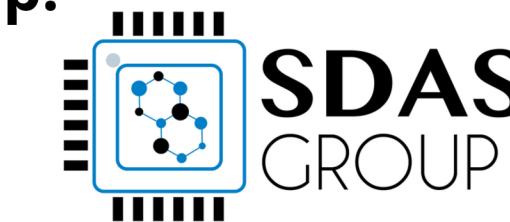


# PFE Progress Follow-Up

## Salma OUMOUSSA

# Reminders

- I am currently undertaking my internship at UM6P College of Computing, in collaboration with the SDAS Group.



- I am being supervised by Professor Diego Peluffo, who can be reached via [diego.peluffo@sdas-group.com](mailto:diego.peluffo@sdas-group.com).
- My research project is centered around **crop classification**, which involves developing and applying machine learning models to classify different crop types based on satellite imagery and time-series data. The aim is to improve the accuracy and efficiency of crop identification in agricultural fields

# Agenda

## Technical-Focused Work

1. Dataset Overview for Crop Classification
2. Data Preprocessing
3. Model Architecture for Crop Classification  
(EarlyRNN)
4. Training Performance and Results Metrics

## Research-Focused Work

1. Research direction
2. Current stage

# This work is within the scope of this project

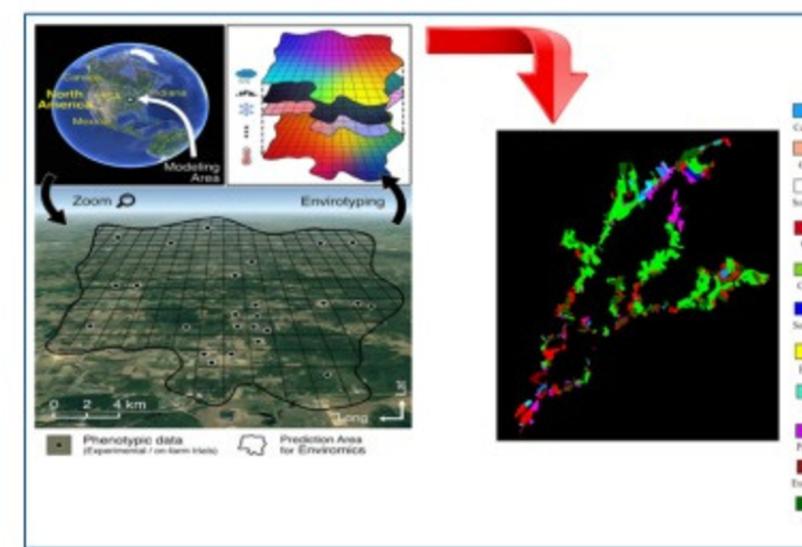


## **CropID:** Precision crop classification platform for tailored fertilizer recommendations through data integration

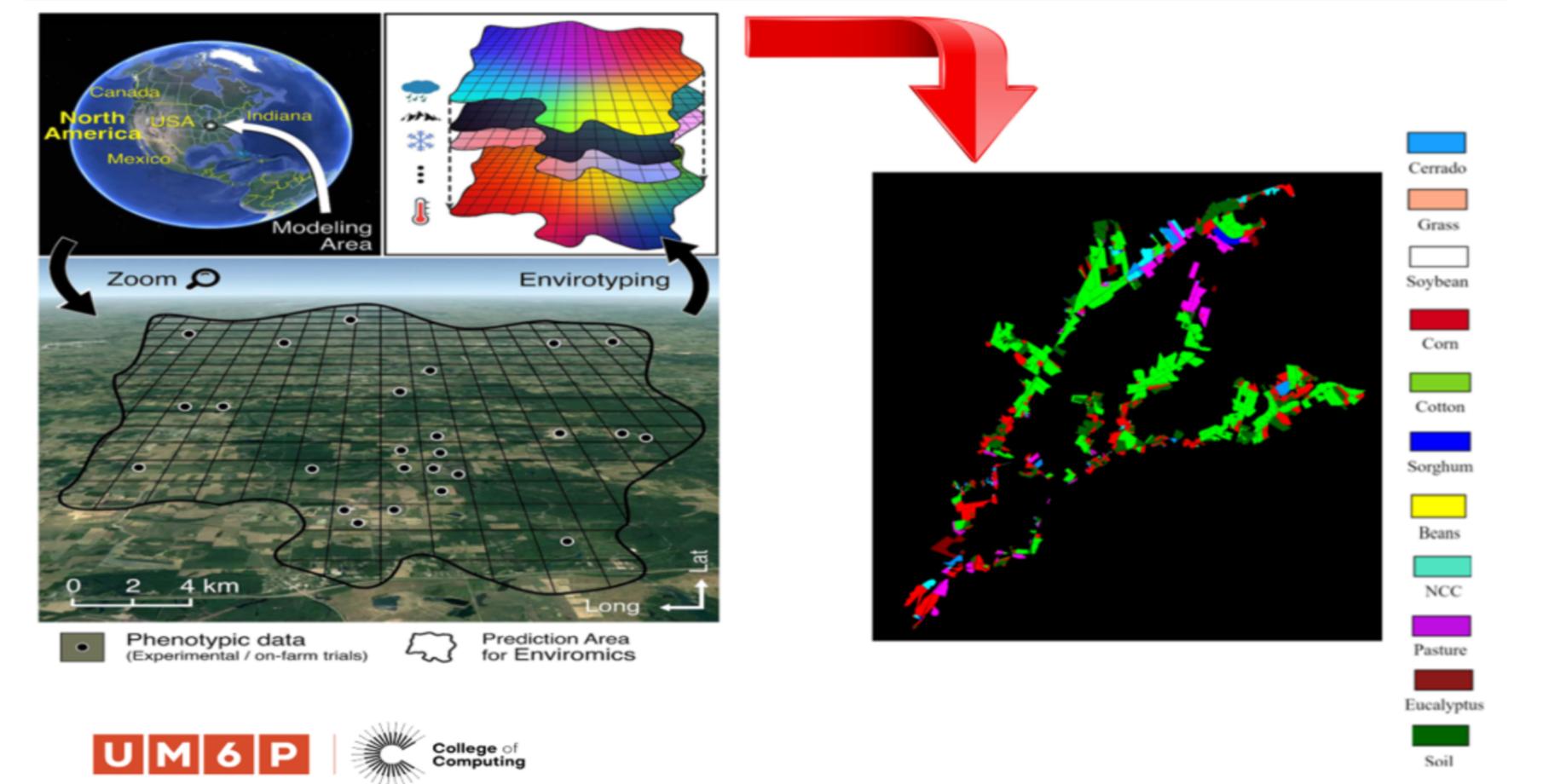
Project proposal –

Version: Dec 23, 2024

Accord spécifique N° FN25  
entre OCP NUTRICROPS et Université Mohammed VI Polytechnique



The **CropID** project integrates multi-modal data sources—including **remote sensing**, **soil health metrics**, and **crop phenology**—to accurately **classify and/or predict crop traits/varieties** and optimize agricultural productivity. A key aspect of the project is delivering precise fertilizer use recommendations, ensuring nutrients are applied at the right time and in the right amounts, tailored to specific crop needs and growth stages, thereby maximizing efficiency and sustainability.





# Technical-Focused Work

# Dataset Overview for Crop Classification

The BavarianCrops dataset is a satellite-based time-series collection designed to support crop classification. It includes temporal patterns of spectral reflectance across agricultural parcels in Bavaria, Germany. This dataset is used for training and evaluating early classification models that detect crop types based on time-series data.

## Data Characteristics:

- **Number of samples:** 16,600
- **Sequence length:** Fixed at 70 time steps
- **Number of features per sample:** 13 (Sentinel-2 spectral bands)
- **Training Set:** 13,280 samples
- **Test Set:** 3,320 samples

# Dataset Overview for Crop Classification

## Mathematical Representation

$$\mathbf{X} \in \mathbb{R}^{N \times T \times D}, \quad N = 16,600, \quad T = 70, \quad D = 13$$

Where:

- $N$  is the number of samples,
- $T$  is the sequence length after preprocessing (fixed at  $T = 70$ ),
- $D$  is the number of features (13 Sentinel-2 spectral bands).

## Example matrix :

$$\mathbf{X} = \begin{bmatrix} 0.609 & 0.004 & 0.182 & 0.159 & 0.573 & 0.478 & 0.514 & 0.526 & 0.532 & 0.540 & 0.522 & 0.538 & 0.228 \\ 0.458 & 0.025 & 0.274 & 0.214 & 0.414 & 0.346 & 0.365 & 0.363 & 0.372 & 0.388 & 0.359 & 0.392 & 0.127 \\ \vdots & \vdots \\ 0.191 & 0.001 & 0.107 & 0.066 & 0.146 & 0.109 & 0.100 & 0.112 & 0.130 & 0.141 & 0.134 & 0.151 & 0.056 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} & \text{Feature 1} & \text{Feature 2} & \text{Feature 3} & \cdots & \text{Feature 13} \\ x_1 & 0.609 & 0.004 & 0.182 & \cdots & 0.228 \\ x_2 & 0.458 & 0.025 & 0.274 & \cdots & 0.127 \\ x_3 & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_N & 0.191 & 0.001 & 0.107 & \cdots & 0.056 \end{bmatrix}$$

# Dataset Overview for Crop Classification

## Crop Types and Dataset Composition

The dataset consists of 7 distinct crop types, each represented by an integer label:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad y_i \in \{0, 1, \dots, 6\}$$

—————>

0 → Meadow
1 → Summer Barley
2 → Corn
3 → Winter Wheat
4 → Winter Barley
5 → Clover
6 → Winter Triticale

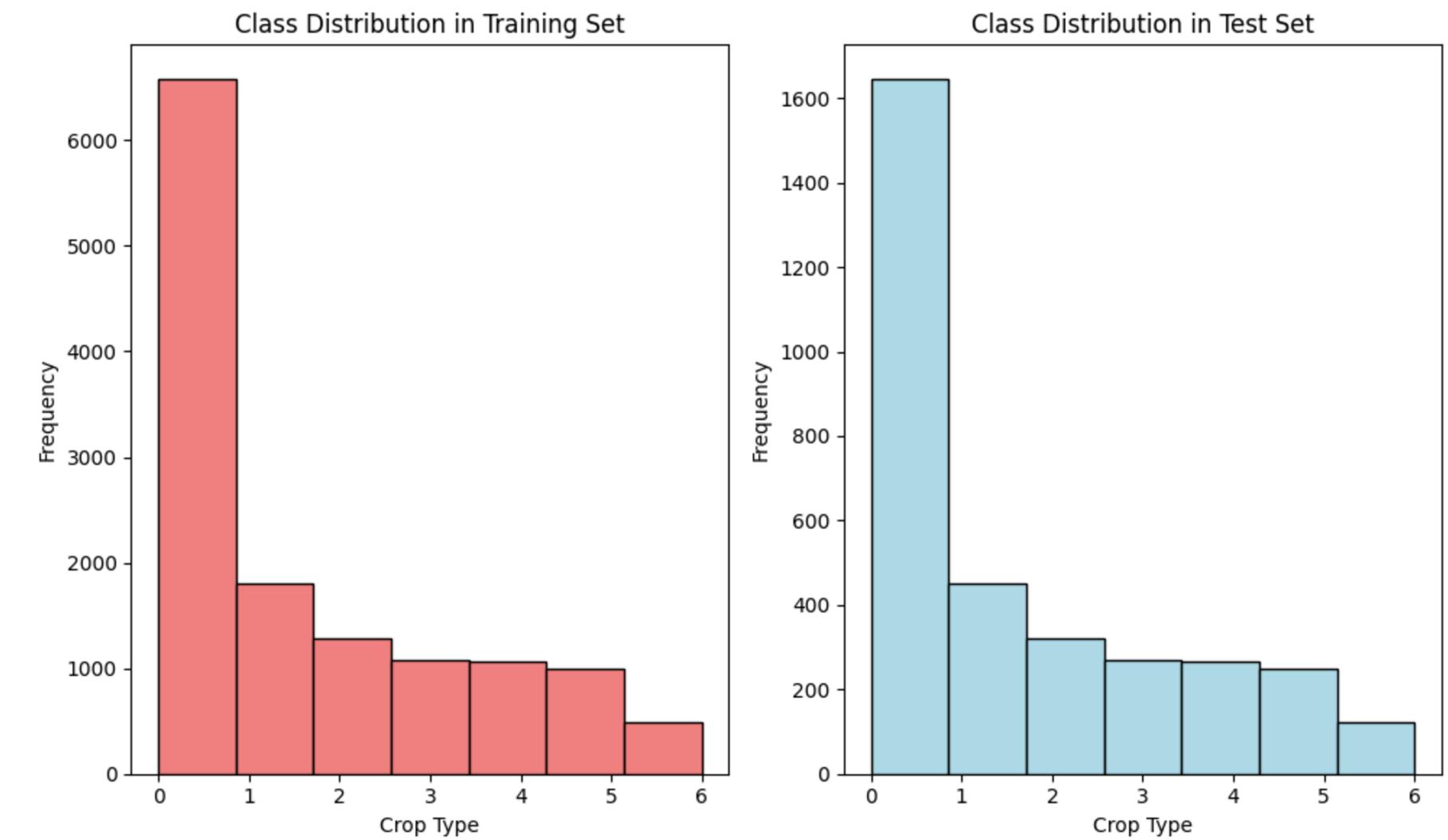
# Dataset Overview for Crop Classification

## Distribution of Crop Types in Training and Test Sets: Imbalance and Variability

The dataset is imbalanced, with certain crop types dominating the samples. This imbalance varies between the training and test sets.

Table 1: Crop label mappings and sample distribution.

Crop Type	Label	Train Samples	Test Samples
Meadow	0	6574	1644
Summer Barley	1	1801	450
Corn	2	1279	320
Winter Wheat	3	1072	268
Winter Barley	4	1070	267
Clover	5	1000	250
Winter Triticale	6	484	121



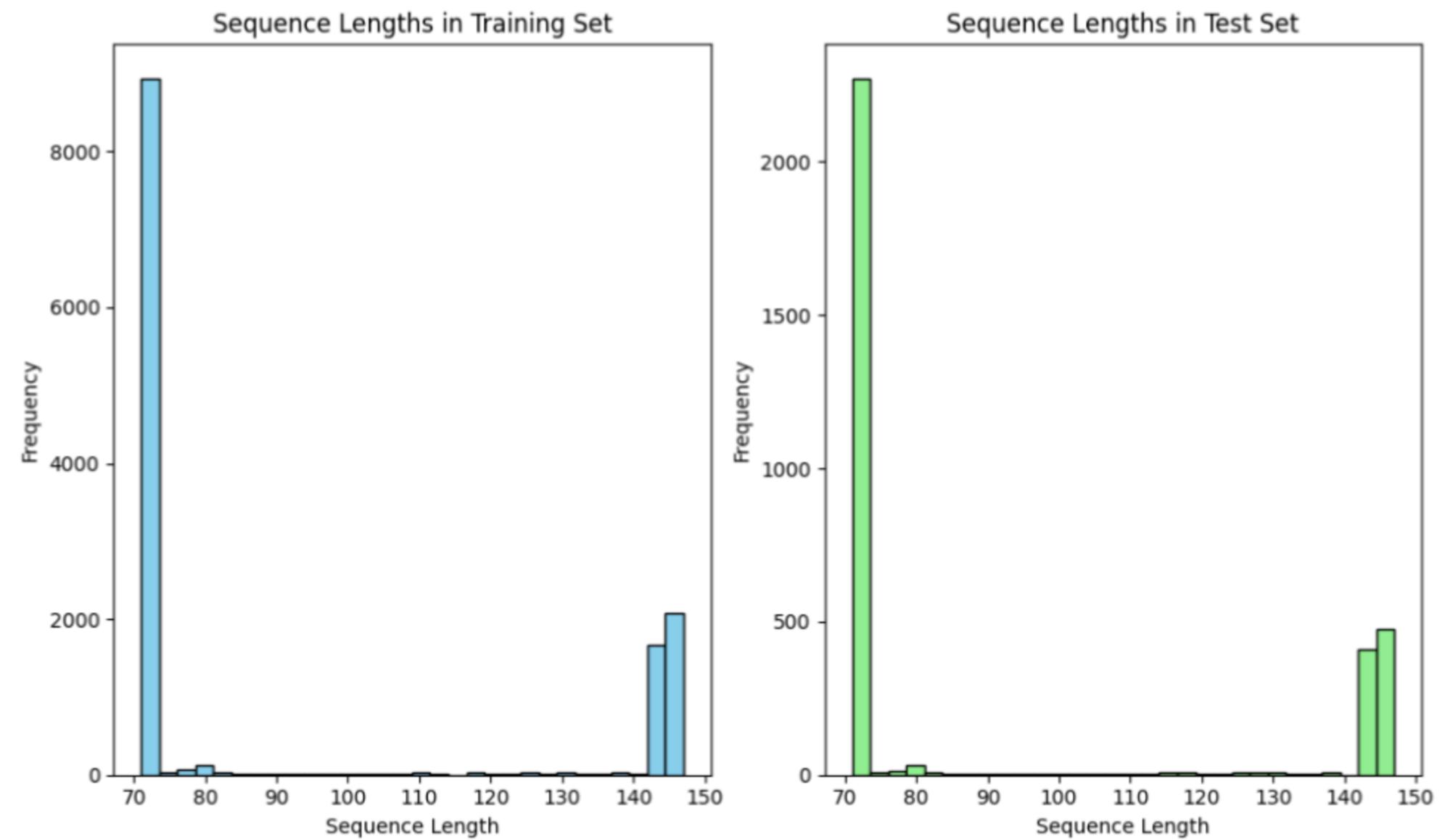
Class distribution in the training and test sets. The imbalance is more noticeable in the training set, where certain crop types are heavily overrepresented

# Dataset Overview for Crop Classification

## Distribution of Sequence Lengths in Training and Test Sets: Variability and Consistency

The dataset has sequences of different lengths, with some being shorter and others longer. This difference is seen in both the training and test sets, although the distributions are slightly different.

- Training sequences: min = 71, max = 147, mean = 93.29
- Test sequences: min = 71, max = 147, mean = 92.27

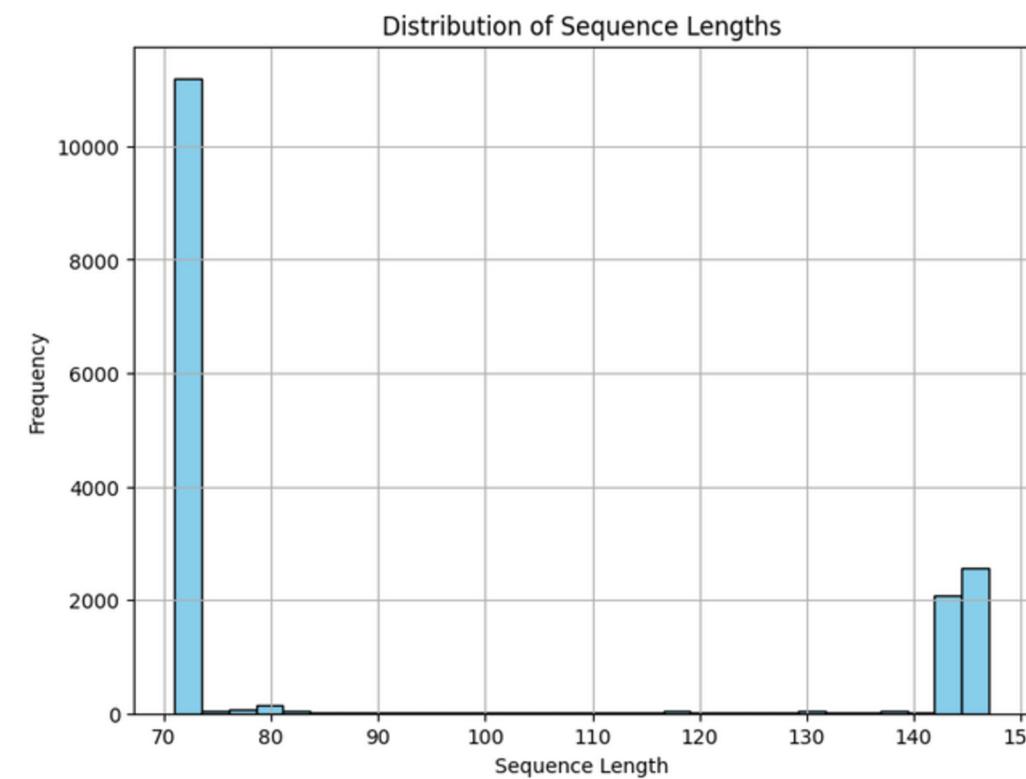


# Dataset Overview for Crop Classification

## Core Challenges

### Variable Sequence Lengths

Time series have varying lengths, making it difficult to use models that require fixed-length inputs



### Class Imbalance

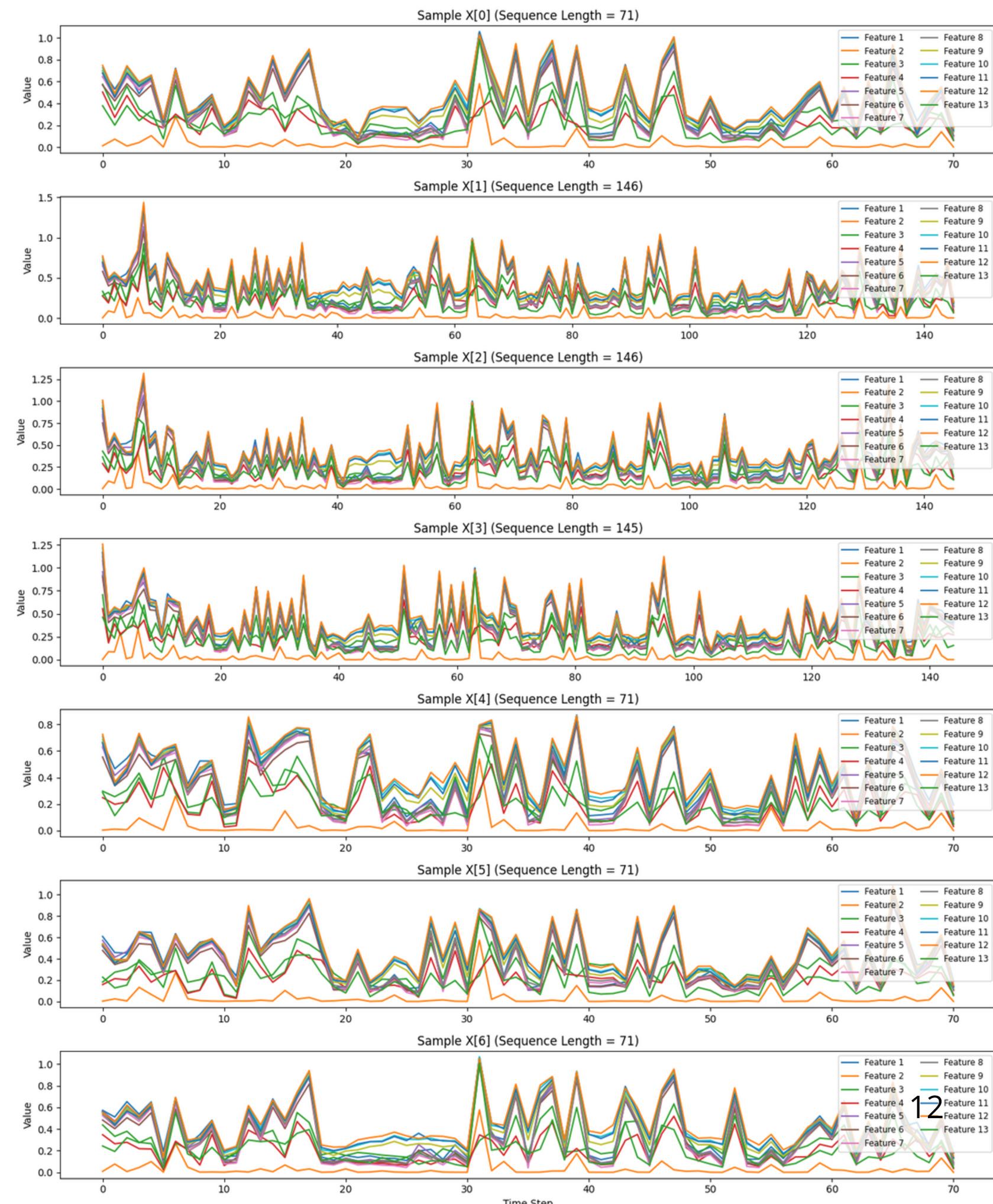
Some crop types are underrepresented, leading to biased model predictions and poor generalization on minority classes.



**Both of these challenges will be addressed during the preprocessing stage to ensure better model performance.**

# Temporal Visualization of Samples

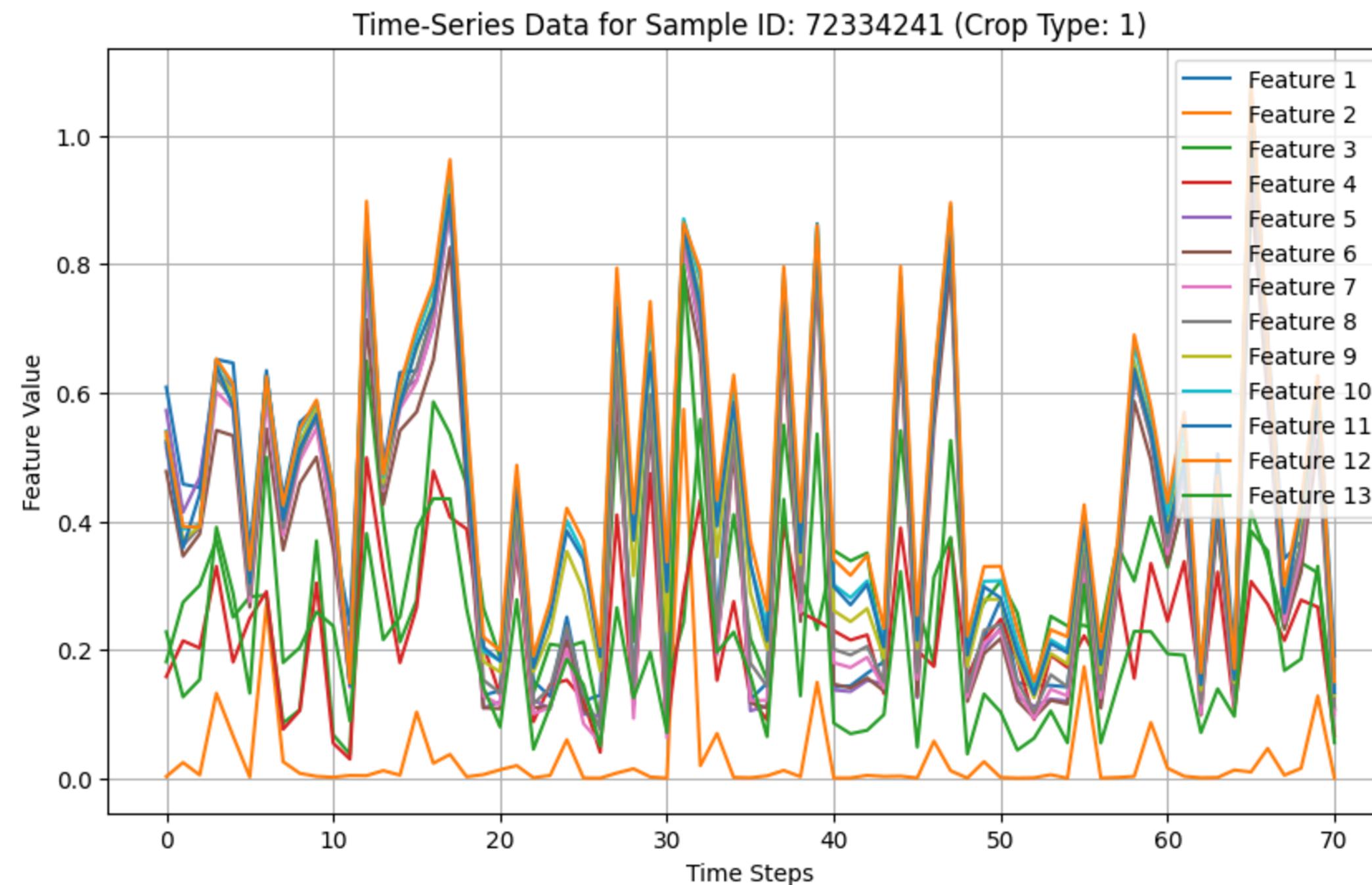
- Time-series feature trends across 7 randomly selected samples from the dataset.
- Each subplot corresponds to one sample, displaying all 13 spectral features over time.
- This visualization highlights the variability in sequence lengths and the temporal dynamics captured in the raw data.



# Temporal Visualization of Samples

## Time-series data for a sample parcel (Crop Type: Summer Barley).

The plot displays the time-series data for a sample parcel (Crop Type: Summer Barley). The data consists of 13 features, each representing a different spectral band of the crop field. The x-axis represents the time steps, ranging from 0 to 70, while the y-axis shows the normalized feature values at each time step.



# Data Preprocessing: Sequence Length Handling

## Sequence Length Handling

**Problem:** The dataset contains sequences of varying lengths, ranging from 71 to 147 time steps. This variability is problematic for machine learning models, which require uniform input lengths.

**Solution:** To ensure uniformity, we perform the following preprocessing steps:

- **Padding:** Sequences shorter than 70 time steps are padded with zeros to a fixed length of 70 time steps.
- **Trimming:** Sequences longer than 70 time steps are truncated, keeping only the first 70 time steps.

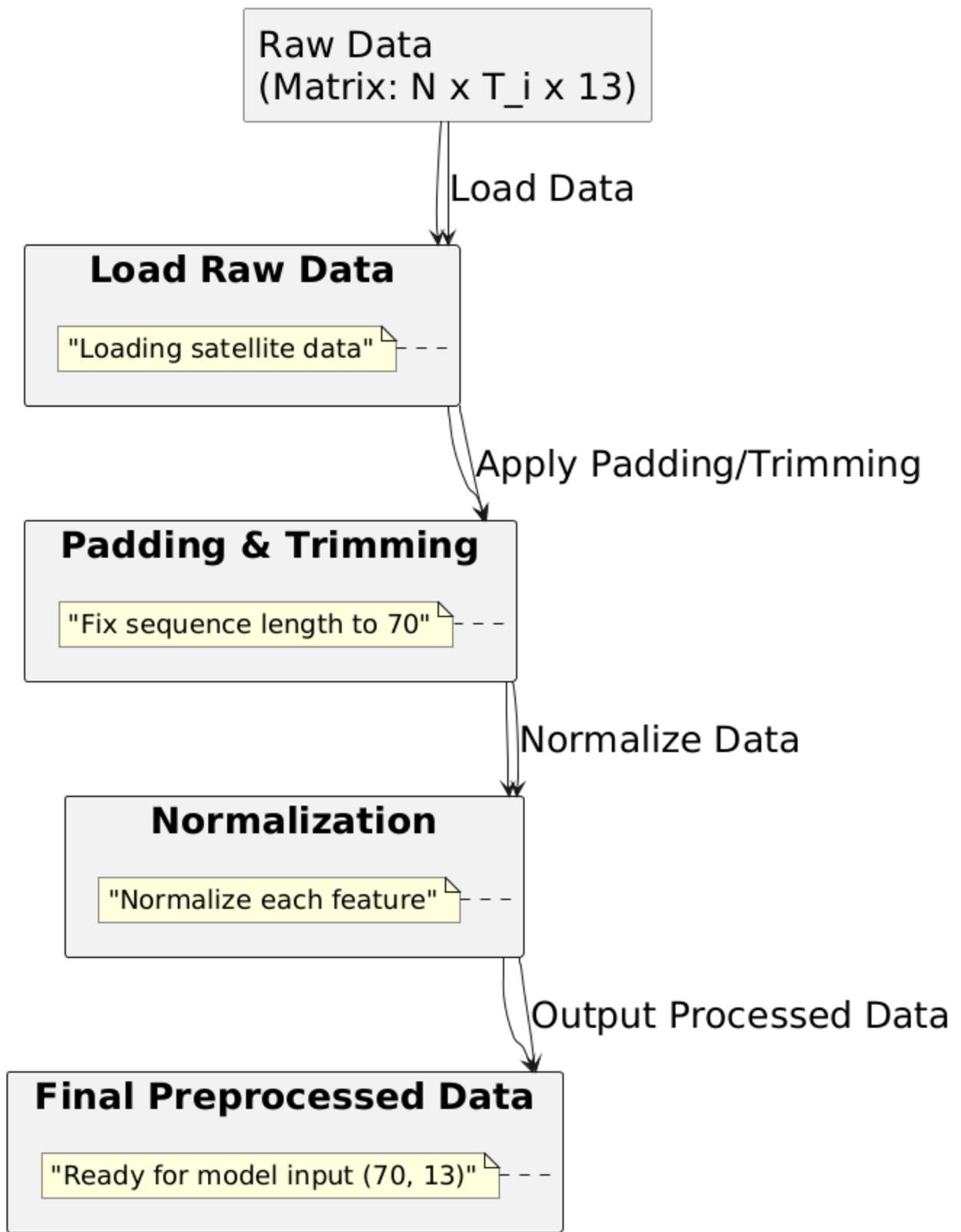
**Effect:** This preprocessing step ensures that all input data has consistent dimensions, enabling efficient training of the model with fixed input sizes.

# Data Preprocessing: Handling Class Imbalance

## Handling Class Imbalance

- Problem:** The dataset is imbalanced, with some crop types (e.g., "Meadow") being overrepresented, while others (e.g., "Winter Triticale") are underrepresented. This imbalance can cause the model to be biased towards the majority classes.
- Solution:** To address the imbalance, we apply class weights in the loss function:  
**Class Weights:** We compute weights inversely proportional to the class frequency, giving more importance to underrepresented classes during training. This helps the model focus more on the minority classes.
- Effect:** By applying class weights, we ensure that the model does not become biased toward the majority class, and learns to classify all crop types more equally.

# Recap | Data preprocessing



# Dataset Overview for after preprocessing

## Mathematical Representation

- Each section represents one crop with its respective number of samples (6574 for Crop 0, 1801 for Crop 1, etc.).
- Each crop section contains 70 time steps, with each time step having 13 features.

Crop 0 (Meadow)	Feature 1	Feature 2	...	Feature 13
$x_1$	0.609	0.004	...	0.228
$x_2$	0.458	0.025	...	0.127
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$x_{6574}$	0.191	0.001	...	0.056
Crop 1 (Summer Barley)	Feature 1	Feature 2	...	Feature 13
$x_1$	0.550	0.023	...	0.389
$x_2$	0.433	0.019	...	0.322
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$x_{1801}$	0.220	0.002	...	0.065
Crop 2 (Corn)	Feature 1	Feature 2	...	Feature 13
$x_1$	0.513	0.018	...	0.426
$x_2$	0.432	0.024	...	0.356
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$x_{1279}$	0.184	0.003	...	0.112
Crop 3 (Winter Wheat)	Feature 1	Feature 2	...	Feature 13
$x_1$	0.567	0.025	...	0.439
$x_2$	0.492	0.020	...	0.378
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$x_{1072}$	0.234	0.007	...	0.190
Crop 4 (Winter Barley)	Feature 1	Feature 2	...	Feature 13
$x_1$	0.518	0.022	...	0.426
$x_2$	0.445	0.016	...	0.366
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$x_{1070}$	0.198	0.003	...	0.137
Crop 5 (Clover)	Feature 1	Feature 2	...	Feature 13
$x_1$	0.523	0.021	...	0.439
$x_2$	0.485	0.018	...	0.395
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$x_{1000}$	0.209	0.002	...	0.152
Crop 6 (Winter Triticale)	Feature 1	Feature 2	...	Feature 13
$x_1$	0.552	0.021	...	0.421
$x_2$	0.487	0.019	...	0.393
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$x_{484}$	0.192	0.003	...	0.101

# Dataset Overview for after preprocessing

Meadow	Summer Barley	Corn	Winter Wheat	Winter Barley	Clover	Winter Triticale		
x <sub>1</sub>	x <sub>1</sub>	x <sub>1</sub>	x <sub>1</sub>	x <sub>1</sub>	x <sub>1</sub>	x <sub>1</sub>		
x <sub>2</sub>	x <sub>2</sub>	x <sub>2</sub>	x <sub>2</sub>	x <sub>2</sub>	x <sub>2</sub>	x <sub>2</sub>		
x <sub>3</sub>	x <sub>3</sub>	x <sub>3</sub>	x <sub>3</sub>	x <sub>3</sub>	x <sub>3</sub>	x <sub>3</sub>		
Culture / Échantillon		Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	...	Feature 13
Crop 0 (Meadow) - 6574 échantillons								
x <sub>1</sub>	0.497	0.944	0.635	0.755	0.564	...	0.195	
x <sub>2</sub>	0.896	0.092	0.817	0.475	0.053	...	0.624	
x <sub>3</sub>	0.895	0.858	0.357	0.088	0.851	...	0.192	
Crop 1 (Summer Barley) - 1801 échantillons								
x <sub>1</sub>	0.522	0.571	0.569	0.755	0.729	...	0.914	
x <sub>2</sub>	0.510	0.121	0.492	0.369	0.381	...	0.824	
x <sub>3</sub>	0.181	0.298	0.398	0.134	0.422	...	0.931	
Crop 2 (Corn) - 1279 échantillons								
x <sub>1</sub>	0.560	0.270	0.625	0.762	0.468	...	0.587	
x <sub>2</sub>	0.931	0.786	0.375	0.447	0.039	...	0.707	
x <sub>3</sub>	0.061	0.905	0.318	0.277	0.702	...	0.219	
Crop 3 (Winter Wheat) - 1072 échantillons								
x <sub>1</sub>	0.567	0.147	0.201	0.426	0.044	...	0.048	
x <sub>2</sub>	0.731	0.896	0.702	0.606	0.964	...	0.586	
x <sub>3</sub>	0.879	0.489	0.035	0.162	0.978	...	0.736	
Crop 4 (Winter Barley) - 1070 échantillons								
x <sub>1</sub>	0.702	0.598	0.160	0.245	0.583	...	0.342	
x <sub>2</sub>	0.222	0.063	0.158	0.958	0.154	...	0.694	
x <sub>3</sub>	0.049	0.661	0.678	0.890	0.823	...	0.357	
Crop 5 (Clover) - 1000 échantillons								
x <sub>1</sub>	0.392	0.544	0.008	0.607	0.306	...	0.510	
x <sub>2</sub>	0.050	0.353	0.852	0.460	0.513	...	0.077	
x <sub>3</sub>	0.532	0.004	0.048	0.944	0.986	...	0.365	
Crop 6 (Winter Triticale) - 484 échantillons								
x <sub>1</sub>	0.356	0.022	0.275	0.148	0.466	...	0.550	
x <sub>2</sub>	0.802	0.583	0.402	0.026	0.535	...	0.455	
x <sub>3</sub>	0.931	0.647	0.848	0.180	0.637	...	0.137	

# Model Architecture for Crop Classification (EarlyRNN)

- EarlyRNN model for time-series crop classification using satellite data.
- LSTM-based architecture with a stopping mechanism for early prediction.

## Model Components:

- Input Transformation Layer:
  - LayerNorm: Normalizes the input features.
  - Linear Transformation: Projects input to 64 dimensions.
- LSTM Backbone:
  - 2-layer LSTM network for sequential data processing.
- Classification Head:
  - Outputs class probabilities for each of the 7 crop types.
- Stopping Decision Head:
  - Decides whether to make an early prediction or continue processing.

# Model Architecture for Crop Classification (EarlyRNN)

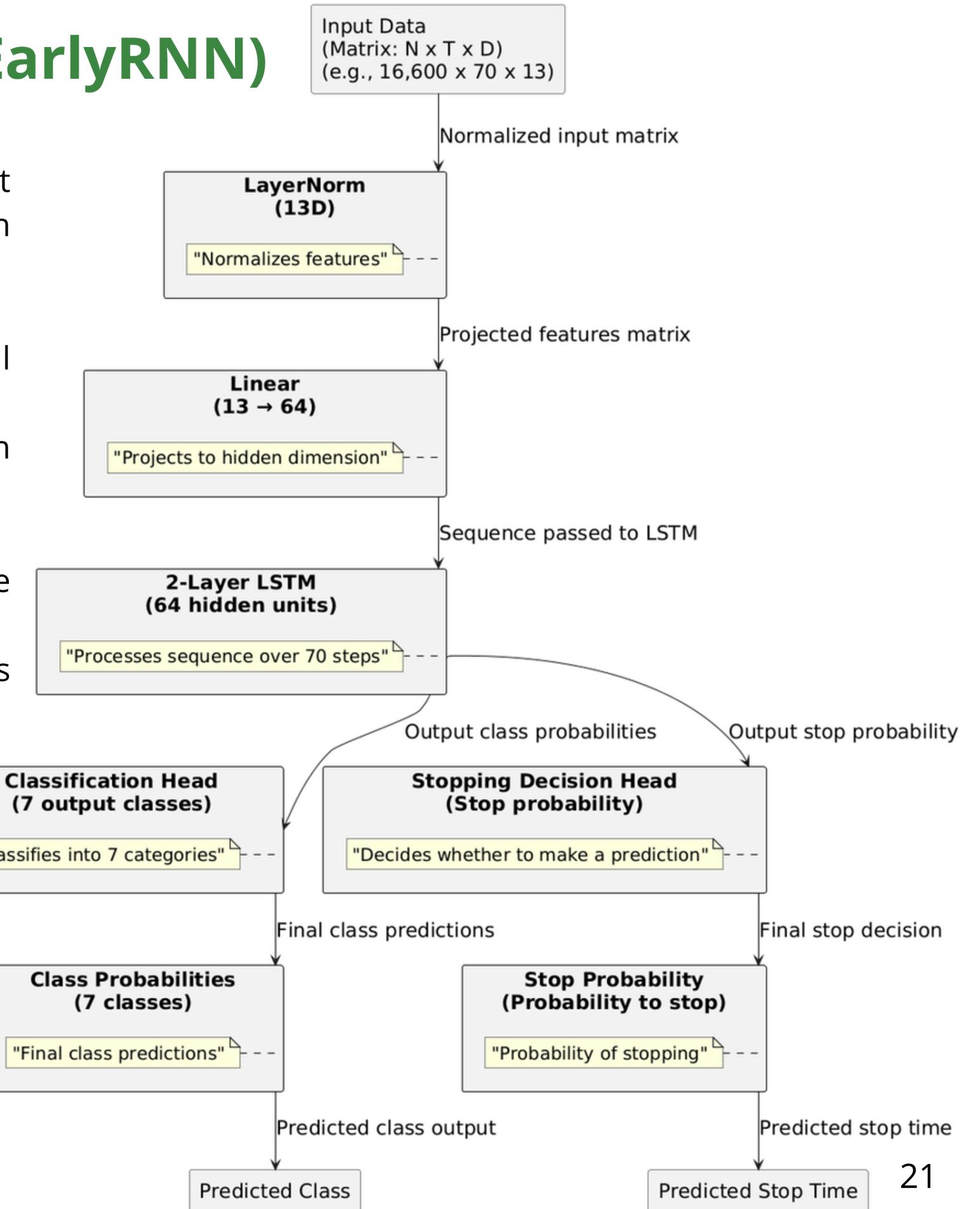
```
● ● ●

1 EarlyRNN(
2     (intransforms):
3         Sequential(
4             (0): LayerNorm((13,), eps=1e-05, elementwise_affine=True)
5             (1): Linear(in_features=13, out_features=64, bias=True)
6         )
7     (backbone):
8         LSTM(64, 64, num_layers=2, bias=False, batch_first=True, dropout=0.2)
9     (classification_head):
10        ClassificationHead(
11            (projection):
12                Sequential(
13                    (0): Linear(in_features=64, out_features=7, bias=True)
14                    (1): LogSoftmax(dim=2)
15                )
16        )
17    (stopping_decision_head):
18        DecisionHead(
19            (projection):
20                Sequential(
21                    (0): Linear(in_features=64, out_features=1, bias=True)
22                    (1): Sigmoid()
23                )
24        )
25    )
26 )
```

# Model Architecture for Crop Classification (EarlyRNN)

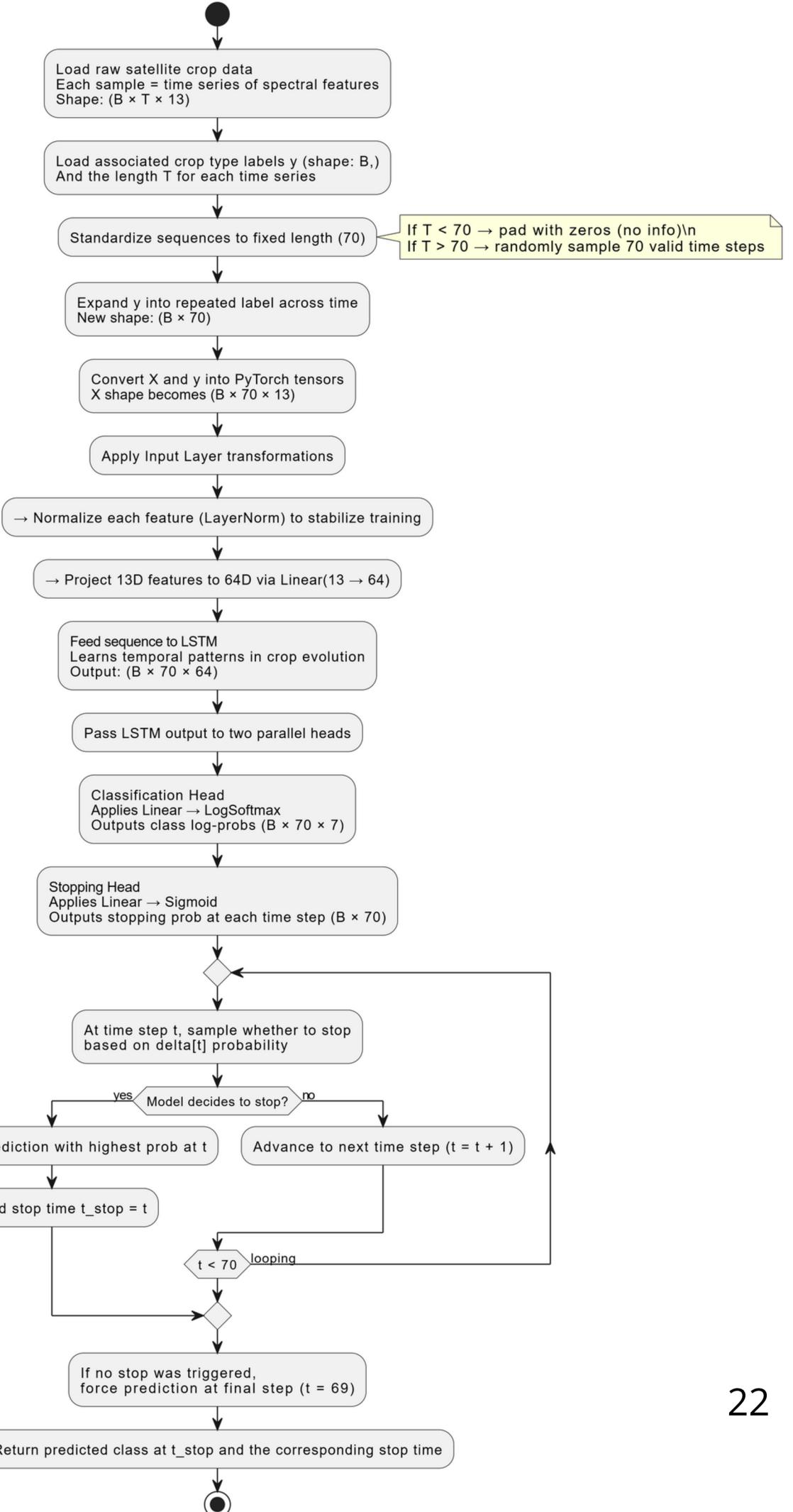
The EarlyRNN architecture is designed for early time-series classification. It processes time-series data from remote sensing (e.g., crop classification) through several key steps:

- 1. Input Normalization:** Normalizes features across the 13 spectral bands.
- 2. Linear Transformation:** Projects the 13 input features into a 64-dimensional space.
- 3. 2-Layer LSTM:** Processes the sequence of features over 70 time steps to learn temporal patterns.
- 4. Classification Head:** Outputs class probabilities for the 7 crop types.
- 5. Stopping Decision Head:** Determines if the model should stop early and make a prediction.
- 6. Final Output:** Predicts the class and stop time based on the highest class probability and stop probability.



# Data Preprocessing and Model Flow

- Overview of the data preprocessing and model flow, illustrating the steps from data loading, preprocessing, LSTM processing, and dynamic stopping mechanism to final predictions.



# Model Training

## Loss Function:

- EarlyRewardLoss: Balances classification accuracy and early predictions, rewarding early correct predictions.

## Optimizer:

- AdamW: Optimizer with weight decay for improved generalization.

## Hyperparameters:

- Learning Rate: 0.001 (with possible scheduler)
- Weight Decay: 0.01 (for regularization)

Batch Size: 64 (balance between memory and computation)

## Evaluation Metrics:

- Accuracy: Measures overall classification success.
- Precision, Recall, F1: Performance for each crop type.
- Earliness Metric: Time step at which the model predicts.
- Balanced Accuracy: Accounts for class imbalance.

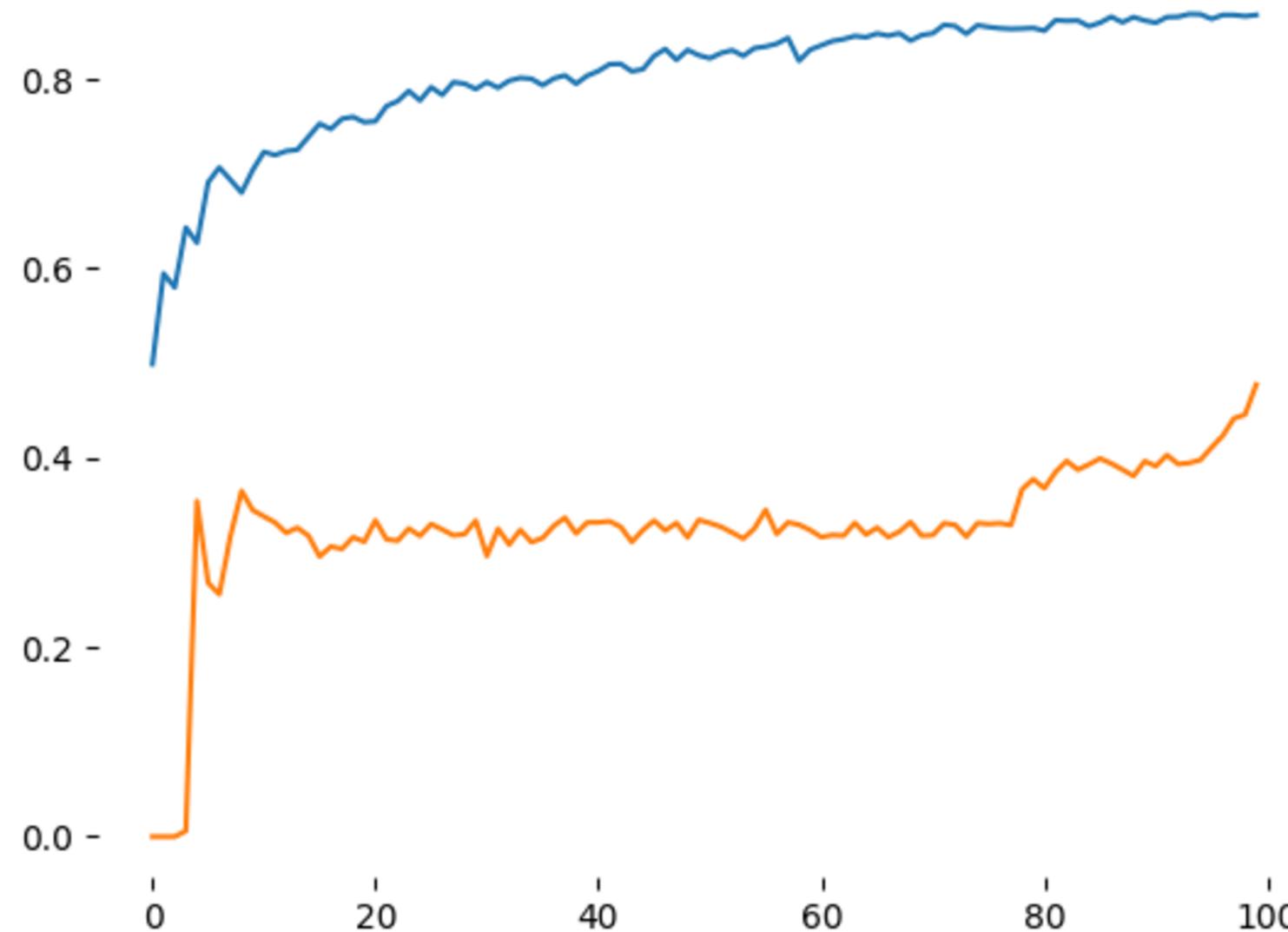
## Validation Strategy:

- Hold-out validation: Splitting by parcel IDs for spatial independence.

## Overfitting Prevention:

- Early Stopping: Stops training when validation loss stagnates.
- Dropout Layers: Used within the LSTM backbone.
- Weight Decay: Regularizes the model further.

## Training Performance:



- The **blue curve** represents accuracy, showing a steady increase, indicating model improvement.
- The **orange curve** represents earliness, remaining relatively stable but rising slightly.
- This suggests the model is learning effectively while maintaining early classification stability.

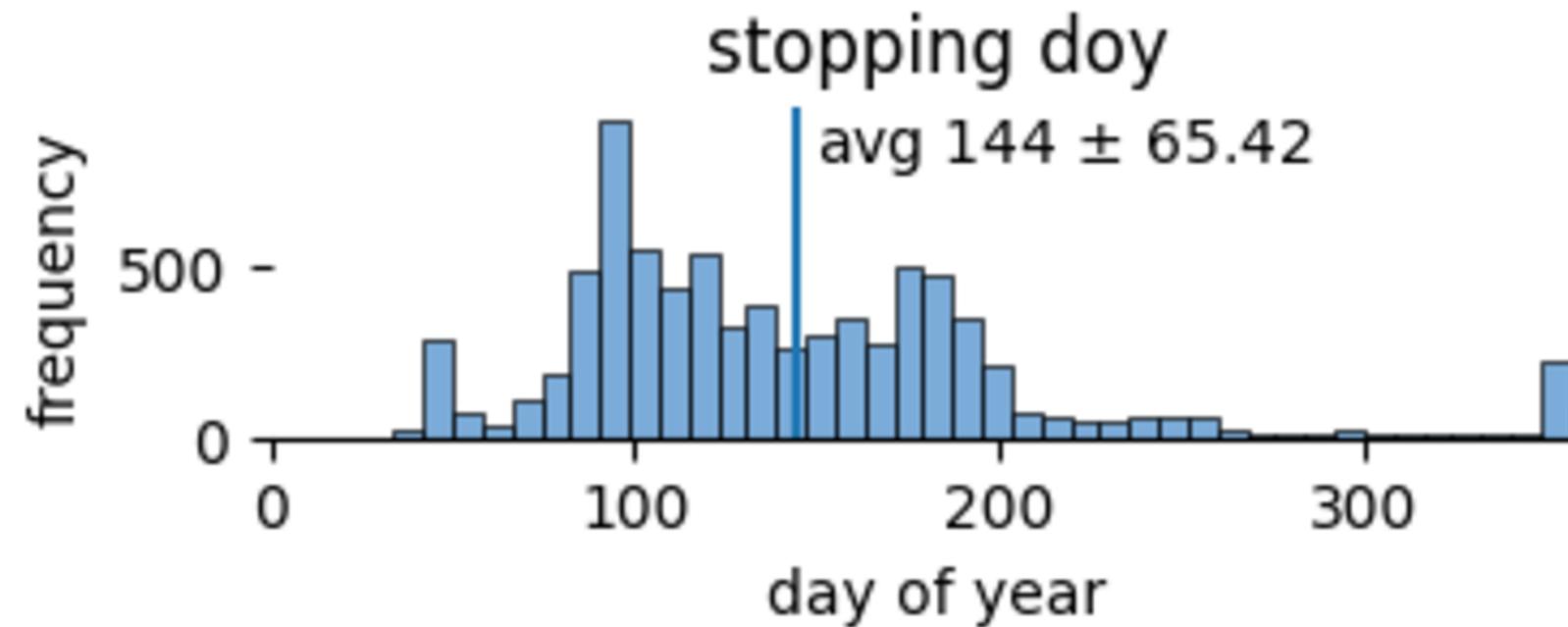
# Metrics

## Accuracy

The model correctly predicts 86% of the crop samples overall.

The recall values show that the model performs well for meadow (high recall), while performing worse for winter triticale (low recall).

## Stopping Day of Year



		true							recall
		meadow	s. barley	corn	wheat	w. barley	clover	triticale	
predicted	meadow	4k	31	21	12	11	107	16	1.0
	s. barley	- 21	820	41	10	2	26	6	0.82
predicted	corn	- 20	5	623	1	2	15	0	0.97
	wheat	- 33	14	11	334	30	3	119	0.77
predicted	w. barley	- 27	12	2	53	328	11	53	0.62
	clover	- 222	14	4	3	2	136	0	0.14
predicted	triticale	- 33	3	7	99	20	1	168	0.07
	meadow	-	-	-	-	-	-	-	0.0

The model mostly classifies crops around day 144(mid-growing season).

- Standard deviation of 65.42 days shows variability in classification timing.
- A peak around 100–150 days suggests the model gains confidence mid-season.

## Technical-Focused Work

- Created a website for the project : (*LIVE NOW AT: <https://crop-ai-research.vercel.app/>*)  
Developed a web interface to showcase the project's objectives and functionalities.  
Integrated a training interface for the EarlyRNN model, allowing me to train and evaluate the model in real-time.  
The website provides insights into the crop classification model, visualization tools, and training configurations.



# Research-Focused Work



# Research-Focused Work

- **Defined research direction**  
**Crop classification using multimodal data + deep learning**
- **Decided on PRISMA-based LR**  
**committed to a review using PRISMA**  
**identified 3 core research questions (ML vs DL vs Multimodal, effectiveness, future)**
- **Built a search strategy**  
**I crafted Boolean search strings across IEEE, WOS, Scopus**  
**Organized keywords into groups (AI tools, crop data types, fusion methods)**
- **Started analyzing literature**  
**Used VOSviewer for keyword co-occurrence**  
**Identified performance trends (CNN > ML; Multimodal fusion = boost in accuracy)**
- **Wrote parts of the paper**  
**Drafted LR intro, and some related work sections**

 **what did I ?**

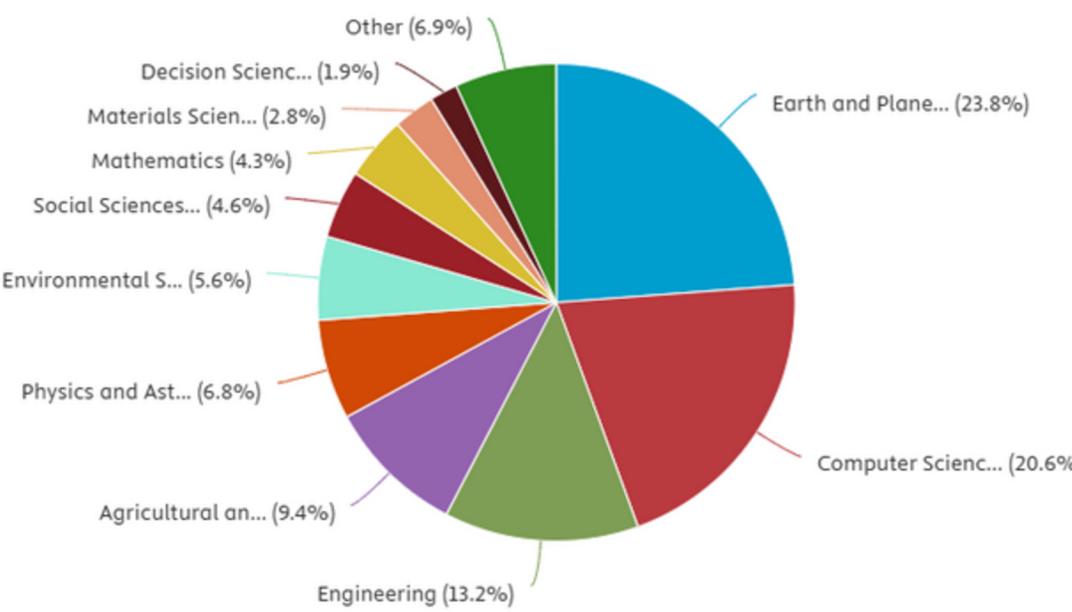
**retrieved the top 500 research papers from Scopus using the following broad search string**

**TITLE-ABS-KEY ( ( "crop classification" OR "crop mapping" ) AND ( "machine learning"  
OR "deep learning" OR "CNN" OR "ConvLSTM" ) AND ( "remote sensing" OR "UAV" OR  
"satellite imagery" ) ) AND PUBYEAR > 2019 AND PUBYEAR < 2026.**

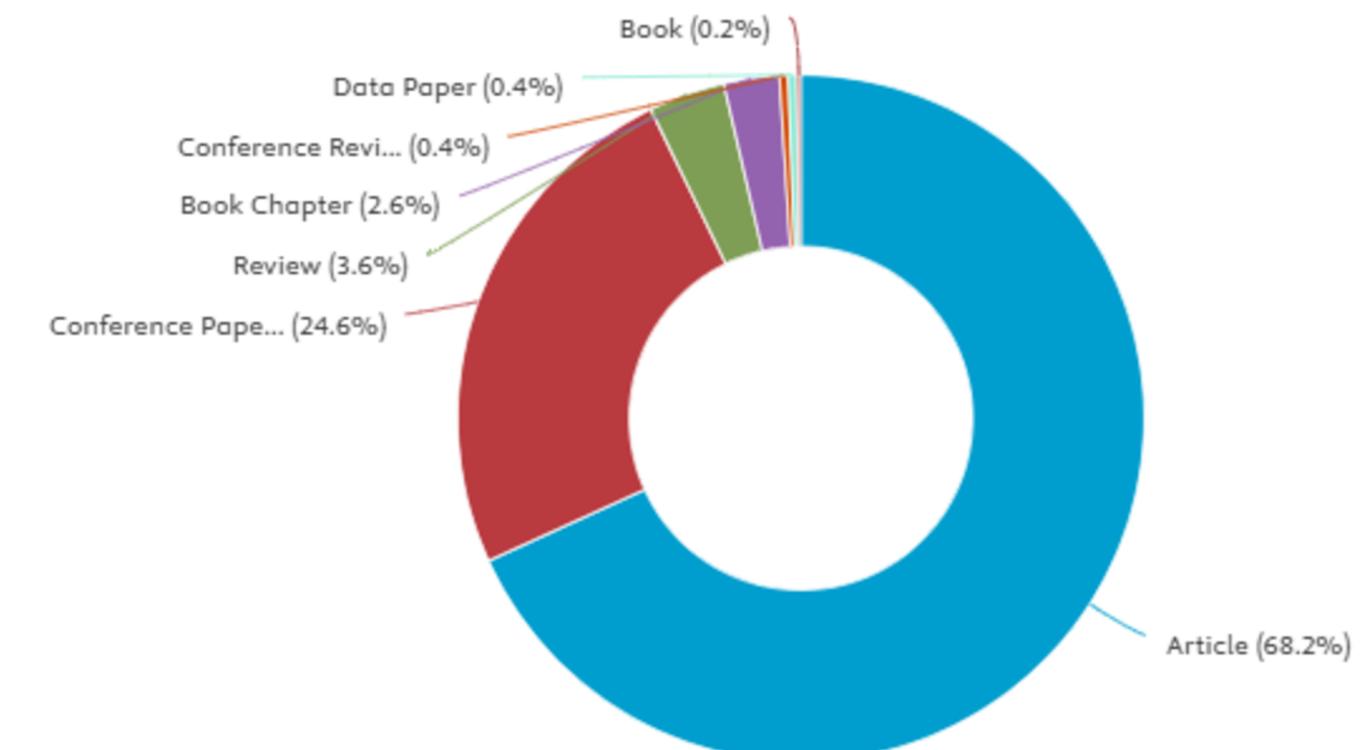
**This query specifically targeted recent literature (2020–2025) focused on crop classification using machine learning and remote sensing technologies.**

# 💡 General insights into the selected papers

Documents by subject area



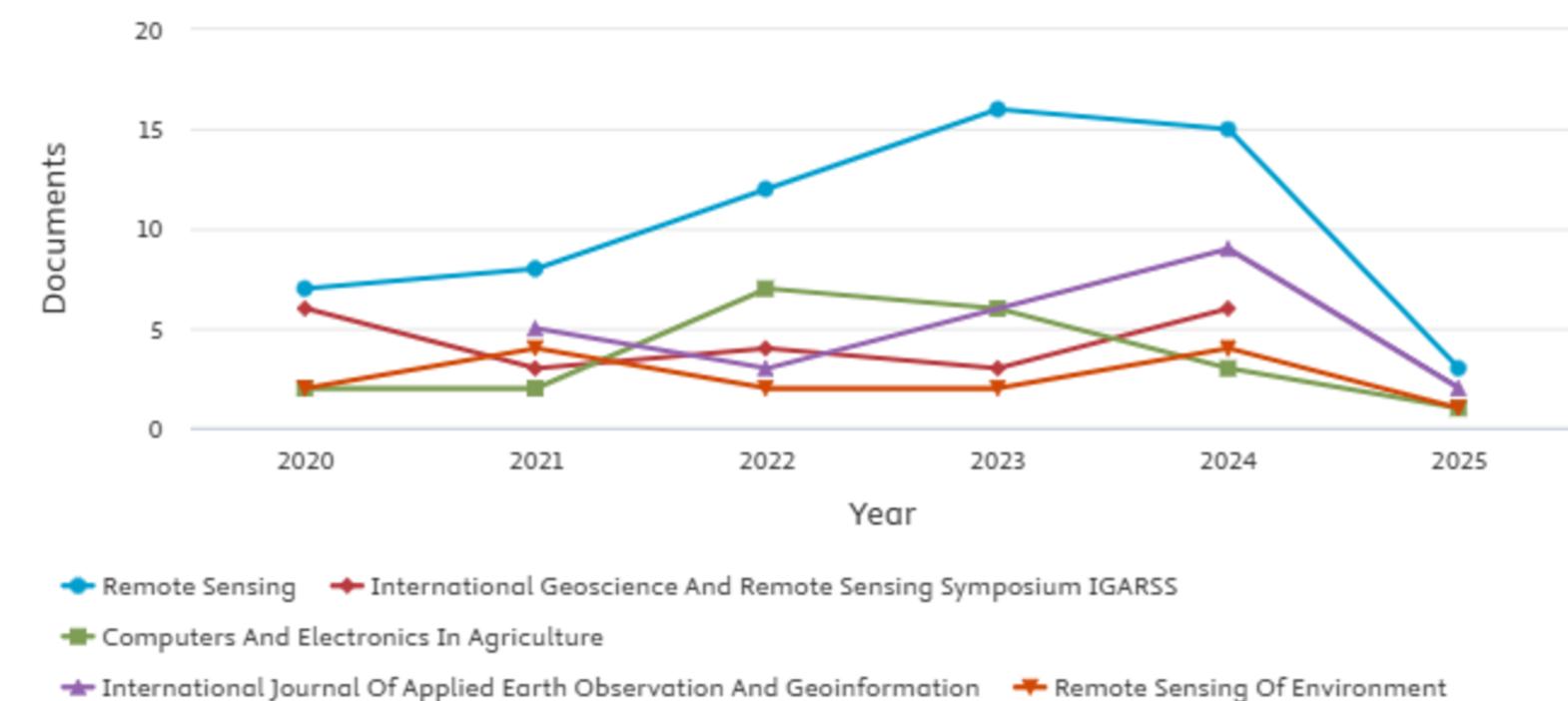
Documents by type



Documents per year by source

Compare the document counts for up to 10 sources.

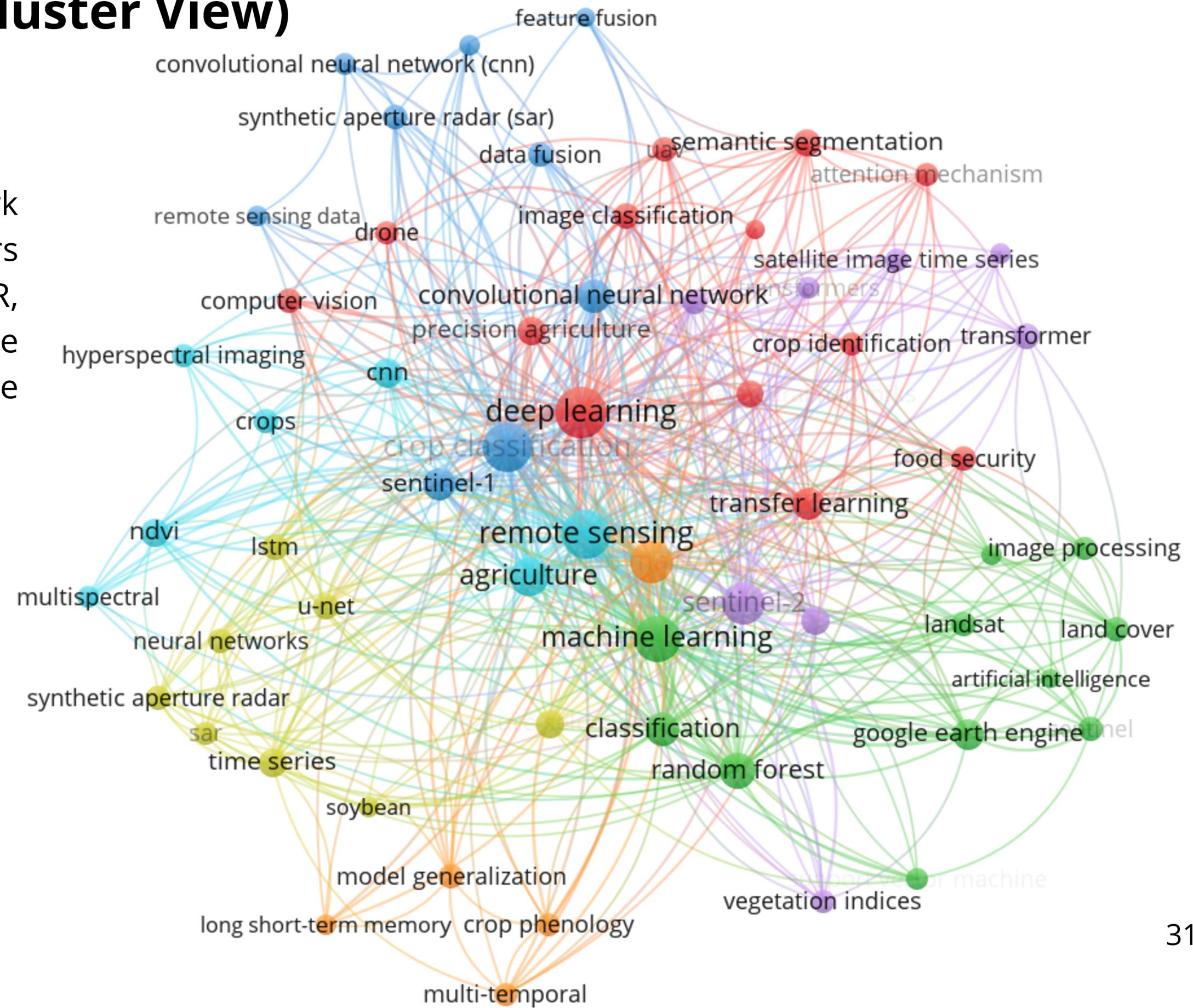
[Compare sources and view CiteScore, SJR, and SNIP data](#)



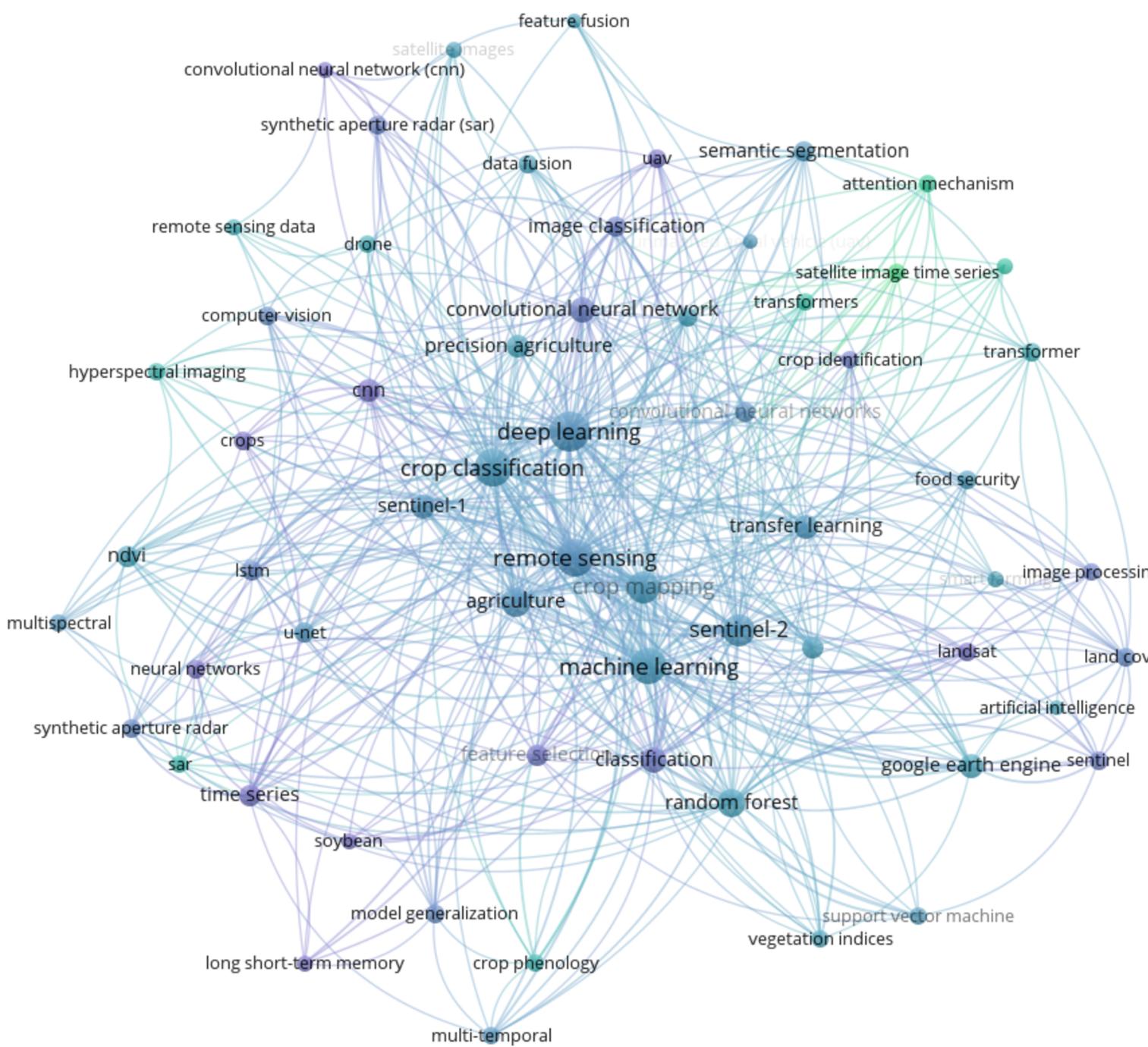
# 1. Network Visualization (Cluster View)

[CLICK ME](#)

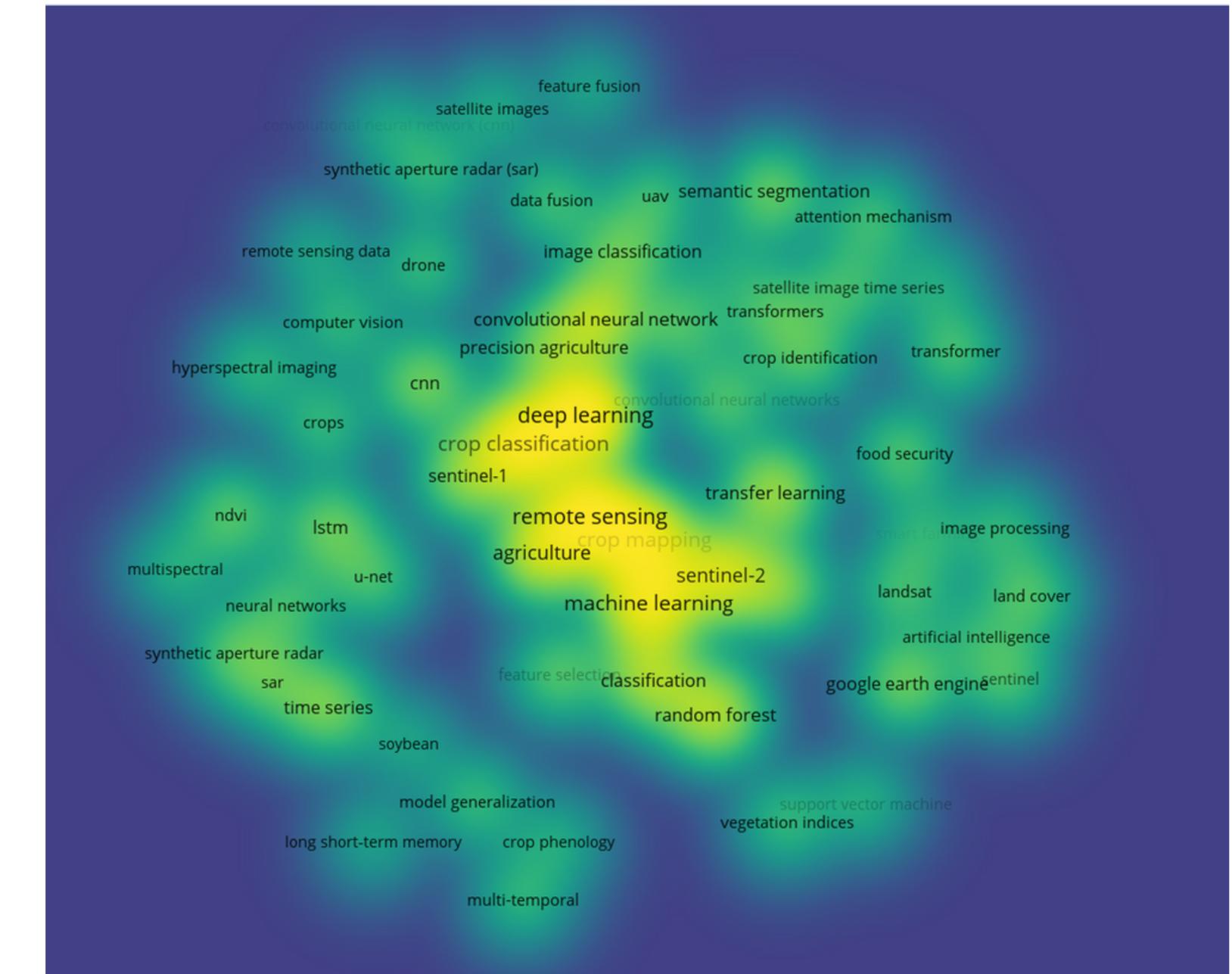
The VOSviewer co-occurrence network revealed tightly connected keyword clusters corresponding to key themes in our SLR, such as deep learning models, satellite imagery, and precision agriculture techniques



## 2. Overlay Visualization (Temporal Evolution)



### **3. Density Visualization (Hotspot View)**



**The list of main keywords was redundant in the subject, so they have been reorganized into four distinct groups to minimize repetition and enhance clarity.**

Theme	Purpose (Linked to RQs)	Keywords
1. Core Concepts	Define the main research domain (applies to all RQs)	Crop classification, Crop type mapping, Crop mapping, Crop monitoring, Precision agriculture, Land cover mapping
2. ML/DL Methods	Capture machine learning & deep learning techniques (RQ1)	Machine learning, Deep learning, Random forest, Support vector machine, Decision tree, K-nearest neighbors, CNN, Convolutional neural network, RNN, Recurrent neural network, LSTM, ConvLSTM, Transformer, Autoencoder
3. Multimodal Techniques	Detect approaches using multiple data sources (RQ2)	Multimodal, Multimodal learning, Data fusion, Sensor fusion, Multisource, Multi-source integration, Multi-input models
4. Data Sources	Enable comparison of input modalities (RQ1, RQ2)	Remote sensing, Satellite imagery, Hyperspectral imagery, Multispectral data, Sentinel-2, Landsat, MODIS, UAV imagery, Drone imagery, Soil data, Climate data, In-situ data, Ground-based sensors

**We have a list of 256 possible research strings, which is too extensive to process. I am currently working on narrowing down the most relevant search string and reviewing the corresponding papers.**

## Conclusion

The internship is progressing well, with steady progress made in various aspects of the project. The work on enhancing the architecture has been successful so far, with several improvements already implemented to boost model performance. The current approach has shown promising results, but there is still room for further optimization and exploration.

## Perspectives

- Continue enhancing the model architecture for better performance.
- Test the model on additional datasets to assess its generalization.
- Explore and compare alternative machine learning approaches.
- Optimize the model for real-world applications and real-time data updates.