Enhancing Crop Classification in Emilia-Romagna (Italy) Using Transformer-Based Multi-Source Data Fusion with Thermal Observations

Ytian Qi [1], Emanuele Mandanici[1], Mohamed Helmy[1], Francesca Trevisiol[2], Gabriele Bitelli[1]

[1] Dept. of Civil, Chemical, Environmental and Materials Engineering (DICAM), Survey and Geomatics Laboratory (LARIG), University of Bologna, Bologna, Italy – yitian.qi@unibo.it
[2] CIMA Research Foundation, Via A. Magliotto 2, 17100 Savona, Italy– Francesca.trevisiol@cimafoundation.org

Keywords: Crop Classification, Thermal Data, Transformer, Deep Learning, Sentinel-1, Sentinel-2

Abstract

This study explores the potential of integrating multi-source remote sensing data—including Sentinel-1 synthetic aperture radar (SAR) imagery, Sentinel-2 optical imagery, and Landsat 8 thermal data—for crop classification in Emilia-Romagna (Northern Italy). Using satellite imagery and agricultural surveys, we constructed a temporal dataset covering 2020 with 27 biweekly time steps. After filtering out underrepresented crop types with insufficient samples for machine learning training, nine crop types remained. We implemented four deep learning models using TensorFlow: Dense Neural Network (DNN), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and Transformer. Our results indicate that removing underrepresented crops significantly improves classification performance, leading to an overall accuracy of approximately 91%. Incorporating Landsat 8 thermal data further enhanced accuracy, with the Transformer model achieving a peak accuracy of 92.08%. A crop-specific analysis revealed that temperature observations notably improved classification for crops with distinct thermal signatures (e.g., sugar beets, corn), whereas limited improvement was observed for spectrally similar cereals (e.g., wheat, barley). Overall, the Transformer model demonstrated exceptional ability in capturing spatial-temporal dependencies in multivariate time-series data. These findings underscore the advantages of integrating multi-source satellite data including thermal infrared and leveraging attention-based neural networks for large-scale agricultural monitoring and resource management.

1. Introduction

Crop Accurate crop classification is a key pillar of agricultural monitoring, allowing for accurate prediction of yield, resource allocation, and evidence-based policy development to solve global food security issues (FAO, 2020; Zhang et al., 2024). Satellite image time series (SITS) remain essential for applications from land cover mapping (Karra et al., 2021) to disaster monitoring (Liu et al., 2023), but many conventional classification approaches struggle to capture the spectral and temporal variability that varies widely among crops in heterogeneous agro-ecological regions (Li et al., 2024; Mathur & Bhattacharya, 2023). Supervised methods that depend on annotated samples result in prohibitive expenses for large-scale applications, although this problem is partially addressed through spatially explicit active learning approaches that specifically optimize sample selection (Kaijage et al., 2024).

The fusion of multi-source satellite data has emerged as a transformative solution to these limitations. Sentinel-1's Synthetic Aperture Radar (SAR) provides cloud-insensitive structural and soil moisture data, while Sentinel-2's medium-resolution optical imagery enables detailed spectral analysis (Drusch et al., 2012; Feng et al., 2024). Complementing these, Landsat 8's thermal observations reveal phenological stages and crop stress signals. Recent studies demonstrate that integrating SAR, optical, and thermal data significantly enhances classification accuracy by overcoming sensor-specific constraints (Phiri et al., 2020; Qi et al., 2023). For example, early-season crop classification frameworks leveraging multi-sensor time-series data (e.g., RCM, Sentinel-1/2) achieve 85% accuracy by iteratively updating predictions with new imagery (Fei, 2024), while fusion strategies on Chongming Island achieved rice extraction precision exceeding 93% using combined SAR-optical features (Chang, 2024).

The advent of transformer-based models has further revolutionized remote sensing analytics. Originally developed for natural language processing (NLP), transformers employ self-attention mechanisms to capture long-range dependencies in spatial-temporal data, outperforming convolutional neural networks (CNNs) in modelling sequential crop phenology (Aleissaee et al., 2023; Khan et al., 2022). The ability of Transformer models to process multi-temporal sequences is well-suited to the dynamic nature of agricultural landscapes. This is supported by (Li et al., 2024), who developed AgriST-Trans, a self-supervised Transformer model pre-trained on Sentinel-2 time series data. AgriST-Trans effectively extracts spatiotemporal crop growth patterns without requiring extensive labeled data, demonstrating the potential of Transformer-based approaches for agricultural applications. Similarly, domain-adaptive models like MDACCN integrate optical and SAR time series to maintain >87% accuracy when transferred between geographically distinct regions (Feng et al., 2024). These architectures excel in capturing subtle phenological differences critical for distinguishing spectrally similar crops, as demonstrated by transformer-based frameworks achieving 85-90% accuracy in classifying 36 land covers using fused Sentinel-1/2 data (Qi et al., 2023).

Despite these advances, challenges persist in scaling models for heterogeneous landscapes and optimizing cross-domain generalizability (Maraveas, 2024; Saini et al., 2024). This study evaluates a transformer-based crop classification framework using fused multi-source satellite data (Sentinel-1/2, Landsat 8) against traditional machine learning and CNN benchmarks. Building on spatiotemporal sample migration techniques (Zhang et al., 2024) and self-supervised pre-training paradigms (Li, 2024), we assess temporal generalizability, feature discriminability, and computational efficiency. By addressing gaps in scalable crop mapping for regions with limited labelled data, our work aims to advance robust solutions for global agricultural monitoring.

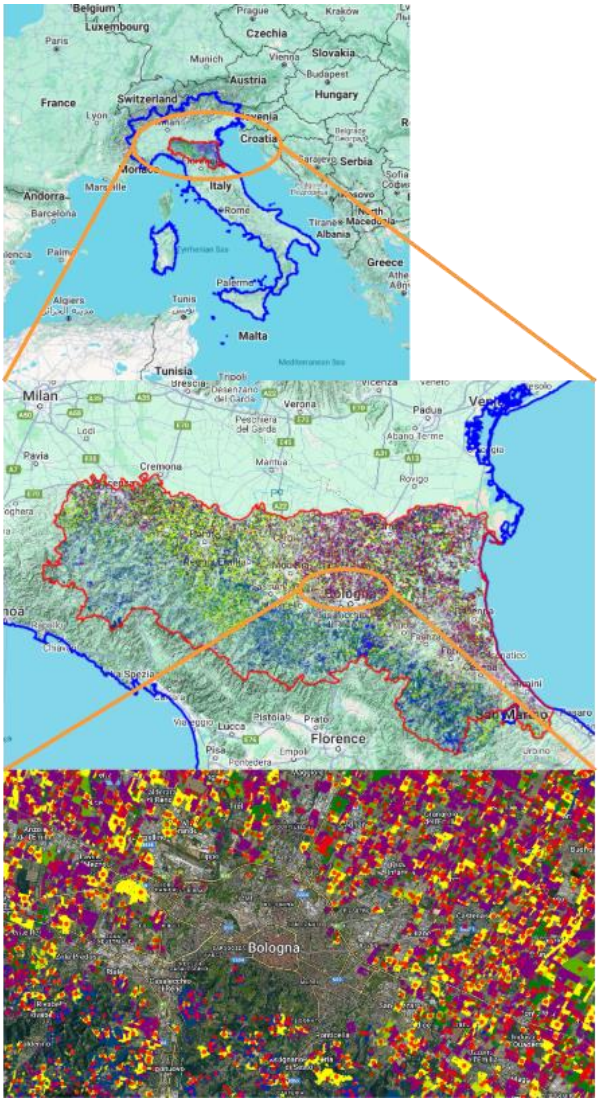

Figure 1. The study area Emilia-Romagna with crop cover

We studied the Emilia-Romagna region (Figure 1), which is the main agricultural region of northern Italy. This area was chosen for this study because of its flat terrain and distinct borders of farmland, which makes it a good region for remote sensing and agricultural research, bringing machine learning into practice. In this region, the geometric regularity of the farmland, often due to historical reasons, provides clear field boundaries and crop pattern detection in satellite imagery, allowing accurate analysis and classification.

Data utilized in this study was obtained from agricultural data made available by the Emilia-Romagna Agency of Agriculture. Compiled in 2020, this database was composed based on voluntary reporting from local farmers, providing us the highest quality ground truth data possible for crop types. This dataset originally consisted of around 160,000 unique individual farmland parcels in the region. Such parcels smaller than 3,000 square meters were omitted to improve data accuracy and specificity. These filtering steps discarded broken or irregular plots that would add noise to the machine learning models. After this preprocessing step, we ended up with a dataset of 60,000 parcels, with each representing a unique agricultural field.

Features of Dataset:

• Geometric Regularity: The flat land of Emilia-Romagna means that the shapes of parcels of farmland are clear and consistent, easily recognizable in satellite images.

• Crop Diversity: The dataset includes multiple crops, indicating the agricultural diversity of the area. This diversity allows for thorough testing and validation of machine learning models across various crops.

• High Quality Ground Truth: The dataset relies on self-reported data collected from local farmers, ensuring maximal accuracy in crop type labels. It is used for the training and evaluation of supervised learning models.

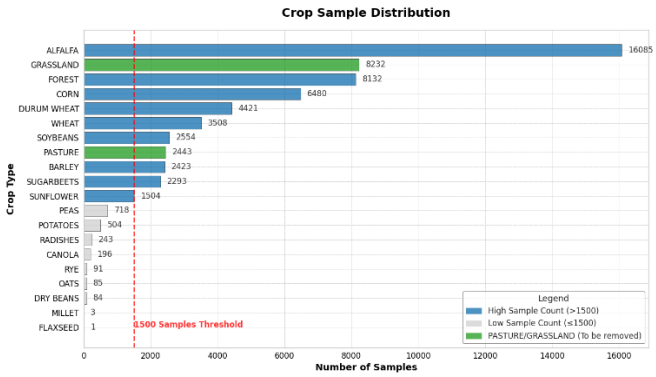

Figure 2. Crop sample distribution in Emilia-Romagna

To ensure sufficient sample size for each crop type, a threshold was set requiring a minimum of 1,500 samples per crop. Crop types with fewer than 1,500 samples were excluded from the main dataset. However, a separate dataset containing these excluded samples was retained as a control group for comparative analysis. After applying this filtering process, the final dataset consisted of 47,400 sample points, covering nine crop types, as shown in the figure 2 above. Each sample point was the geometric centre of the corresponding agricultural parcel.

Crop category	Sample count
Alfalfa	16085
Forest	8132
Cron	6480
Durum wheat	4421
Wheat	3508
Soybeans	2554
Barley	2423
Sugarbeets	2293
Sunflower	1504
Total	47400

Table 1. Distribution of crop samples after filtering in Emilia-Romagna

It also shows basic statistics after filtering, where the total number of samples collected across nine major crop categories and types meets the results of at least 1,500 per crop type (Table 1). This threshold ensures that the dataset is large enough to

enable machine learning model training, validation, and testing, minimizing the risk of data imbalance and enhancing classification reliability. Covering 9 types of crops with 47,400 samples, the dataset shows strong potential for improving model performance through model generalizability and accuracy.

1.1 SAR Data

The Sentinel-1 satellites (which are part of the European Copernicus program) are outfitted with C-band Synthetic Aperture Radar (SAR) sensors that allow for imaging regardless of whether it is day or night, or if there are cloudy skies. The SAR provides a range of detected spatial resolution of 10 m (the data are also accessible via the Google Earth Engine platform). Two satellites of the Sentinel-1 constellation follow a near-polar orbit, with 175 orbits in a 12-day repeat cycle (Torres et al., 2012).

The SAR instrument on Sentinel-1 has three main functioning acquisition modes: Interferometric Wide Swath (IW), Extra Wide Swath (EW), and Stripmap (SM), which are distinguished based on swath widths and spatial resolutions. In this study, the SAR data from the IW mode was used, which is characterized by 250 km swath and 10 m spatial resolution. Since SAR data can penetrate cloud cover, it is more beneficial than optical data for crop classification. It also provides significant information such as soil moisture and vegetation structure, which improves the classification accuracy of the crops.

1.2 Optical Multispectral Data and Indexes

Sentinel-2, also belonging to the Copernicus programme, has a multispectral imager (MSI) onboard to capture data in 12 spectral bands in the visible and shortwave infrared (SWIR) spectrum. The resolution of these bands varies from 10 to 20 to 60 meters (for Geostationary) depending on the band. One numerical crop classification study, in which all 12 bands were fused and used to improve the spectral richness of the dataset, was found in this study.

In addition, some vegetation indexes were computed from the Sentinel-2 data to facilitate classification. Vegetation indexes are indexes such as Normalized Difference Vegetation Index (NDVI), Normalized Difference Moisture Index (NDMI), Enhanced Vegetation Index (EVI), Soil-Adjusted Vegetation Index (SAVI), etc. They offer essential information regarding vegetation health in terms of water content and structural properties, greatly enhancing the performance of the classification model (Trevisiol et al., 2023).

1.3 Thermal Data

Apart from Sentinel satellite data, this analysis also relied on thermal data from Landsat 8, which houses the Thermal Infrared Sensor (TIRS). For example, Band 10 (ST_B10) among the thermal bands is designed for land surface temperature (LST) at the thermal infrared 10.60–11.19 μm (Salih et al., 2018). This band is useful for showing surface heating dynamics directly and is a good predictor of land surface temperature.

Thermal data from Band 10 is essential to detect spatial temperature differences that can be correlated with soil moisture levels and plant transpiration. When integrated effectively, it represents an important factor of differentiation in crop classification as it enhances the recognition of heat stress and irrigation dynamics and provides supplementary information to optical and SAR data.

1.4 Methodology

The methodology depicted in Figure 3 follows a systematic pipeline from data collection, classification, and model evaluation. The Sentinel-1 (SAR), Sentinel-2 (spectral), and Landsat 8 (thermal) data for 2020 were collected through Google Earth Engine (GEE) with a biweekly temporal resolution, which results in 27 observations for each location. Preprocessing was applied to the datasets, including cloud, snow, and shadow masking for optical images (this process can avoid overestimating the spectral reflectance of a specific pixel that has been affected by clouds (Khai et al., 2022), radiometric calibration (for SAR data) (the SAR data originally measured the backscattered waves that were then confirmed by radiometric correction (Ma et al., 2023; Pirotti et al., 2023), speckle filtering (which refers to the granular compositions of the SAR image), and co-registration (georeferencing to spatially align all datasets).

Complex Data Layers (CDLs) were generated by aggregating the preprocessed data, including Sentinels 1, 2, and Landsat 8 imagery on a biweekly basis, which were then merged to create a Data Cube, which is an image in which the X and Y dimensions represent spatial positions, and the Z-axis captures temporal shifts throughout the year. Township-level real crop data collected from official agricultural records were used for the extraction of sample points, with each sample corresponding to the geometric center of a farmland parcel.

The Feature Matrix was created based on values extracted from all 27 data layers, including SAR backscatter (VV, VH), spectral reflectance (12 Sentinel-2 bands), four vegetation indices (NDVI, NDMI, SAVI, EVI), and LST (Landsat 8 Band 10). To stabilize model training, the matrix was normalized.

Implementation:
The four deep learning models—Dense Neural Network (DNN), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and Transformer—were implemented in TensorFlow and were used for classification. 70% of the dataset was used for training, followed by 15% for validation and 15% for testing for each model. Model performance was evaluated on the test dataset by overall classification accuracy after training. Fig. 3 is a flowchart, summarizing data collection, preprocessing, modeling, and evaluation. It merges multi-source remote sensing data and deep learning models to obtain better performance of crop classification in the study area.

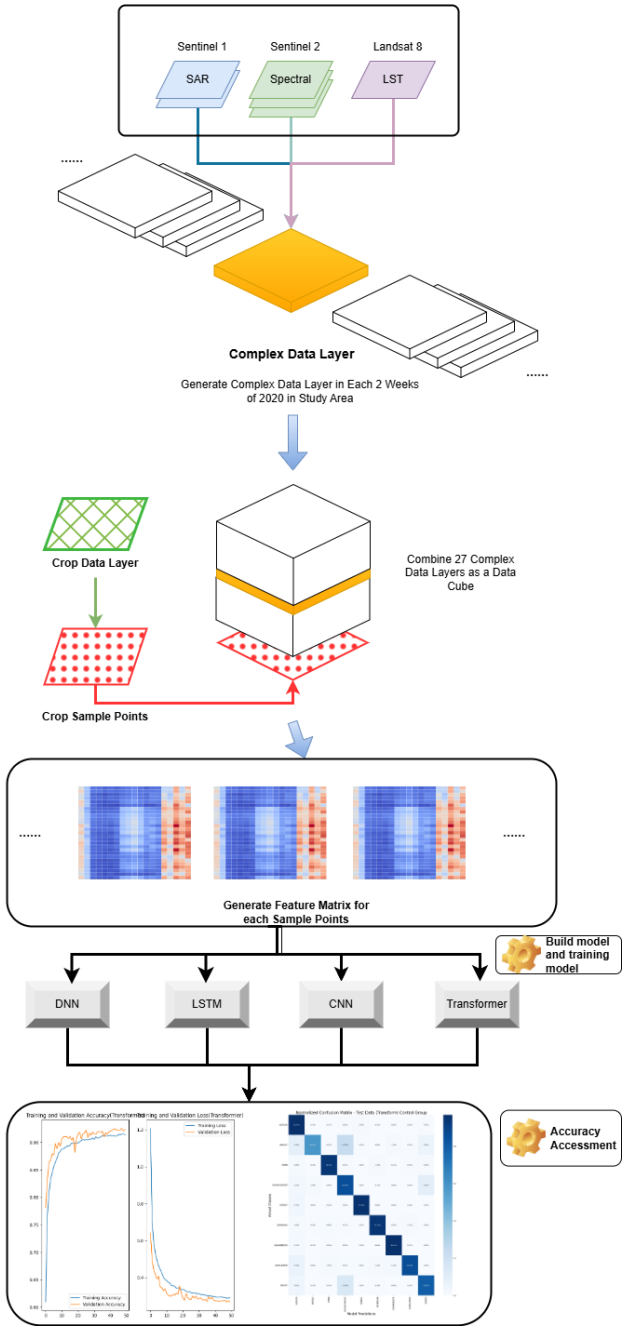

Figure 3. Workflow of the crop classification process

1.4.1 Feature Matrix

In the study area, as illustrated below, we randomly selected a sample point where biweekly SAR data, multispectral data, thermal data, and vegetation indices were processed. Data in each band were normalized to bring all values into a range of zero to one. The final output was the feature matrix that was generated from this process.

The feature matrix combines programmatic features from various inputs:

Two SAR bands (VV and VH) from Sentinel-1. 12 spectral bands from Sentinel-2 (wavelengths span from 443.9 nm to 2202.4 nm). Thermal data from Landsat 8 (Band 10, Land Surface Temperature, LST), which is at a wavelength of 10.60–11.19 μm.

Four vegetation indices: Normalized Difference Vegetation Index (NDVI), Normalized Difference Moisture Index (NDMI), Enhanced Vegetation Index (EVI), and Soil-Adjusted Vegetation Index (SAVI)—calculated from Sentinel-2 data to augment the standard input matrix.

The two-week sampling of the data allows us to have 27 time steps across the entire year, which gives the formulation of a multivariate time series dataset for every sample point. Then, this dataset was converted into a feature matrix in which the vertical axis consists of the 27 biweekly time steps and the horizontal axis consists of SAR bands, spectral bands, thermal data, and vegetation indices values.

The Final Feature Matrix consists of 27 rows and 19 columns:

- 2 columns of SAR data (Sentinel-1 with VV and VH)
- 12 columns of Sentinel-2 spectral data
- 1 thermal data column from Landsat 8 (Band 10 – LST)
- 4 vegetation indices (NDVI, NDMI, EVI, SAVI) columns.

The feature matrix, combining the various feature maps of four different modalities (SAR, spectral, thermal, and vegetation index), along with its time-series information, is expected to enhance the accuracy and robustness of crop classification.

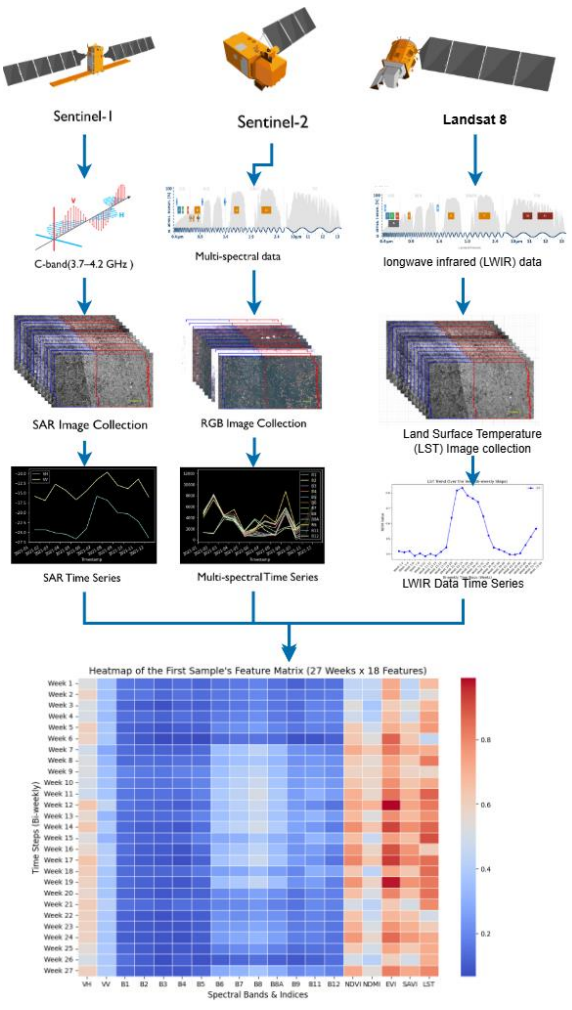

Figure 4. Workflow for generating the feature matrix

1.4.2 Transformer

As shown in Figure 5 which is adapted from Vaswani et al. (2017), the Transformer model follows an encoder-decoder architecture, where the encoder processes the input sequence and transforms it into a continuous representation $z = (z_1, ..., z_n)$. The decoder then utilizes this representation to generate the output sequence in an autoregressive manner, where previously generated tokens serve as additional inputs for predicting subsequent elements (Vaswani et al., 2017).

The core of the Transformer model is its use of self-attention mechanisms and feed-forward networks within both the encoder and decoder layers. Unlike recurrent models, the Transformer uses multi-head self-attention, allowing the model to capture dependencies between all elements in the sequence simultaneously. Additionally, residual connections and layer normalization enhance training stability and convergence efficiency. The output of each sub-layer is computed as:

$$LayerNorm(x + Sublayer(x))$$

where Sublayer(x) represents the transformation applied by the respective sub-layer.

In this study, the input to the Transformer is a multivariate time series consisting of 27 time steps and 19 variables, extracted from the feature matrix. The Transformer architecture for this task consists of 6 encoder layers and 6 decoder layers, where all sub-layers output vectors have a dimensionality of 19, matching the 19 variables in the feature matrix. The model's final output is a vector of length 9, representing the predicted probability of the different crop classes of Table 1.

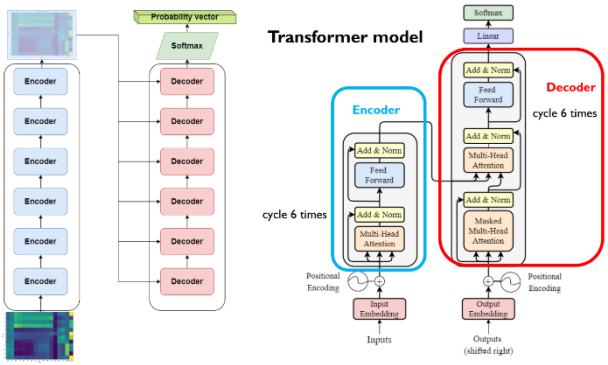

Figure 5. The Transformer model

The Self-Attention mechanism, as depicted in Figure 6 which is adapted from Vaswani et al. (2017), computes three matrices: Q (query), K (key), and V (value). The input is a feature matrix of dimensions 27×19, where 27 represents biweekly time steps, and 19 corresponds to feature variables for each sample point or the output from the previous encoder block. The Q, K, and V matrices are derived by applying learnable weight matrices W_Q, W_K, W_V to the input matrix X through linear transformations. Each row in X, Q, K, and V represents a specific time step in the sequence.

The Self-Attention mechanism computes attention scores by taking the dot product between each row vector in Q and K, which quantifies the similarity between different time steps. To prevent excessively large values, these scores are scaled by dividing by the square root of the key vector dimension $\sqrt{d_k}$, where d_k is the dimensionality of the key vectors. The resulting matrix, QK^T, has dimensions 27×27, encapsulating the attention relationships across all time steps.

Next, the softmax function is applied to QK^T, normalizing the scores into attention coefficients that indicate the relative significance of each time step. Finally, the softmax matrix is multiplied by V, producing the output matrix Z, which aggregates the weighted information from all time steps and serves as the input for subsequent processing layers.

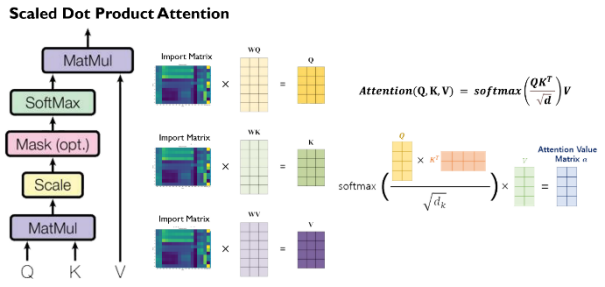

Figure 6. Scaled dot product attention

With reference to Figure 7 which is adapted from Vaswani et al. (2017), the Multi-Head Attention is an ensemble of multiple parallel Self-Attention layers. Multi-Head Attention leverages h distinct Self-Attention processes on the identical input X, generating h output matrices Z as opposed to one. As an example, in the case that h = 8, the same input traverses eight independent Self-Attention layers as seen in the image. Then these h output matrices are concatenated through the feature dimension.

The concatenated matrix is processed through a linear transformation layer and then the output matrix Z is generated. Importantly, the dimensions of the output matrix Z are identical to the input matrix X to maintain consistency in the network. Multi-Head Attention allows the model to attend to different aspects of the input sequence by splitting the attention operation into multiple heads.

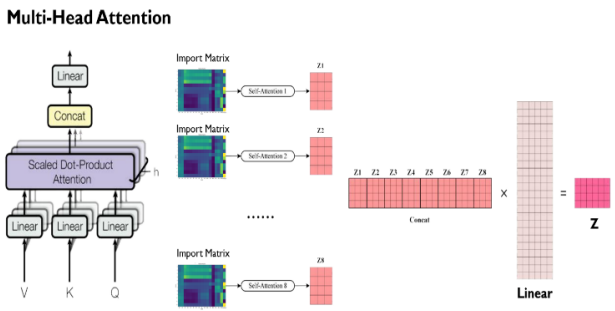

Figure 7. Multi-Head attention

2. Results and Discussion

In this study, two control groups and one experimental group were established to evaluate the impact of different datasets and data fusion methods on crop classification:

- **Control Group 1** used the unfiltered dataset, which included crop types with fewer than 1,500 samples.

This group utilized only SAR data from Sentinel-1 and spectral data from Sentinel-2.

- **Control Group 2** used the filtered dataset, which excluded crop types with fewer than 1,500 samples, but also relied solely on Sentinel-1 and Sentinel-2 data.
- **Experimental Group** used the filtered dataset, integrating SAR data from Sentinel-1, spectral data from Sentinel-2, and thermal data from Landsat 8. This comprehensive dataset was processed using four deep learning models: Dense Neural Network (DNN), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and Transformer to evaluate and compare the performance across different architectures.

Data	DNN	LSTM	CNN	Transformer
unfiltered data	78.34%	77.69%	79.59%	80.82%
S1&S2	91.34%	91.58%	91.62%	91.65%
S1&S2&L8	91.85%	91.94%	91.95%	92.08%

Table 2. Accuracy table of 4 deep learning models

The result of the experiment is illustrated in Table 2, which includes three various datasets for training and testing of the four deep learning models: DNN, LSTM, CNN, and Transformer.

In the first line of the table, we can check the unfiltered dataset (1: least samples more than 1,500); the other lines represent crops with less than 1,500 samples. The second one is the filtered dataset using Sentinel-1 (SAR) and Sentinel-2 (spectral) data. The third row of the filtered dataset is now composed of Sentinel-1, Sentinel-2, and Landsat 8 thermal data integrated together. If we combine less bright and dim signs, recognition accuracy is relatively low for all models when the dataset is not filtered. The results were as follows: The Transformer was the most accurate at 80.82%, followed by CNN (79.59%), DNN (78.34%), and LSTM (77.69%).

These lower accuracy rates imply that including crop types with small sample sizes degrades model performance as there isn't enough data to learn from. When the filtered dataset containing data from both Sentinel-1 and Sentinel-2 was applied, a marked enhancement was seen for all models: The Transformer model once again provided top accuracy (91.65%), followed closely by CNN (91.62%), LSTM (91.58%), and DNN (91.34%) yielding comparably strong results.

This shows that if we take away underrepresented crop types, it allows the model to generalize better with the dataset, thus gaining a higher accuracy overall. Lastly, the best accuracy for Landsat 8 thermal data integrated with the filtered dataset was noticed from the Transformer model (92.08%), higher than any other model. The LSTM achieved 91.94%, the DNN achieved 91.85%, and the CNN achieved 91.51%.

Cyclical variations of the correlation coefficient: Results indicated that the thermal data performance was better at identifying the taxonomic grouping than the SAR and spectral data, thus, the thermal data have diverse information in addition to SAR and spectral data, consequently providing complementary insight in increasing classification accuracy.

Overall, the Transformer model consistently outperformed the other models, even more with additional thermal data. This suggests that using several data sources may improve classification performance, while the self-attention mechanism of the Transformer model can potentially improve the capture of complex patterns available in multivariate time series data.

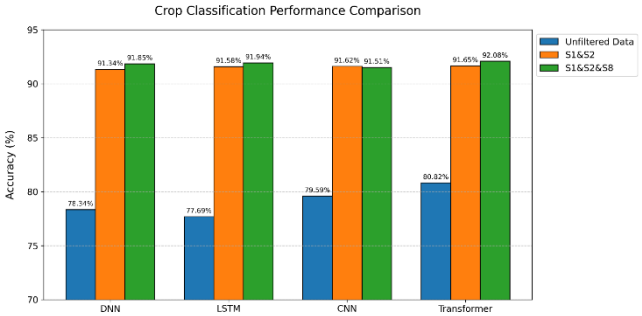

Figure 8. crop classification performance comparison

Figure 8 illustrates the crop classification accuracy of the four deep learning models, namely DNN, LSTM, CNN, and Transformer, using the three different datasets of Table 2:

- The unfiltered dataset (blue)
- The filtered dataset with Sentinel-1 and Sentinel-2 (orange)
- The filtered dataset with Sentinel-1, Sentinel-2, and Landsat 8 thermal (green) data

As can be seen in the chart, Transformer outperforms all other methods in every case and achieves the highest accuracy (92.08%) when data is filtered and the Landsat 8 thermal band is included. CNN, LSTM, and DNN show comparable improvements, both on switching from the unfiltered dataset to the filtered and additionally from adding thermal data.

These visual trends are consistent with the quantitative results indicating that filtering the dataset markedly improves classification performance and that integrating Landsat 8 thermal data bolsters crop type discrimination capabilities even further. The Transformer model outperformed traditional methods and other models, indicating its appropriateness for multivariate time series derived from multiple satellite sources.

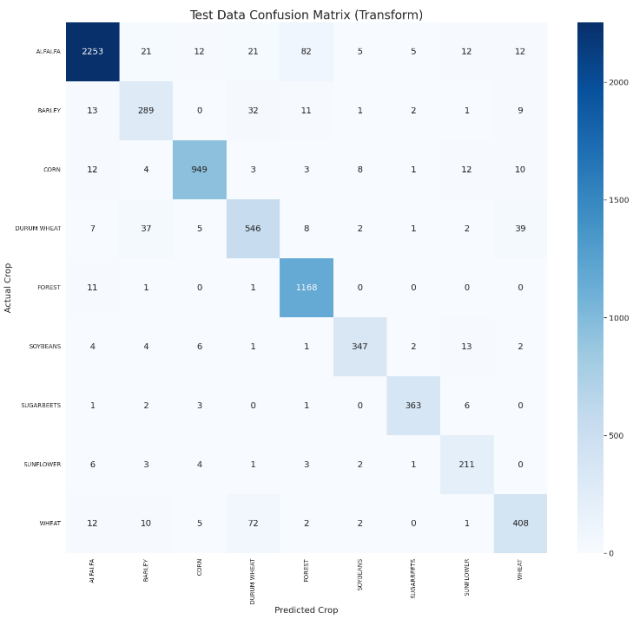

Figure 9. Confusion matrix transformer model with 9 crop types

The result (Confusion matrix of Transformer model) indicate that distinguishing accuracy varies with different crop types. It is still not very effective at separating them from other crops (Figure 9). The Transformer model performed best in this series of experiments on sugar beets (96.54%) and forest (98.9%); it's difficult to tell separately for either of these crops any variety that comes close in spectral features, so unusual as these factors likely make recognition more convenient for them.

Like corn (94.71%) and alfalfa (92.98%), both may further emerge when grains mature and also have very good depictive power over time due to this characteristic being affected, probably inappropriate anyhow. Moreover, the classification accuracy remained fairly good for soybeans (96.54%) and sunflower (91.34%), despite showing that while the model can indeed distinguish these crops, when they are in similar spectral states at certain growth stages, it may make one small mistake.

Meanwhile, this model showed lower levels of accuracy (84.39% for durum wheat, 79.69% for wheat, and 80.73% for barley) when classifying cereals such as wheat or barley. However valid these inferences may be at large-scale landscape level, certainly, misclassification still occurs occasionally between these similar types of crops today using current spectral and temporal features.

Figure 9's classification rates are relatively poor among these similar types of crops, which suggests that the model is of limited utility. There appears also (as shown most clearly in Figure 9) to be little information from thermal data to help separate these crops. This indicates that crop classification in this category can be improved, for example, by incorporating additional spectral indices or by stating time periods as features of lightness values rather than merely distinguishing one temporal segment from another.

Table 3 presents a comparison of the accuracy achieved by four deep learning models—DNN, LSTM, CNN, and Transformer—across nine crop types using the S1, S2, and Landsat 8 data fusion.

The Transformer model consistently shows high accuracy across most crop types, outperforming other models for Forest (98.9%), Sunflower (91.34%), and Soybeans (91.32%). This result indicates that the self-attention mechanism of the Transformer effectively captures complex temporal and spectral relationships in the data.

Crop Type	DNN (%)	LSTM (%)	CNN (%)	Transformer (%)
ALFALFA	95.79	95.17	95.21	92.98
BARLEY	81.84	83.52	74.02	80.73
CORN	95.81	94.11	93.61	94.71
DURUM WHEAT	72.95	79.6	78.98	84.39
FOREST	97.21	95.17	96.95	98.9
SOYBEANS	87.37	90.79	90	91.32
SUGARBEETS	95.48	96.81	96.81	96.54
SUNFLOWER	83.98	89.18	85.71	91.34
WHEAT	81.25	78.12	82.23	79.69

Table 3. Accuracy comparison of deep learning models for crop classification using Multi-Source satellite data

DNN is performing well for the crop types of Corn (95.81%), Sugar beets (95.48%), and Alfalfa (95.79%), which further affirms its capability of dealing with structured input data. In contrast, both Durum Wheat (72.95%) and Wheat (81.25%) had markedly lower precision, which can be due to misclassification caused by spectral similarity.

As an architecture that is meant to learn temporal dependencies, LSTM gives good results on Sugar beets (96.81%) and Corn (94.11%), suggesting that in particular, this architecture benefits from biweekly time series. However, the feature struggles for Wheat (78.12%) and Barley (83.52%) compared to the other crops. Utilizing CNN's capability of extracting global spatial-temporal patterns from the time-series matrix, it reaches a high accuracy of 96.81% for Sugar beets and 96.95% for Forest. On Barley (74.02%), however, its performance is lower than the other models, indicating that this model might have less ability to identify crops with a subtle temporal signature.

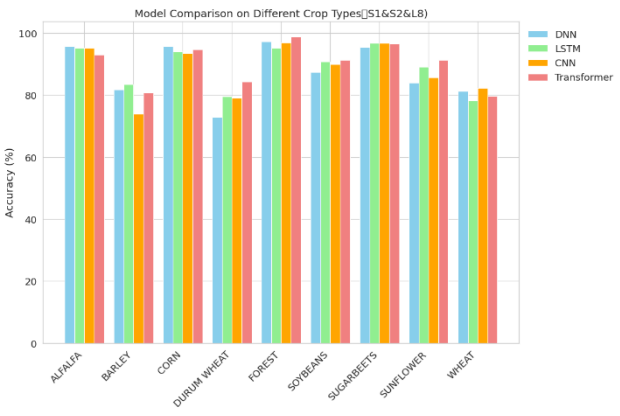

Figure 10. Accuracy comparison of deep learning models for crop classification using Multi-Source data fusion (S1, S2, L8)

Figure 10 highlights the Transformer model as the most robust for crop classification using the Sentinel-1, Sentinel-2, and Landsat-8 fused dataset. Thermal data from Landsat 8 provides valuable information, improving classification accuracy,

particularly for crops with distinct temperature signatures. While all models show competitive performance, the Transformer’s ability to integrate information across time and features gives it a significant advantage.

3. Conclusion

In this study, the effect of dataset preprocessing and multi-source information fusion on crop classification accuracy is examined using four deep learning models—Deep Neural Network (DNN), Long-Short Term Memory (LSTM), Convolutional Neural Network (CNN), and Transformer. The control groups (unfiltered as opposed to filtered dataset) and an experimental group (filtered dataset plus additional thermal information from Landsat 8) allowed us to distinguish between the impact of data quality and multi-source combination.

A few main findings emerged:

First, data filtering aids classification rates greatly. When the unfiltered dataset—which contains crop types present in only a few instances—was employed, all models only achieved relatively lower accuracy, with 80.82% as the maximum (reached by the Transformer model). After discarding minor crop types, overall accuracy went up significantly (everything over 91% in all models). This result illustrates the necessity of providing a large enough training sample size for each crop class—limited training samples can interfere with deep neural networks' ability to comprehend nuanced spatiotemporal characteristics.

Second, adding thermal information from Landsat 8 to the filtered dataset improved crop classification performance even more. By incorporating thermal data from Landsat 8 alongside SAR data from Sentinel-1 and spectral data from Sentinel-2, all three machine learning models leveraged complementary crop characteristics that were previously underutilized. For this reason, the Transformer model scored highest as usual (92.08%), just apart from DNN, LSTM, and CNN. This confirms that integrating thermal data with spectral and SAR information significantly enhances crop classification accuracy, particularly for crops with overlapping spectral signatures..

Third, similar to other architectures utilizing all data sources, the Transformer model consistently demonstrated superior performance. Its self-attention mechanism looks especially well placed for transforming complex relationships in multivariate time series data. By considering weightings between different time steps or bands of spectral data, the Transformer can effectively synthesize information across SAR, optical, and thermal channels. While DNN, LSTM, and CNN benefited significantly from data filtering and fusion, their performance remained below that of the Transformer model.

Moreover, the examination of the classification performance for each crop type indicated that crops with differentiated temporal or thermal signals, such as Forest, Sugarbeets, Corn, and Alfalfa, attained elevated accuracy levels (mostly above 90%).

Wheat, Barley, and Durum Wheat, however, were harder to discriminate between rice crops due to similarities in their growth cycles and overlapping spectral signatures. Thermal data provided some degree of differentiation but had little effect on these closely related types of cereal.

Our results demonstrate the importance of strong preprocessing (i.e., removing underrepresented crops), combining different satellite data (SAR, optical, and thermal), and the use of deep learning architectures such as the Transformer to achieve optimal levels of crop classification performance.

Another aspect that has been addressed in this study is the potential of multi-source and multi-temporal remote sensing data for agricultural monitoring, allowing for a better understanding of which model architectures and data fusion techniques lead to the best performance results.

Further research could investigate using more sensors or data at a higher temporal resolution to better differentiate more difficult-to-resolve crops. The proposed methodologies and insights gleaned from this work can help practitioners and researchers better design accurate, scalable systems involved in precision agriculture, yield prediction, and resource management.

4. References

Aleissaee, A. A., Kumar, A., Anwer, R. M., Khan, S., Cholakkal, H., Xia, G., & Khan, F. S. (2023). Transformers in Remote Sensing: a survey. Remote Sensing, 15(7), 1860. https://doi.org/10.3390/rs15071860

Chang, Y., Zhang, L., & Zheng, B. (2024, July 7). Comparison of methods for crop classification and rice extraction on Chongming Island based on Sentinel-1 and Sentinel-2 data. IEEE Conference Publication. IEEE Xplore. https://ieeexplore.ieee.org/document/10641149

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

FAO. (2020). Crop Monitoring for Improved Food Security. Food and Agriculture Organization of the United Nations. https://www.fao.org/publications

Fei, C., Li, Y., McNairn, H., & Lampropoulos, G. (2024, July 7). Early-season crop classification utilizing time series-based deep learning with multi-sensor remote sensing data. IEEE Conference Publication. IEEE Xplore. https://ieeexplore.ieee.org/document/10642827

Feng, L., Gui, D., Han, S., Qiu, T., & Wang, Y. (2024). Integrating optical and SAR time series images for unsupervised domain adaptive crop mapping. Remote Sensing, 16(8), 1464. https://doi.org/10.3390/rs16081464

Feng, L., Han, X., Hu, C., & Chen, X. (2016). Four decades of wetland changes of the largest freshwater lake in China: Possible linkage to the Three Gorges Dam? Remote Sensing of Environment, 176, 43–55. https://doi.org/10.1016/j.rse.2016.01.011

Feng, S., Zhao, J., Liu, T., Zhang, H., Zhang, Z., & Guo, X. (2019). Crop type identification and mapping using machine learning algorithms and Sentinel-2 Time Series data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(9), 3295–3306. https://doi.org/10.1109/jstars.2019.2922469

Ienco, D., Gaetano, R., Dupaquier, C., & Maurel, P. 2017. Land cover classification via multitemporal spatial data by deep recurrent neural networks. IEEE Geoscience and Remote Sensing Letters, 14(10), 1685–1689. https://doi.org/10.1109/lgrs.2017.2728698

Kaijage, B., Belgiu, M., & Bijker, W. 2024. Spatially Explicit Active Learning for Crop-Type Mapping from Satellite Image Time Series. Sensors, 24(7), 2108. https://doi.org/10.3390/s24072108

Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J. C., Mathis, M., & Brumby, S. P. 2021. Global land use / land cover with Sentinel 2 and deep learning. 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, Pp. 4704-4707, July, 2021. https://doi.org/10.1109/igarss47720.2021.9553499

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. 2022. Transformers in Vision: a survey. ACM Computing Surveys, 54(10s), 1–41. https://doi.org/10.1145/3505244

Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. 2017. Deep learning classification of land cover and crop types using remote sensing data. IEEE Geoscience and Remote Sensing Letters, 14(5), 778–782. https://doi.org/10.1109/lgrs.2017.2681128

Li, H. 2024. AgriST-Trans: A self-supervised Transformer pre-trained model for crop classification based on time-series remote sensing. IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/10660832

Liu, P., Liu, Y., Guo, X., Zhao, W., Wu, H., & Xu, W. 2023. Burned area detection and mapping using time series Sentinel-2 multispectral images. Remote Sensing of Environment, 296, 113753. https://doi.org/10.1016/j.rse.2023.113753

Maraveas, C. 2024. Image analysis Artificial intelligence Technologies for plant phenotyping: current state of the art. AgriEngineering, 6(3), 3375–3407. https://doi.org/10.3390/agriengineering6030193

Mathur, I., & Bhattacharya, P. 2023. From pixels to patterns: review of remote sensing techniques for mapping shifting cultivation systems. Spatial Information Research, 32(2), 131–141. https://doi.org/10.1007/s41324-023-00547-9

Misra, G., Cawkwell, F., & Wingler, A. 2020. Status of Phenological Research Using Sentinel-2 Data: A Review. Remote Sensing, 12(17), 2760. https://doi.org/10.3390/rs12172760

Phiri, D., Simwanda, M., Salekin, S., Nyirenda, V., Murayama, Y., & Ranagalage, M. 2020. Sentinel-2 Data for Land Cover/Use Mapping: A Review. Remote Sensing, 12(14), 2291. https://doi.org/10.3390/rs12142291

Pirotti, F., Adedipe, O., Leblon, B., 2023. Sentinel-1 Response to Canopy Moisture in Mediterranean Forests before and after Fire Events. *Remote Sensing* 15, 823. https://doi.org/10.3390/rs15030823

Qi, Y., Bitelli, G., Mandanici, E., & Trevisiol, F. 2023. Application of Deep Learning Crop Classification Model Based on Multispectral and SAR Satellite Imagery. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVIII-1/W2-2023, 1515–1521. https://doi.org/10.5194/isprs-archives-xlviii-1-w2-2023-1515-2023

Saini, N., Dhir, R., & Kaur, K. 2024. Crop classification using deep learning on Time series SAR Images: a survey. In Lecture notes in networks and systems (pp. 1–10). https://doi.org/10.1007/978-981-99-7814-4_1

Salih, M. M., Jasim, O. Z., Hassoon, K. I., & Abdalkadhum, A. J. 2018. Land Surface Temperature Retrieval from LANDSAT-8 Thermal Infrared Sensor Data and Validation with Infrared Thermometer Camera. International Journal of Engineering & Technology, 7(4.20), 608. https://doi.org/10.14419/ijet.v7i4.20.27402

Trevisiol, F., Mattivi, P., Mandanici, E., & Bitelli, G. 2024. Cross-sensors comparison of popular vegetation indexes from Landsat TM, ETM+, OLI, and Sentinel MSI for time-series analysis over Europe. IEEE Journals & Magazine. https://ieeexplore.ieee.org/abstract/document/10360186

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., Traver, I. N., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., L’Abbate, M., Croci, R., Pietropaolo, A., . . . Rostan, F. 2012. GMES Sentinel-1 mission. Remote Sensing of Environment, 120, 9–24. https://doi.org/10.1016/j.rse.2011.05.028

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. 2017. Attention is All you Need. arXiv (Cornell University), 30, 5998–6008. https://arxiv.org/pdf/1706.03762v5

Yuan, Y., & Lin, L. 2020. Self-Supervised Pretraining of transformers for satellite Image Time Series classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 474–487. https://doi.org/10.1109/jstars.2020.3036602

Zhang, G., Chen, W., Li, G., Yang, W., Yi, S., & Luo, W. 2019. Lake water and glacier mass gains in the northwestern Tibetan Plateau observed from multi-sensor remote sensing data: Implication of an enhanced hydrological cycle. Remote Sensing of Environment, 237, 111554. https://doi.org/10.1016/j.rse.2019.111554

Zhang, H., Lou, Z., Peng, D., Zhang, B., Luo, W., Huang, J., Zhang, X., Yu, L., Wang, F., Huang, L., Liu, G., Gao, S., Hu, J., Yang, S., & Cheng, E. 2024. Mapping annual 10-m soybean cropland with spatiotemporal sample migration. Scientific Data, 11(1). https://doi.org/10.1038/s41597-024-03273-5