, where *pos* is the position in the time series, and i is the dimension.

$$\begin{aligned} PE_{pos,2i} &= \sin(pos/10000^{2i/d_{model}}) \\ PE_{pos,2i+1} &= \cos(pos/10000^{2i+1/d_{model}}) \end{aligned} \tag{8}$$

5.6 LITE

The LITE model has been developed recently with the aim of reducing the number of parameters while maintaining the performances. According to (Ismail-Fawaz et al., 2023), the LITE model is 2.78 times faster than InceptionTime while consuming 2.79 times less C02 and Power with only 2.34% of InceptionTime's number of parameters. This is therefore interesting to try this new model made for TSC. The architecture is presented on Figure 14. In addition to the bottleneck convolutions which are known from the InceptionTime model to reduce the number of parameters, the LITE model also adds a DepthWise Separable Convolution (DWSC), which is divided into the DepthWise convolution and the PointWise one. This separable convolution uses fewer multiplications for a smaller number of parameters to learn compared to standard convolution.

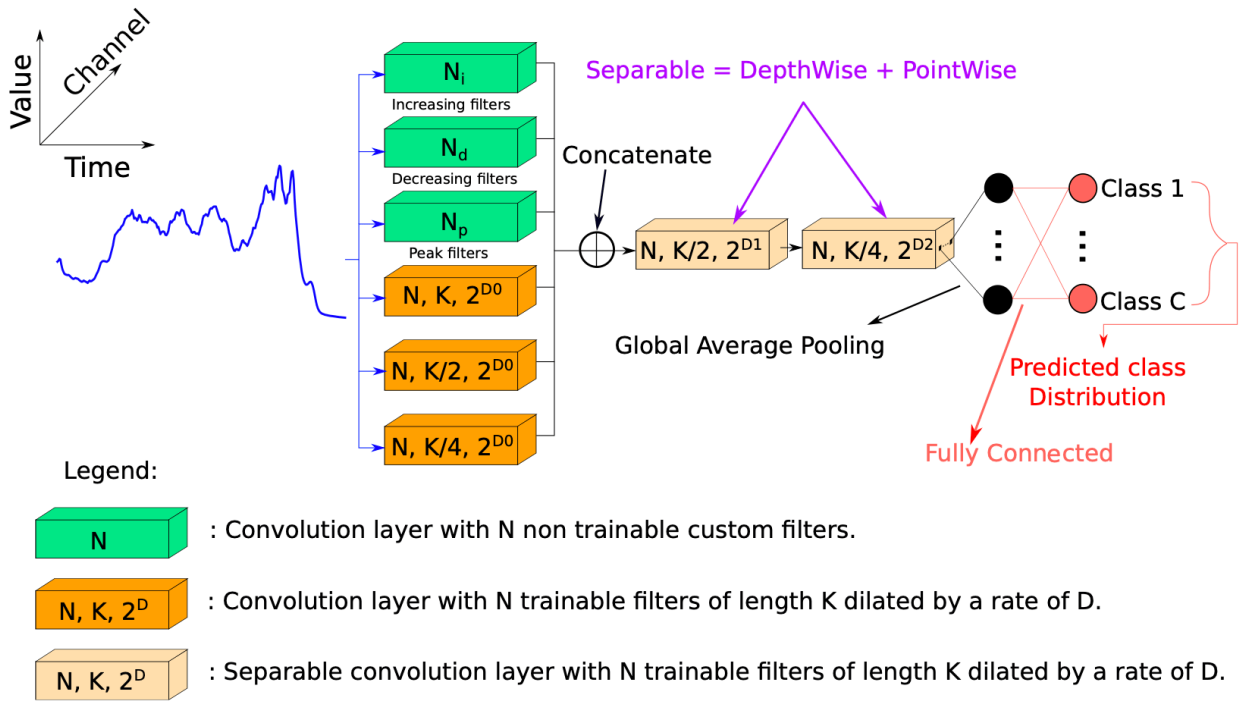


Figure 14 – LITETime architecture, taken from Ismail-Fawaz et al. (2023)

5.7 Sensor fusion

5.7.1 Convolutional models

The innovation of using these state-of-the-art models is feature fusion. The 2 sensors S2 and PS are complementary, both in spectral and in temporal diversity. The first one is richer in spectral bands, while the second one is richer in the number of dates. Features fusion is a way to describe better the tree species phenology. While machine learning is making data fusion before the model or probabilities fusion after it, deep learning has the advantage of improving the learning process by making a fusion of the features during the process (Wenger et al., 2022b,a).

For models using convolutions layers, i.e. InceptionTime, Hybrid InceptionTime and LITE, the sensor fusion is done at the end of the architecture, just after the average pooling. The model method is shown in Figure 15. Two models are running in parallel, one for each sensor, and only the fully connected layer is done using both sensors' outputs to predict the class.

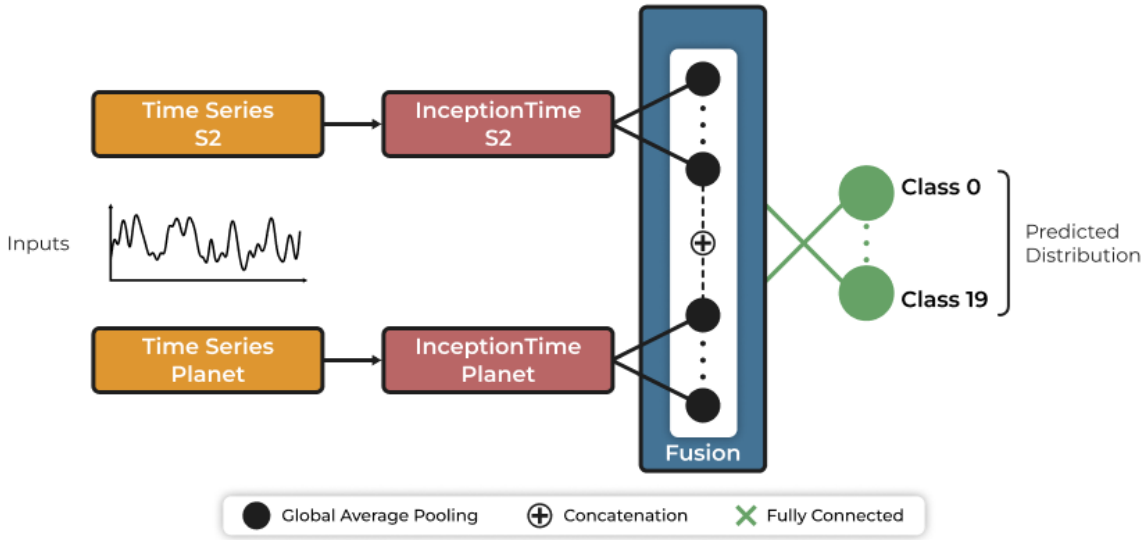


Figure 15 – Sensor fusion architecture for InceptionTime model. Same architecture for Hybrid and LITE fusion

5.7.2 Transformers

For the Transformer model, only the encoder is used for classification, the decoder is used for generative tasks (Zerveas et al., 2021). However, the fusion of S2 and PS sensors is not an easy task, especially for merging inside the encoder. The hypothesis made is that the Transformer will perform better with a more complete time series as an input. There is an issue, the sensors do not have the same number of spectral bands and, thereby, not the same input size. To deal with it, a convolution layer has been applied to the 10 spectral bands of S2 in order to reduce from 10 to 4 spectral bands (Figure 16) and then merge the 2 sensors into one single input.

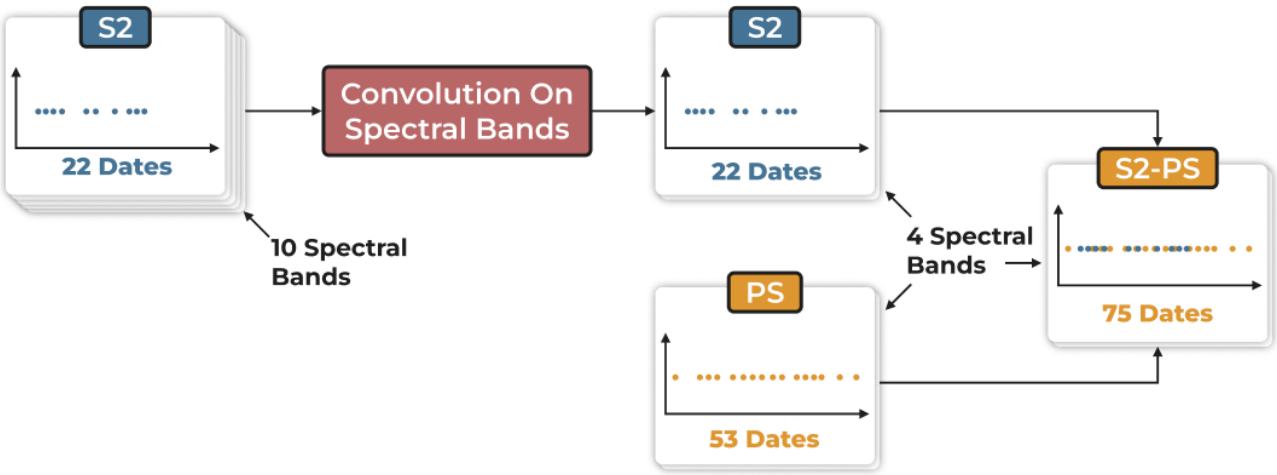


Figure 16 – Convolution on S2 spectral band to reduce from 10 to 4

The dates have been interleaved to keep a chronological order, shown in Figure 17. The input is then composed of 22+53=75 dates with 4 spectral bands. The model remains the same for one or two sensors.

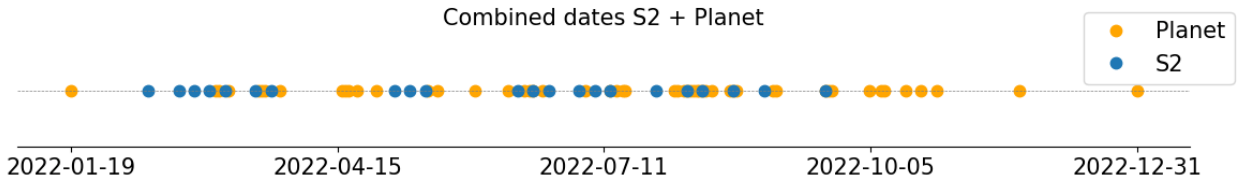


Figure 17 – Combined dates from S2 and PS

5.8 Fine-tuning

Deep learning is certainly a step forward with a lot of applications, but it remains a costly discovery, in terms of CO2 and power consumption, especially for small datasets or for personal use. Thereby, new methods are developed in order to use already trained models on smaller datasets. The main asset is to avoid long and consuming training parts by directly using a pre-trained model. For instance, there are widely known deep learning models trained for translation among many languages. It is then possible to use these models on a quite small and different dataset and still get relevant results. There are 2 well-known methods in order to make it: the inference and the fine-tuning. Inference consists of applying a pre-trained model on a new dataset, without changing any parameters or weights. On the contrary, fine-tuning will contain a small training part. Because of the optimal parameters and weights from the pre-trained model, the training set can be smaller, it has been divided by 4 in this study. It is only done to adapt the best parameters to the new dataset. A visual representation of the fine-tuning process is shown on Figure 18. The model is first pre-trained on Strasbourg (the source dataset), and then the corresponding weights are fine-tuned on Nancy (the target dataset). Nancy's normalisation has been applied with the minimum and maximum values from Nancy's dataset because the results where better and more adapted to this new dataset.

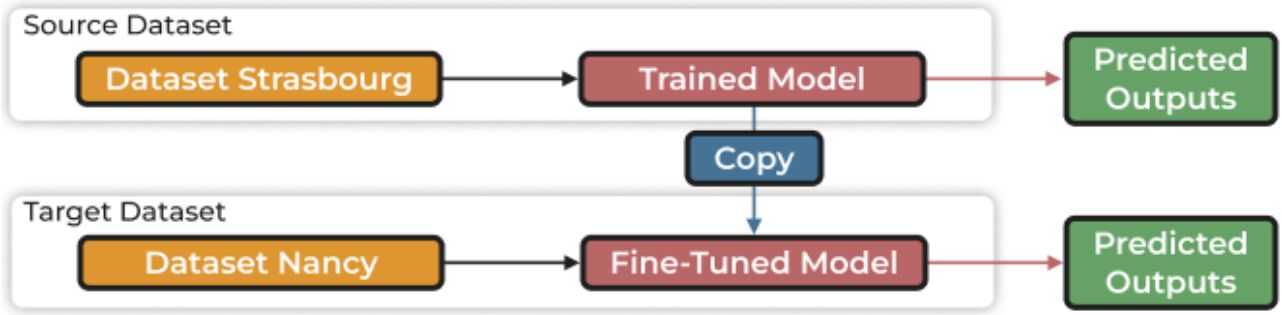


Figure 18 – Fine-tuning process from Strasbourg to Nancy, inspired by Ismail Fawaz et al. (2018)

5.9 Deep learning model evaluation methods

5.9.1 Training and validation loss plot

As mentioned previously, during each split, the model is learning over 30 or 60 epochs. After every epoch, the training and validation losses are computed and are expected to decrease over the training epochs. The loss is a way to evaluate qualitatively the hyperparameters and the overfitting. Indeed, if the model is constantly learning, the training loss will endlessly decrease, but on the other hand, the validation loss will decrease while remaining higher than the training one, until it increases again because of overfitting. Looking at the training and validation losses is then useful to check how the model is learning and if it is overfitting.

5.9.2 Confusion matrix and metrics definitions

Confusion matrix

As an indicator, the confusion matrix is a widely used tool to evaluate the model quantitatively. It gives the correct and wrong classifications per species as a tree number or a percentage of trees. Each row and each column represent a species, in the same order from top to bottom and from left to right. The rows are the true labels, while the columns are the predicted labels. The diagonal covers the number of right classifications. Everywhere else indicates the number of wrong classifications. There are 4 terms to know beforehand:

- True Positive (TP): Correctly identifying a positive case
- True Negative (TN): Correctly identifying a negative case
- False Positive (FP): Incorrectly identifying a negative case as positive
- False Negative (FN): Incorrectly identifying a positive case as negative

Table 2 gives a simple example of these previous definitions.

		Predicted labels	
		Cancer = Yes	Cancer = No
True labels	Cancer = Yes	**True Positive**	**False Negative**
	Cancer = No	**False Positive**	**True Negative**

Table 2 – Simplified confusion matrix

Metrics definition: Accuracy, F1-scores, Precision and Recall

There are distinct metrics to evaluate the model scores, defined in equations 9.

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP+FP} \\ \text{Recall} &= \frac{TP}{TP+FN} \\ \text{Accuracy} &= \frac{TP+TN}{TP+FP+TN+FN} \\ \text{F1-score} &= 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision+Recall}} \end{aligned} \tag{9}$$

The precision measures the proportion of TP among all the positive predictions made by the model, it is higher if the model detected a lot of cases. The recall measures the proportion is TP among all the actual positive individuals, it is lower if the model missed a lot of cases. Accuracy is the proportion of correct predictions made by the model among the entire dataset. The F1-score balances the precision and recall, it is their harmonic mean. While precision and recall are interesting to look at for each species, accuracy and F1-score provide summarised information about the model performance in classification.

6 Results

6.1 Global results for convolutional models

The convolutional models include InceptionTime, Hybrid InceptionTime and LITE. Different scenarios have been tested, with and without sensor fusion. The influence of kernel size and interpolation with Savitsky-Golay smoothing are studied and the results are shown in this part, focusing on accuracy. In the next tables, the highest accuracy is highlighted in dark green, the second highest in light green and the lowest in red.

6.1.1 Influence of kernel size

First, the scenarios with widely used parameters from the bibliography and their accuracy are presented in table 3, without interpolation and with a kernel size $k_2 = \{10, 20, 40\}$. The accuracy is computed with its standard deviation, as the mean between the 5 splits accuracies. The best results are happening with sensor fusion. The Hybrid model has the best accuracy with 69% of correct classifications (dark green), followed closely by the InceptionTime model with 67% (light green). The others have lower accuracies, especially LITE with 45% (red). Besides, the sensor fusion is always improving the results, and it is particularly efficient for LITE which gained around 10% of accuracy with both sensors and went over the 50% mark.

$k_2 = \{10, 20, 40\}$	S2	Planet	S2-Planet
InceptionTime	0.6292 +/- 0.0043	0.6221 +/- 0.0131	0.6768 +/- 0.0139
Hybrid InceptionTime	0.6305 +/- 0.0037	0.6242 +/- 0.0103	**0.6913 +/- 0.0034**
LITE	0.4574 +/- 0.0244	0.4875 +/- 0.0251	0.5612 +/- 0.025

Table 3 – Accuracy for scenarios on 3 convolutional models with $k_2 = \{10, 20, 40\}$

In order to see the influence of the kernel size, scenarios with smaller filters $k_1 = \{2, 4, 8\}$ have been tried. The results are shown in Table 4. Accuracies are close to the previous ones, but a little lower. The Hybrid model accuracy decreased so that best model is InceptionTime here. The results remain almost identical for InceptionTime, while the smaller filters influence the Hybrid and LITE through lower accuracies. LITE is not competitive anymore.

$k_1 = \{2, 4, 8\}$	S2	Planet	S2-Planet
InceptionTime	0.6315 +/- 0.0044	0.6400 +/- 0.0041	**0.6660 +/- 0.0035**
Hybrid InceptionTime	0.6048 +/- 0.0057	0.6073 +/- 0.0149	0.6320 +/- 0.0631
LITE	0.3648 +/- 0.0196	0.3720 +/- 0.009	0.4366 +/-0.0132

Table 4 – Accuracy for scenarios on 3 convolutional models with $k_1 = \{2, 4, 8\}$

Smaller filters are not improving the results for this study case. Therefore, the next scenarios will only use the classic filter lengths $k_2 = \{10, 20, 40\}$.

Finally, the training and validation losses for the Hybrid model with fusion are shown in appendix C Figure 26. The losses are two downward slopes over the epochs, while the validation loss remains higher than the training one as expected from a deep neural network. They are reaching a plateau, without increasing at the end.

6.1.2 Influence of interpolation with Savitsky-Golay smoothing

The interpolation followed a Savitsky-Golay smoothing is supposed to have a better representation of the tree phenology with more dates. The filter lengths are the classic ones, and the results are shown in table 5. The accuracies are lower than without interpolation. The results with sensor fusion are still higher, and the Hybrid model remains above 60%, whereas the others have dropped under. Besides, the PS results are higher around 4% than the S2 ones.

interpol+smoothing	S2	Planet	S2-Planet
InceptionTime	0.5434 +/- 0.0199	0.5779 +/- 0.0124	0.5880 +/- 0.0648
Hybrid InceptionTime	0.5431 +/- 0.014	0.5847 +/- 0.0144	**0.6235 +/- 0.0052**
LITE	0.3520 +/- 0.0147	0.3755 +/- 0.0265	0.4262 +/- 0.0201

Table 5 – Accuracy for scenarios on 3 convolutional models with interpolation and smoothing

The linear interpolation with smoothing is not improving the models. The best model accuracy using convolutional filters remains for the classic filter length with a fusion sensor and without interpolation.

6.2 Global results of Transformers

Different scenarios have been tested using the Transformers model, with varying parameters. Only the model and attention sizes are part of the scenarios. The Q and V values depend on the model size, and the other parameters remain fixed.

6.2.1 Influence of model size and *attention* size

The size of the model can be 64 or 128 according to empirical tests because the model is not learning with higher values. It could be because the dataset is small in time. The attention size can be 8, 16 or none as explained previously. None is then 22 for S2, 53 for PS and 75 for both. The results of these scenarios are shown in Table 6.

dmodel	attention size	S2	Planet	S2-Planet
64	8	0.6253 +/- 0.0062	0.6212 +/- 0.0070	0.6224 +/- 0.0068
64	16	0.6218 +/- 0.0080	0.6232 +/- 0.0053	0.6341 +/- 0.0068
64	none	0.6187 +/- 0.0147	0.6163 +/- 0.0048	**0.6373 +/- 0.0044**
128	8	0.6140 +/- 0.0114	0.6202 +/- 0.0088	0.5660 +/- 0.1172
128	16	0.6242 +/- 0.0072	0.6242 +/- 0.0131	0.6042 +/- 0.0359
128	none	0.6258 +/- 0.0068	0.6111 +/- 0.0278	0.6040 +/- 0.0387

Table 6 – Accuracy for scenarios on transformer model

The first thing to notice is that all the accuracies are very close to each other. The second is that $dmodel = 64$ has the best accuracy with no attention size, and $dmodel = 128$ has the last one with an attention size of 8. Except for this last result, all the accuracies are above 60%. Besides, even if they are generally better with a sensor fusion than without, as it is with the convolutional models, the gap is almost none. However, the results do not seem stable. By running a few time the model with the same parameters, the results can change up to around 5%, which makes it difficult to argue with a better model.

6.2.2 Attention maps

The very new specificity of Transformers is the *attention map* as possible additional output. They are a precious tool to understand how the model is learning the features of time series and why it has wrong classifications. However, the understanding of attention maps is almost unknown for classification tasks using time series. It is a recent discovery and very unknown from large public.

An attention map highlights the key moments in the decision-making process for each object class. Rows and columns represent the time series from top to bottom and left to right. The pixel value is the attention weight assigned by the model. The higher the value is, the more intense the attention is on these dates. It means that these dates are more important than the others in the decision-making during the classification process. The attention map can be captured for each head and each layer in the encoder process of the Transformer. A mean value is then given for all the species. However, it is more interesting to look at the attention maps for each species because information can be taken to explain why the model has wrong classifications. More than that, it can be relevant to have attention maps per species for the correct and the wrong classifications separately because the information given could be different for each case. The hypothesis is that a species wrongly classified is making links between dates corresponding to another species specificity for correct classification. Even though the mechanism is way more complex, this is a first intuitive guideline.

Figure 27 in Appendix D shows all the attention maps per species for the correct and wrong classifications, with a model size of 64 and an attention size of 16. Even if no attention size gives better results, these results are not very stable as mentioned before, and the attention size of 16 narrows the filed of analysis. The scales are purposely not normalised by assuming that the main point is the see the highest attention for each species and not the highest among all the species.

The attention maps are all looking very similar, often with a higher attention at the beginning and the end of the time series. There is surprisingly a high noticeable point around date 20 for most of the species. Besides, a lot of vertical lines are visible, and sometimes a few light pixels close to the diagonal. In addition, there are not important visible differences between correct and incorrect maps.

6.3 Overview of best models accuracy

Figure 19 summarises a comparison of the 4 model's accuracies using one sensor and their fusion. The best parameters are chosen for each model. Each model has the best accuracy with a sensor fusion. The Hybrid model has the higher accuracy, followed closely by the InceptionTime and the Transformer models. LITE remains above 50% with fusion, but its accuracy is not as much competitive.

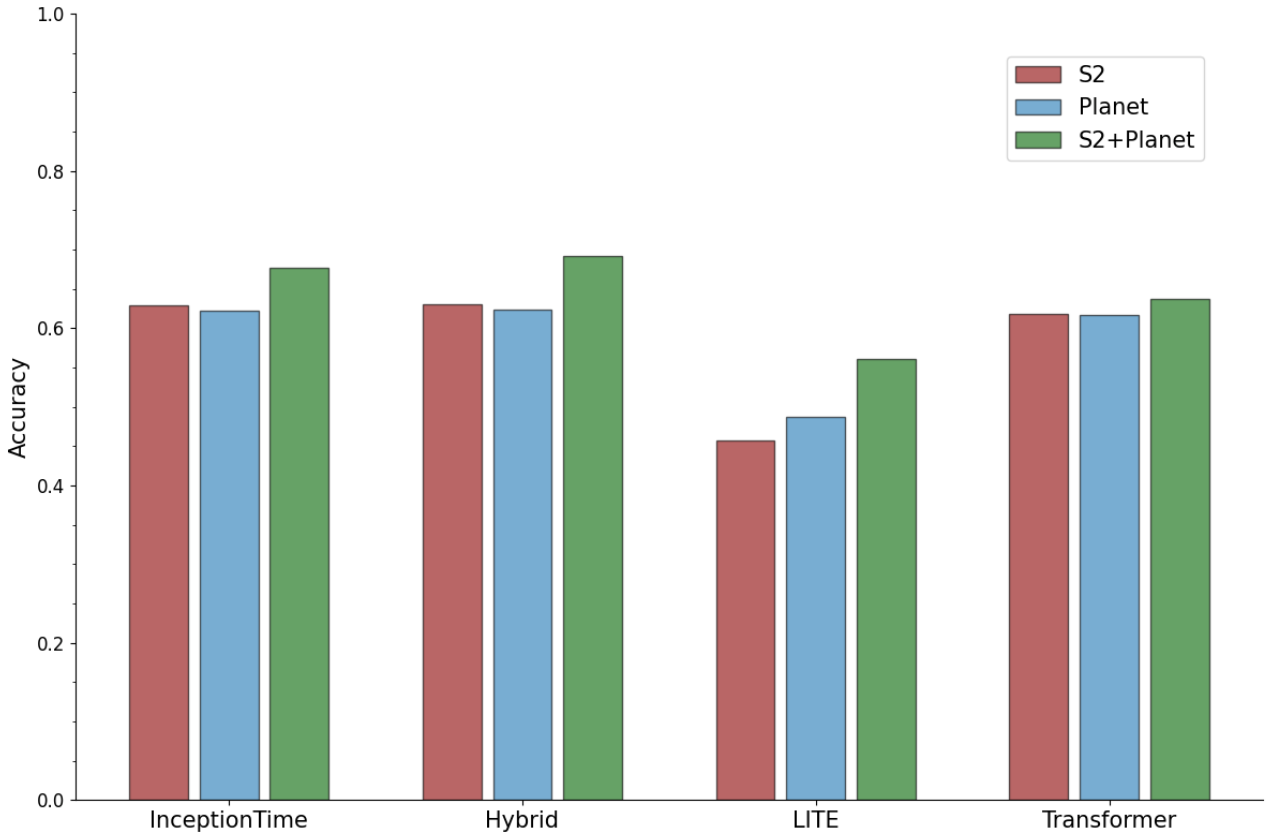


Figure 19 – Comparison of the best model's accuracies with one and two sensors

6.4 Urban tree species classification

In this section are presented the best model results for each tree species and a model comparison.

6.4.1 F1-scores comparison per species

The F1-scores are represented for each species from the Hybrid and the InceptionTime models in Figure 20 for both sensors. The scales are starting at 0.4 for greater understanding. The two models are competitive, they have a close F1-score for each species, and almost all the species are above 50% for both models. However, only 2 species have a higher F1-score for InceptionTime, which are the *Pinus nigra* and the *Alnus glutinosa*. Besides, the *Platanus x acerifolia* has from far the highest F1-score in both models, and it is the most represented species. It is also noticeable that the least represented species do not have the lowest F1-scores. For instance, the second highest score is for the *Styphnolobium japonicum*, the 4th species with the fewest number of trees.

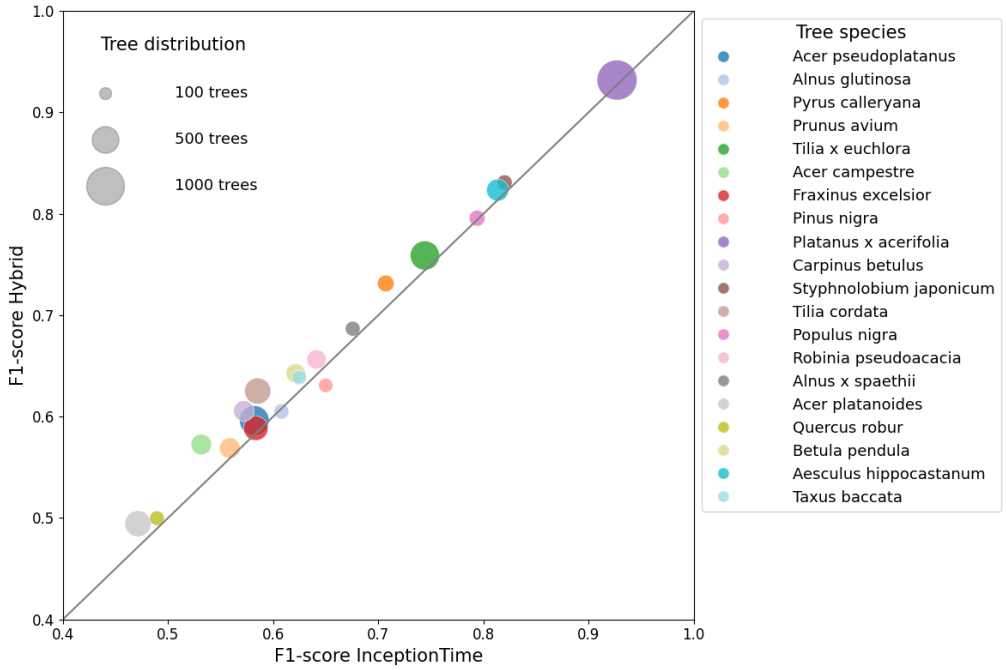


Figure 20 – F1-score comparison for H-Inception and InceptionTime models per species

The F1-scores are then compared for Hybrid and Transformer models in Figure 21. All the species have a higher F1-score for the Hybrid model. Besides, the 2 last scores, *Quercus robur* and *Acer platanoides* are here further away from 50% for Transformers, creating a bigger gap between the species scores for this model. However, Transformer still manages to get a correct score for the under-represented species and is competitive in that way.

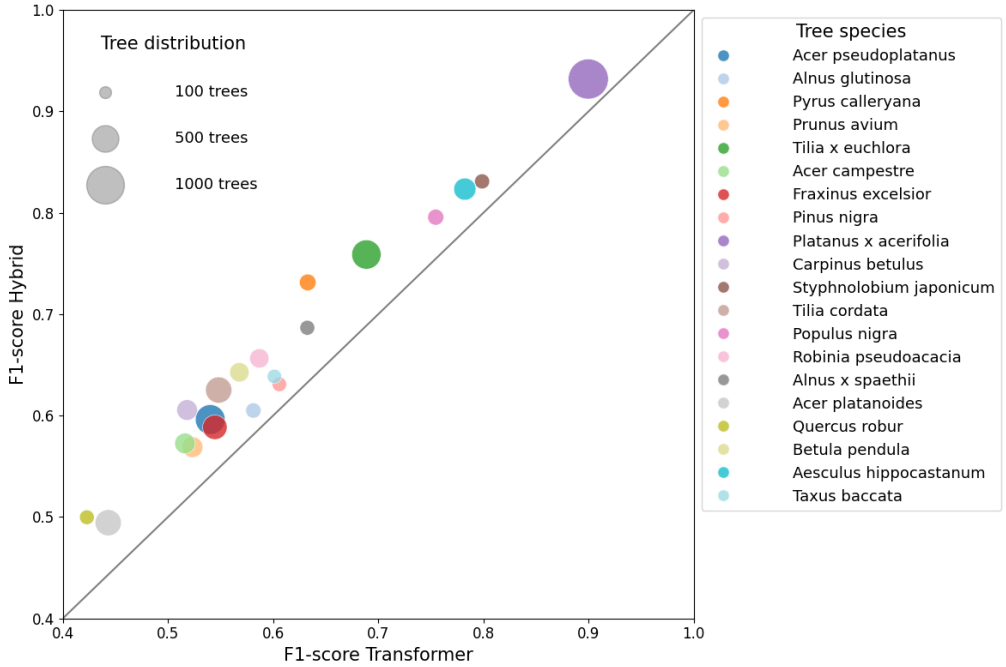


Figure 21 – F1-score comparison for H-Inception and Transformer models per species

6.4.2 Confusion matrix for Hybrid model

The confusion matrix in percentage is presented for the best model, Hybrid InceptionTime, in Figure 22. The percentage of correct classification can be seen on the diagonal, where the higher numbers are. The other numbers are errors in the classification. They are low everywhere, except for a few pixels. For instance, 11% of *Acer platanoides* trees have been classified as *Acer pseudoplatanus*, and the same the other way. 10% of *Tilia cordata* trees have been classified as *Tilia x euchlora* and the same the other way. Finally, the species *Planatus x acerifolia* has the best and very high number of correct classifications with 93%.

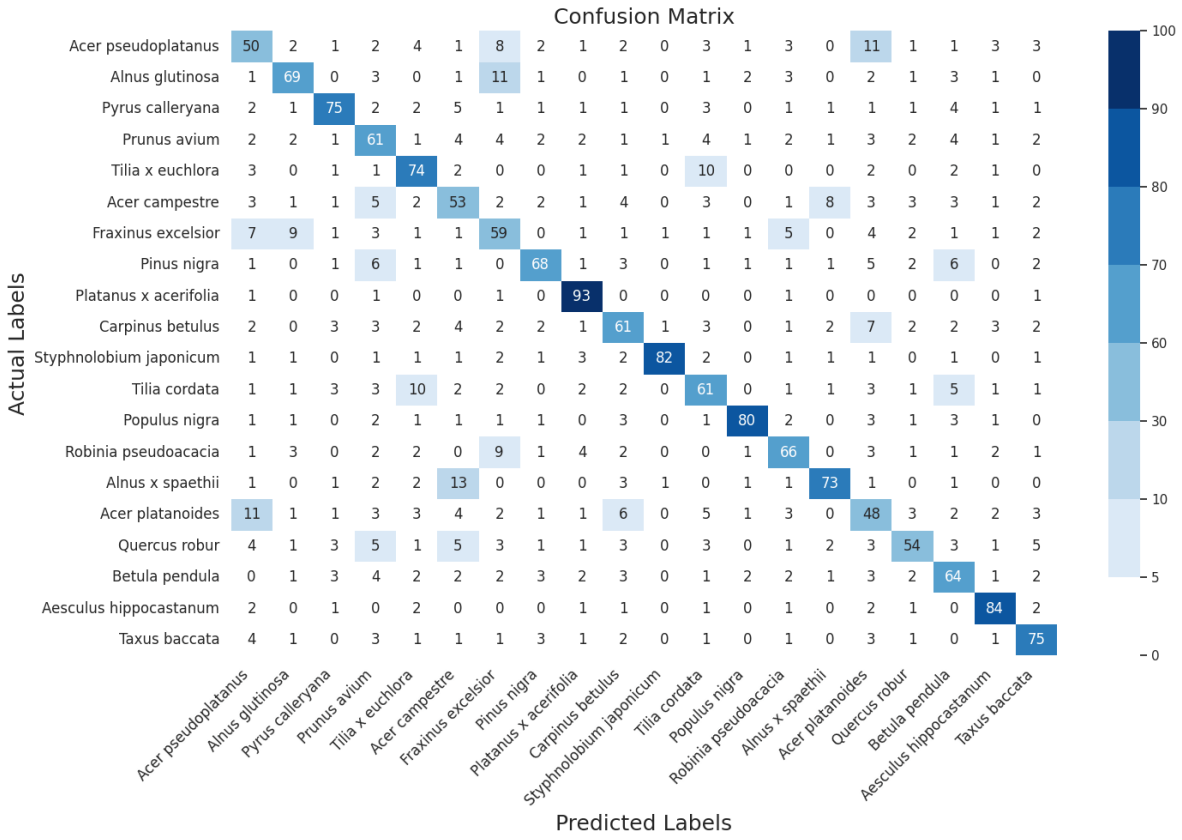


Figure 22 – Confusion matrix for Hybrid InceptionTime model

6.4.3 Thematic observation of classifications

The results of the process can be observed on a map. The aim is to understand if the geographical arrangement of trees has an influence on the model prediction, as well as their spacing. Figure 23 is mapping the urban trees in Strasbourg for the Hybrid model, and highlights the correct and wrong classifications. There are visually more correct than wrong predictions, which is consistent with the accuracy. Two focuses on various areas are shown on the right side. The trees in the streets near Republic Place are well-aligned and more separated, the predictions are particularly accurate. On the contrary, the trees in the Orangerie Park are closer to each other, not especially aligned and the classifications are weaker.

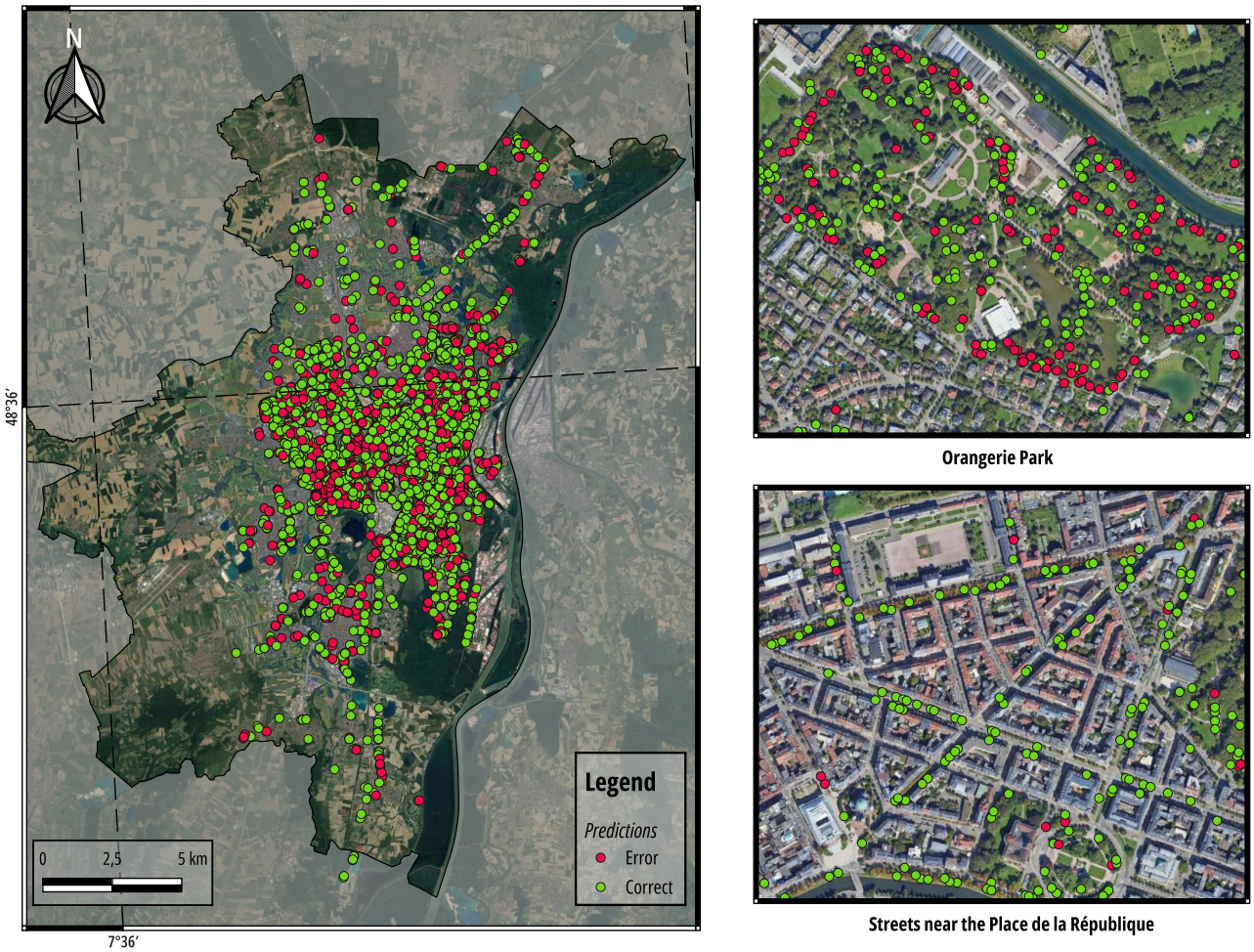


Figure 23 – Qualitative results for the test set over Strasbourg for H-Inception model. On the right side of the map, two focuses over the Orangerie park and the Place de la République

6.5 Fine tuning over Nancy

The inference and fine-tuning have been tried on Nancy with the best model's parameters trained on Strasbourg: Hybrid InceptionTime, kernel size of $k_2 = \{10, 20, 40\}$. Table 7 is showing the results. The inference model is not learning on Nancy dataset at all, whereas the fine-tuning network produces an accuracy of 56%, which is around 13% under the one for Strasbourg, but the network is still learning above 50%. Besides, the whole training on Nancy has been tried as well. Although the dataset is smaller for Nancy, the model is learning up to 70% which surpasses Strasbourg accuracy with the same parameters.

Hybrid InceptionTime	S2-Planet
Inference	0.0722 +/- 0.0000
Fine-tuning	0.5600 +/- 0.0098
All training	**0.7055 +/- 0.0046**

Table 7 – Accuracy for inference, fine-tuning and all training on Nancy

7 Discussion

7.1 Convolutional models

The results showed that a sensor fusion is improving the accuracy for all the convolutional models. Indeed, S2 and PS complement each other, in terms of spectral bands and number of dates. Besides, the Hybrid InceptionTime model furnishes the best results even if the gap with classic InceptionTime is small. It confirms the hypothesis that the hand-crafted features can detect up and down slopes and peaks. However, smaller filters do not improve the results, which confirms what Fawaz et al. (2020) explained.

On the other hand, LITE has the lowest accuracies in general. Indeed, this model has been created with the aim of reducing the number of parameters while maintaining the results. However, the classification task is complex, as confirmed by the t-SNE study. This could explain why this model does not have the best results, although it exceeds 50% of accuracy with a classic filter length with both sensors.

In addition, linear interpolation with Savitsky-Golay smoothing is far from competitive. It could be because the interpolation is linear, while the tree phenologies are not. The network could then not understand the spectral variations, even with the smoothing. It can also be the reason why results are better for PS than for S2 because it has more dates and so the interpolation with smoothing is closer to the real time series. Moreover, the smoothing is made to reduce the noise while maintaining the feature's characteristics, but it could have smoothed also important details which are different from the noise.

Finally, the decreasing losses in Figure 26 are consistent with deep learning expectations. Even if the validation loss remains above the training one as it is usually, it is not increasing at the end, which shows that there is no overfitting.

7.2 Transformers

First of all, the results are very close with one or two sensors. It could be because the fusion is made directly on the dataset before the model. The hypothesis of gaining more information with 75 dates does not seem correct. Another type of sensor fusion could be thought out in the future. Besides, the results appear to be a little bit better with a model size of 64 and also with a higher attention size, probably because they are closer to the number of dates. However, these are only hypotheses because the results may not be stable over several runs.

As Transformers are quite recent for the application of TSC, the analyses are not an easy task, especially regarding the attention maps. For now, they are not understood or explained correctly. The fact that all the maps look similar could be because tree phenology is similar between species, or because the classification task is very complex, or just because attention models can not learn clear distinctions between species. It was expected to have higher attention

contrasts to be put in parallel with the tree's phenology to highlight links between dates (Liu et al., 2021). Transformers do not look adapted for TSC in this study case. Attention maps are very useful for prediction tasks. For instance, it gives information about each word's importance in the prediction of the other words. However, there is no prediction in classification tasks.

7.3 Tree species classification analysis

In Figure 20 and 21, almost all the species have higher F1-scores for the Hybrid model, which is consistent with the fact that the Hybrid model is the best one in accuracy. This model is improving the general results but not significantly compared to InceptionTime. The most representative species *Platanus x acerifolia* has the highest F1-score for both models because the network has learned intensively about this species' features (Krawczyk, 2016). Indeed, it has more than 7500 trees in the original dataset, while the second species has around 4500 trees, way less. This result is consistent with the confusion matrix in Figure 22 where there is 93% of correct classifications. However, the fact that the least represented species do not have the lowest F1-score proves that the imbalanced between the classes has been corrected efficiently with the weights assigned inversely to the frequency (Audebert et al., 2018). On the other hand, species have lower scores for Transformer (Figure 21), which is consistent with the general accuracy scores. In both cases, *Acer platanoides* has one of the lowest F1-scores. By looking at the confusion matrix, this species has only 48% of good classifications, and 11% of this species trees have been classified as *Acer pseudoplatanus*. Indeed, these 2 species are close to each other and have similar phenologies. For the same reason, *Tilia x euchlora* and *Tilia cordata* are misunderstood.

Moreover, the map of urban trees in Strasbourg in Figure 23 has good classifications in the streets near Republic Place because the trees are aligned and separated. First, the space between them is allowing the pixels to have pure values and not a combination with neighbour trees. Besides, the tree species in these streets are almost all the same. On the other hand, there are more wrong classifications in the Orangerie Park. Indeed, in this case the trees are close to each other, which can create a blend of value for each tree and confuses the neural network. In addition, the species in this park are very different from one tree to its neighbour. It might be easier for the network to classify well-spaced trees. Besides, the general environment can have an influence on the behaviour of the neighbour tree phenology (Franco, 1986).

7.4 Fine tuning over Nancy

The inference method did not work on Nancy. It could probably be a problem with one of the choices made for pre-processing or training, which has not been identified yet. It can be linked to the dataset itself, the choice of samples or something else.

However, the fine-tuning did obtain promising results with a very small training dataset. Intuitively, the minimum and maximum values of Strasbourg should be used to normalise Nancy dataset because it allows consistence in the model processing. However, the model did not learn using the values from Strasbourg, but it did using Nancy's values. These values are more fitting with Nancy's dataset and the small training after that helped the model to quickly adapt and perform. This result showed the interest of fine-tuning with a small training part.

Conclusion

Through this research internship, the objective was to implement new deep-learning models from the state-of-the-art and develop a method for using sensor fusion. More than improving results for Time Series Classification on urban tree species, the long-term guideline is to re-use these new methods in other cities using fine-tuning to have a bigger impact.

Methodological issues have been focused on sensor fusion. On convolutional models, it has been efficiently applied and improved results. On Transformer model, it has been applied beforehand as a first way of trying fusion with this new type of model architecture.

The key findings and the main contributions of this research are the very high and competitive results of the Hybrid InceptionTime model through the hand-crafted filters improving detection of growth, decline and peaks. The InceptionTime model remains competitive, while the LITE has for now lower results. Transformers are not surpassing the other methods and does not seem as adapted as the others for TSC.

In the future, a geographical stratification can be tried into set splits (Maxwell et al., 2021). Indeed, the high species density per pixel can create a potential bias between the different sets. A map of use land can help to stratify the data. However, this perspective can create an imbalance in the dataset distribution between the train and test sets for the spatial organisation. For instance, trees in streets are often aligned and trees in parks are often in groves. It can lead to excessive focus on a type of spatial organisation in the training part and then lower scores for the under-represented type in the train set. Moreover, the tree crowns could be better defined and add the altitude component using LIDAR source of data.

Finally, it could be useful in the future to analyse the errors, by looking at the tree phenology and by examining the tree spatial organisation and localisation. For instance, producing a Gradient-weighted Class Activation Mapping (Grad-CAM) could help to analyse the tree phenology, while analysing the Local Climate Zones (LCZ) could be useful for the spatial organisation. All these perspectives are related to the climate stake and the necessity to act on several sides including the urban vegetation which is at the core of ecosystem services.

Bibliography

Shahzad Ali, Abdul Basit, Muhammad Umair, and Jian Ni. Impacts of climate and land coverage changes on potential evapotranspiration and its sensitivity on drought phenomena over South Asia. *International Journal of Climatology*, 44(3):812–830, March 2024. ISSN 0899-8418, 1097-0088. doi: 10.1002/joc.8357. URL https://rmets.onlinelibrary.wiley.com/doi/10.1002/joc.8357.

Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1):53, March 2021. ISSN 2196-1115. doi: 10.1186/s40537-021-00444-8. URL https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00444-8.

Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140:20–32, June 2018. ISSN 09242716. doi: 10.1016/j.isprsjprs.2017.11.011. URL https://linkinghub.elsevier.com/retrieve/pii/S0924271617301818.

Clément Bressant, Pierre-Alexis Herrault, and Anne Puissant. Fine-Scale Phenology of Urban Trees From Satellite Image Time Series: Toward a Comprehensive Analysis of Influencing Factors. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:11685–11706, 2024. ISSN 1939-1404, 2151-1535. doi: 10.1109/JSTARS.2024.3411304. URL https://ieeexplore.ieee.org/document/10552035/.

Jin Chen, Per. Jönsson, Masayuki Tamura, Zhihui Gu, Bunkei Matsushita, and Lars Eklundh. A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter. *Remote Sensing of Environment*, 91(3):332–344, June 2004. ISSN 0034-4257. doi: 10.1016/j.rse.2004.03.014. URL https://www.sciencedirect.com/science/article/pii/S003442570400080X.

Hamish Clarke, Brett Cirulis, Trent Penman, Owen Price, Matthias M. Boer, and Ross Bradstock. The 2019–2020 Australian forest fires are a harbinger of decreased prescribed burning effectiveness under rising extreme conditions. *Scientific Reports*, 12(1):11871, July 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-15262-y. URL https://www.nature.com/articles/s41598-022-15262-y.

Copernicus. Copernicus: June 2024 marks 12th month of global temperature reaching 1.5°C above pre-industrial | Copernicus, july 2024. URL https://climate.copernicus.eu/copernicus-june-2024-marks-12th-month-global-temperature-reaching-15degc-above-pre-industrial#.

Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic Head: Unifying Object Detection Heads with Attentions, 2021. URL https://arxiv.org/abs/2106.08322. Version Number: 1.

Clément Dechesne, Clément Mallet, Arnaud Le Bris, and Valérie Gouet-Brunet. Semantic segmentation of forest stands of pure species combining airborne lidar data and very high resolution multispectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 126:129–145, 2017. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2017.02.011. URL https://www.sciencedirect.com/science/article/pii/S0924271616302763.

Angus Dempster, François Petitjean, and Geoffrey I. Webb. ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495, September 2020. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-020-00701-z. URL http://arxiv.org/abs/1910.13051. arXiv:1910.13051 [cs, stat].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018. URL https://arxiv.org/abs/1810.04805. Version Number: 2.

Manuel Esperon-Rodriguez, Mark G. Tjoelker, Jonathan Lenoir, John B. Baumgartner, Linda J. Beaumont, David A. Nipperess, Sally A. Power, Benoît Richard, Paul D. Rymer, and Rachael V. Gallagher. Climate change increases global risk to urban forests. *Nat. Clim. Change*, 12:950–955, October 2022. ISSN 1758-6798. doi: 10.1038/s41558-022-01465-8. URL https://www.nature.com/articles/s41558-022-01465-8#citeas.

European Comission. Wildfires in the Mediterranean: EFFIS data reveal extent this summer, August 2024. URL https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/wildfires-mediterranean-effis-data-reveal-extent-summer-2023-09-08_en. [Online; accessed 18. Aug. 2024].

European Commission. Increasing tree coverage to 30% in European cities could reduce deaths linked to urban heat island effect, August 2023. URL https://environment.ec.europa.eu/news/increasing-tree-coverage-30-european-cities-could-reduce-deaths-linked-urban-heat-island-effect-2023-06-21_en. [Online; accessed 19. Aug. 2024].

Johann Faouzi. Time Series Classification: A Review of Algorithms and Implementations. In Jorge Rocha, Cláudia M. Viana, and Sandra Oliveira, editors, *Time Series Analysis - Recent Advances, New Perspectives and Applications*. IntechOpen, March 2024. ISBN 978-0-85466-053-7 978-0-85466-052-0. doi: 10.5772/intechopen.1004810. URL https://www.intechopen.com/chapters/1185930.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. InceptionTime: Finding AlexNet for Time Series Classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, November 2020. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-020-00710-y. URL http://arxiv.org/abs/1909.04939. arXiv:1909.04939 [cs, stat].

Matheus Pinheiro Ferreira, Daniel Rodrigues Dos Santos, Felipe Ferrari, Luiz Carlos Teixeira Coelho Filho, Gabriela Barbosa Martins, and Raul Queiroz Feitosa. Improving urban tree species classification by deep-learning based fusion of digital aerial images and LiDAR. *Urban Forestry & Urban Greening*, 94:128240, April 2024. ISSN 16188667. doi: 10.1016/j.ufug.2024.128240. URL https://linkinghub.elsevier.com/retrieve/pii/S1618866724000384.

M. Franco. The influence of neighbours on the growth of modular organisms with an example from trees. *Phil. Trans. R. Soc. Lond. B*, 313(1159):209–225, August 1986. ISSN 2054-0280. doi: 10.1098/rstb.1986.0034. URL https://royalsocietypublishing.org/doi/10.1098/rstb.1986.0034.

Government of Canada. Canada's record-breaking wildfires in 2023: A fiery wake-up call, August 2024. URL https://natural-resources.canada.ca/simply-science/canadas-record-breaking-wildfires-2023-fiery-wake-call/25303. [Online; accessed 18. Aug. 2024].

Sean Hartling, Vasit Sagan, and Maitiniyazi Maimaitijiang. Urban tree species classification using UAV-based multi-sensor data fusion and machine learning. *GIScience & Remote Sensing*, 58(8):1250–1275, November 2021. ISSN 1548-1603, 1943-7226. doi: 10.1080/15481603.2021.1974275. URL https://www.tandfonline.com/doi/full/10.1080/15481603.2021.1974275.

Dino Ienco, Raffaele Gaetano, Claire Dupaquier, and Pierre Maurel. Land cover classification via multitemporal spatial data by deep recurrent neural networks. *IEEE Geoscience and Remote Sensing Letters*, 14(10):1685–1689, 2017. doi: 10.1109/LGRS.2017.2728698. URL https://ieeexplore.ieee.org/document/8006221.

Intergovernmental Panel On Climate Change (Ipcc). *Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 1 edition, June 2023. ISBN 978-1-00-932584-4. doi: 10.1017/9781009325844. URL https://www.cambridge.org/core/product/identifier/9781009325844/type/book.

Ipcc. *Global Warming of 1.5°C: IPCC Special Report on Impacts of Global Warming of 1.5°C above Pre-industrial Levels in Context of Strengthening Response to Climate Change, Sustain-*

able Development, and Efforts to Eradicate Poverty. Cambridge University Press, 1 edition, June 2022. ISBN 978-1-00-915794-0 978-1-00-915795-7. doi: 10.1017/9781009157940. URL https://www.cambridge.org/core/product/identifier/9781009157940/type/book.

Ali Ismail-Fawaz, Maxime Devanne, Jonathan Weber, and Germain Forestier. Deep Learning For Time Series Classification Using New Hand-Crafted Convolution Filters. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 972–981, Osaka, Japan, December 2022. IEEE. ISBN 978-1-66548-045-1. doi: 10.1109/BigData55660.2022.10020496. URL https://ieeexplore.ieee.org/document/10020496/.

Ali Ismail-Fawaz, Maxime Devanne, Stefano Berretti, Jonathan Weber, and Germain Forestier. LITE: Light Inception with boosTing tEchniques for Time Series Classification. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, October 2023. doi: 10.1109/DSAA60987.2023.10302569. URL https://ieeexplore.ieee.org/document/10302569.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Transfer learning for time series classification. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1367–1376, Seattle, WA, USA, December 2018. IEEE. ISBN 978-1-5386-5035-6. doi: 10.1109/BigData.2018.8621990. URL https://ieeexplore.ieee.org/document/8621990/.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, July 2019. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-019-00619-1. URL https://link.springer.com/10.1007/s10618-019-00619-1.

N. Karasiak, D. Sheeren, M. Fauvel, J. Willm, J.-F. Dejoux, and C. Monteil. Mapping tree species of forests in southwest France using Sentinel-2 image time series. In *2017 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, pages 1–4, Brugge, Belgium, June 2017. IEEE. ISBN 978-1-5386-3327-4. doi: 10.1109/Multi-Temp.2017.8035215. URL http://ieeexplore.ieee.org/document/8035215/.

Y. Kobayashi, G.G. Karady, G.T. Heydt, and R.G. Olsen. The Utilization of Satellite Images to Identify Trees Endangering Transmission Lines. *IEEE Transactions on Power Delivery*, 24 (3):1703–1709, July 2009. ISSN 0885-8977, 1937-4208. doi: 10.1109/TPWRD.2009.2022664. URL http://ieeexplore.ieee.org/document/5071199/.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Prog. Artif. Intell.*, 5(4):221–232, November 2016. ISSN 2192-6360. doi: 10.1007/s13748-016-0094-0. URL https://link.springer.com/article/10.1007/s13748-016-0094-0?TB_

iframe=true&error=cookies_not_supported&code=a3e33168-782e-41e5-8585-e731754069d2#citeas.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, May 2015. ISSN 1476-4687. doi: 10.1038/nature14539. URL https://www.nature.com/articles/nature14539#citeas.

Minghao Liu, Shengqi Ren, Siyuan Ma, Jiahui Jiao, Yizhou Chen, Zhiguang Wang, and Wei Song. Gated Transformer Networks for Multivariate Time Series Classification, March 2021. URL http://arxiv.org/abs/2103.14438. arXiv:2103.14438 [cs].

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Ivo Augusto Lopes Magalhães, Osmar Abílio de Carvalho Júnior, Osmar Luiz Ferreira de Carvalho, Anesmar Olino de Albuquerque, Potira Meirelles Hermuche, Éder Renato Merino, Roberto Arnaldo Trancoso Gomes, and Renato Fontes Guimarães. Comparing machine and deep learning methods for the phenology-based classification of land cover types in the amazon biome using sentinel-1 time series. *Remote Sensing*, 14(19), 2022. ISSN 2072-4292. doi: 10.3390/rs14194858. URL https://www.mdpi.com/2072-4292/14/19/4858.

Aaron E. Maxwell, Timothy A. Warner, and Luis Andrés Guillén. Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—part 2: Recommendations and best practices. *Remote Sensing*, 13(13), 2021. ISSN 2072-4292. doi: 10.3390/rs13132591. URL https://www.mdpi.com/2072-4292/13/13/2591.

Mohammad Amin Morid, Olivia R. Liu Sheng, and Joseph Dunbar. Time series prediction using deep learning methods in healthcare. *ACM Trans. Manage. Inf. Syst.*, 14(1), jan 2023. ISSN 2158-656X. doi: 10.1145/3531326. URL https://doi.org/10.1145/3531326.

Robbe Neyns, Kyriakos Efthymiadis, Pieter Libin, and Frank Canters. Fusion of multi-temporal PlanetScope data and very high-resolution aerial imagery for urban tree species mapping. *Urban Forestry & Urban Greening*, 99:128410, September 2024. ISSN 16188667. doi: 10.1016/j.ufug.2024.128410. URL https://linkinghub.elsevier.com/retrieve/pii/S1618866724002085.

T. R. Oke. *Boundary layer climates*. Routledge, London, second edition edition, 1987. ISBN 978-0-203-40721-9. OCLC: 51200739.

Claudia Paris, Giulio Weikmann, and Lorenzo Bruzzone. Monitoring of agricultural areas by using Sentinel 2 image time series and deep learning techniques. In Lorenzo Bruzzone, Francesca Bovolo, and Emanuele Santi, editors, *Image and Signal Processing for Remote*

Sensing XXVI, volume 11533, page 115330K. International Society for Optics and Photonics, SPIE, 2020. doi: 10.1117/12.2574745. URL https://doi.org/10.1117/12.2574745.

Charlotte Pelletier, Geoffrey I. Webb, and François Petitjean. Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series. *Remote Sensing*, 11(5):523, January 2019. ISSN 2072-4292. doi: 10.3390/rs11050523. URL https://www.mdpi.com/2072-4292/11/5/523. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

Nathalia Philipps, Pierre Kastendeuch, Olivier Montauban, and Georges Najjar. Rôle de la végétation et de la géométrie urbaine dans la variabilité spatio-temporelle de l'îlot de chaleur urbain: cas de la ville de Strasbourg. 2020. URL https://www.researchgate.net/publication/365993562_ROLE_DE_LA_VEGETATION_ET_DE_LA_GEOMETRIE_URBAINE_DANS_LA_VARIABILITE_SPATIO-TEMPORELLE_DE_L%27ILOT_DE_CHALEUR_URBAIN_CAS_DE_LA_VILLE_DE_STRASBOURG.

Daniel Scheffler, André Hollstein, Hannes Diedrich, Karl Segl, and Patrick Hostert. AROSICS: An Automated and Robust Open-Source Image Co-Registration Software for Multi-Sensor Satellite Data. *Remote Sensing*, 9(7):676, July 2017. ISSN 2072-4292. doi: 10.3390/rs9070676. URL https://www.mdpi.com/2072-4292/9/7/676. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

Karen C. Seto, Burak Güneralp, and Lucy R. Hutyra. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proceedings of the National Academy of Sciences*, 109(40):16083–16088, 2012. doi: 10.1073/pnas.1211658109. URL https://www.pnas.org/doi/abs/10.1073/pnas.1211658109.

Sebastian Stroud, Julie Peacock, and Christopher Hassall. Vegetation-based ecosystem service delivery in urban landscapes: A systematic review. *Basic and Applied Ecology*, 61:82–101, 2022. ISSN 1439-1791. doi: https://doi.org/10.1016/j.baae.2022.02.007. URL https://www.sciencedirect.com/science/article/pii/S1439179122000184.

André Stumpf, David Michéa, and Jean-Philippe Malet. Improved Co-Registration of Sentinel-2 and Landsat-8 Imagery for Earth Surface Motion Measurements. *Remote Sensing*, 10(2):160, January 2018. ISSN 2072-4292. doi: 10.3390/rs10020160. URL https://www.mdpi.com/2072-4292/10/2/160.

Chenxi Sun, Shenda Hong, Moxian Song, and Hongyan Li. A review of deep learning methods for irregularly sampled medical time series data, 2020. URL https://arxiv.org/abs/2010.12493.

Irfan Ullah, Sourav Mukherjee, Sidra Syed, Ashok Kumar Mishra, Brian Odhiambo Ayugi, and Saran Aadhar. Anthropogenic and atmospheric variability intensifies flash drought episodes

in South Asia. *Communications Earth & Environment*, 5(1):267, May 2024. ISSN 2662-4435. doi: 10.1038/s43247-024-01390-y. URL https://www.nature.com/articles/s43247-024-01390-y.

I UN-DESA. 2018 revision of world urbanization prospects, united nations department of economic and social affairs. 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].

Zhiguang Wang, Weizhong Yan, and Tim Oates. Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline, December 2016. URL http://arxiv.org/abs/1611.06455. arXiv:1611.06455 [cs, stat].

Romain Wenger, Anne Puissant, Jonathan Weber, Lhassane Idoumghar, and Germain Forestier. Multimodal and Multitemporal Land Use/Land Cover Semantic Segmentation on Sentinel-1 and Sentinel-2 Imagery: An Application on a MultiSenGE Dataset. *Remote Sensing*, 15(1):151, December 2022a. ISSN 2072-4292. doi: 10.3390/rs15010151. URL https://www.mdpi.com/2072-4292/15/1/151.

Romain Wenger, Anne Puissant, Jonathan Weber, Lhassane Idoumghar, and Germain Forestier. U-Net feature fusion for multi-class semantic segmentation of urban fabrics from Sentinel-2 imagery: an application on Grand Est Region, France. *International Journal of Remote Sensing*, 43(6):1983–2011, March 2022b. ISSN 0143-1161, 1366-5901. doi: 10.1080/01431161.2022.2054295. URL https://www.tandfonline.com/doi/full/10.1080/01431161.2022.2054295.

Romain Wenger, Clément Bressant, Lucie Roettelé, Germain Forestier, and Anne Puissant. Improving Urban Tree Species Classification with High Resolution Satellite Imagery and Machine Learning. *ResearchGate*, 2024. doi: 10.13140/RG.2.2.10960.11521. URL https://rgdoi.net/10.13140/RG.2.2.10960.11521.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A Transformer-based Framework for Multivariate Time Series Representation Learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2114–2124, Virtual Event Singapore, August 2021. ACM. ISBN 978-1-4503-8332-5. doi: 10.1145/3447548.3467401. URL https://dl.acm.org/doi/10.1145/3447548.3467401.

Xiaoyong Zhang, Weiwei Jia, Yuman Sun, Fan Wang, and Yujie Miu. Simulation of spatial and temporal distribution of forest carbon stocks in long time series—based on remote sensing and deep learning. *Forests*, 14(3), 2023. ISSN 1999-4907. doi: 10.3390/f14030483.

Appendices

A Information about the species and spectral bands

Species	Percentage of trees in Strasbourg	Percentage of species in Nancy
Platanus x acerifolia	16,7	13,2
Acer pseudoplatanus	9,2	9,3
Tilia x euchlora	8,9	8
Tilia cordata	7,1	5,7
Acer platanoides	7	17,8
Fraxinus excelsior	6,2	7,8
Aesculus hippocastanum	5,1	2,6
Prunus avium	4,4	3,9
Carpinus betulus	4,4	7,2
Acer campestre	4,4	3,4
Betula pendula	3,8	3,4
Robinia pseudoacacia	3,8	5,7
Pyrus calleryana	2,9	3,6
Populus nigra	2,6	2,5
Alnus glutinosa	2,4	1,1
Styphnolobium japonicum	2,3	0
Alnus x spaethii	2,3	0,8
Quercus robur	2,2	1,7
Taxus baccata	2,2	0,2
Pinus nigra	2,1	2
Total number of trees	**45084**	**25020**

Table 8 – Percentage of trees per species in Strasbourg and Nancy

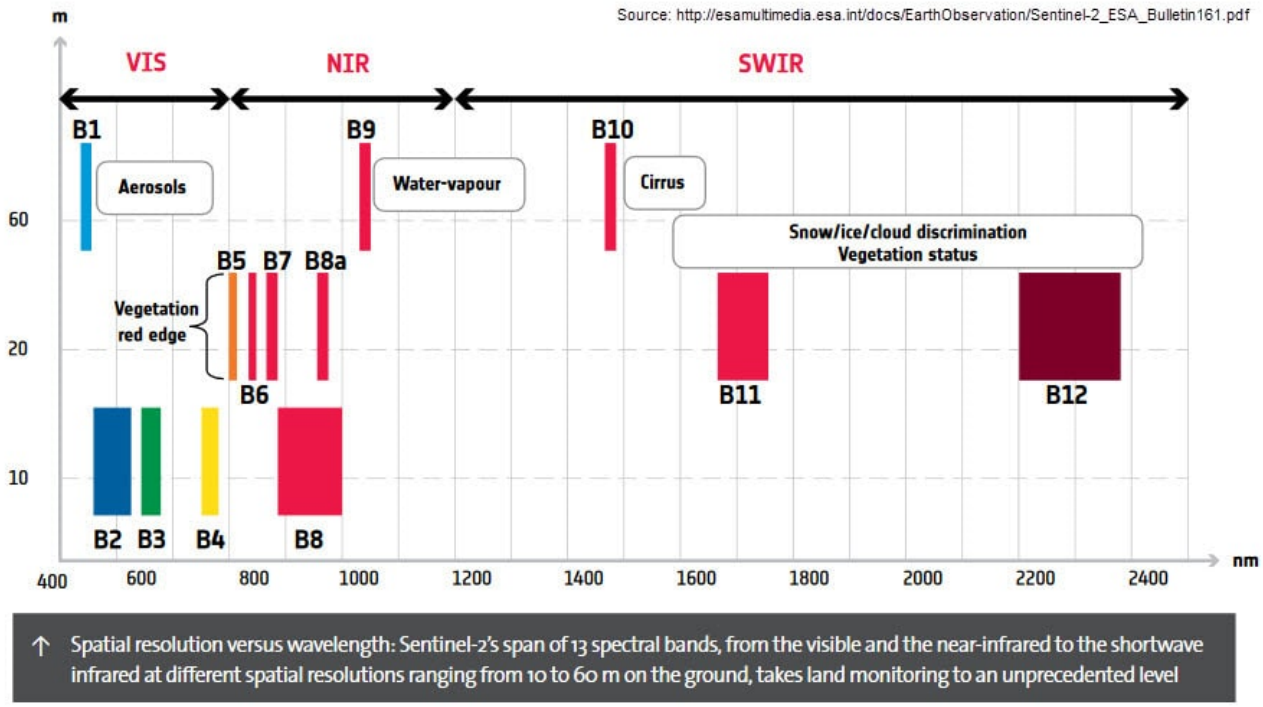


Figure 24 – Sentinel-2 spectral bands, bands at 10 to 20m resolution are used: B2, B3, B4, B5, B6, B7, B8, B8a, B11 and B12

Band No	Band Name	Spectral Range [µm]
B1	Blue	440 – 510
B2	Green	520 – 590
B3	Red	630 – 685
B4	Red-Edge	690 – 730
B5	Near Infra-red	760 – 850

Figure 25 – PlanetScope spectral bands, bands B1, B2, B3 and B4, are used

B Removed trees in Nancy area

Species	Number of trees	Number of removed trees S2	Percentage of removed trees S2 (%)
Acer campestre	903	28	3.1
Acer platanoides	4789	294	6.1
Acer pseudoplatanus	2679	348	**13.0**
Aesculus hippocastanum	651	11	1.7
Alnus glutinosa	285	0	0
Alnus x spaethii	215	6	2.8
Betula pendula	889	28	3.1
Carpinus betulus	2009	170	8.5
Fraxinus excelsior	2057	19	0.9
Pinus nigra	519	7	1.3
Platanus x acerifolia	3360	35	1.0
Populus nigra	630	6	1.0
Prunus avium	993	2	0.2
Pyrus calleryana	908	6	0.7
Quercus robur	439	8	1.8
Robinia pseudoacacia	1491	42	2.8
Taxus baccata	60	0	0
Tilia cordata	1513	75	5.0
Tilia x euchlora	2272	279	**12.3**
Total	26662	1364	5.1

Table 9 – Removed trees per species for S2

Species	Number of trees	Number of removed trees PS	Percentage of removed trees PS (%)
Acer campestre	903	16	1.8
Acer platanoides	4789	52	1.1
Acer pseudoplatanus	2679	6	0.2
Aesculus hippocastanum	651	1	0.2
Alnus glutinosa	285	0	0
Alnus x spaethii	215	0	0
Betula pendula	889	4	0.4
Carpinus betulus	2009	34	1.7
Fraxinus excelsior	2057	79	3.8
Pinus nigra	519	0	0
Platanus x acerifolia	3360	25	0.7
Populus nigra	630	2	0.3
Prunus avium	993	8	0.8
Pyrus calleryana	908	5	0.6
Quercus robur	439	7	1.6
Robinia pseudoacacia	1491	29	1.9
Taxus baccata	60	0	0
Tilia cordata	1513	9	0.6
Tilia x euchlora	2272	0	0
Total	26662	277	1.0

Table 10 – Removed trees per species for PS

C Loss plot for the Hybrid model

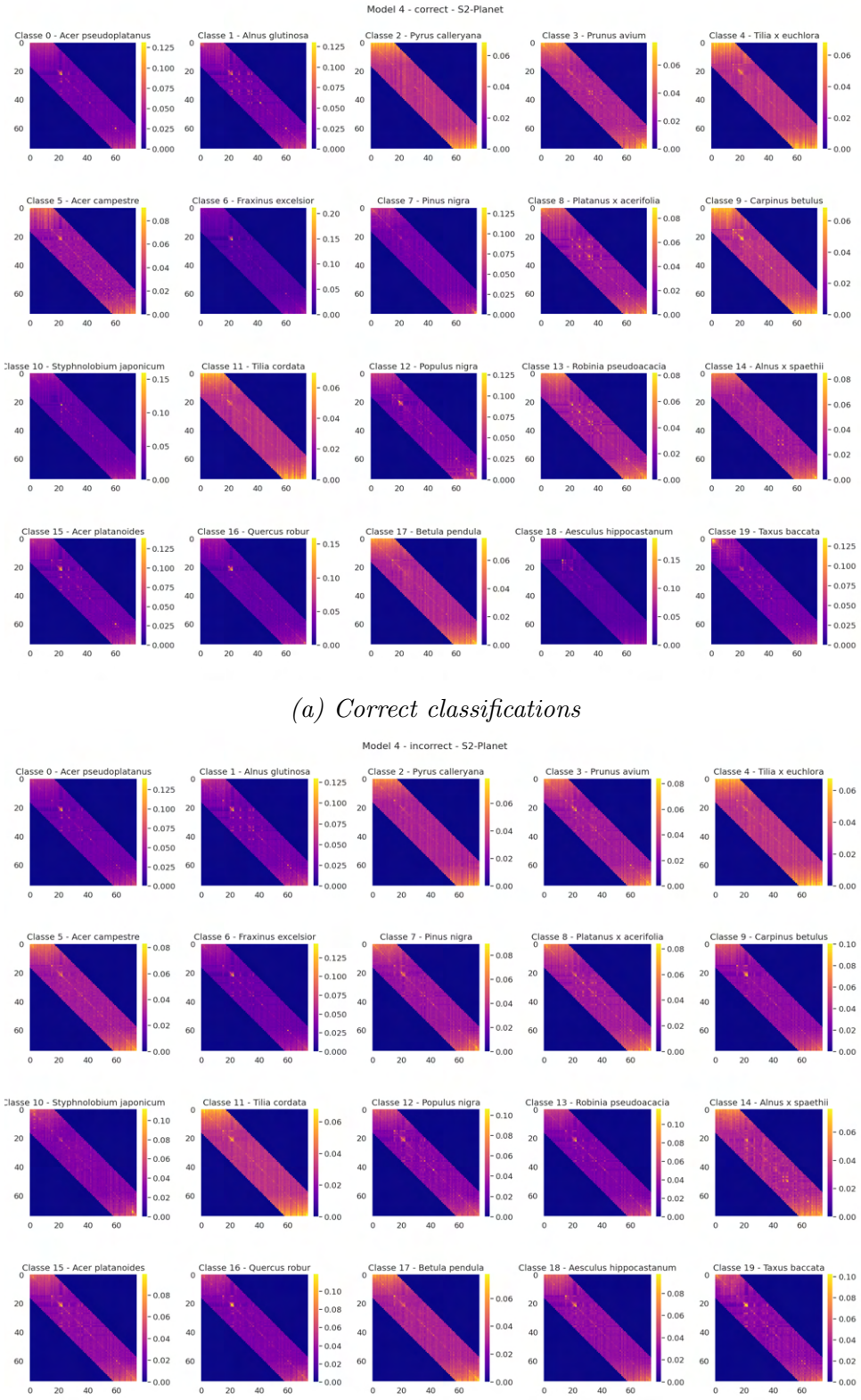
Figure 26 – Training and validation losses over epochs for Hybrid model with both sensors

D Computer programs

All the developed programs are available on GitHub to this adress:

https://github.com/latilmarie/Projet-fin-etudes

E Attention maps

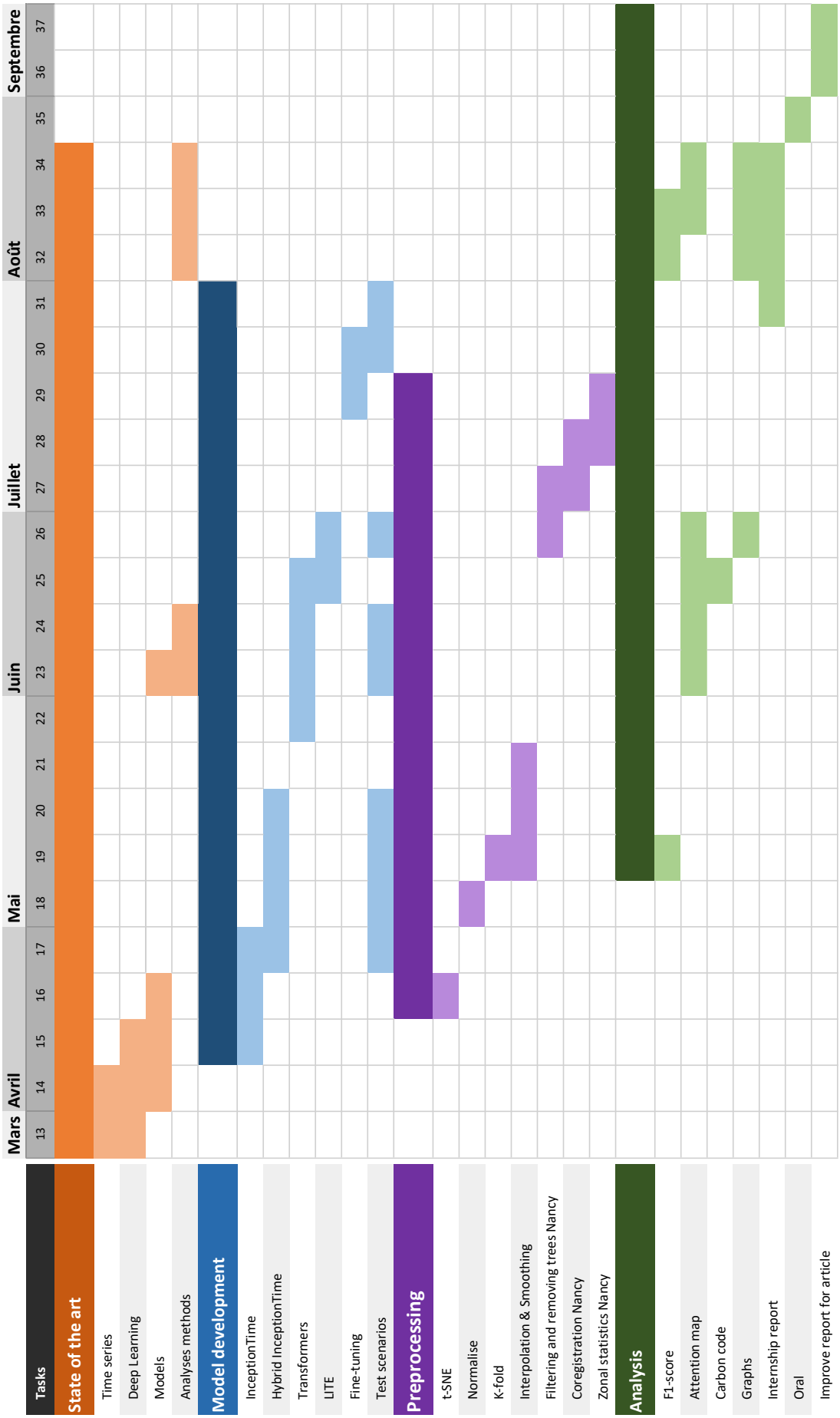


(a) Correct classifications

(b) Wrong classifications

Figure 27 – Attention maps per species for correct and wrong classifications, attention size = 16 and dmodel = 64

F Gantt diagram



G Fiche archive (C)

1. Nom du stagiaire : Marie Latil

2. Filière : SICOM (Signal, Image, Communication and Multimedia)

3. Année universitaire : 2023-2024

4. Titre du stage : Deep learning for urban tree species classification.
   Période : du 25 mars au 13 September 2024

5. Structure d'accueil : LIVE UMR 7362 (Laboratoire Image, Ville et Environnement)
   Adresse : 3 Rue de l'Argonne - 67000 Strasbourg - FRANCE

Figure 28 – Logo du LIVE

6. Responsables du stage :
   Anne Puissant, Professeur : anne.puissant@live-cnrs.unistra.fr

7. Tuteur école :
   Dawood Al Chanti, Professeur associé : dawood.al-chanti@grenoble-inp.org

8. Descriptif du stage validé par le CRE :
   L'étude de la végétation urbaine est cruciale pour la gestion durable des espaces verts, la régulation du climat urbain, la biodiversité, et le bien-être des habitants. Ce stage de recherche vise à utiliser des méthodes d'apprentissage profond (deep learning) en combinant différents jeux de données satellitaires (Sentinel-2, Planet, etc) pour produire des cartographies des essences d'arbres sur toute la zone urbaine (Strasbourg). Il s'agit donc d'un stage en lien direct avec la filière SICOM, dont les enjeux environnementaux raisonnent avec mon projet professionnel ainsi que les attentes d'un futur ingénieur provenant de l'ENSE3.

9. Moyens mis à disposition par le laboratoire :
   Bureau, ordinateur fixe Linux, cluster à distance pour tourner en GPU, disque dur avec les données satellitaires sur Nancy.
   Encadrement présent : Post-doctorant (Romain Wenger), Professeur (Germain Forestier), Informaticien (Marc Fleck).

Deep learning for urban tree species classification

Marie LATIL

Résumé

Dans un contexte de changement climatique, les villes se réchauffent à grande vitesse, alors que la végétation et en particulier les arbres sont capables de contrer ces effets négatifs en tant qu'acteurs de services écosystémiques. L'utilisation des images satellitaires pour produire des séries temporelles est très largement répandue pour des tâches de classification. Certains capteurs envoient hebdomadairement des images gratuites de haut résolution spatiale pouvant être combinées avec des méthodes d'apprentissage profond pour de la classification d'essences d'abres. Cette étude teste des modèle d'apprentissage profond à l'état-de-l'art pour de la classification sur 20 essences d'arbre urbains sur la ville de Strasbourg. En créant une fusion des 2 capteurs Sentinel-2 and PlanetScope, les nouveaux modèles développés prennent en compte des sources de données multiples et multitemporelles. Le modèle Hybrid InceptionTime génère la plus haute accuracy avec 69% d'arbres correctement classifiés. La généralisation de ce modèle appliqué par fine-tuning sur une nouvelle source de données à Nancy produit des résultats prometteurs avec 56% de classifications correctes.

Mots-clés : essences d'arbres, séries temporelles, apprentissage profond, Sentinel-2, PlanetScope

Abstract

In the context of climate change, urban areas are warming up, while vegetation and especially trees are able to balance the negative effects, acting as ecosystem service providers. Using satellite images to provide time series is widely spread for classification tasks. Some sensors furnish free weekly high spatial resolution images and can be combined with deep-learning methods to classify tree species. This study tests state-of-the-art deep-learning models on 20 urban tree species classification in Strasbourg. By creating a fusion of the 2 sensors Sentinel-2 and PlanetScope, the newly developed models are taking into account multi-source and multi-temporal data sources. The Hybrid InceptionTime model generates the highest accuracy with 69% of correct tree classifications. The generalisation of this model applied with fine-tuning on a new dataset in Nancy is providing promising results with 56% of correct classifications.

Keywords: tree species, time-series, deep learning, Sentinel-2, PlanetScope