KU LEUVEN

Integration of spatial-temporal context in remote sensing image classification

Zhiqi Wang

Thesis submitted for the degree of Master of Science in Artificial Intelligence, option Engineering and Computer Science

Thesis supervisor:
Prof. Stef Lhermitte

Assessor:
Dr. Stien Heremans
Prof. dr. Matthew Blaschko

Academic year 2022-2023

Table of Contents

List of figures

List of tables

Abbreviations

ML	Machine learning
DL	Deep learning
RF	Random forest
LSTM	Long-short time memory
ConvLSTM	Convolutional long-short time memory
NDVI	Normalized Difference Vegetation Index
CNN	Convolutional neural network
RNN	Recurrent Neural Networks
RS	Remote sensing
DEM	Digital elevational model
RMSE	Root mean square error
MSE	Mean square error
MAE	Mean absolute error

Abstract

In recent years, with the development of remote sensing technology, more real-time and rich data sources have been provided for monitoring vegetation dynamics and abundance over a wide range and over a long period of time. However, the interpretation of remote sensing data is challenging due to the complexity and variability of the natural environment. Advanced models such as machine learning can automatically learn complex patterns and characteristics from large amounts of data, providing more accurate and general prediction results, and are therefore ideal for monitoring vegetation dynamics that are sensitive to climate and the surrounding environment.

This study presents a comparison of the performance of four machine learning models (Random Forest, LSTM, ConvLSTM, and Transformer) for predicting Normalized Difference Vegetation Index (NDVI) values using remote sensing images. The intention is to investigate opportunities for spatial-temporal learning in remote sensing data analysis using machine learning techniques. Models were trained on data from a range of European ecosystems and land cover types, derived from the EarthNet2021 challenge platform. The initial evaluation focused on the ability of each model to capture temporal information, assessing their predictive ability across various time steps. This was followed by an investigation into how the models processed spatial-temporal information when expanded from a center pixel to a 10x10 pixel area. The aim was to ascertain whether an increase in spatial coverage could effectively enhance predictive ability.

Results revealed that the models exhibited similar performance in single-step time series forecasting, with the Random Forest model delivering the quickest and simplest results (R^2 = 0.888, training time = 264 seconds). However, all four models demonstrated instability in the multi-step recursive forecasting of time series. Compared with the true NDVI value, the Random Forest model showed a notable decrease in prediction accuracy beyond the second step, with the Mean Absolute Error (MAE) at the first step increasing by 250% relative to the initial step. For the rest of the three models, the MAE of the predicted NDVI value increased by approximately 0.02 at each step for all models.

By contrast, when observation data was broadened to encompass both time series and adjacent spatial information, the ConvLSTM and Transformer models which are good at dealing with spatial-temporal relations displayed significant performance improvement. They achieved high accuracy in multi-step recursive prediction (MAE <0.075, RMSE <0.125 at the 10th step), thereby underlining their effectiveness in capturing temporal and spatial dependencies in NDVI data.

These findings highlight the intricacy of NDVI predictions and underline the strong correlation between machine learning methods and the small-scale prediction of NDVI spatial-temporal relationships. The disparity in the performance of different models emphasizes the necessity of selecting an appropriate model to fulfill the specific requirements of the prediction task. For single-step predictions or tasks where temporal dependencies are less significant, traditional machine learning models like Random Forest may be optimal. Conversely, LSTM models may be suitable for immediate prediction tasks, while ConvLSTM or Transformer models appear more apt for long-term spatial-temporal predictions.

Notably, the Transformer model exhibited robustness comparable to the classic spatial-temporal model ConvLSTM in the NDVI prediction task, highlighting its potential for environmental applications. This reaffirms that for spatial-temporal data, the choice of machine learning models can significantly influence prediction accuracy. Comprehensive NDVI forecasts using these models can thus facilitate effective future vegetation dynamics forecasts, contributing to precision in forestry and agricultural planning.

Keywords: *remote sensing; spatial-temporal data; machine learning model; NDVI*

1. Introduction

The development of remote sensing has ushered in a new era for Earth observation, characterized by an unprecedented wealth of real-time, highly accurate, and expansive spatial-temporal data. However, inherent issues such as atmospheric effects, including cloud occlusion and light refraction, invariably result in substantial amounts of missing data, thereby affecting the long-term sustained observations and research of the land surface (Schowengerdt, 2006). This study posits that the introduction of advanced machine learning techniques, coupled with the integration of relevant weather and geographical variables, may yield significantly more accurate results for land surface monitoring such as vegetation dynamic observation and land classification that depend on time correlation and are sensitive to the surrounding environment.

Remote sensing products like Landsat (Wulder et al., 2016) and Sentinel-2 (Drusch et al., 2012) present an ideal resource for models designed to analyze regional effects of climate change, they present extensive utility in the field of vegetation dynamic monitoring due to their broad data coverage and the capacity to record a variety of climate data in real-time. Although they provide a rich data source for monitoring vegetation health and abundance over large areas and time periods, interpretation of remote sensing data can be challenging due to the complexity and variability of natural environments. Since the distribution of vegetation is highly sensitive to climatic conditions and the surrounding environment, being predominantly influenced by factors such as light, temperature, and water, the ability to predict vegetation dynamics while integrating climate and topography factors holds immense potential (Kladny et al., 2022). One promising strategy is to predict the evolution of land surfaces involves a fusion of cues from dense satellite time-series imagery and weather data. This approach acknowledges the fact that land surface observations are influenced by multiple local factors such as climate, topography, soil, etc. Further enhancement of the spatial resolution of satellite imagery may be achieved by incorporating additional inputs, thereby potentially augmenting the accuracy of these forecasts (Diaconu, Saha, Gunnemann, et al., 2022).

The Normalized Difference Vegetation Index (NDVI) was used in the study as an indicator to monitor vegetation dynamics and it was combined with relevant local weather (such as temperature and rainfall) and topography to predict future NDVI values, aiming to monitor the changes in local vegetation dynamics over time (Ahmad et al., 2023). Central to this study is the NDVI, it is the most used vegetation index for remote sensing imagery and serves as an indicator of spatial-temporal image classification. NDVI not only allows for the distinction of vegetation types but also enables the tracking of vegetation coverage changes over time, emphasizing its critical role in the spatial-temporal image classification effect (OpenAI, 2023). In addition, the data can be easily accessed through a variety of open remote sensing products, such as sentinel-2 and MODIS, which facilitate the collection and processing of large amounts of remote sensing data.

In recent years, ML methods have enjoyed considerable success in environmental fields. They benefit from the automatic learning of complex patterns as well as large amounts of data and have already been evident in applications of weather forecasting and vegetation dynamic monitoring. Such as Artificial neural networks (ANN) which are good at nonlinear processes and recurrent neural networks (RNN) which are good at processing time series (Xie et al., 2008). Nevertheless, current ML methods predominantly adopt a pixel-by-pixel analysis approach, neglecting the spatial context and thereby imposing limitations on the analysis of large-scale satellite images (Ahmad et al., 2023). Given that vegetation dynamics of interest in this study are easily influenced by the surrounding environment and require a time period to adapt to environmental changes, it becomes imperative to consider the spatial-temporal dependence of vegetation systems (Akbar et al., 2019). Thus, this study seeks to explore the potential of spatial-temporal learning in remote sensing data analysis using machine learning (ML) methods, with the aim of identifying predictive models of broader applicability. Through careful evaluation, four models of random forest, Long Short-Term Memory (LSTM), Convolutional

LSTM (ConvLSTM), and Transformer were selected to carry out the research of this study.

For prediction in the time series, the widely used Random Forest model and the Long Short-Term Memory (LSTM) model are both renowned for their superior performance in time series prediction. Random Forest was chosen as a nonlinear algorithm due to its excellent computational scalability for very large datasets but can be easily substituted by other nonlinear machine learning techniques, such as neural networks or kernel methods (Breiman, 2001). For spatial-temporal data prediction, the Convolutional LSTM (ConvLSTM) model, a powerful deep learning model, is well-regarded for its ability to learn effective spatial-temporal features, thus being capable of encoding spatial information efficiently while retaining the ability to process temporal data (Reddy & Prasad, 2018) (SHI et al., 2015b). The Transformer model addresses these inefficiencies by relying on a mechanism called attention, which allows the model to focus on different parts of the input sequence and aim to deal with long-range dependencies more efficiently. Its parallel processing power speeds up the training process and makes the model scalable to large datasets (Bello et al., 2019). Despite the theoretical promise, the current application of Transformers in remote sensing is limited, and more research is needed to fully understand.

Throughout the study, a comprehensive review of the relevant literature was conducted, outlining the state-of-the-art and identifying gaps that the study aims to address. First, the study reviewed the literature on remote sensing, NDVI, and their applications in agriculture, forestry, land cover mapping, and ecosystem monitoring to help gain a deeper understanding of the importance of NDVI as a vegetation index and the potential applications of this study. And further emphasize the importance of NDVI observations and how they serve as an indispensable tool for monitoring and predicting vegetation dynamics. When it comes to machine learning, the literature on deep learning methods and their applications in remote sensing were reviewed. A comparative analysis of Random Forest, LSTM, ConvLSTM, and Transformer was conducted to clarify their respective advantages and disadvantages in processing spatial-temporal data. As well as their applicability to different types of predictions and the potential benefits of using these models for NDVI prediction. In addition, remote sensing data and related climate data for training the model were collected and data pre-processed. After this, the models are built and trained, and the ability of each model to capture temporal and spatial information, as well as predictive power at different time steps, is evaluated and compared.

In conclusion, this study highlights the importance of NDVI observations, the application of traditional and advanced machine learning models, the use of spatial-temporal observation models, and the choice of specific models to address the complexities of environmental studies. We are eager to provide a nuanced understanding of how these aspects contribute to the robustness of prediction models in the context of remote sensing and climate change, hopefully providing new insights into the use of remote sensing data and machine learning models in environmental research.

2. Literature review

2.1 Remote sensing & NDVI

The assessment and understanding of vegetation dynamics are of utmost importance for a plethora of applications from agricultural management to climate change research (Fensholt & Proud, 2012a). Vegetation distribution is inherently influenced by factors such as sunlight, temperature, and moisture availability, making it highly sensitive to climate and environmental conditions. Over recent years, remote sensing techniques have gained significant popularity in the realm of large-scale vegetation monitoring, offering expansive coverage and the capability to record diverse climatic data in real-time (Mulla, 2013).

$$NDVI = \frac{NIR - Red}{NIR + Red} \quad (1)$$

The Normalized Difference Vegetation Index (NDVI) has been widely accepted as an efficient index for quantifying vegetation cover and dynamics at extensive spatial and temporal scales (Pettorelli et al., 2005). This acceptance can be attributed to its ability not only to discriminate vegetation types but also to monitor changes in vegetation coverage over time. Besides, the data can easily be accessed through multiple remote sensing products, like sentinel-2 and MODIS. NDVI values across multiple spectral bands and time periods and can reflect the proportion of solar radiation absorbed by plants in photosynthesis. It serves as an indicator of greenness and photosynthetic activity in vegetation and can be easily calculated at red and infrared (NIR) wavelengths shown in Equation (1), with values typically ranging between -1 and +1.

NDVI can be roughly divided into water bodies, buildings, bare land, grasslands, sparse vegetation, and dense vegetation, and the specific ranges of each category can be referred to in the table below (Table 1) (Akbar et al., 2019). Higher values signify denser and healthier vegetation (Tucker, 1979), for example, forest areas generally have higher NDVI values than agricultural or urban areas, while wetlands and shrublands have intermediate NDVI values. NDVI has demonstrated its utility in various areas, however, considering the delay in vegetation's response to environmental shifts due to ecosystem resilience and resistance characteristics, NDVI data often need to be combined with other remote sensing and ground-based data sources. This combination enhances the comprehensiveness and accuracy of predictions, enabling researchers to gain a broader and real-time understanding of temporal and spatial changes in vegetation cover and ecosystems.

Table 1 - NDVI classification

Class	***NDVI Range***
Water	-0.28 – 0.015
Build-up	0.015 – 0.14
Barren Land	0.14 – 0.18
Shrub and Vegetation	0.18 – 0.27
Sparse Vegetation	0.27 – 0.36
Dense Vegetation	0.36 – 0.74

Rainfall serves as a key determinant of vegetation growth and health by influencing vegetation type, distribution, and scale (Knapp et al., 2015). Hence, alterations in rainfall patterns directly or indirectly impact vegetation dynamics. Temperature, on the other hand,

regulates the physiological and biochemical processes integral to vegetation growth, establishing it as another crucial factor in vegetation dynamics studies (Chuine, 2000). Several studies have documented a significant influence of temperature on NDVI (Zhang, Jiang, et al., 2022) (Marzban et al., 2018). In another hand, the Digital Elevation Model (DEM) provides topographic information that can help understand the vegetation patterns due to its impact on microclimate and soil conditions (Zhang, He, et al., 2022). As a result, incorporating NDVI, local rainfall, temperature, and DEM data can help create more accurate vegetation forecasts.

The versatility of NDVI is evident in its wide range of applications, from agriculture and forestry to land cover mapping and ecosystem monitoring. For instance, in agriculture, where it can offer valuable insights into future farming planning and yield estimation, thereby enhancing agricultural productivity and sustainability (Insua et al., 2019). Similarly, in forestry, NDVI can help in assessing forest health, thereby aiding in forest management decisions (Prăvălie et al., 2022) (Maselli, 2004). In the context of land cover mapping, NDVI plays a significant role in differentiating between various land cover types, facilitating more accurate land use classifications (Xie et al., 2008). Ecosystem monitoring, particularly in the context of climate change, is another area where NDVI has proved invaluable. By tracking changes in vegetation health and coverage, NDVI can help in detecting signs of ecosystem disturbance and degradation (Pettorelli et al., 2005) (Fensholt & Proud, 2012b).

In conclusion, NDVI, with its wide-ranging applications and proven effectiveness, has established itself as an essential tool for vegetation monitoring and dynamics studies. By integrating NDVI with other critical parameters like rainfall, temperature, and DEM, we can significantly enhance the accuracy and comprehensiveness of vegetation forecasts, leading to improved agricultural practices, more informed land use decisions, and a deeper understanding of our changing ecosystems. This literature review validates the selection of these four input variables and lays a solid foundation for the subsequent analysis.

2.2 Machine Learning Approaches for Predicting Vegetation Dynamics

The advent of machine learning (ML) has ushered in a new era of data analysis, enabling the automated learning of complex patterns and features from large datasets. These advanced technologies have been increasingly utilized in various domains, and their applications in environmental science have been rapidly expanding (He et al., 2016). Given the inherently nonlinear relationship between climate and vegetation, ML algorithms present a promising avenue for predicting vegetation dynamics. For example, random forests have been used for land cover classification and vegetation health prediction (Belgiu & Drăguţ, 2016). LSTM networks have been used to forecast time series data, including NDVI (Reddy & Prasad, 2018). ConvLSTM is a variant of LSTM designed for spatial data and has also been used for environmental applications (SHI et al., 2015a). Transformer models, although more commonly associated with natural language processing, have recently been used for time series forecasting in the environmental domain (Vaswani et al., 2017). Therefore, this section will focus on the literature related to the four chosen ML models: Random Forest, LSTM, ConvLSTM, and Transformer, and will mainly explore their potential in predicting vegetation dynamics.

2.2.1 RF

Random Forest (RF) is a supervised learning algorithm known for its ability to handle high-dimensional data and nonlinear relationships between features and target (Belgiu & Drăguţ, 2016), and is widely known for its robustness and simplicity (Breiman, 2001). It works by forming a set of decision trees, each of which votes for the output of the most common class, thereby minimizing prediction error (Breiman, 2001). Random forests have demonstrated proficiency in a range of application domains and are frequently used in Earth observation

applications, including land cover classification and vegetation health prediction (Rodriguez-Galiano et al., 2012). Its robustness to overfitting and noise makes it an attractive model for remote sensing applications. However, despite RF's advantages, it can have difficulty capturing temporal dependencies in the data and may require a large amount of training data for optimal performance.

2.2.2 LSTM

While traditional ML methods have been successful in fitting observations, they often fail to capture temporal dependencies or spatial relationships. To overcome this problem, this study considers deep learning (DL) models that can learn hierarchical representations of data that capture spatial and temporal dependencies (Goodfellow et al., 2016). Among DL models, LSTMs are a variant of Recurrent Neural Networks (RNNs) designed to solve the vanishing gradient problem by retaining memory cells that can store information over time, popular for their ability to handle long-term dependencies in time series data (Hochreiter & Schmidhuber, 1997). LSTMs have demonstrated their effectiveness in a variety of applications, including remote sensing data analysis (Bermúdez et al., 2017), and have also been successfully used to predict time series data, including NDVI (Reddy & Prasad, 2018). However, despite their impressive capabilities, LSTMs can require large amounts of training data to achieve optimal performance, and the training process can be computationally expensive.

2.2.3 ConvLSTM

Despite the success of LSTM in handling temporal dependencies, it lacks the ability to effectively capture spatial features in image data. Likewise, while convolutional neural networks (CNNs) excel at image classification tasks by capturing the complex relationships between pixels and their surroundings, they do not handle temporal dependencies well. ConvLSTM is a fusion of LSTM and CNN that overcomes these limitations by simultaneously learning the spatial and temporal features of the data. This makes it ideal for analyzing remote sensing data, which often contain spatial and temporal information (SHI et al., 2015a). ConvLSTM has found many environmental applications, including the fields of remote sensing time series analysis and land cover classification (Diaconu, Saha, Gunnemann, et al., 2022). The research by Shi et al. (2015) first used ConvLSTM for precipitation nowcasting, and its application to remote sensing time series analysis has appeared in many works. For example, Yuan et al. (2020) adopted ConvLSTM for semi-supervised time series land cover classification, successfully capturing spatial-temporal features.

Despite its advantages, ConvLSTMs also have some limitations. For example, in a study by Moskolaï et al. (2020), the performance of ConvLSTM in predicting the next frame of Sentinel-1 time series data decreases as the sequence length increases. And like other LSTM networks, it requires a lot of training data, and the training is computationally intensive. However, with its ability to learn hierarchical spatial and temporal representations, ConvLSTM remains a strong candidate for predicting changes in NDVI.

2.2.4 Transformer

Although ConvLSTM has been shown to effectively capture temporal and spatial dependencies in data, it suffers from computational inefficiency. The Transformer model, introduced by Vaswani et al. (2017), addresses these inefficiencies by relying on a mechanism called attention. This mechanism allows the model to focus on different parts of the input sequence when making predictions, allowing the model to deal with long-range dependencies more efficiently. Its parallel processing power speeds up the training process and makes the model scalable to large datasets (Bello et al., 2019). Although Transformer models were originally developed for natural language processing tasks and have been widely used in this

field, their application to time series forecasting and remote sensing data is an emerging field of research (Bello et al., 2019), including in the field of environment (Vaswani et al., 2017).

Transformer models process all elements of a sequence simultaneously, which improves their ability to learn complex patterns in data, even with relatively short time steps. This parallel processing also speeds up the training process and makes Transformers more scalable for large datasets, a common feature of remote sensing data. However, despite their theoretical promise, the current evidence base for the application of Transformers in remote sensing is limited, and more research is needed to fully understand the strengths and limitations of Transformers in this regard. In addition, due to the complexity of the model, Transformer requires a lot of computing resources for training, which may not be applicable in all scenarios (OpenAI, 2023).

In summary, the chosen models Random Forest, LSTM, ConvLSTM, and Transformer, each bring unique strengths to the task of predicting NDVI. Random Forest's ability to capture nonlinear relationships and handle high-dimensional data, LSTM and ConvLSTM's handling of temporal and spatial dependencies, and Transformer's computational efficiency and attention mechanism make them suitable choices for this study. Taking advantage of these strengths, this study aims to develop an efficient and robust system to predict vegetation dynamics.

2.3 Data sources and preprocessing

Data preprocessing is an important step in any remote sensing analysis, as it greatly affects the accuracy and validity of subsequent analysis. The data used in this study is from the EarthNet2021x dataset from the platform Earthnet challenge (Requena-Mesa et al., 2021). The EarthNet2021x dataset is a large-scale dataset of multispectral remote sensing imageries and corresponding ground truth labels covering various regions of the world and contains approximately 32,000 samples. This includes high-resolution satellite imagery (RGB, near-infrared and short-wave infrared) from multiple sensors, including Landsat-8, Sentinel-2, and MODIS, and combines with related environmental data from the gridded observational dataset (E-OBS) from multiple locations around Europe (Copernicus Climate Change Service, 2020), such as precipitation, sea-level pressure, and temperature (minimum, maximum, and average) (Kladny et al., 2022). Besides, this dataset also contains the digital elevation model (DEM) to help provide corresponding local geographic information.

This dataset also did the preprocessing, which not only saves a lot of time compared to analyzing open satellite data but also results in more standardized high-quality satellite images (Requena-Mesa et al., 2021). The preprocessing includes 1) Image registration, which is used to register satellite imagery to a common coordinate system, ensuring images are spatially comparable. 2) Removal clouds and shadows, because clouds and shadows can significantly affect the quality and accuracy of remote sensing data. 3) Image selection, the selection is based on criteria such as clouds and shadows, image resolution, etc., aims to select the best quality image for each location and period. 4) Feature extraction, extract weather variables like temperature, and precipitation from E-OBS (in situ datasets) and integrate them with imagery (OpenAI, 2023) (Diaconu, Saha, Günnemann, et al., 2022).

Overall, the EarthNet2021x dataset has been heavily preprocessed to ensure a high-quality dataset to aid machine learning and remote sensing applications. The figure below (Figure 1) shows the quality of the preprocessed EarthNet2021x dataset. A sample randomly selected from the data shows RGB satellite images (left) and NDVI images (right) of the same date and location.

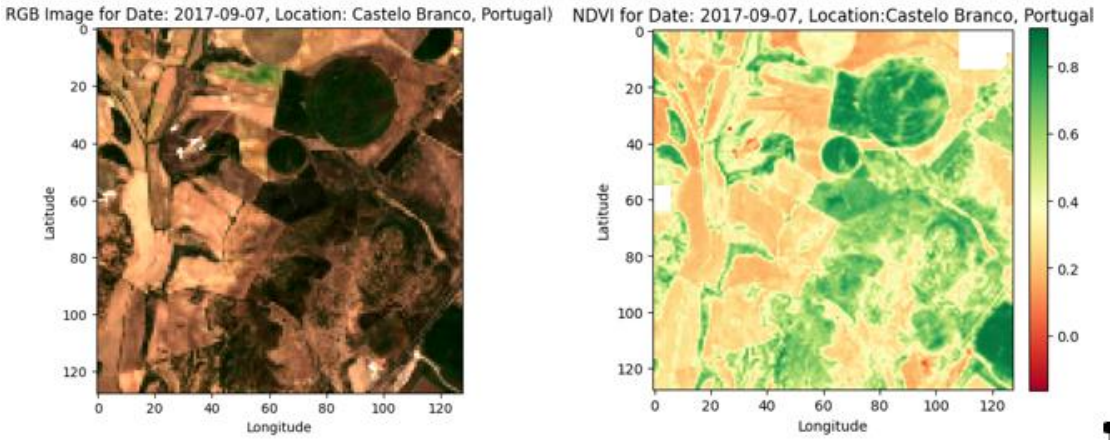

Figure 1 - The RBG image (left) and NDVI image (right) for a sample from the Earthnet2021x dataset

3. Material and methods

3.1 Methodology

This study aims to compare the performance of various Machine Learning (ML) models Random Forest, LSTM (Long Short-Term Memory), ConvLSTM (Convolutional LSTM), and Transformer in forecasting time series values derived from remote sensing images. The central hypothesis posits that certain models are better suited for accurately predicting the Normalized Difference Vegetation Index (NDVI), a critical indicator in agriculture, forestry, and land use.

We adopt an exploratory research approach, aiming to identify the most effective ML models for NDVI prediction and understand how these models process temporal and spatial information. The use of time series forecasting in this study is driven by the inherent temporal aspect of the NDVI and weather data. The methodology employs a sliding window method to capture temporal dependencies and applies two types of forecasting techniques: single-step and recursive multi-step forecasting. Single-step forecasting helps to understand the basic time series forecasting capabilities of each ML model. Multi-step recursive forecasting requires the model to predict multiple steps in the future based on the forecast data, so the complexity of forecasting is greatly increased. Multi-step forecasting tests the model's ability to handle temporal and spatial dependencies in forecasting scenarios, which are very challenging and more similar to real-world scenarios.

Based on this, the research was then divided into four experiments (Figure 2), each examining model performance under different forecasting tasks, making use of both temporal and spatial data. The study's importance lies in its potential to provide insights into efficient ML model selection for NDVI forecasting, contributing to the broader research area of efficient land use management and remote sensing.

Experiment 1 (S1): Single-Pixel Time Series Forecasting

This experiment constituted the baseline for this study. By focusing on a single (center) pixel and predicting NDVI using time series data, we aim to see how each model performs when given a simple single pixel time series. It helps to understand the basic time series forecasting capabilities and efficiency of each ML model and lays the foundation for more complex tasks. The central pixel is extracted from a sequence of satellite images, and the time series data of NDVI, rainfall, and average temperature for this pixel serve as observed variables, the models are trained to predict the NDVI at the next timestep (t).

Experiment 2 (E2): Spatial-temporal Single-Pixel Forecasting

This experiment extends the range of observations from the center pixel to a 10x10 pixel area adjacent to the center pixel and predicts the NDVI of the center pixel on the time series as in E1. By increasing the complexity dimension of the data by including spatial data, we tested the model's ability to handle the additional complexity and analyzed how the introduction of spatial dependencies affected predictive performance. The observation variables include spatial-temporal NDVI and DEM, along with time series data of rainfall and temperature. Only ConvLSTM and Transformer models are used here because of their inherent ability to handle spatial-temporal data, and it is assumed that both should outperform models that cannot handle spatial dependencies.

Experiment 3 (E3): Averaged Spatial-temporal Forecasting

This experiment aims to adapt spatial-temporal tasks to models (random forests and LSTMs)

that cannot directly handle such data. By averaging the spatial data, NDVI and DEM values were converted back into a time series format that these models can use. The goal of this experiment is to see if it is possible to preserve and exploit spatial dependencies, even in the format used to fit some models after simple processing such as averaging, and to compare their performance with ConvLSTM and Transformer models under similar conditions.

Experiment 4 (E4): Spatial-temporal Sequence Forecasting

This experiment pushes the model further by requiring the two spatial-temporal models, ConvLSTM and Transformer, to be trained on the same spatial-temporal data as in E2, and to predict spatial-temporal NDVI data sequences. This increases the complexity of the prediction task because it is expected that the model can capture the temporal and spatial dependencies between different pixels in the future sequence, which is a challenge to the model's predictive ability.

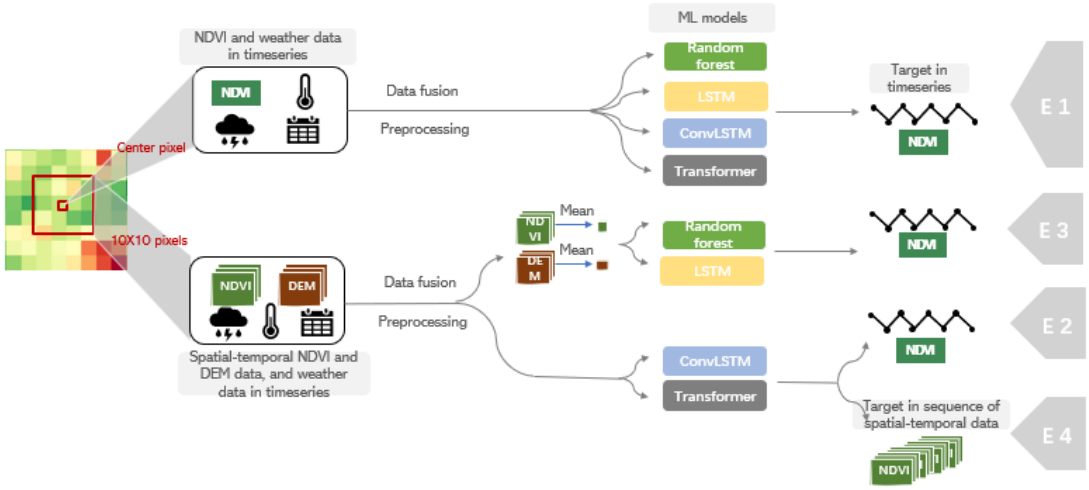

Figure 2 - The schematic illustration of the four experiments on examining different ML model's performance under different forecasting tasks

In all four experiments, single-step forecasting predicts the NDVI at the time (t) based on previous NDVI values and weather parameters (Figure 3). Recursive multi-step forecasting, on the other hand, shifts input data after each prediction, including the latest forecast as part of the subsequent input (Figure 4). For example, in a 5-step forecast, if the sliding window contains 20 elements, the input for the 5th step prediction includes the last 16 actual values and the 1st, 2nd, 3rd, and 4th predicted values. This method is applied to E1, E2, and E3 to assess the accuracy of the forecasting methods and the robustness of the models for multiple-step prediction.

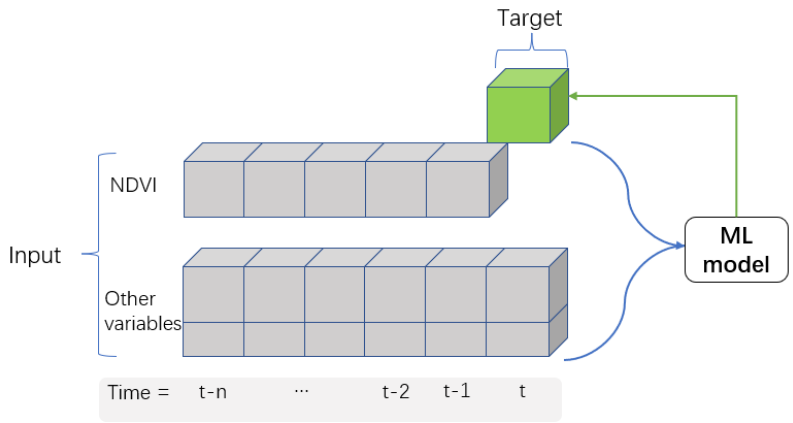

Figure 3 - The schematic illustration of the single-step forecasting for the NDVI through ML models

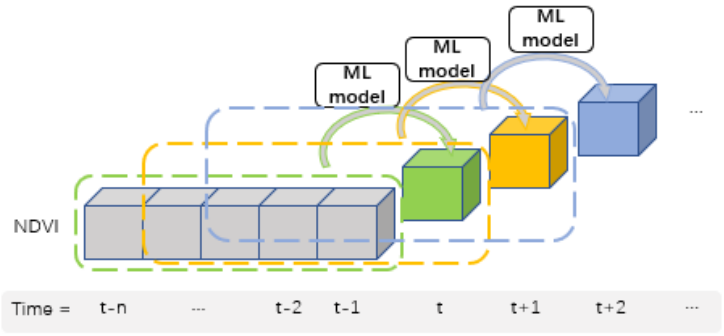

Figure 4 - The schematic illustration of the recursive multiple-steps forecasting for the NDVI through ML models

All the models in this study are developed and trained using Google Colab, a cloud-based platform that offers an interactive environment for machine learning and data science projects. This allows for the efficient execution of models without a substantial load on the local system. The study employs various software libraries and tools like Python-based ML frameworks to train and evaluate ML models, and scikit-learn for Random Forest, Keras for LSTM, ConvLSTM, and Transformer models. In addition, Pandas and NumPy libraries handle data manipulation tasks, including data cleaning, transformation, and analysis. Matplotlib and Seaborn libraries are used to visualize the performance and results of the models, providing valuable insights into the effectiveness of each ML model. The Random Forest model is built using the RandomForestRegressor class from the Scikit-learn library, a popular choice for high-level machine-learning tasks. TensorFlow's Keras library serves as the foundation for creating the LSTM, ConvLSTM, and Transformer models. Keras simplifies the process of building and training these models, offering modularity and flexibility. LSTM and ConvLSTM models are implemented using the Sequential model class and LSTM, ConvLSTM2D, and Dense layers from Keras. For the Transformer model, components from TensorFlow's keras library such as MultiHeadAttention and LayerNormalization are utilized.

Overall, this study will provide a robust comparison of the predictive performance of four ML models when forecasting NDVI values from remote sensing images, offering significant insights for agriculture and land use management.

3.2 Data collection and preprocessing

3.2.1 Data collection

The Earthnet2021X dataset served as the primary data source for this study. It is suitable for training deep-learning neural networks due to its extensive collection of samples. The dataset spans from 2016 to 2020 and contains roughly 32,000 samples (Requena-Mesa et al., 2021). The land coverage of this dataset includes but is not limited to various types of landforms, vegetation, and human settlements on the Earth's surface. Each sample consists of Sentinel-2 satellite imagery at 20-meter resolution, 128x128 pixel resolution, weather-related variables (such as precipitation, sea-level pressure, and temperature minimum, maximum, and average), and digital elevation models (DEM). The weather-related variables are sourced from the E-OBS dataset, which collects data from observation stations across Europe (Copernicus Climate Change Service, 2020). And this dataset separates each tile into smaller "minicubes", representing small spatial areas over specific time points in a three-dimensional manner (2D spatial + 1D temporal) (Diaconu, Saha, Günnemann, et al., 2022).

This dataset was selected due to its comprehensive preprocessing of the data, saving time and yielding standardized, high-quality satellite images. Preprocessing steps include image

registration for spatial comparability, cloud and shadow removal, and feature extraction of weather variables from E-OBS. After careful consideration, the Normalized Difference Vegetation Index (NDVI), rainfall, temperature, and DEM were selected as input variables, with NDVI data serving as output targets. NDVI was calculated from satellite imagery, combined with the climate variables and geographic information in the relevant location could significantly aid in predicting vegetation dynamics on the spatial-temporal scale.

Data access and extraction were performed using Colab. The Earthnet2021X dataset was downloaded from the Earthnet Challenge 2021 platform, the NDVI was then calculated based on the wavelength information contains in the dataset provided by the Sentinel-2 satellite products. Because the NDVI is recorded every 5 days, rainfall and temperature are daily data, so to simplify calculations, rainfall, and temperature data were aggregated into 5-daily data, thus, keeping all the needed features in the dataset synched as 30 timesteps.

The data extraction was carried out after the aimed dataset and all the features were available, it was divided into two parts according to the requirements of the experiments. The first dataset contains the center pixel of each satellite image that was extracted to obtain the NDVI, rainfall, and temperature values and date and then saved these features in individual tables. The second dataset contains the adjacent 10X10 pixels expended from the center pixel from each satellite image, provided NDVI, DEM, rainfall, and temperature values. This dataset introduced NDVI and DEM as three-dimensional variables (longitude, latitude, and time), with rainfall, temperature, and date remaining as time series, and saved these features in individual tables as well. The following table (Table 2) lists information on the collected data, which was ready for the next step preprocessing and then training and evaluating the models according to the four experiment needs that were proposed by this study.

Table 2 - The information of the collected input and output data

	Raw Data	***Features***	***Data***	***Timesteps***	***Resolution***	***Dimension***
Input	**Center pixel** (Length: 240810)	NDVI	5-Daily	30	20m X 20m	1D (time)
		Rainfall	5-Daily	30	In-situ	1D
		Temperature	5-Daily	30	In-situ	1D
	10X10 pixels (Length: 240810)	NDVI	5-Daily	30	20m X 20m	3D (Longitude, latitude, time)
		DEM	5-Daily	30	-	3D
		Rainfall	5-Daily	30	In-situ	1D
		Temperature	5-Daily	30	In-situ	1D
Output	**Center pixel**	NDVI	5-Daily	30	20m X 20m	1D/3D

3.2.2 Preprocessing

Data preprocessing is an important step in any remote sensing analysis, as it greatly affects the accuracy and validity of subsequent analysis. In this step, the feature fusion was first carried out, which aim to combination of various features to improve the accuracy of the predictive model, then the data was cleaned to prepare for the interpolation. Next, the interpolation is applied aim to fill in the missing data, and last, data was split into training, validation, and testing data, and then normalized for the training and evaluate the ML models.

For feature fusion, a table was created to merge all the extracted values of the features based

on the date, sample ID, and time step (30 timesteps). Thus, each row in the new merged table represented the feature values of one sample in a single time step.

Data cleaning was crucial to ensure data quality for training the ML models. In this study, samples with more than 30% missing NDVI values were removed. Also, samples with more than 10% NDVI values less than zero were excluded, as these values often indicated cloud cover or water bodies, which would affect model performance and introduce noise. And convert the NDVI values of the remaining samples containing NDVI less than 0 to NaN. The purpose is to keep the NDVI values of the data set within the range of 0-1, which is convenient for the model to focus on learning vegetation patterns.

Next, the Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) method was applied for filling the missing value of NDVI, rainfall, and temperature datasets. This is an advanced polynomial interpolation method, that can produce a smooth and continuous curve and provide more realistic estimates. It can handle unevenly spaced or non-monotonic data better than seasonal mean imputation, and it is computationally more efficient and stable than splines or polynomials (OpenAI, 2023). The figure below (Figure 5) shows the changes in the time series of the three features of NDVI, rainfall, and temperature before and after interpolation. The lack of data in NDVI is relatively common, which is a challenge that cannot be underestimated for the predictive ability of the models proposed in this study.

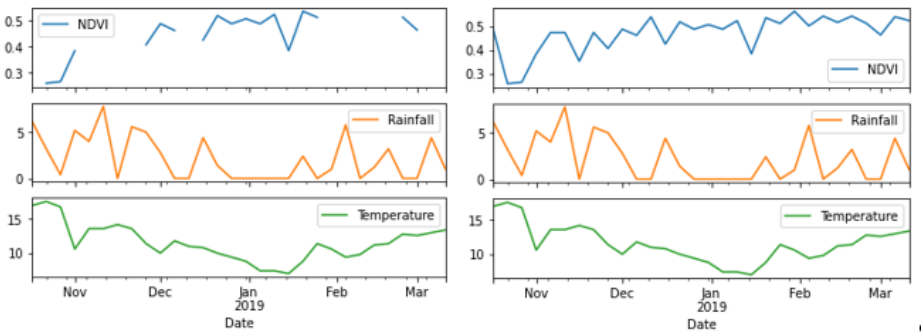

Figure 5 - Compare the NDVI, rainfall and temperature data in timeseries for a single sample before (left) and after (right) the interpolation

Lastly, the datasets were randomly split into 70% for training, 20% for validation, and 10% for testing to evaluate model performance. Splitting was performed on the sample ID basis to avoid partitioning the time steps of one sample into different sets. Next, the data were normalized using the Minmax Scaler, the scaler was first fitted on the training data and transformed the training data, then subsequently transformed the validation and testing data using the same scaler. The MinMax Scaler scaled and translated each feature individually to a value between 0 and 1. It works well on data that is not normally distributed and is particularly effective for image applications (OpenAI, 2023). The following tables (Table 3) summarize the procedures for the data collection and preprocessing for this study.

Table 3 - Data collection and preprocessing procedures

Process	*Description*
Data Collection	Data from Earthnet2021X, extract center pixel and 10X10 pixels from all minicubes separately as two datasets.
Feature Extraction	Extracted the NDVI, DEM (only for multiple pixels), rainfall and temperature from center pixel dataset and 10x10 pixels dataset separately.
Data Cleaning	Removed minicubes with >30% missing NDVI values and with >10% of NDVI values <0.
Missing Value Filling	Applied the PCHIP method to fill gaps in NDVI, DEM, rainfall, and temperature datasets.
Data Splitting	Randomly split the datasets into 70% training, 20% validation, and 10% testing datasets.
Normalization	Normalized the data using the minmax scaler based on the training set.

3.3 Modeling method

The modeling method involves the application of four models, including Random Forest, LSTM, ConvLSTM, and Transformer model. These models were chosen for their proven effectiveness in dealing with time series data like ours. And ConvLSTM and Transformer models are also very good at processing spatial-temporal data (SHI et al., 2015b) (Vaswani et al., 2017).

The purpose of modeling in this study is to train four models to predict the NDVI value at the 21st time step, respectively, given the NDVI data of each previous 20 timesteps and the rainfall and temperature data of 21 timesteps. In terms of the modeling process, each model goes through the same general steps: data preprocessing, model building and training, and model evaluation. During preprocessing, the input and output sequences were created for training, validation, and test data. These series are created based on a sliding window approach, capturing temporal dependencies in the data. In detail, first, the data was split into overlapping windows of a specific size (20 timesteps), then used the sequence of data in each window to predict the NDVI in the next time step. And repeated this process for each unique sample in the dataset to ensure that our model learns from a wide range of studies in time mode. The model was then trained using the training data and validated using the validation data. After training, the model was evaluated on the testing data using mean squared error (MSE) as a performance metric.

3.4 Build the machine learning models

3.4.1 Pre-processing of model data

Knowing that the data set of this study contains 30 timesteps, each sample in the data set was isolated and treated as a separate entity, which means that each sample has its own NDVI, DEM, rainfall, and temperature in the time series data. Before training the model, we created a model architecture that can predict the future NDVI value of each sample based on its past value. To achieve this, a sliding window approach was applied using a window of a certain size (20 timesteps) to capture time dependencies in the data. Then we created input and output sequences for training, validation, and testing data, and reshaped the input data into 3D tensors with three dimensions (number of samples, number of time steps, and number of features) to fit the model. This approach enables the model to predict the next NDVI value using information from every 20 timesteps in the past. Since there was a total of 30 timesteps, each sample can run 20 times and get 10 predictions.

- **Center pixel data preprocessing**

In center-pixel data preprocessing, the input data shape was (n, 20, 3), where 3 represents the number of input features (NDVI, rainfall, and temperature). The shape of the output was (n,) because the output feature indicated the next step NDVI value.

- **Multi-pixels data preprocessing**

For the preprocessing of multi-pixel data, the approach was very similar to the single-pixel case, it just needed to further consider the spatial correlation between different pixels. Compared with the center pixel data approach, the changes are that first the DEM feature was added as an input, and second, that the data augmentation was applied to create 100 copies of 'Rainfall' and 'Temperature' to ensure that all features had the same shape. In addition, to fit the input data to train the ConvLSTM and Transformer models, the input data was reshaped from a 3D to a 5D tensor (number of samples, number of timesteps, length, width, number of features). In this case, the input shape was (n, 20, 10, 10, 4) where (10,10) represents the spatial dimensions that indicate a 10x10 grid of data, and 4 represents the number of input features (NDVI, DEM, rainfall, and temperature). The shape of the output was (n,10,10,1) because the output feature is the next step NDVI values in the 10X10 grids.

Next, for the Random Forest and LSTM model under the multi-pixel case, we first took the mean value of all input and output features of the multi-pixel datasets. Then creating the approach of input and output sequence was the same as the center pixel case, thus the shape of the input was 3D and the output was 1D, this was due to the applied average approach converting all feature values into time series data.

3.4.2 Build and train the models

- **Center pixel data training model**

Building and training our models was carried out once the center pixel data was ready, the models include Random Forest, LSTM, ConvLSTM, and Transformer models.

For the Random Forest model, the input data was reshaped from (n, 20, 3) to (n, 60) to fit the model structure. Then the model was trained with 200 estimators and a maximum depth of 10, the estimator and maximum depth represent the number of trees in the random forest and the maximum depth of each tree, respectively.

The LSTM model is built with a three-layer architecture, includes an LSTM layer with 128 units, and "tanh" as the activation function. This was followed by two dense layers with 64 and 32 units respectively, using the "relu" activation function because it helps the model learn complex patterns without suffering from the vanishing gradient problem. And the final layer was a dense layer with 1 unit for regression output. Lastly, L1 and L2 regularizes were used in LSTM layers to prevent overfitting.

The ConvLSTM model contains two ConvLSTM2D layers, the first with 64 filters and the second with 32, the filters are used to transform the input data. A kernel size of (3,3) is a common choice, and each layer is followed by a dropout layer for regularization to control overfitting.

Finally, the Transformer model was built using a self-attention mechanism with multi-head attention, which allows the model to focus on different locations, thereby being able to capture various aspects of the input (Vaswani et al., 2017). The model implemented here consists of an encoder with multi-head self-attention, positional encoding, and a feed-forward network. The Transformer encoder Layers include a multi-head attention layer and a feed-forward neural

network, and the global average pooling layer was used to reduce the spatial dimension. The last layer was a dense layer with 1 unit for regression output.

Except for Random Forest, the other three deep learning models are trained using the Adam optimizer with a learning rate of 0.0005. The LSTM model trained for 200 epochs, ConvLSTM and Transformer model each trained for 100 epochs, and the early stop callbacks were applied to all models to prevent overfitting.

- **Multi-pixels data training model**

Besides single-pixel data, multi-pixel data was used to train the four models. For the Random Forest and LSTM models, since all input and output of them were averaged from the multi-pixel data, the model structure was the same as for the single-pixel case.

However, for the ConvLSTM and Transformer models, to adapt to the processing of 5D spatial-temporal data, we doubled the filters contained in the two ConvLSTM2D layers in the ConvLSTM model, that is, the first and second layers had 128 and 64 filters respectively. The kernel size and activation function remain the same, but an additional Dense layer with 32 units with "relu" activation function was added, corresponding to the complexity of features that can be learned.

For the Transformer model, a 2D convolution operation was first applied to the input sequence and then reshaped to 3D to prepare for the Transformer encoder. This was followed by applying a Transformer encoder, then averaging the encoded output, passing through a dropout and a dense layer, reshaping to the desired 4D output value, passing through another dense layer, and finally reshaping again to match the target output shape.

The application of optimizer, number of epochs and early stop callbacks was same as the center pixel data training models.

3.4.3 Model adjustments

The model tuning was carried out after training the model. Hyperparameters play a critical role in the performance of machine learning algorithms, the tune of the hyperparameters of models could optimize the model performance and speed up the training process (Hutter et al., 2015). The adjustments were made based on how the models performed on the validation data, through comparing the MSE results, we could find the most desired models' structures. The lower the MSE, the better the model performance. Also, it is better to choose a model whose MSE on the training set is slightly smaller than that on the validation set, which was one of the ways to prevent overfitting.

The way to choose appropriate hyperparameters for the models is discussed here. Dropout is the percentage of neurons that are ignored during training, it is commonly used to prevent overfitting and to improve the generalization ability of the model (OpenAI, 2023). Here, it was set to 0.1 for the ConvLSTM model and 0.05 for the Transformer model. The Adam optimizer was used, it is known for its fast convergence and robustness to noisy gradients. The mean square error (MSE) loss function is chosen, because it penalizes large errors more than small errors, and is popular for dealing with regression problems (Hutter et al., 2015). Next, the early stopping was applied for all deep learning models, and the patience was set to 10 or 15, it is another method to prevent overfitting, enabling saving the best weights if the model's performance on the validation set does not improve after the set epochs. The batch size represents the number of samples that are processed together in each training iteration, here it was set to 64 or 128 to balance the computational requirements and the accuracy of the gradient estimation (OpenAI, 2023). The learning rate as a hyperparameter determines the step size of each iteration while moving towards the minimum value of the loss function. It is one of the hardest and most important hyperparameters to set, too high will make learning skip minima,

but too low will take too much time to converge or get stuck in undesired local minima (Hutter et al., 2015). After iterative tuning, the chosen learning rate here was 0.0005 for the three deep learning models.

After visualizing the predictions of the respective models, the model structure mentioned in the table below was confirmed as the model with the best simulation results (Table 4). The next step is to extract the model-predicted test data and compare it with the original target test set to evaluate model performance.

Table 4 - The most desirable model settings after the model adjustments

Model	*Input Shape*	*Dropout Rate*	*Optimizer*	*Learning Rate*	*Loss Function*	*Batch Size*	*Epochs*
LSTM *(Single pixel)*	(20, 4)	-	Adam	0.0005	MSE	32	200
ConvLSTM *(Single pixel)*	(20, 1, 1, 4)	0.1	Adam	0.0005	MSE	64	100
Transformer *(Single pixel)*	(20, 1, 1, 4)	0.05	Adam	0.0005	MSE	64	100
ConvLSTM *(10X10 pixels)*	(20, 10, 10, 4)	0.1	Adam	0.0005	MSE	128	100
Transformer *(10X10 pixels)*	(20, 10, 10, 4)	0.05	Adam	0.0005	MSE	64	100

3.5 Evaluating accuracy

To gauge the efficiency of the model, we're utilizing four statistical quantitative parameters: the coefficient of determination (R^2), the root mean square error (RMSE), the mean absolute error (MAE), and processing time.

3.5.1 Coefficient of determination (R^2)

The coefficient of determination, or R^2, is derived from squaring the correlation coefficient (R). It signifies the degree of the linear association between the actual and estimated data and can be employed to ascertain how much variation is instigated by two factors (D. N. Moriasi et al., 2007). Values of R^2 range from 0 to 1, with values closer to 1 indicating a stronger relationship between the two variables being analyzed. Generally, an R^2 value exceeding 0.5 is considered to demonstrate a satisfactory correlation. The formula for R^2 is expressed in Equation (2):

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \quad (2)$$

3.5.2 Root mean square error (RMSE)

RMSE is a measure that exhibits the average discrepancy between the actual value and the model's outcome. It assesses the variance between the true output and the model's forecasted output value, training stops once the RMSE starts increasing, implying that the minimum square error value or the test sets square error has been achieved. Considering the calculation time, the training process is usually stopped once the good patterns exceed 98% (Kisi, 2013). The formula for RMSE is given in Equation (3):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - x_i)^2}{N}} \quad (3)$$

3.5.3 Mean absolute error (MAE)

MAE, similar to RMSE, is a widely used error metric in the assessment of models. It quantifies the error in terms of the unit (or squared unit) of the component under investigation (D. N. Moriasi et al., 2007), assisting in interpreting the results. In general, an MAE value that is approximately half the standard deviation of the measured data is deemed to be low. MAE computes the average absolute differences between predicted and actual values; it gives a linear score where all individual differences equally impact the average. The formula for MAE is provided in Equation 4:

$$MAE = \frac{\sum_{i=1}^{n} abs(y_i - x_i)}{n} \quad (4)$$

3.5.4 Processing speed

When evaluating a machine learning model, apart from traditional accuracy, etc., an important aspect that is often overlooked is its computational efficiency, which is the time it takes to train the model. Processing speed as an evaluation metric is especially important in real-world scenarios where timely results are required or system resources are constrained. It provides important insights into the computational cost of the model (Pouyanfar et al., 2018). For situations where frequent model updates or very large data are used, the efficiency of a model depends not only on its predictive power but also on its speed and required computing resources. Deep learning models tend to have higher accuracy but also require more computing resources and longer training time. Conversely, simpler models such as Random Forest or Linear Regression models can be trained faster, but they may not be able to effectively capture complex patterns in the data. Ideally, a good model should guarantee both high accuracy and relatively fast training speed (OpenAI, 2023) (Pouyanfar et al., 2018).

4 Results

4.1 The model performance visualization on single-step forecasting

First, the ability of the model in single-step prediction would be viewed by comparing the model predicted value and the actual value on the scatter plot and timeseries plot. This evaluation was applied to the first three scenarios to evaluate four machine learning models - random forest, LSTM, ConvLSTM and Transformer performance in predicting Normalized Difference Vegetation Index (NDVI).

4.1.1 Time Series Forecasting Comparison (E1, E2, E3)

- **Scatter plot for center pixel input (E1** - Random Forest, LSTM, ConvLSTM, Transformer model**)**

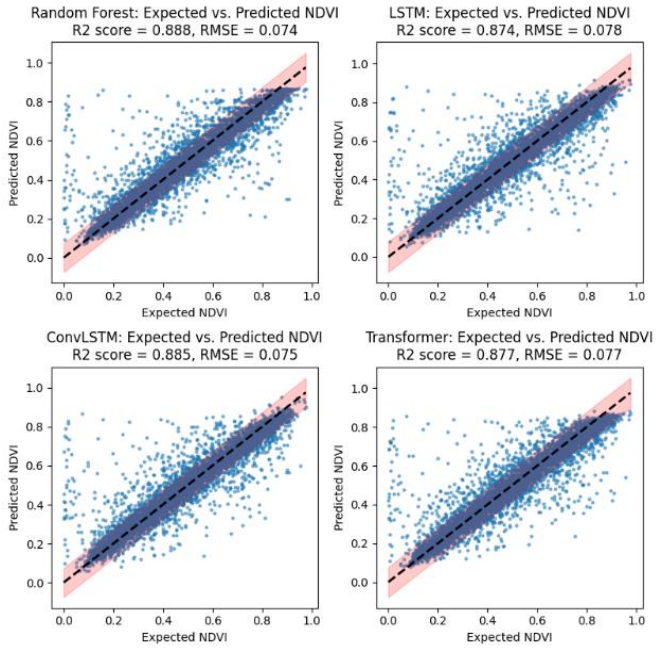

Figure 6 - Comparative scatter plot of testing data for single pixel prediction (E1)

In the first experiment (E1), all models were trained to predict NDVI from time series data of center pixels. Visual inspection of the scatterplots comparing the true NDVI to the predicted NDVI for each model shows that the forecasts are generally quite accurate (R^2 was around 0.88), with a large proportion of the data points clustered around a one-to-one line, indicating there was a high correlation between the predicted and true values. Furthermore, most of the data points fell within the region of the red-shaded deviation line, confirming that most forecasts do not deviate significantly from actual values.

Analyzing the performance metrics further, the Random Forest model seems to slightly outperform the LSTM, ConvLSTM, and Transformer models. It had the highest R^2 score of 0.888, indicating that the model explained 88.8% of the variability in the dependent variable that could be predicted from the independent variables. The model also had the lowest RMSE value of 0.074, which indicates that, on average, the Random Forest model's predictions deviate

from the actual values by 0.074 units. However, it was found that when the true NDVI was zero, the predicted NDVI was not equal to zero, but spanned the entire NDVI range, this mismatch could result from model limitations, noise, etc.

Although the Random Forest model slightly outperformed the other models, it is important to note that the differences in R^2 and RMSE scores between the models were all within 5%, suggesting that all four models provide reasonably good performance for this NDVI single step forecasting task.

- **Scatter plot for multiple pixels input (E3 -** Random Forest, LSTM model**)**

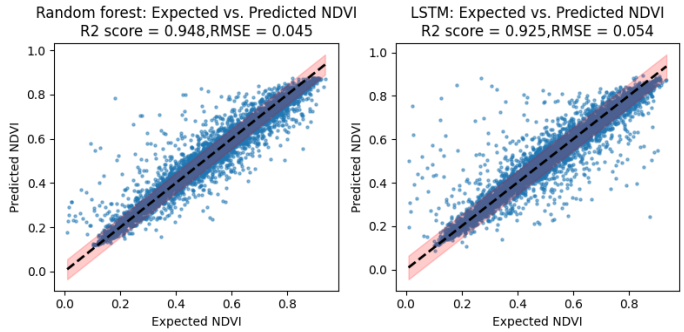

Figure 7 - Comparative scatter plot of testing data for multiple pixel prediction (E3)

- **Scatter plot for multiple pixels input (E2 –** ConvLSTM, Transformer model**)**

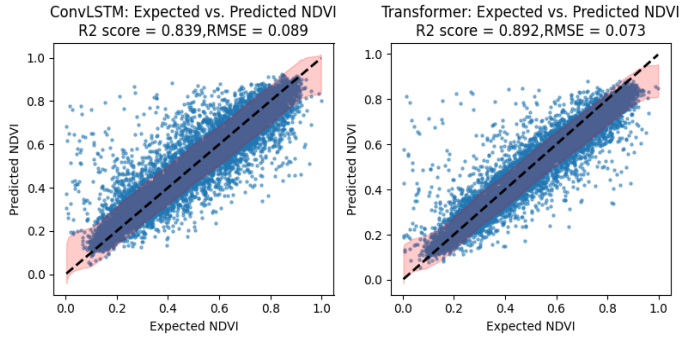

Figure 8 - Comparative scatter plot of testing data for multiple pixel prediction (E2)

In the second experiment (E2), the ConvLSTM and Transformer models were adapted to utilize spatial-temporal data from a 10x10 pixel grid that expanded from the center pixel, while in the third experiment (E3), the Random Forest and LSTM models used the average NDVI value of this grid. These models perform well (R^2 = 0.948 for the Random Forest model, R^2= 0.925 for the LSSTM model, R^2 = 0.839 for the ConvLSTM model, and R^2 = 0.892 for the Transformer model) in general by viewing how closely the data points cluster around the one-vs-line.

But the ConvLSTM and Transformer models had more points outside the bias range above the one-vs-line than below it, suggesting that these models tend to overestimate NDVI. And despite the extra spatial information, the performance of the ConvLSTM and Transformer models degrades relative to the Random Forest and LSTM models. This decline, represented by lower R^2 scores and higher RMSE values, may be attributed to factors such as differential handling of spatial information, and possible model overfitting due to the increased complexity of input data. Interestingly, the random forest and LSTM models using the mean of 10x10 pixels as input significantly reduced their misprediction of higher NDVI values when the expected NDVI was zero. This shows that using spatial information even in a simple way (as an average)

improves the performance of these models.

Overall, it is important to note that despite these differences, all models exhibit robust performance with R^2 scores consistently above 0.8, and RMSE consistently lower than 0.1.

- **NDVI in timeseries for center pixel input (**E1 - Random Forest, LSTM, ConvLSTM, Transformer model**)**

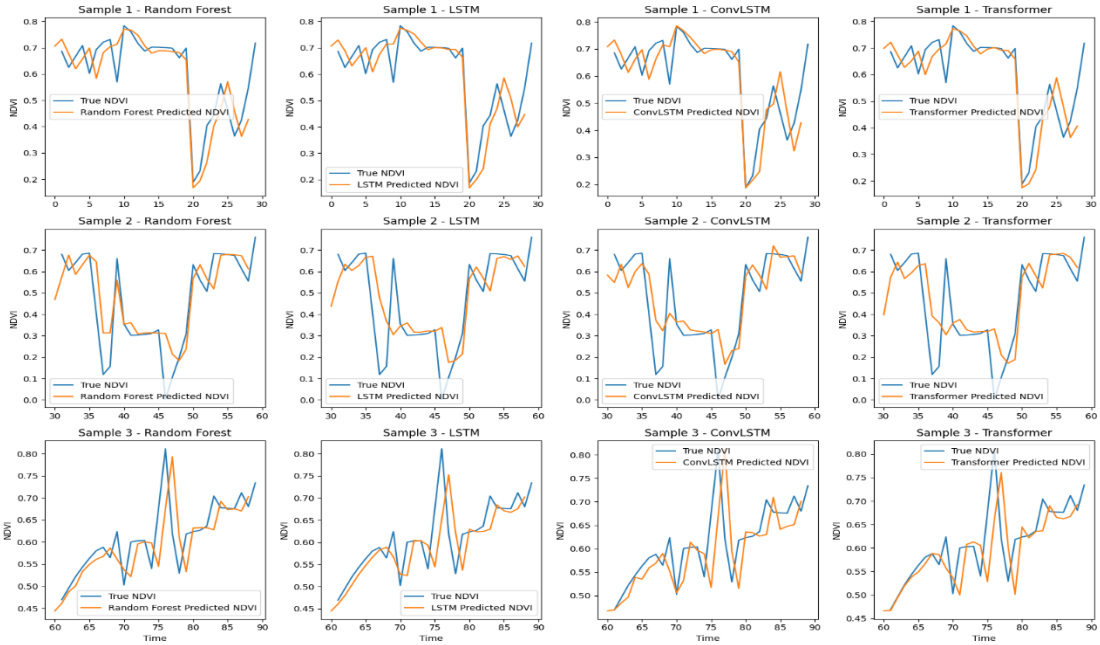

Figure 9 - Randomly selected samples of representative models' performance for center pixel prediction

- **NDVI in timeseries for multiple pixels input (E3 -** Random Forest, LSTM model & **E2 –** ConvLSTM, Transformer model**)**

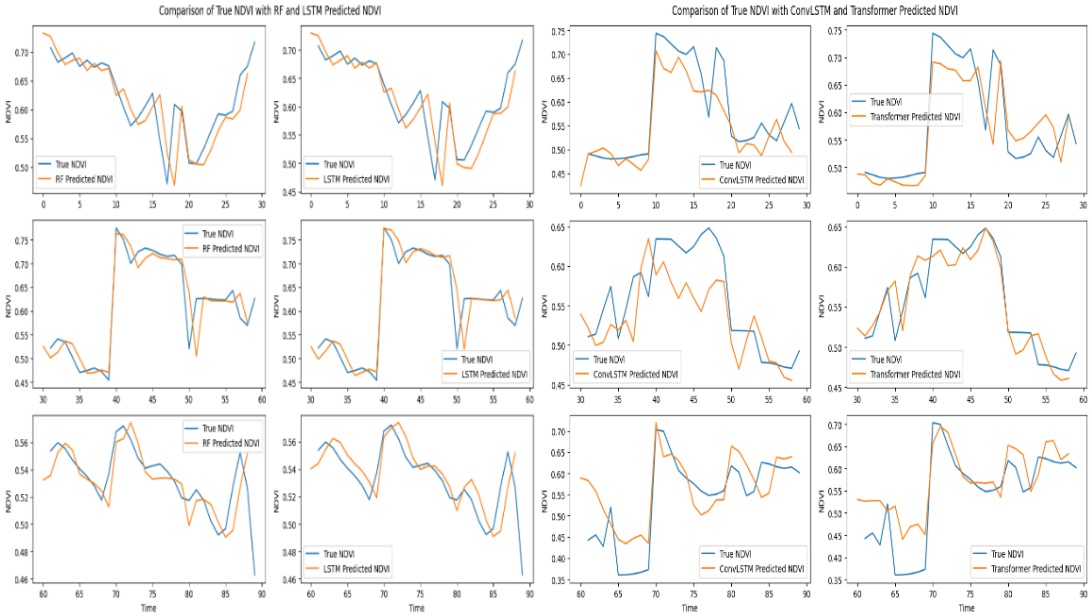

Figure 10 - Randomly selected samples of representative models' performance for multiple pixel predictions; Left E3 (Random Forest and LSTM, and right E2 (ConvLSTM and Transformer).

Upon examination of the time series plots depicting the predicted NDVI values from the four models with the true NDVI values, a broad performance trend becomes apparent. All models have successfully captured the underlying temporal changes in NDVI values on a macro scale, this indicates the models were capable of understanding and simulating the general trends and periodicities in the data.

In situations where the NDVI values exhibited a significant variation (e.g., a change of approximately 0.5 units between two consecutive time periods), the prediction value performed exceptionally well, and matched the observed values closely. This may indicate that the models are well-equipped to recognize and react to substantial shifts in the data. However, the models' proficiency appears to wane somewhat when the change in NDVI values between two consecutive time steps is minimal (within a range of about 0.2 units). In these instances, the predicted values from the models can sometimes lag behind the true values by one to two timesteps. This observation points towards the models' tendency to react slower to more subtle changes in the data, potentially because of their inherent learning processes and tuning parameters.

Some interesting observations can be made by comparing the prediction trends between the four models. The Random Forest model appeared most adept at capturing peak values in the NDVI series. This may be due to the model's robustness to noise and ability to model non-linear relationships. LSTM, ConvLSTM, and Transformer models, on the other hand, exhibited characteristics like occasionally missing smaller peak values in the data. Which is worth noticing, both the ConvLSTM and Transformer models exhibited a tendency to underestimate NDVI values. This could potentially be attributed to the structure and learning mechanisms of these models, especially in their handling of complex spatial-temporal dependencies.

In summary, all models broadly capture NDVI changes over time, although they exhibited different sensitivities to peaks and small changes in the series. Therefore, the choice of the model may depend on the specific requirements of the forecasting task at hand. Further research and hyperparameter tuning may improve the model's ability to recognize more subtle changes and reduce the observed forecast lag.

4.1.2 Spatial-Temporal Sequence Forecasting (E4 – ConvLSTM, Transformer model)

In this section of the study, we aim to evaluate the spatial-temporal predictive capability of ConvLSTM and Transformer models through experiment 4 (E4). Because the two models in this experiment use the spatial-temporal data of the learning input to predict the spatial-temporal NDVI of the next step, they provide a great chance to explore the ability of the two models to learn spatial relationships. By comparing the model's predicted NDVI maps with the true NDVI maps, we can visually gauge the models' performance in terms of representing spatial heterogeneity in NDVI. To carry out this analysis, three samples were randomly selected from the testing dataset and plotted as images, expected to interpret a general trend of the whole predicted NDVI value. For each image, the NDVI maps are visualized using a color gradient to represent the range of NDVI values, where the intensity of color reflects the magnitude of NDVI.

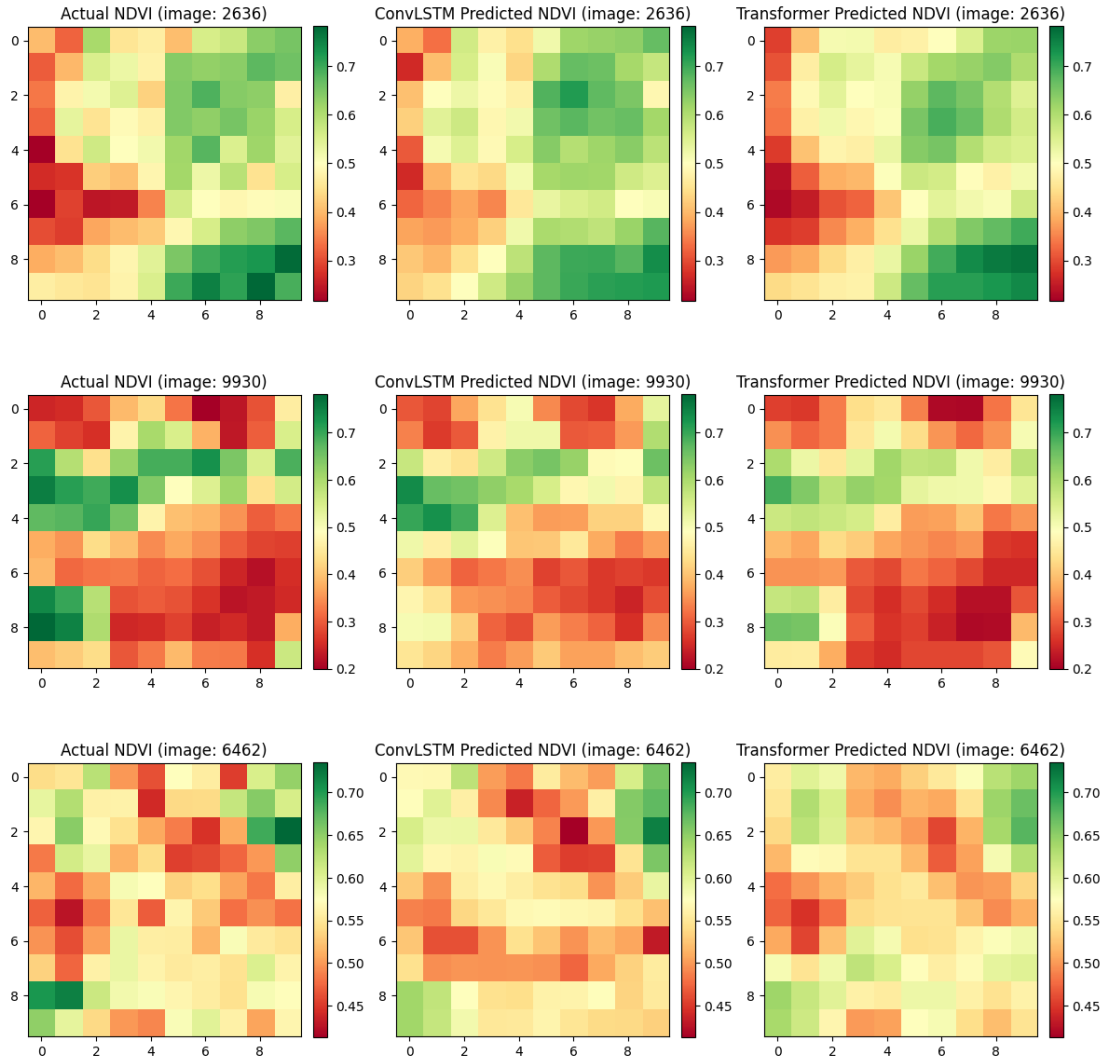

Figure 11 - Randomly selected samples of models' performance in E4 - ConvLSTM and Transformer model

From the results of the visual comparison, the predicted values of the ConvLSTM and Transformer models generally showed an NDVI distribution similar to the real NDVI value image. Particularly, they excelled in areas where pixel values exhibited substantial variation. This observation aligns with the models' inherent capacity to decipher complex spatial-temporal patterns.

However, consistent with our prior observations, both models exhibited a tendency to overestimate lower NDVI values and underestimate higher ones. Upon closer examination of the NDVI scale, it is evident that the model's accuracy in predicting higher NDVI pixel values surpasses that of smaller NDVI values, both in terms of spatial distribution and magnitude. This could stem from the models' training processes, where they have been exposed to more samples of higher NDVI values, thereby refining their ability to predict these values more accurately.

A subtle difference between the two models emerges when we consider the prediction of distinct pixel values within a small spatial range. In such cases, the Transformer model displays superior capability to ConvLSTM, which occasionally overlooks such fine-scale differences. This discrepancy could be attributed to the self-attention mechanism in the Transformer model, which allows it to capture relationships between pixels irrespective of their spatial proximity. Conversely, ConvLSTM's convolutional operation might be more inclined to prioritize local

spatial dependencies, thereby potentially missing out on fine-scale contrasts.

4.2 The model performance visualization on multi-step forecasting

The second phase of the study delved into the predictive power of the four models in the context of multi-step forecasting, with a particular focus on the next 10 steps. The rationale behind using a recursive multi-step forecasting setup is that it mimics a real-world environment where forecasts are based on past observations as well as the model derived from their previous predictions.

To compare the ability of each model (Random Forest, LSTM, ConvLSTM, and Transformer) to accurately predict the NDVI value at future time points, this study randomly selected 3 samples, and compared the NDVI values at the future steps 1st, 5th, and 10th, respectively. Three selected future time points provided an opportunity to investigate the model's ability to capture short-term (step 1st), medium-term (step 5th), and long-term (step 10th) changes in NDVI data. Random selection of samples ensures that the analysis includes a broad and representative set of scenarios, enhancing the robustness and generalizability of the results. For clarity and better visualization, each sample is plotted individually. From this we obtained a series of graphs intended to provide a visual metric to assess a model's ability to predict future NDVI values, thereby helping to identify the most suitable model for multi-step NDVI forecasting.

4.2.1 The NDVI prediction in 1st timesteps in the future (E1, E2, E3)

- **Prediction based on center pixel input (**E1 - Random Forest, LSTM, ConvLSTM, Transformer model**)**

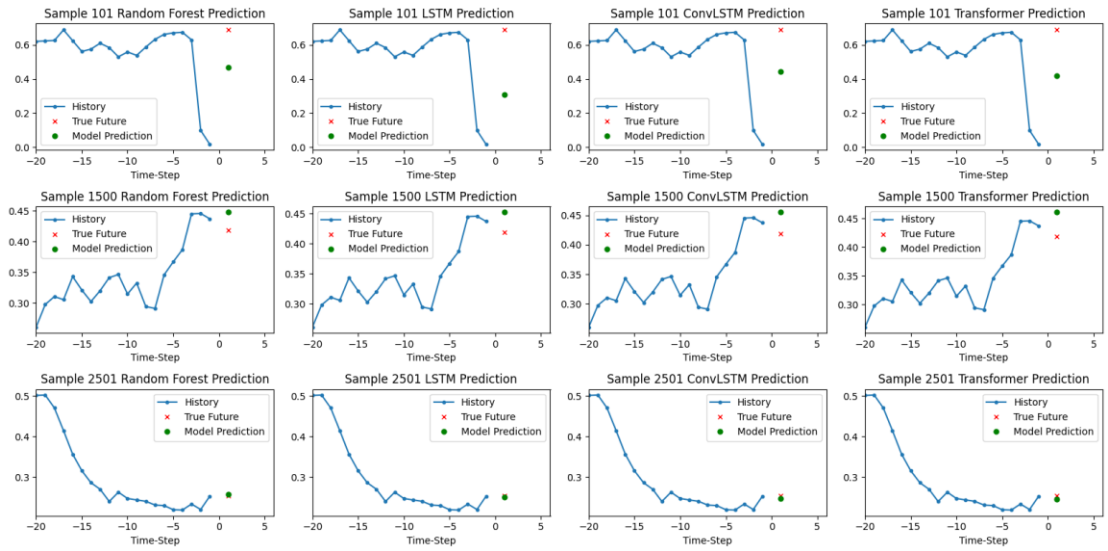

Figure 12 - Randomly selected samples of representative models' single pixel prediction in E1 for 1 timestep in the future

- **Prediction based on multiple pixels input (E3 -** Random Forest, LSTM model & **E2 –** ConvLSTM, Transformer model**)**

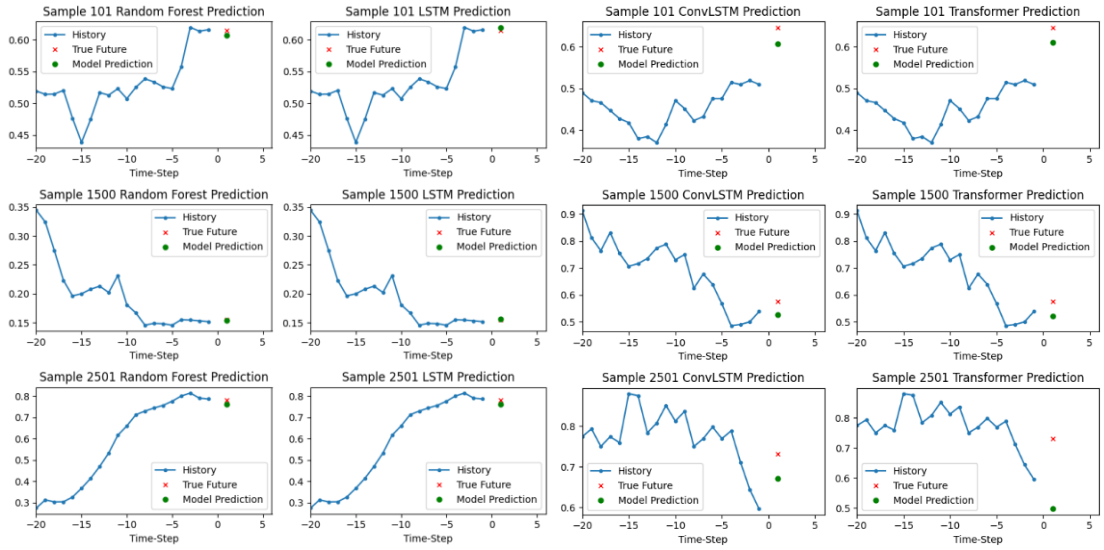

Figure 13 - Randomly selected samples of representative models' multiple pixel prediction in E3 (left 2) and E2 (right 2) for 1 timestep in the future

4.2.2 The NDVI prediction in 5 timesteps in the future

- **Prediction based on center pixel input (**E1 - Random Forest, LSTM, ConvLSTM, Transformer model**)**

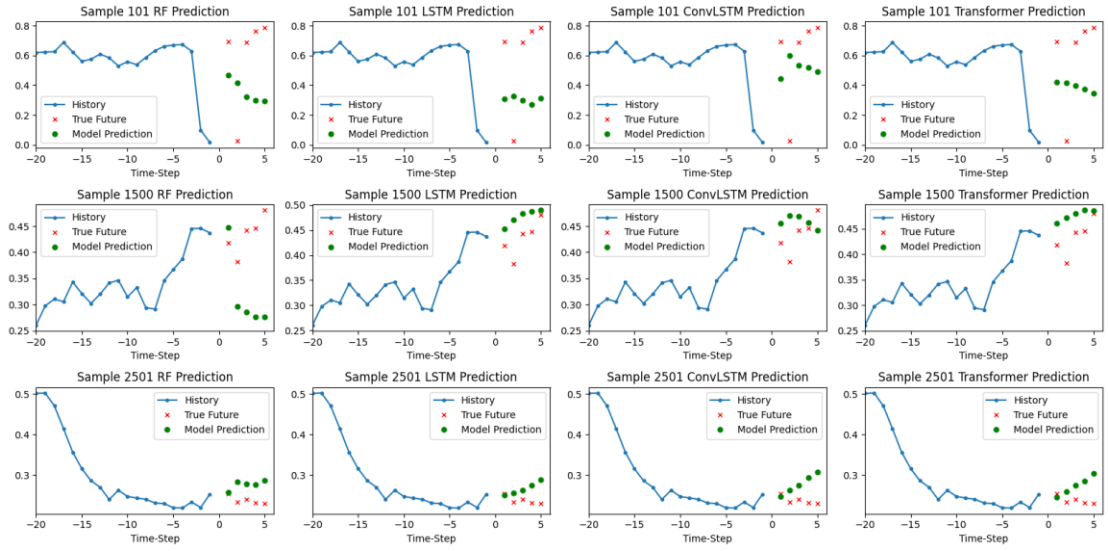

Figure 14 - Randomly selected samples of representative models' single pixel prediction in E1 for 5 timesteps in the future

- **Prediction based on multiple pixels input (E3 -** Random Forest, LSTM model & **S2 –** ConvLSTM, Transformer model**)**

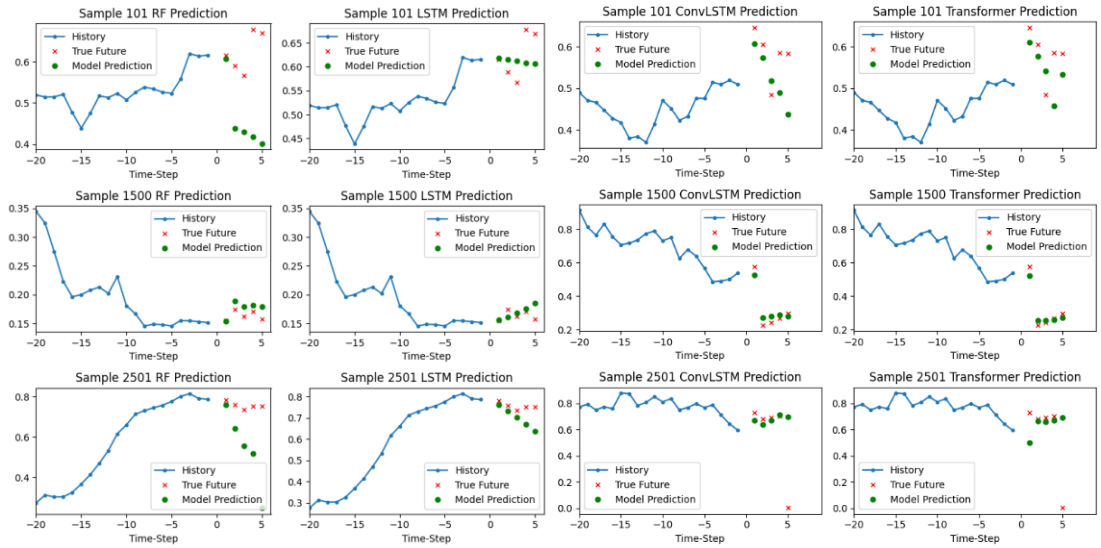

Figure 15 - Randomly selected samples of representative models' multiple pixel prediction in E3 (left 2) and E2 (right 2) for 5 timesteps in the future

4.2.3 The NDVI prediction in 10 timesteps in the future

- **Prediction based on center pixel input (**E1 - Random Forest, LSTM, ConvLSTM, Transformer model**)**

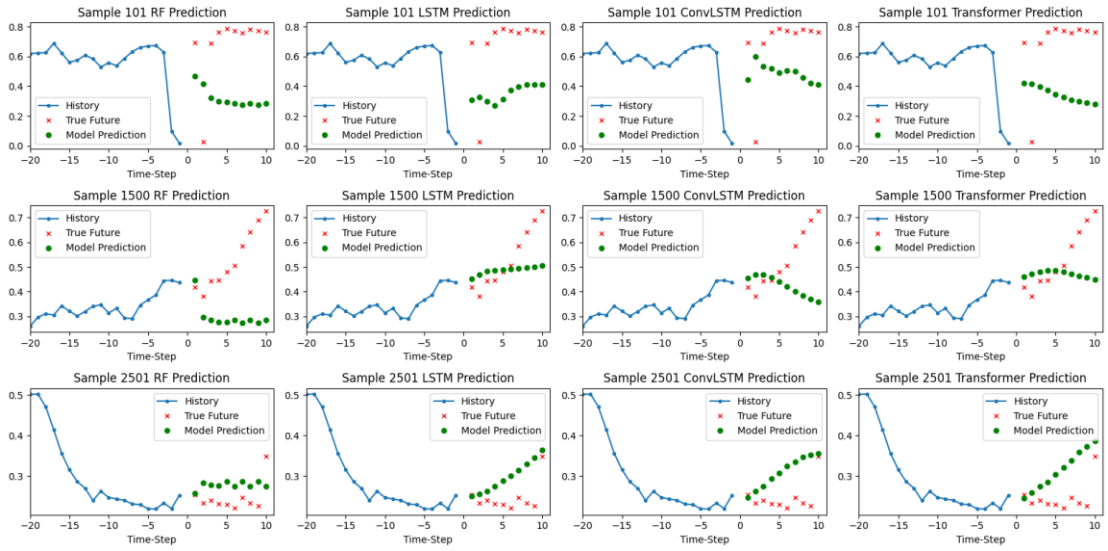

Figure 16 - Randomly selected samples of representative models' single pixel prediction in E1 for 10 timesteps in the future

- **Prediction based on multiple pixels input (S3 -** Random Forest, LSTM model & **S2 –** ConvLSTM, Transformer model**)**

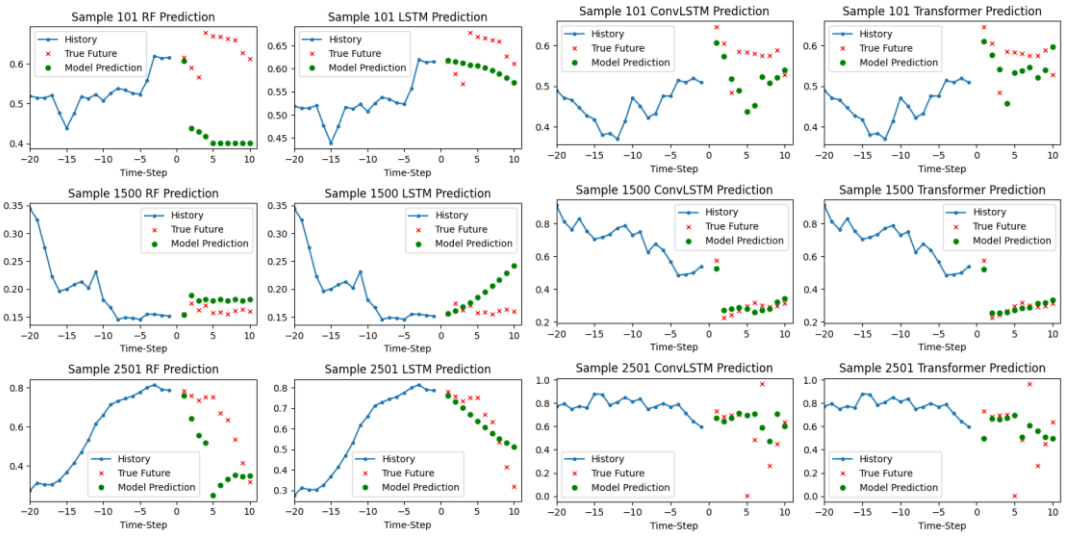

Figure 17 - Randomly selected samples of representative models' multiple pixel prediction in E3 (left 2) and E2 (right 2) for 10 timesteps in the future

Upon close observation of the prediction graphs for the four models, variations between models were noticeable, underscoring the differentiated forecasting competencies of these models over future steps.

Initially, at the first step of the forecast, the models exhibited similar predictive accuracy across both single-pixel and multi-pixel inputs. However, as the forecasting progressed to the five timesteps, discernible divergences emerge, particularly with single-pixel input. All models with single-pixel input exhibited an increasing discrepancy between the predicted and true NDVI values, with Random Forest particularly displaying pronounced deviation. The increased deviation with advancing prediction steps could be indicative of the model's inadequacy in capturing the temporal dynamics in NDVI values, particularly over extended periods.

In contrast, upon examination of multi-pixel inputs at the 5 timesteps forecasting, ConvLSTM and Transformer demonstrated marked robustness in prediction accuracy, while the performance of Random Forest and LSTM remained largely unchanged. This divergence in performance can be attributed to the ability of ConvLSTM and Transformer to utilize spatial information from multi-pixel input, whereas LSTM and Random Forest relied on average values, hence being oblivious to spatial context.

Continued to the 10 timesteps forecasting, the performance of models with single-pixel input echoed their trends in 5 timestep forecasting, with the gap between predicted and true values expanding further. It is worth noting that the models, excluding Random Forest, accurately predicted the NDVI values for the first and tenth steps, possibly owing to deep learning models' adeptness in capturing pronounced variations in the time series. This observation might hint at an underlying strategy of these models, where they interpolate smoother values between time points with significant variations. Additionally, the multi-step predictions showed a tendency to regress towards the mean NDVI value (~0.48), a manifestation of the models' potential bias towards average conditions.

In sharp contrast, ConvLSTM and Transformer models demonstrated remarkable robustness in multi-step forecasting, with minimal deviations from true values, even up to the tenth step. This could be due to their inherent architectural superiority in handling sequence data, particularly for multi-step forecasts. Convolutions in ConvLSTM allow it to capture spatial patterns, while the attention mechanisms in Transformers provide it the capability to model long-term dependencies. The poor performance of Random Forest in this task can be attributed to its fundamental limitations in handling temporal sequence data.

4.3 Model evaluation

4.3.1 Model performance on predicting the NDVI in single timestep

In this section, we quantitatively compared the performance of four distinct machine learning models—Random Forest, LSTM, ConvLSTM, and Transformer—in predicting the Normalized Difference Vegetation Index (NDVI) based on both single-pixel and multi-pixel inputs. Model performance was evaluated using four metrics: the coefficient of determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and training time. The following table (Table 5) compares in detail the performance of each model measured by performance metrics under different input tasks.

Table 5 - Comparison of the model performance in single-pixel and multi-pixel input tasks

Input NDVI	***Model***	***Model performance***			
		R^2	*RMSE*	*MAE*	*Training Time (second)*
Single pixel	Random forest	0.888	0.074	0.040	246
	LSTM	0.874	0.078	0.042	5442
	ConcLSTM	0.885	0.075	0.040	5442
	Transformer	0.877	0.077	0.042	2035
Multiple pixels	Random forest	0.947	0.045	0.027	351
	LSTM	0.925	0.054	0.029	3009
	ConcLSTM	0.825	0.093	0.059	5160
	Transformer	0.882	0.076	0.048	587

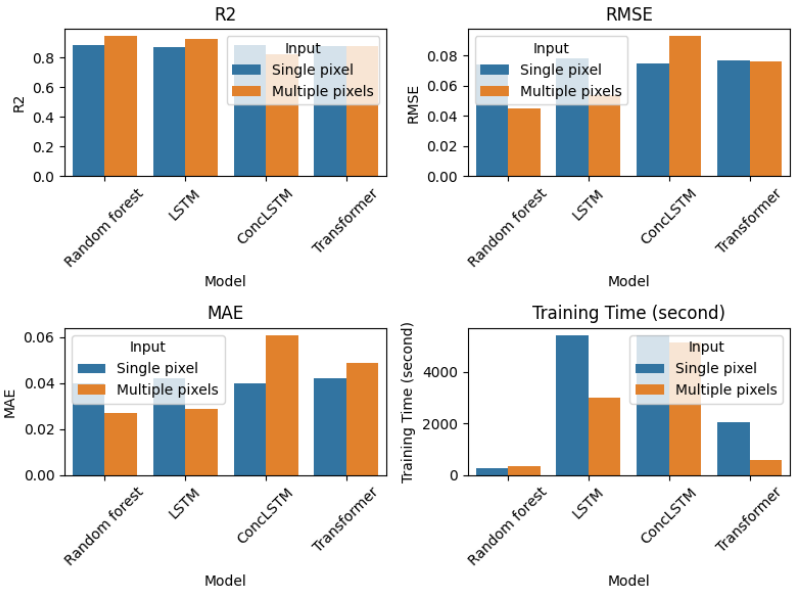

Figure 18 - Visualization of the R2, RMSE, MAE and Training time for both single pixel and multiple pixel prediction of all the representative models

When predicting NDVI based on single pixel input, the Random Forest model demonstrated superior performance across all metrics. It achieved the highest R^2 score of 0.888, suggesting it explained the most variability in the data. It also reported the lowest RMSE (0.074) and MAE (0.040), indicating the smallest average errors. Moreover, with a training time of 246

seconds, the Random Forest model was considerably more computationally efficient than the LSTM, ConvLSTM, and Transformer models, which required 5442, 5442, and 2035 seconds respectively.

In the case of predicting NDVI based on multiple pixel inputs, the Random Forest model again emerged as the top performer. It yielded the highest R^2 score (0.947), indicating exceptional predictive accuracy, and reported the lowest RMSE (0.045) and MAE (0.027) values, reflecting small prediction errors. And it is still the model with most computationally efficient, requiring 351 seconds for training, the second fast model was the Transformer model which requires 581 seconds. Interestingly, the ConvLSTM and Transformer model showed a marked decline in performance with multiple pixel input, especially ConvLSTM achieving the lowest R^2 score (0.825) and highest RMSE (0.093) and MAE (0.059) values, implying issues with model fit and prediction error. Despite the shorter training time (587 seconds), the Transformer model's performance on multiple pixel input was inferior to that of the Random Forest and LSTM models, although it showed improved performance compared to the ConvLSTM model. This suggests a possible trade-off between model accuracy and computational efficiency.

In summary, while all models demonstrated robust performance, the Random Forest model consistently outperformed the others. The diversity of model performance under the multiple pixel input underscores the importance of considering trade-offs between predictive accuracy, error minimization, and computational requirements when choosing machine learning models for NDVI prediction.

4.3.2 Model performance on predicting the NDVI in multi-timesteps

In this section, the performance of multi-step forecasting models is evaluated using quantitative evaluations of the mean absolute error (MAE) and root mean square error (RMSE) of the forecasting steps. This is done for single-pixel and multi-pixel inputs to the respective models. To illustrate these findings, the MAE and RMSE values for each model are plotted over the prediction steps. These line graphs (Figure 19) effectively capture the performance trends of each model, allowing a clear visual comparison of their relative accuracy and error progression in the prediction steps.

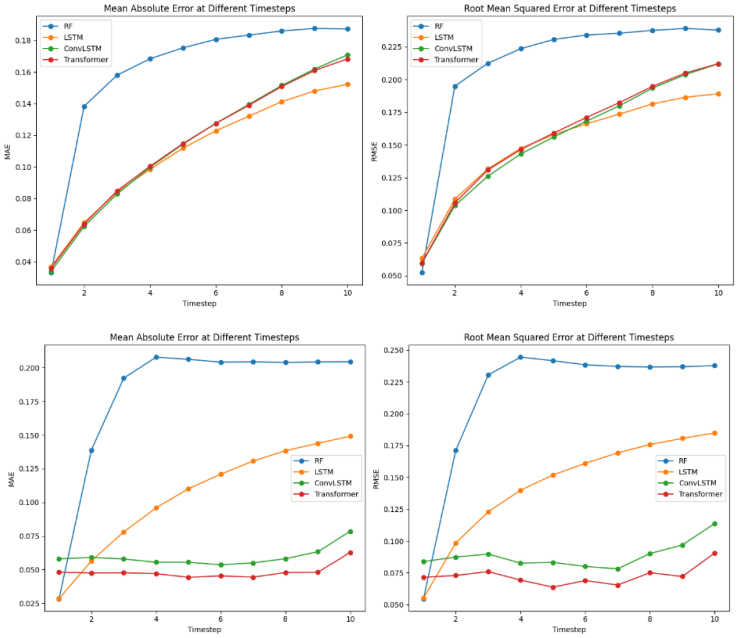

Figure 19 - Visualization of the RSME and MAE for single pixel (up) and multi-pixel (down) prediction of the representative models at each timestep

Starting with the single-pixel input, it was observed that the three models (LSTM, ConvLSTM, Transformer) performed close, with MAE values ranging from 0.03 to 0.16 and RMSE values ranging from 0.05 to 0.21, except for Random Forest. It is worth noting that LSTM slightly outperforms ConvLSTM and Transformer in terms of MAE and RMSE. The close performance of these models indicates their respective ability to accurately predict NDVI values using only temporal information. However, random forests showed an unusual pattern, and while showing comparable accuracy to the other models in the first step predictions, the error in MAE increased significantly from 0.03 to 0.14 in the second step, and thereafter as the number of steps advanced gradually increased. A similar trend was observed for the RMSE values. This behavior highlights the limitations of models that can effectively handle multi-step forecasting tasks when only relying on temporal information.

Considering the multi-pixel input, it is clear from the MAE and RMSE values of the ten prediction time steps that the error rate of the LSTM model increases significantly as the number of prediction steps increases. This gradual but consistent increase in error suggests that the LSTM model, while fairly accurate for immediate predictions, gradually loses its predictive power over time. MAE increased from 0.028 to 0.149 and RMSE increased from 0.055 to 0.185, emphasizing the sensitivity of LSTMs to the number of forecasting steps in the time series. On the other hand, the random forest model shows the same clear trend as for the single-pixel input. Despite starting with a low error rate comparable to that of an LSTM model in the first prediction step, the random forest's error rate rises significantly in subsequent steps. This rapid increase in error rate shows that random forest models do not perform well in multi-step forecasting tasks, especially in maintaining accuracy beyond the immediate forecast step. In contrast, the ConvLSTM and Transformer models exhibit more stable error rate patterns. Both models exhibit relatively small fluctuations in MAE and RMSE values during the prediction step. For example, the MAE of ConvLSTM varies between 0.054 and 0.079, while for Transformer it remains between 0.044 and 0.063, and the RMSE values also remain within a relatively narrow range. The robustness of prediction accuracy over time, and their ability to utilize spatial-temporal information more efficiently, reiterates the importance of incorporating spatial context to enhance NDVI predictions.

Taken together, these observations reveal the strengths and limitations of each model in handling multi-step NDVI predictions. While LSTM performed better for immediate predictions, ConvLSTM and Transformer proved to be more reliable for long-term predictions. Random forest model shows clear weaknesses in multi-step forecasting, highlighting the importance of choosing a model according to the nature and requirements of the forecasting task.

4.3.3 Model performance on predicting the NDVI in spatial aspects

In this section, we further analyzed the performance of the ConvLSTM and Transformer models in capturing the spatial variations of NDVI by classifying the predicted NDVI values into five distinct classes - Build-up (NDVI: 0.015-0.14), Barren Land (NDVI: 0.14-0.18), Shrub and Grassland (NDVI: 0.18-0.27), Sparse Vegetation, and Dense Vegetation. This classification assesses the models' capabilities in discerning distinct land cover types as represented by the different NDVI classes, which allowed computing error metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Accuracy for each NDVI class. The accuracy metric provided a novel perspective on the models' performance by evaluating how often the predicted NDVI value deviated from the actual NDVI value by no more than a predefined tolerance of 10%.

The pie charts (Figure 20) provided an immediate snapshot of the distribution of the actual and predicted NDVI classes, thereby revealing any discrepancies between the true and predicted land cover types. The bar plots of the error metrics (Figure 21) offered an insightful

comparison of the ConvLSTM and Transformer models across the different NDVI classes.

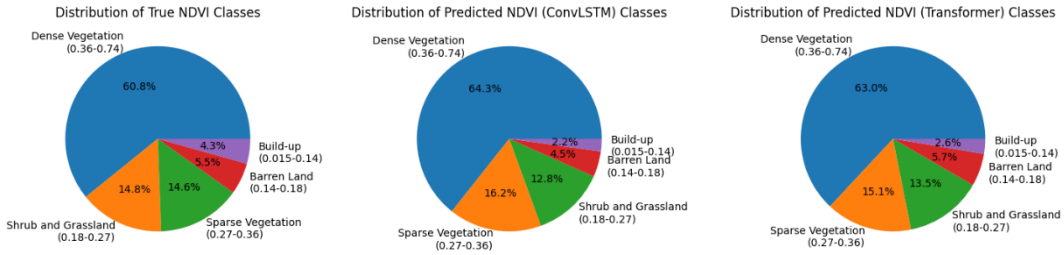

Figure 20 - Predicted and true NDVI values in E4 divided per NDVI class

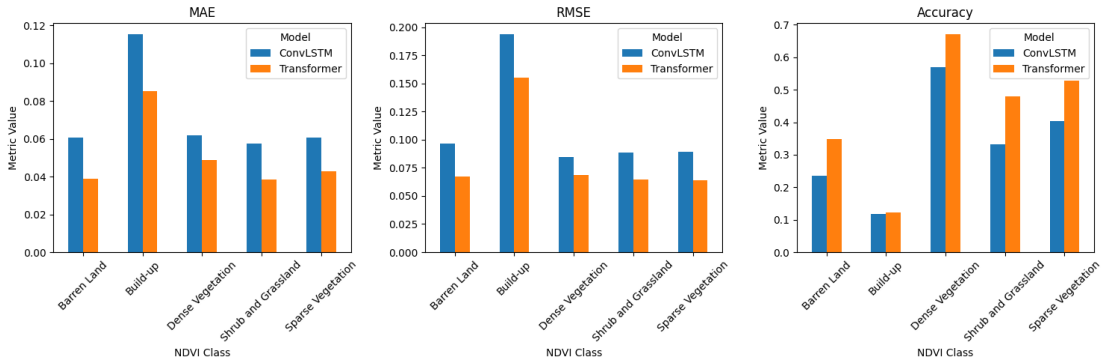

Figure 21 - Visualization of the MAE, RMSE and accuracy of the ConvLSTM and Transformer model in E4

First, across all NDVI classes, the Transformer model consistently outperformed the ConvLSTM model in terms of MAE, MSE, and RMSE, indicating that the Transformer model was generally more precise in predicting the NDVI values. Second, the Transformer model also demonstrated higher accuracy across all classes, suggesting that it was more robust in tolerating small errors. Particularly, the accuracy of the Transformer model was notably higher for classes with higher NDVI values, pointing to the model's superior performance in predicting denser vegetation classes. The results thus revealed that while both models were adept at capturing the broad spatial trends of NDVI, the Transformer model was superior in terms of precision, accuracy, and ability to discern distinct land cover types. However, the results also underscored the necessity for further refinement of these models, especially in improving their performance for classes with lower NDVI values.

Next, the models' performances across different NDVI classes were interpreted as well. The build-up has the lowest NDVI values and represents urban or built-up lands with minimal vegetation. The two models had relatively high MAE, and MSE here than other classes, and the accuracy values for both models were quite low (below 15%), indicating challenges in precisely predicting the NDVI for this class. Barren lands are non-vegetated areas such as sand dunes or rocky terrains. For this class, both models showed relatively low accuracy than other classes (around 35% compared to about 24% for ConvLSTM). Shrub and Grassland represent areas with sparse vegetation, for this class, the Transformer model had a relatively higher accuracy (around 48% compared to about 33% for ConvLSTM). It may be due to the model's ability to better capture the temporal patterns in vegetation growth and die-off in these environments. Sparse Vegetation represents areas with more substantial vegetation cover than shrub and grasslands but less dense vegetation. The Transformer model outperformed the ConvLSTM model across all metrics, and it had a higher accuracy of approximately 53% compared to 40% for ConvLSTM. This might indicate the Transformer model's superior ability to handle

transitions between different vegetation types. The Transformer model was more accurate and robust in predicting the NDVI for this class, with an accuracy of approximately 67% compared to about 57% for ConvLSTM. This may reflect the Transformer model's stronger capability in modeling the complex interactions among climate, soil, and biological factors that drive vegetation dynamics in these areas.

In conclusion, the Transformer model consistently outperformed the ConvLSTM model across all NDVI classes. And both models demonstrated better performance in predicting NDVI for areas with denser vegetation (Shrub and Grassland, Sparse Vegetation, Dense Vegetation) compared to less vegetated areas (Build-up, Barren Land). This may be due to the higher signal-to-noise ratio in the NDVI data for more vegetated areas, which makes the temporal patterns more discernable and thus easier to predict.

5 Discussion & Recommendations

5.1 The model performance discussion

5.1.1. Key Findings

In this study, we have comparatively evaluated the effectiveness of four distinct Machine Learning models - Random Forest, LSTM, ConvLSTM, and Transformer - for predicting NDVI values derived from remote sensing images. The methodology was structured across four different scenarios, focusing on two key predictive aspects: single-step and multi-step forecasting. It can be found that single-step forecasting performance was relatively similar across all models, with Random Forest being the fastest and simplest (R^2 =0.888, Training Time is 10 times faster than the other three models). However, there were clear differences in multi-step prediction capabilities. Both ConvLSTM and Transformer models maintained high accuracy levels for the future 10 steps （MAE <0.075, RMSE <0.125), whereas the Random Forest model experienced a significant drop in accuracy after the initial prediction step (MAE from 0.025 in the first step drop to 0.135 in the second step). The performance of the LSTM model is slightly better than that of the Random Forest, as the predictive ability gradually decreases with the increase in the number of steps. Thus, for single-pixel multi-step predictions, LSTM, ConvLSTM, and Transformer models demonstrated superior performance over the Random Forest model.

Next, this study ventured into challenging terrain, aiming to compare the performance of traditional machine learning models in multi-step forecasting tasks (Moskolaï et al., 2020). This drastic difference in performance between models underscores the effectiveness of ConvLSTM and Transformer models in capturing temporal dependencies in NDVI data. They were distinctive as they integrated spatial dependencies into their predictive models, this ability to extract and process spatial information was reflected in the stability of their predictions over time (OpenAI, 2023). In particular, the Transformer model displayed outstanding robustness in NDVI prediction tasks (Within ten steps of forecasting, MAE only increased from 0.05 to 0.065). This shows comparable capabilities to the most popular ConvLSTM model in spatial-temporal forecasting, emphasizing the potential of the Transformer model in environmental monitoring and prediction.

5.1.2 Interpretation of Results

Our findings underscore the complexity of NDVI prediction and the important role of machine learning models in decoding this complexity. The variation in the performance of different models illustrates the importance of selecting an appropriate model aligning with the specific requirements of the forecasting task. LSTM models may be a fitting choice for immediate prediction tasks, while ConvLSTM or Transformer models might be more suitable for longer-term predictions. Traditional machine learning models like Random Forests might be best utilized for single-step predictions or tasks where temporal dependencies are less crucial.

The performance difference between the models can be largely attributed to their underlying architecture and learning mechanisms. Random Forests, as traditional machine learning models, operate on an ensemble learning approach. This method doesn't account for temporal dependencies between observations, which is vital in time series forecasting tasks, impairing their performance in multi-step prediction tasks. In stark contrast, LSTM, ConvLSTM, and Transformer models are designed to capture these temporal dependencies, making them better suited for such tasks. LSTMs leverage a form of 'memory' to learn from sequences of data, thus performing well for immediate predictions by capturing long-term time

dependencies (Hochreiter & Schmidhuber, 1997). The performance of LSTMs, however, degrades over multiple time steps due to the increasing complexity of recursively predicting future values based on their predictions. ConvLSTM, a derivative of LSTM, adds a convolutional structure to the input and output transformations, enabling it to capture spatial dependencies in the data, along with temporal ones. This capability likely drives ConvLSTM's robust performance in multi-step predictions (SHI et al., 2015b) (Lin et al., n.d.). Transformer models utilize self-attention mechanisms to consider entire data sequences when making predictions. Unlike LSTMs that process data sequentially, Transformer's attention mechanism offers a more global view of the data, potentially augmenting its predictive performance (Vaswani et al., 2017). And even though Transformer is not specifically designed for spatial-temporal prediction like ConvLSTM, it shows amazing robustness for recursive prediction at multiple time steps.

In comparison to Previous Research, our study significantly expands upon existing research by conducting a comparative analysis of four distinct machine learning models for NDVI prediction. Previous studies have been narrowly focused on a single model, leaving a gap in understanding how different models perform in both single-step and multi-step forecasting scenarios. Our comprehensive analysis fills this gap by highlighting the strengths and limitations of each model in these contexts. This comparative approach can guide future research in model selection and development for NDVI prediction tasks.

5.1.3 Limitations

Despite the findings mentioned above, this study does have limitations that should be considered when interpreting the results. In single-step predictions, ConvLSTM and Transformer models didn't outperform simpler models such as Random Forests and LSTMs in all tasks, as demonstrated by their poorer performance in terms of R2 and RMSE. The reasons for this poorer performance include overfitting due to complexity and imbalance in the data, noise in the data due to inadequate preprocessing, and challenges associated with the training and tuning of these complex models.

Moreover, ConvLSTM and Transformer models tend to underestimate larger NDVI values and overestimate smaller NDVI values when processing multi-pixel inputs. This discrepancy could be linked to the learning mechanism of these models, their complexity, and the choice of loss functions. This leads to the performance of the two in single-step forecasting being not as good as the simpler random forest and LSTM models, but this disadvantage can be easily ignored due to the excellent performance in multi-step forecasting (OpenAI, 2023). Future studies might need to employ different loss functions, use regularization techniques, or perform data augmentation to better represent extreme NDVI values in the training set. Also, tuning hyperparameters and adjusting the model architecture may help improve model performance in these scenarios. Also, because the data used in this study mixed a variety of climate and geographic types on a global scale, thus the performance of these models across different geographical regions and climates remains to be examined, which can add depth to our understanding of NDVI prediction.

In conclusion, our research underscores the importance of integrating both temporal and spatial dimensions in NDVI predictions. Models that succeed in integrating these dimensions, such as ConvLSTM and Transformer, outperform those that cannot, emphasizing the importance of the appropriation ate choice of predictive models for complex multidimensional tasks such as NDVI prediction.

5.2 The future expectations

The insights from this study bear significant implications for NDVI forecasting, environmental science, and the application of machine learning models in these areas. The

results emphasize the importance of incorporating both temporal and spatial information in NDVI prediction tasks. The superior performance of LSTM, ConvLSTM, and Transformer models over traditional ones like Random Forests in specific scenarios showcases the potential of these advanced models in enhancing prediction accuracy. This, in turn, contributes to efficient land use management and climate change mitigation strategies (Rodriguez-Galiano et al., 2012). Furthermore, this study expands the application of machine learning in environmental science, setting a foundation for future work in this field (OpenAI, 2023).

According to the results, there are several potential avenues for future exploration from this study. A significant point of interest lies in improving the robustness of machine learning models, particularly for multi-step NDVI prediction. It might be beneficial to explore the utility of hybrid models that amalgamate the strengths of different models or experiment with more complex training strategies to boost model performance over longer prediction intervals. Additionally, extending the comparison to other machine learning models and exploring alternative training methods like transfer learning could further enhance our understanding and optimization of NDVI prediction tasks.

While our study provides important insights into the effectiveness of various machine learning models for NDVI prediction, there remains scope for model enhancements. Future efforts should pay more attention to temporal dependencies and spatial information when selecting or designing machine learning models. Besides, despite our significant findings, this study does have limitations that future research could address. First of all, more than 15% of the samples have more than 50% missing values in the NDVI time series, which increases the difficulty of the model prediction. This research believes that this will make the choice of interpolation method very important. In this study, the PCHIP method that only considers temporal correlation is chosen when interpolating spatial-temporal data. In the future, the Autoregressive Integrated Moving Average model (ARIMA) or Krig method (OpenAI, 2023) that can consider both temporal and spatial dependencies will greatly improve the quality of the data. Moreover, this study only considered the forecasting in the future 10 timesteps, further exploration of model robustness over longer prediction periods and under different environmental conditions would enhance the versatility and practicality of these models.

The findings from our research can be applied in a variety of areas. The insights into the relative strengths and weaknesses of the different machine learning models for NDVI prediction can be used to improve land use management and climate change mitigation strategies. By incorporating these models into these areas, we could facilitate more accurate predictions of vegetation dynamics and thus make more informed decisions on land use and climate change mitigation. As the demand for accurate NDVI predictions continues to grow, our findings will help researchers and practitioners make more informed decisions about the best models and approaches to employ, which lead to better decision-making and more effective management of our natural resources (Fensholt & Proud, 2012a) (Pettorelli et al., 2005).

6 Bibliography

1 Ahmad, R., Yang, B., Ettlin, G., Berger, A., & Rodríguez-Bocca, P. (2023). A machine-learning based ConvLSTM architecture for NDVI forecasting. *International Transactions in Operational Research*, *30*(4), 2025–2048. https://doi.org/10.1111/itor.12887

2 Akbar, T. A., Hassan, Q. K., Ishaq, S., Batool, M., Butt, H. J., & Jabbar, H. (2019). Investigative Spatial Distribution and Modelling of Existing and Future Urban Land Changes and Its Impact on Urbanization and Economy. *Remote Sensing*, *11*(2), 105. https://doi.org/10.3390/rs11020105

3 Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

4 Bello, I., Zoph, B., Vaswani, A., Shlens, J., & Le, Q. V. (2019). *Attention Augmented Convolutional Networks*. 3286–3295. https://openaccess.thecvf.com/content_ICCV_2019/html/Bello_Attention_Augmented_Convolutional_Networks_ICCV_2019_paper.html

5 Bermúdez, J. D., Achanccaray, P., Sanches, I. D., Cue, L., Happ, P., & Feitosa, R. Q. (2017). Evaluation of recurrent neural networks for crop recognition from multitemporal remote sensing images. *Anais Do XXVII Congresso Brasileiro de Cartografia*, 800–804.

6 Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

7 Chuine, I. (2000). A Unified Model for Budburst of Trees. *Journal of Theoretical Biology*, *207*(3), 337–347. https://doi.org/10.1006/jtbi.2000.2178

8 Copernicus Climate Change Service. (2020). *E-OBS daily gridded meteorological data for Europe from 1950 to present derived from in-situ observations* [Data set]. ECMWF. https://doi.org/10.24381/CDS.151D3EC6

9 D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, & T. L. Veith. (2007). Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Transactions of the ASABE*, *50*(3), 885–900. https://doi.org/10.13031/2013.23153

10 Diaconu, C.-A., Saha, S., Gunnemann, S., & Xiang Zhu, X. (2022). Understanding the Role of Weather Data for Earth Surface Forecasting using a ConvLSTM-based Model. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1361–1370. https://doi.org/10.1109/CVPRW56347.2022.00142

11 Diaconu, C.-A., Saha, S., Günnemann, S., & Zhu, X. X. (2022). *Understanding the Role of Weather Data for Earth Surface Forecasting Using a ConvLSTM-Based Model*. 1362–1371. https://openaccess.thecvf.com/content/CVPR2022W/EarthVision/html/Diaconu_Understanding_the_Role_of_Weather_Data_for_Earth_Surface_Forecasting_CVPRW_2022_paper.html

12 Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., & Bargellini, P. (2012). Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment*, *120*, 25–36.

https://doi.org/10.1016/j.rse.2011.11.026
13 Fensholt, R., & Proud, S. R. (2012a). Evaluation of Earth Observation based global long term vegetation trends—Comparing GIMMS and MODIS global NDVI time series. *Remote Sensing of Environment*, *119*, 131–147. https://doi.org/10.1016/j.rse.2011.12.015
14 Fensholt, R., & Proud, S. R. (2012b). Evaluation of Earth Observation based global long term vegetation trends—Comparing GIMMS and MODIS global NDVI time series. *Remote Sensing of Environment*, *119*, 131–147. https://doi.org/10.1016/j.rse.2011.12.015
15 Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
16 He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. 770–778. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html
17 Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
18 Hutter, F., Lücke, J., & Schmidt-Thieme, L. (2015). Beyond Manual Tuning of Hyperparameters. *KI - Künstliche Intelligenz*, *29*(4), 329–337. https://doi.org/10.1007/s13218-015-0381-0
19 Insua, J. R., Utsumi, S. A., & Basso, B. (2019). Estimation of spatial and temporal variability of pasture growth and digestibility in grazing rotations coupling unmanned aerial vehicle (UAV) with crop simulation models. *PLOS ONE*, *14*(3), e0212773. https://doi.org/10.1371/journal.pone.0212773
20 Kisi, O. (2013). Modeling of Dissolved Oxygen in River Water Using Artificial Intelligence Techniques. *Journal of Environmental Informatics*, 92–101. https://doi.org/10.3808/jei.201300248
21 Kladny, K.-R., Milanta, M., Mraz, O., Hufkens, K., & Stocker, B. D. (2022). *Deep learning for satellite image forecasting of vegetation greenness* [Preprint]. Plant Biology. https://doi.org/10.1101/2022.08.16.504173
22 Knapp, A. K., Hoover, D. L., Wilcox, K. R., Avolio, M. L., Koerner, S. E., La Pierre, K. J., Loik, M. E., Luo, Y., Sala, O. E., & Smith, M. D. (2015). Characterizing differences in precipitation regimes of extreme wet and dry years: Implications for climate change experiments. *Global Change Biology*, *21*(7), 2624–2633. https://doi.org/10.1111/gcb.12888
23 Lin, Z., Li, M., Zheng, Z., Cheng, Y., & Yuan, C. (n.d.). *Self-Attention ConvLSTM for Spatiotemporal Prediction*.
24 Marzban, F., Sodoudi, S., & Preusker, R. (2018). The influence of land-cover type on the relationship between NDVI–LST and LST-Tair. *International Journal of Remote Sensing*, *39*(5), 1377–1398. https://doi.org/10.1080/01431161.2017.1402386
25 Maselli, F. (2004). Monitoring forest conditions in a protected Mediterranean coastal area by the analysis of multiyear NDVI data. *Remote Sensing of Environment*, *89*(4), 423–433. https://doi.org/10.1016/j.rse.2003.10.020
26 Moskolaï, W., Abdou, W., Dipanda, A., & Kolyang, D. T. (2020). *Application of LSTM architectures for next frame forecasting in Sentinel-1 images time series* (arXiv:2009.00841). arXiv. https://doi.org/10.48550/arXiv.2009.00841
27 Mulla, D. J. (2013). Twenty five years of remote sensing in precision agriculture:

Key advances and remaining knowledge gaps. *Biosystems Engineering*, *114*(4), 358–371. https://doi.org/10.1016/j.biosystemseng.2012.08.009

28 Pettorelli, N., Vik, J. O., Mysterud, A., Gaillard, J.-M., Tucker, C. J., & Stenseth, N. Chr. (2005). Using the satellite-derived NDVI to assess ecological responses to environmental change. *Trends in Ecology & Evolution*, *20*(9), 503–510. https://doi.org/10.1016/j.tree.2005.05.011

29 Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M.-L., Chen, S.-C., & Iyengar, S. S. (2018). A Survey on Deep Learning: Algorithms, Techniques, and Applications. *ACM Computing Surveys*, *51*(5), 92:1-92:36. https://doi.org/10.1145/3234150

30 Prăvălie, R., Sîrodoev, I., Nita, I.-A., Patriche, C., Dumitraşcu, M., Roşca, B., Tişcovschi, A., Bandoc, G., Săvulescu, I., Mănoiu, V., & Birsan, M.-V. (2022). NDVI-based ecological dynamics of forest vegetation and its relationship to climate change in Romania during 1987–2018. *Ecological Indicators*, *136*, 108629. https://doi.org/10.1016/j.ecolind.2022.108629

31 Reddy, D. S., & Prasad, P. R. C. (2018). Prediction of vegetation dynamics using NDVI time series data and LSTM. *Modeling Earth Systems and Environment*, *4*(1), 409–419. https://doi.org/10.1007/s40808-018-0431-3

32 Requena-Mesa, C., Benson, V., Reichstein, M., Runge, J., & Denzler, J. (2021). *EarthNet2021: A large-scale dataset and challenge for Earth surface forecasting as a guided video prediction task* (arXiv:2104.10066). arXiv. http://arxiv.org/abs/2104.10066

33 Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *67*, 93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002

34 Schowengerdt, R. A. (2006). *Remote Sensing: Models and Methods for Image Processing*. Elsevier.

35 SHI, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., & WOO, W. (2015a). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in Neural Information Processing Systems*, *28*. https://proceedings.neurips.cc/paper/2015/hash/07563a3fe3bbe7e3ba84431ad9d055af-Abstract.html

36 SHI, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., & WOO, W. (2015b). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in Neural Information Processing Systems*, *28*. https://proceedings.neurips.cc/paper/2015/hash/07563a3fe3bbe7e3ba84431ad9d055af-Abstract.html

37 Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, *8*(2), 127–150. https://doi.org/10.1016/0034-4257(79)90013-0

38 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

39 Wulder, M. A., White, J. C., Loveland, T. R., Woodcock, C. E., Belward, A. S.,

Cohen, W. B., Fosnight, E. A., Shaw, J., Masek, J. G., & Roy, D. P. (2016). The global Landsat archive: Status, consolidation, and direction. *Remote Sensing of Environment*, *185*, 271–283. https://doi.org/10.1016/j.rse.2015.11.032

40 Xie, Y., Sha, Z., & Yu, M. (2008). Remote sensing imagery in vegetation mapping: A review. *Journal of Plant Ecology*, *1*(1), 9–23. https://doi.org/10.1093/jpe/rtm005

41 Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., & Zhang, L. (2020). Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, *241*, 111716. https://doi.org/10.1016/j.rse.2020.111716

42 Zhang, Y., He, Y., Li, Y., & Jia, L. (2022). Spatiotemporal variation and driving forces of NDVI from 1982 to 2015 in the Qinba Mountains, China. *Environmental Science and Pollution Research*, *29*(34), 52277–52288. https://doi.org/10.1007/s11356-022-19502-6

43 Zhang, Y., Jiang, X., Lei, Y., & Gao, S. (2022). The contributions of natural and anthropogenic factors to NDVI variations on the Loess Plateau in China during 2000–2020. *Ecological Indicators*, *143*, 109342. https://doi.org/10.1016/j.ecolind.2022.109342

44 OpenAI. (2023). ChatGPT (April 20 version) [Large language model]. https://chat.openai.com/

Use of ChatGPT (or any other AI Writing Assistance) – Form to be completed

Student name: **Zhiqi Wang**

Student number: **r0822618**

Please indicate with "X" whether it relates to a course assignment or to the master thesis:

O This form is related to a **course assignment**.

Course name: .....

Course number: .....

X This form is related to **my Master thesis**.

Title Master thesis: Integration of spatial-temporal context in remote sensing image classification

Promotor: Prof. Stef Lhermitte

Please indicate with "X":

O **I did not use** ChatGPT or any other AI Writing Assistance.

X I did use AI Writing Assistance. In this case **specify which one** (e.g. ChatGPT/GPT4/...):

GPT4 and GPT-3.5

.....

Please indicate with "X" (possibly multiple times) in which way you were using it:

X Assistance purely with the language of the paper

- *Code of conduct*: This use is similar to using a spelling checker

X As a search engine to learn on a particular topic

- *Code of conduct*: This use is similar to e.g. a google search or checking Wikipedia. Be aware that the output of Chatbot evolves and may change over time.

X For literature search

- *Code of conduct*: This use is comparable to e.g. a google scholar search. However, be aware that ChatGPT may output no or wrong references. As a student you are responsible for further checking and verifying the absence or correctness of references.

O **For short-form input assistance**

- *Code of conduct*: This use is similar to e.g. google docs powered by generative language models

X To let generate programming code

- *Code of conduct*: Correctly mention the use of ChatGPT and cite it. You can also ask ChatGPT how to cite it.

X To let generate new research ideas

- *Code of conduct*: Further verify in this case whether the idea is novel or not. It is likely that it is related to existing work, which should be referenced then.

O **To let generate blocks of text**

- *Code of conduct*: Inserting blocks of text without quotes from ChatGPT to your report or thesis is not allowed. According to Article 84 of the exam regulations in evaluating your work one should be able to correctly judge on your own knowledge.

In case it is really needed to insert a block of text from ChatGPT, mention it as a citation by using quotes. But this should be kept to an absolute minimum.

O **Other**

- *Code of conduct*: Contact the professor of the course or the promotor of the thesis. Inform also the program director. Motivate how you comply with Article 84 of the exam regulations. Explain the use and the added value of ChatGPT or other AI tool: ….

Further important guidelines and remarks

- ChatGPT cannot be used related **to data or subjects under NDA agreement.**
- ChatGPT cannot be used related **to sensitive or personal data due to privacy issues**.
- **Take a scientific and critical attitude** when interacting with ChatGPT and interpreting its output. Don't become emotionally connected to AI tools.
- As a student you are responsible to comply with Article 84 of the exam regulations: your report or thesis should reflect your own knowledge. Be aware that plagiarism rules also apply to the use of ChatGPT or any other AI tools.
- **Exam regulations Article 84**: "Every conduct individual students display with which they (partially) inhibit or attempt to inhibit a correct judgement of their own knowledge, understanding and/or skills or those of other students, is considered an irregularity which may result in a suitable penalty. A special type of irregularity is plagiarism, i.e. copying the work (ideas, texts, structures, designs, images, plans, codes , ...) of others or prior personal work in an exact or slightly modified way without adequately acknowledging the sources. Every possession of prohibited resources during an examination (see article 65) is considered an irregularity."
- **ChatGPT suggestion about citation**: "Citing and referencing ChatGPT output is essential to maintain academic integrity and avoid plagiarism. Here are some guidelines on how to correctly cite and reference ChatGPT in your Master's thesis: 1. Citing ChatGPT: Whenever you use a direct quote or paraphrase from ChatGPT, you should include an in-text citation that indicates the source. For example: (ChatGPT, 2023). 2. Referencing ChatGPT: In the reference list at the end of your thesis, you should include a full citation for ChatGPT. This should include the title of the AI language model, the year it was

published or trained, the name of the institution or organization that developed it, and the URL or DOI (if available). For example: OpenAI. (2021). GPT-3 Language Model. https://openai.com/blog/gpt-3-apps/ 3. Describing the use of ChatGPT: You may also want to describe how you used ChatGPT in your research methodology section. This could include details on how you accessed ChatGPT, the specific parameters you used, and any other relevant information related to your use of the AI language model. Remember, it is important to adhere to your institution's specific guidelines for citing and referencing sources in your Master's thesis. If you are unsure about how to correctly cite and reference ChatGPT or any other source, consult with your thesis advisor or a librarian for guidance."

Additional reading

ACL 2023 Policy on AI Writing Assistance: https://2023.aclweb.org/blog/ACL-2023-policy/

KU Leuven guidelines on citing and referencing Generative AI tools, and other information:https://www.kuleuven.be/english/education/student/educational-tools/generative-artificial-intelligence

Dit formulier werd opgesteld voor studenten in de Master of Artificial intelligence. Ze bevat een code of conduct, die we bij universiteitsbrede communicatie rond onderwijs verder wensen te hanteren. Deze template samen met de code of conduct zal in de toekomst nog verdere aanpassingen behoeven. Het schept alvast een kader voor de 2de en de 3de examenperiode van 2022-2023.