Temporal Attention Gates in Time Series Crop Classification: An Effective Tool?

Harry Fyjis-Walker

December 11, 2025

GitHub: COMP0173_CW2_Crop_Classification

1 Replication of Baseline AI Methodology

The chosen baseline for this coursework was the Attention U-Net model developed in John, D. and Zhang, C. (2022): *An attention-based U-Net for detecting deforestation within satellite sensor imagery.*[1] The paper evaluates the utility of this architecture for semantic segmentation in the context of deforestation detection, benchmarking the model on three open-source Sentinel-2-based datasets:

(i) The 3-band (RGB) Amazon Rainforest dataset (Bragagnolo *et al.* (2019) [2]

(ii) 4-band Amazon Forest image dataset (Bragagnolo *et al.* (2021) [3]

(iii) 4-band Atlantic Forest image dataset (Bragagnolo *et al.* (2021) [3]

To evaluate the performance of Attention U-Net, the authors train from scratch four additional models: U-Net, Residual U-Net, ResNet50 with a SegNet backbone, and FCN32 with a VGG16 backbone. The workflow is clearly documented on the project's Github page. The original code is detailed in Experimentation.ipynb, which divides the process into sections: Load Packages, Functions, Ingestion and Processing of Datasets, Models, Training, and Metrics Computation.

Here, the baseline methodology is replicated for both U-Net and Attention U-Net models on the RGB dataset, the code for which is available here (for the RGB dataset, evaluation is only performed on the validation set in the original publication). The optimal learning rates (LR) and epochs defined in the original publication are preserved, with the Attention U-Net trained with LR=0.0005 over 50 epochs and the U-Net with LR=0.0001 over 30 epochs. The implementation is shown to be highly reproducible, with the weighted Intersection over Union (IoU), Precision, Recall, and F1-score of the replicated U-Net and Attention U-Net baselines all differing from the published results by less than 1% (Table 1).

Dataset		Validation			
		IoU	Precision	Recall	F1-score
U-Net	Original Paper	0.8888	0.9571	0.9473	0.9522
	Reproduced	0.8870	0.9475	0.9461	0.9468
Attention U-Net	Original Paper	0.9028	0.9574	0.9526	0.9550
	Reproduced	0.9036	0.9534	0.9533	0.9534

Table 1: Comparison of published and replicated results for U-Net and Attention U-Net on the 3-band Amazon Rainforest dataset.[1]

The U-Net implementation here involves a 5-level symmetric encoder-decoder structure. The encoder implements successive 3 x 3 convolutions (two per stage via the convBlock structure) and ReLU activations, followed by 2 x 2 max-pooling operations for downsampling. The number of filters is doubled sequentially, from 64 to 1024 in the bottleneck. The four upsampling stages use Conv2DTranpose layers (kernel size 2, stride 2) and concatenate the corresponding feature maps from the encoder. The final output uses a single 1 x 1 convolution with sigmoid activation to produce the segmentation mask.[1]

The Attention U-Net model uses a similar 5-level structure, with encoder filters increasing from 32 to 256 in its core convolutional block. This itself uses two Conv2D layers (kernel size 3, padding 'same', He initialisation) with a ReLU

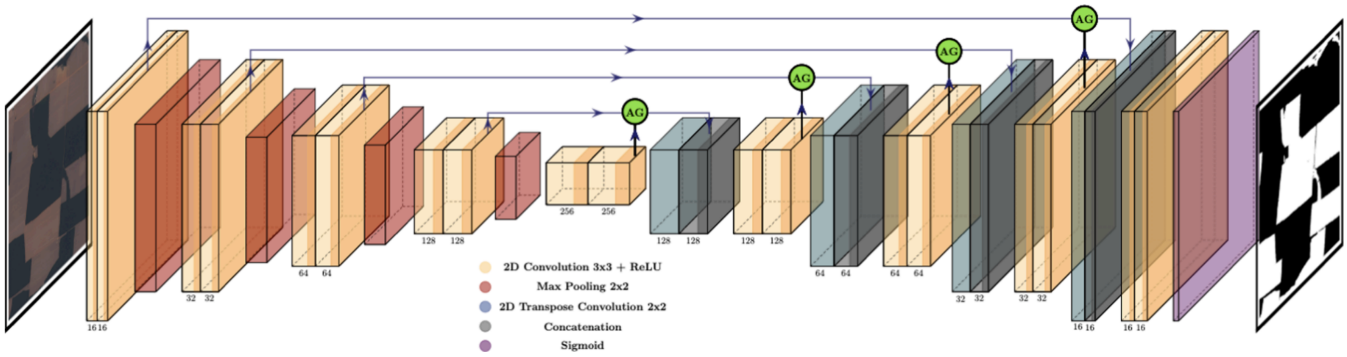

Figure 1: Network architecture diagram for the baseline Attention U-Net, reproduced from John and Zhang (2022).[1]

activation after each convolution. An attention gate is placed on the skip connections, which uses a convolution and sigmoid activation to generate an attention coefficient map that is multiplied with the encoder features before being concatenated with the upsampled decoder features. The final output layer again uses a single 1 x 1 convolution with sigmoid activation.[1] The network architecture diagram from the original paper is reproduced in Figure 1 above.

2 Identification of Contextually Relevant Challenge

2.1 Problem and SDG Alignment

Sub-Saharan Africa (SSA) is estimated to have 332.3 million food insecure people in 2025, the most of any region globally.[4] Here, 80% of the food supply is provided by smallholder farms.[1][5][6] Simultaneously, such farms employ the majority of the labour force and, within rural areas, over 70% are depend fully on small-scale agriculture for income.[7][8] It is widely recognised that the provision of effective support to smallholder farmers is crucial for both poverty (SDG 1) and food insecurity (SDG 2) alleviation in the region.[7]

The challenges faced by small-scale farmers are diverse and complex. Crop yields are limited by lack of access to information on best agronomic practices and are increasingly threatened by extreme climate events such as droughts and outbreaks of pests and diseases.[9] Among the former, surveys conducted by Phiri *et al.* (2017) suggested that the major information needs of smallholders in Mali were crop husbandry and pest and disease control.[10] Rising input costs, restricted market access, and difficulty accessing credit also represent major barriers, as do inadequate transport infrastructure and storage facilities and limited access to modern tools.[11][9][12]

Among these, increasing crop yield is of significant importance. While the climatic and biophysical features of SSA are sufficient to address the growing food demand, yield gaps (the difference between actual and potential yield) are often calculated to be over 70%, compared to below 30% for many crops in Western European countries.[13][14][15][16] Closing these gaps is widely considered a crucial avenue towards increased food security and economic stability.[13]

Here, artificial intelligence has the potential to contribute, already proving valuable in applications such as advising seed choice and forecasting locust swarms.[17][18] Applications at the smallholder level, however, are impeded by an information gap on both field boundaries and crop types. The application of deep learning architectures to remote sensing data represents a promising avenue to closing this gap (e.g. in yield prediction, disease identification, tailored recommendations for seed type and planting time etc.); however, such applications have so far been hindered by (i) small parcel sizes in SSA, which require high resolution satellite imagery for classification, (ii) the irregular nature of field shapes, which vary widely in appearance due to inhomogeneous terrain, and (iii) a lack of ground truth data on parcel shape and crop variety for model training and validation.[19][20] Wang *et al.* (2022) investigated the efficacy of transfer learning - with pre-training on well-labeled French field parcel datasets - combined with higher resolution Airbus SPOT imagery for field delineation in India. Their approach yielded more accurate field delineation in fewer epochs than training directly on the Indian dataset. Simultaneously, the pre-trained model required fewer labels to achieve strong performance, suggesting that it offers a promising approach for addressing (iii).[19]

With respect to both field delineation and crop classification (which are intertwined and complimentary, for example with discrimination between crop types aiding distinction between fields), common practice is the fusion of optical multispectral imagery (Sentinel-2, Landsat) with Synethetic Aperture Radar (SAR), often alongside high-resolution spatial data.[21] More recently, the practice of using time series data rather than single-period classification, which contains useful information on phenological characteristics, has demonstrated improved performance.[22][21]

Within time series analysis, temporal attention mechanisms show promise in enhancing discriminatory phenological characteristics, allowing for more efficient and selective information extraction.[23] As such, in this coursework, I

[1]Definitions vary, but here the Food and Agriculture Organisation define a smallholder farm as one of below 10 hectares in area.[5]

decide to adapt John and Zhang (2022)'s model to investigate the use the U-Net architecture[2] and Temporal Attention Gates in time-series-based crop classification in France using Sentinel 2 L2A imagery. I aim to continue this work to investigate more appropriate models, integration of SAR imagery and spatial information, and, crucially, transferability to smallholder contexts in SSA and Asia.

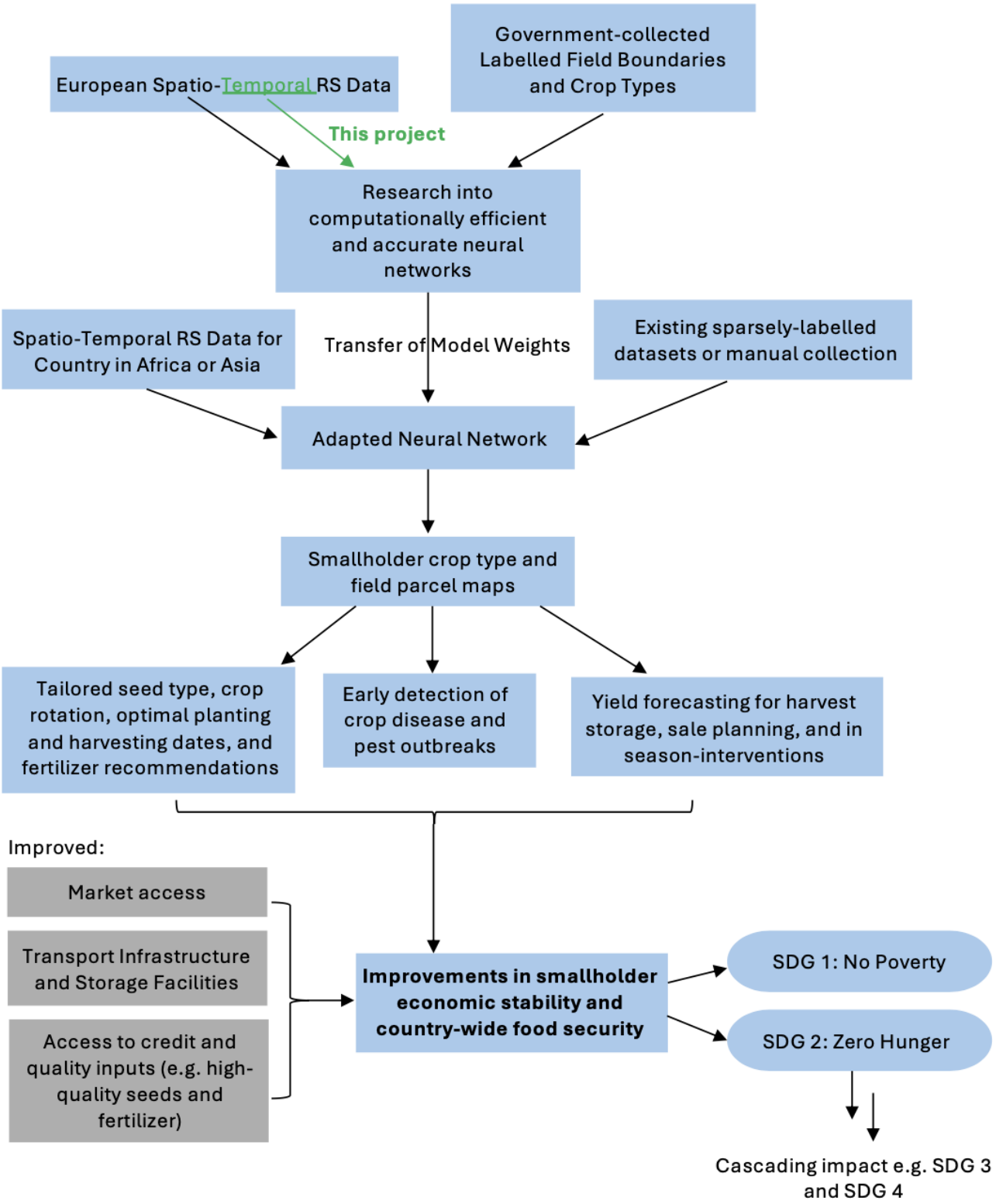

Figure 2: Flowchart illustrating the role of this project in the larger space of smallholder agriculture.

[2]U-Net is typically used for segmentation tasks using spatial data, but has recently seen some success in 1D time-series-based classification (see Zhu *et al.* (2025).[24]

2.2 Ethical Considerations and Sustainability Analysis

The ethical considerations of this project and the wider context of crop-type mapping can be grouped into three main categories.

Firstly, representation bias is a major concern. Crop-type datasets are typically highly imbalanced. Models used to generate predictions on such datasets may assign less importance to minority crop types, amplifying this bias.[25] This risks marginalising farmers producing less common crops. In the case of BreizhCrops, where crop type maps are used to assign EU subsidies, application of biased AI architectures may cause growers of minority crops such as sunflowers and nuts to receive incorrect subsidies. More generally, if AI is to be used to provide agronomic advice (crop husbandry recommendations, crop disease early warning, etc.), suggestions on best practices for minority crop growers may be less reliable than for those cultivating common crop types. Poor advice, and in turn potentially worse crop yields, could have detrimental knock on effects on the food security and economic stability of these individuals. As a result, models of this kind may act as SDG 10 (Reduced Inequalities) inhibitors.[26]

Further, neural networks are computationally demanding architectures.[27] Effective solutions to the issues discussed in 2.1 should be driven by local actors to promote self-sustaining and resilient societies; as such, the potential for transferability to low-resource settings is essential. The high compute requirements of these models may inhibit effective transfer, risking exclusion of local communities from development and decision making processes and, as such, potentially inhibiting Targets 16.7 (Ensure responsive, inclusive and representative decision-making) and 16.8 (Strengthen the participation in global governance).[28]

In a similar vein, neural networks are often regarded as black boxes, where the factors contributing to predictions are obscured.[29] To support policymakers in reliable and fair decision making, robust explainability techniques (for example, integrated gradients or SHAP) must be implemented and evaluated prior to deployment to allow actors to thoroughly investigate potential sources of bias and unreliable rationale.[30]

Privacy and data sovereignty are also key considerations here. Publically available information on field-level data carries risks of competition for resources, potentially leading to over-exploitation, land speculation, or increased conflict among competing stakeholders. In addition, many publically available agricultural datasets are maintained by national governments or private companies. This risks the undermining of local governance and innovation (again inhibiting SDG 16): local institutions and low-income countries may left dependent on data pipelines controlled elsewhere. Transparent datasharing and usage agreements (e.g. aligned with CARE Principles for Indigenous Data Governance) and promotion of local capacity-building for data ownership are essential to promoting longer-term self-sustainability.[31]

Specific ethical consideartions for the BreizhCrops dataset used for this project are discussed in Section 3.2.

2.3 Scalability and Limitations

As Wang *et al.* (2022) note, transfer learning has the potential to enable crop classification and field delineation at scale.[19] However, the approach of using time series data has a number of key limitations. The first limiting factor is temporal noise. The BreizhCrops dataset, like all optical Satellite Imagery Time Series (SITS) datasets, features irregular acquisition dates and cloud contamination.[25] This can be detrimental to model performance, and necessitates error-prone pre-processing interpolation.

Second, generalizability across contexts poses a significant challenge. Model accuracy tends to decrease substantially when deployed across spatially distant regions (spatial transfer) or across different years (temporal transfer).[32] These declines are often attributable to regional variations in meteorological conditions, soil types, and, crucially, phenological timing.[33] The small and irregularly shaped field parcels in sub-Saharan Africa, and often inhomogeneous landscape, further exacerbate this issue.[19] This is also the source of another key limitation in satellite resolution; Wang *et al.* (2022) use Airbus imagery in India to circumvent this issue, but high resolution data is often inaccessible or costly.[19]

Third, the design choice to use 1D parcel-averaged features results in the inevitable neglect of local spatial features. Although this simplification reduces computational load and regularizes the input for time series analysis, it sacrifices spatial context that might be necessary to resolve subtle spectral similarities between different crop types or distinguish fine boundary details.10 Spatio-Temporal models offer a promising route forward, integrating the benefits of both data types.[34]

Finally, the lack of labelled fields in the Global South represents a major barrier to supervised and semi-supervised labelling algorithms. Greater investment into ground-truthing is a necessary step to enabling robust analysis and scalability.[19]

3 Context-Appropriate Alternative Dataset

3.1 Dataset Identification

The BreizhCrops dataset provides a benchmark for crop classification using time-series data, comprising approximately 608,000 field-level Sentinel-2 observations for nine crop varieties in Brittany, France[3] between January 1 and December 31, 2017. The raw data were obtained from the *Institut national de l'information géographique et forestière* (IGN), which provides open-access anonymised parcel shapes and associated crop types.

The observations are subdivided into four regions within Brittany: Côtes-d'Armor (FRH01), Finistère (FHR02), Ille-et-Vilaine (FRH03), and Morhiban (FRH04). Both top- (L1C, 13 spectral bands) and atmospherically corrected bottom-of-atmosphere (L2A, 10 spectral bands) processing levels are included, with reflectance levels averaged over each individual field.[25]

Code	Crop Type
0	barley
1	wheat
2	rapeseed
3	corn
4	sunflower
5	orchards
6	nuts
7	permanent meadows
8	temporary meadows

Table 2: Crop varities included in BreizhCrops and associated class codes.[35]

Atmospherically-corrected satellite data has been shown to improve landcover mapping tasks, and L2A data is largely preferred to L1C in Sentinel-2-based crop classification.[36][24][37] Transferability to other regions has previously been cited as rationale for L1C use [25], and is an essential consideration here (Section 2.1); however, as of January 2017, global L2A Sentinel-2 data is openly available.[38] As such, the BreizhCrops L2A data is selected for investigation in this work.

3.2 Access Process and Ethical Considerations

The IGN instituted key mitigation measures regarding data stewardship: the release involves anonymized parcel geometries and the data is offered under an open license policy.[25] These actions address fundamental principles of privacy ethics and promote transparency, allowing wide academic access while protecting the confidentiality of individual farming units.[39]

However, the nature of the data collection introduces a number of ethical considerations. Firstly, EU farmers are legally mandated to provide information on plot shape, size, and crop type cultivated to enable distribution of crop-based subsidies.[25] This undermines the validity of the consent provided: often reliant on these subsidies to turn a profit, they have little choice but to agree to the sharing of their data. Further, whether there is transparency regarding possible uses of the data provided is unclear. If there is the possibility of an AI system being deployed to make decisions directly impacting a farmer's financial standing, Fairness, Accountability, and Transparency (FAT) must be ensured in the decision-making process. This context mandates that the development and operation of the 1D Attention U-Net must include mandatory bias audits and integrate policies that ensure transparent decision protocols.[40] The selection of an attention-based architecture, therefore, also requires parallel development of tailored Explainable AI methods that can interpret the temporal weights assigned by the attention mechanism to provide clear, auditable explanations for classification outcomes.[41] Third, while the IGN employ teams to partially verify this data, the self-reported nature of the data may lead to errors and inaccuracies that risk introducing bias into the model.[25]

3.3 Preprocessing Methods

As per the recommendation of Rußwurm *et al.* (2020), regions FRH01 and FRH02 are used in training, FRH03 in validation, and FRH04 in testing, avoiding spatial leakage.[25] In the original dataset, the resulting training, validation, and test sets contains 319,258, 166,391, and 122,614 samples respectively. The crop distributions in the respective sets are displayed in Figure 1. Due to RAM limitations and considerations of computational efficiency, subsampling is investigated. Previous studies of this dataset have observed no reduction in training accuracy when subsampling to a total train/validation/test sample size of 28,000.[42] As such, a train/validation split of 22,000/6,000 is employed here.

[3]Mean annual temperatures: 5.6°C (winter) to 17.5° (summer); mean annual precipitation: 650 mm.[25]

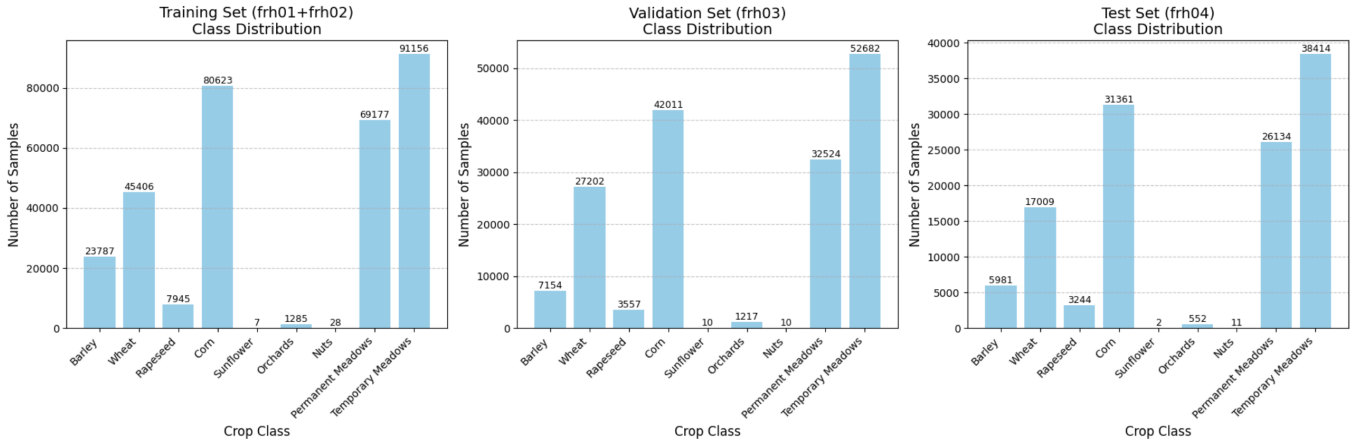

Figure 3: Original distributions in unmodified training, validation, and test sets.

3.3.1 Class Imbalance Mitigation

The datasets suffer from extreme class imbalance, containing three classes with frequencies below 1%, two of which (sunflower and nuts) contain only 7 and 29 samples respectively (Figure 1). Addressing class imbalance requires careful consideration. Previous crop classification studies observe increased bias and reduced overall classification accuracy - as well as low accuracy of categories with few samples - with high sample imbalance (reword), and early testing here demonstrates extremely poor performance on *Orchards* classification.[43][44] However, extensive manipulation of the training set distribution risks information loss from the true data distribution and prediction bias due to misrepresentation of reality in the training set, leading to poor generalisation.[45]

Here, a combination of oversampling, augmentation, and downsampling techniques are employed to robustly balance the training set. The validation and test sets are not included, so as to ensure robust evaluation of performance on real-world, imbalanced data. The validation set is drawn to a size of 6,000 samples via stratified random sampling, while the test set is left unchanged. Early testing (A1) demonstrated the detrimental impact of low sample numbers on the predictability of the *Orchards*, *Sunflower*, and *Nuts*. For the former, Oversampling method to resolve the High-dimensional Imbalanced Timeseries classification (OHIT) is employed. Time series data necessitates oversampling techniques that are faithful to its sequential nature and that preserve correlations between adjacent features within a given time step; OHIT is specifically designed to address these issues in the context of classification tasks with imbalanced data[4] and as such is highly relevant to this task.[46][47] Balancing the need for greater representation of *Orchards* data to bring it into the "learnable" regime and the risk of overfitting with high proportions of synthetic data [48][49], a cap of 50% synthetic data is employed.

Sunflower and *Nuts* lack sufficient data volume to support reliable interpolation.[50][51] All data points in the original training set are kept in the subsampled set and data augmentation is employed in preference to increase diversity of the training set without generating physically implausible data points.[52] The common time series augmentations employed are jittering (adding minor, random noise to the time series spectral values) and scaling (applying random scaling factors to the entire time series, simulating differences in cloud cover influence or atmospheric optical depth).[52] Ideally, the transformations are applied dynamically during the train process - in each epoch, the model encounters a different version of the original samples - to increase the *effective* sample size and prevent overfitting on the limited initial data points. However, here RAM limitations meant that only static augmentation was possible.[53]

In line with the chosen train/val size of 28,000, majority classes are downsampled using random subsampling to approximately 3,200 samples each. *Permanent Meadows* and *Temporary Meadows* showed the strongest confusion with *Orchards* in early testing (A1), so are given a slightly higher allocation to increase diversity in those two classes. The final training set distribution is displayed below (Figure 4).

3.3.2 Feature Engineering and Sequence Length

Vegetation indices (VIs) are widely adopted in remote sensing crop classification to enhance spectral characteristics of crops, enabling improved discrimination between crop types.[54] The optimal combination of such indices is highly dependent on the crop types under investigation; however, the combination of complimentary NDVI and red edge indices is often reported to improve crop classification accuracy.[24][42][55][56] With respect to BreizhCrops, Yuan *et*

[4]OHIT operates by first employing a density-ratio-based shared nearest neighbour clustering algorithm to capture distinct sub-clusters within the high-dimensional feature space of the minority class. For each identified sub-cluster, a shrinkage technique is applied to its covariance matrix, allowing for accurate and reliable estimation of each cluster's unique covariance structure. Finally, the synthetic samples are generated based on a multivariate Gaussian distribution parametrized by the estimated mean and preserved covariance matrix to ensure that the synthetic data adheres to the statistical manifold defined by the original, rare samples, minimizing the risk of introducing falsified temporal dependencies.[46]

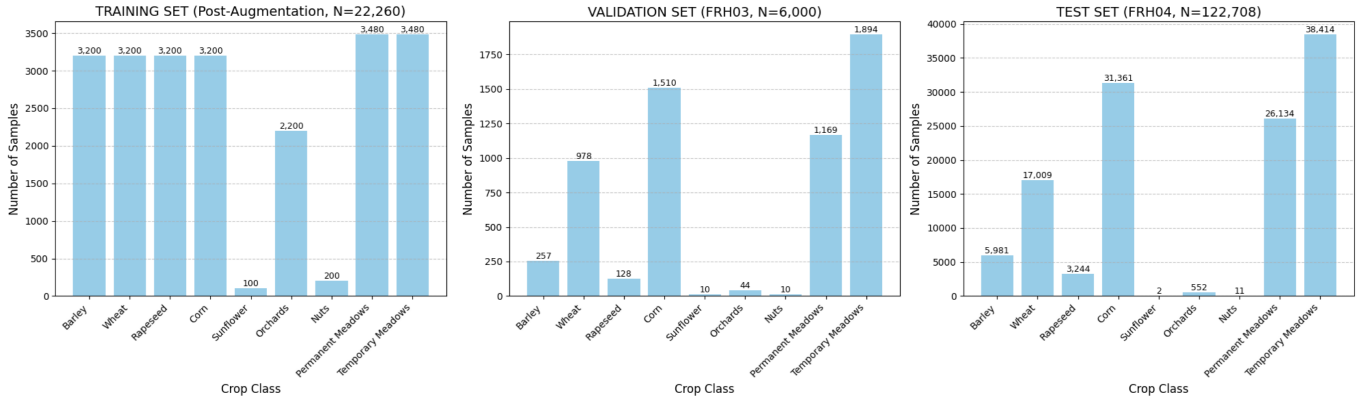

Figure 4: Class distributions in final training, validation, and test sets.

al. (2023) rank feature importance of 13 VIs in crop classification with RandomForest using L1C data and observe consistent improvement in overall accuracy (OA) with additional VIs, and the highest OA with all thirteen.[42] The same authors investigated the impact of sequence length in increments of five units, finding that overall accuracy improved up to a sequence length of 50, followed by a sharp drop at 55, postulated by the authors to be due to a reduction in information on minority classes.[42] Figure 5 displays their results.

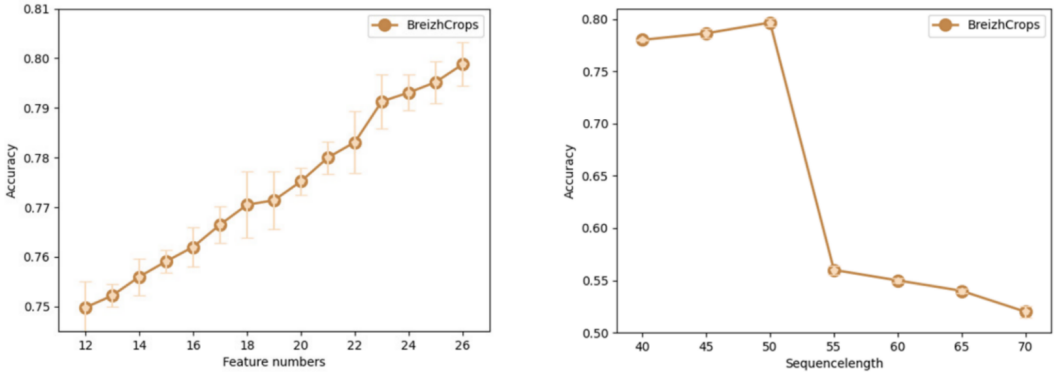

Figure 5: Reproduction of Yuan *et al.* (2023) analysis of the impact of number of VI features and feature length on overall accuracy (OA).[42]

To obtain the baseline feature set, the VIs selected by Yuan *et al.* (2023) are combined with NDRE and MSAVI. The former is regarded as the gold-standard for resisting the saturation common to B4-based indices during peak growth and has demonstrated excellent performance when combined with NDVI, providing stability and discrimination during the high-biomass phase (peak season).[24] The latter offers dynamic minimisation of soil background influence, compared to the fixed soil adjustment factor required by SAVI. 48 is used as an optimal feature length for the baseline model developed in Section 4, abiding by the findings observed by the above authors while ensuring compatibility with the model architecture. The VIs used are detailed in Table 1.[5]

3.3.3 Masking and Interpolation

For the L2A data, Rußwurm *et al.* (2020) obtained 374 images from PEPS, covering the seven S2 tiles for Brittany. PEPS employs the MAJA image processing chain to produce the L2A data.[CITE] This filters out images with cloud-cover over 80% before applying atmospheric, adjacency, and slope corrections.[25]

As the authors note, this process leaves two key dataset characteristics for practioners to address in pre-processing. First, despite the initial filtering of images with greater than 80% cloud cover, residual atmospheric artifacts such as clouds and shadows persist, leading to positive outliers and localised disturbances in the time series reflectance values (see Figure - plot spectral bands). Second, due variable acquisition times across the multiple tiles and the non-uniform removal of cloudy observations across the monitoring period, aggregation of data from overlapping tiles yields irregularly spaced time series for individual field parcels.[25]

To address these issues, various interpolation techniques are considered. Linear interpolation has been shown to provide

[5]It is noted, however, that Yuan *et al.* (2023)'s methodology and results (i) are for L1C data rather than the L2A data used in this work and (ii) give little consideration to computational cost from increased dimensionality, which is an important consideration if these classification models are to be effectively and accesibly adapted for low-resource regions.[57] As such, Section 5 incorporates ablation studies to more accurately inform feature selection and sequence length in the context of BreizhCrops L2A data.

Feature Variables	Calculation Formula	Resolution (m)
NDVI2 [24]	$NDVI2 = \frac{(B8 - B4)}{(B8 + B4 + 0.1)}$	10
BI [23]	$BI = \sqrt{\frac{2 \times B4^2}{B3^2}}$	10
VARI [25]	$VARI = \frac{(B3 - B4)}{(B3 + B4 - B2)}$	10
NDWI [26]	$NDWI = \frac{(B8 - B8a)}{(B8 + B8a)}$	20
IRECI [23]	$IRECI = \frac{(b7 - b4) \times b6}{B5}$	20
MTVI2 [27]	$MTVI2 = \frac{1.5 * (1.2 * (B8 - B3) - 2.5 * (B5 - B3))}{\sqrt{(2 \times B8 + 1)^2 - 6 \times B5 + 5 \times B3 + 0.5}}$	20
RVI [25]	$RVI = \frac{B8}{B4}$	10
GCVI [28]	$GCVI = \frac{B4}{B3} - 1$	10
MNDWI [29]	$MNDWI = \frac{(B3 - B11)}{(B3 + B11)}$	20
EVI [30]	$EVI = \frac{2.5 \times (B8 - B4)}{B8 + 6 \times B4 - 7.5 \times B2 + 1}$	10
SAVI [31]	$SAVI = \frac{1.5 \times (B8 - B4)}{(B8 + B4 + 0.5)}$	10
BCI [32]	$BCI = 0.1360 \times B3 + 0.2611 \times B4 + 0.3895 \times B8$	10
GNDVI [33]	$GNDVI = \frac{(B8 - B3)}{(B8 + B3)}$	10
MSAVI [34]	$\text{MSAVI} = \frac{2B8 + 1 - \sqrt{(2B8 + 1)^2 - 8(B8 - B4)}}{2}$	10
NDRE [35]	$\text{NDRE} = \frac{B8 - B5}{B8 + B5}$	10

Table 3: VI indexes used.[42]

a good trade-off between accuracy and computational efficiency for cloudy S2 pixels, with no substantial improvement observed using higher-order methods such as cubic splines.[58] Savitzky-Golay filtering is widely applied for medium-resolution sensors and effectively smooths NDVI or spectral time series, enabling preservation of phenological dynamics while mitigating noise.[59][60] Harmonic and parametric methods, such as HANTS, assymetric Gaussian, and double logistic can achieve robust smoothing and phenology extraction, but their performance varies with geographic region and thus risk poor generalisability.[61][62]

Given Brittany's mid-latitude agricultural landscape and considerations of computational efficiency and transferability, linear interpolation and Svavitzky-Golay filtering, for gap filling and temporal smoothing respectively, represent appropriate choices here. This converts the irregularly sampled observations into a fixed-length temporal grid (T = 50) and provides a smooth approximation of the expected reflectance trajectory. To retain information about data availability, an imputation mask is generated to record the original observation pattern (1 = measured, 0 = missing). The mask is then mapped to the 50-step temporal grid using nearest-neighbour assignment to preserve its binary structure, and concatenated to the feature vector to allow the model to explicitly account for the reliability of the imputed steps during training.

3.3.4 Standardisation

Sentinel 2 reflectance data is typically non-Gaussian, heavily skewed, and susceptible to noise and outliers.[63] RobustScaler, which centres data on the median and scales based on the interquartile range, is commonly used in remote

sensing tasks owing to its resilience to skew and efficacy in handling outliers and noisy features.[64][65] This approach is adopted here, with RobustScaler applied on the raw spectral bands prior to VI calculation to avoid propagation of noise into the calculated indices.

4 Modifications to Model Architecture

4.1 Justification of Architectural Modifications

4.1.1 Structural Modifications

(i) All core convolutional layers were changed from Conv2D to Conv1D and Conv2DTranspose to Conv1DTranspose, adapting the model from processing spatial features (Height x Width) to temporal features (Time Steps).

(ii) The input shape was changed from a spatial image format to a temporal sequence format ($T_{regularised}$, $F_{augmented}$), representing (50, 25) (Time Steps x Features)

(iii) MaxPooling2D and UpSampling2D were replaced with their one-dimensional counterparts, ensuring operations only affect the time dimension.

(iv) The attention-block (a 2D mechanism) was re-implemented as the TemporalAttentionGate(a 1D mechanism), shifting focus from local spatial context to phenologically relevant time steps.

(v) A custom mechanism using Cropping1D and a STATIC-CROPPING-MAP was introduced to align the varying sequence lengths resulting from pooling on the input T=50 dimension. This fixes the temporal mismatch between the encoder's skip connections and the decoder's upsampled features.

(vi) BatchNormalisation layers are added after every convolution to improve gradient flow and training stability for the 1D time series data.[1]

(vii) The final output layer filter count was changed from 1 (for binary deforestation) to NUM-CLASSES = 9 (for crop classification), and the final output layer activation was changed accordingly from Sigmoid to Softmax to allow multi-class classification

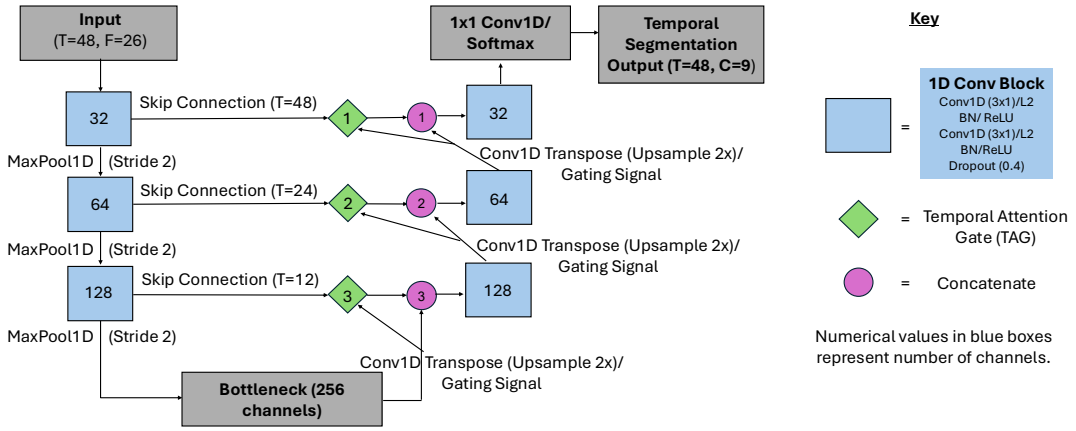

Figure 6: A network architecture diagram of the adapted model.

4.1.2 Loss Function

The imbalanced nature of the validation and testing sets means that many of the loss functions commonly employed in deep learning classification tasks are sub-optimal. For example, Cross-Entropy Loss, employed by John and Zhang (2022) in the baseline methodology [1], assigns equal weights to all classes and can therefore lead to bias towards to majority class since it.[66] Focal Loss and DICE Loss offer suitable alternatives.[67][24] Here, Farhadpour *et al.* (2024)'s guide on loss function and metric selection for imbalanced datasets is relied upon. Here it is observed that for balanced training data and imbalanced validation/testing data, as in this context, Macro-Dice loss exhibited superior performance; it is chosen for use here.[68] It should be noted that class-weighting becomes redundant in this scenario as Dice-loss treats minority and majority classes equally.[69]

4.1.3 Optimiser

While Adam optimiser is used in the baseline architecture and widely employed in crop classification time series tasks [70][71][72][73], AdamW, which modifies its predecessor by decoupling the L2 regularisation from the adaptive gradient rule, is employed here, having been shown to exhibit superior performance both generally and with time series data.[74][75]

4.1.4 Early Stopping Strategy

The model is trained for 50 epochs, incorporating the ReduceLROnPlateau mechanism to dyanmically adjust the learning rate based on the validation set loss to support stable convergence, with patience set to 5 epochs.[24] To improve efficiency upon lack of validation loss improvement, patience for early stopping is set to 15 epochs. [67]

4.1.5 Dropout

John and Zhang (2022) note that while the addition of Dropout is not tested in their experimentation, it has been shown to improve deep learning model performance in multiple domains.[1] Rußwurm *et al.* (2020) observe that 40% dropout is optimal for the BreizhCrops Transformer model, while Gadiraju *et al.* (2020) find the optimal dropouts for DenseNet and ResNet50 on this dataset were 0.2 and 0.5, respecitively.[25][76] Guided by these similar models trained on the same dataset, 40% dropout is chosen here.

4.1.6 Weight Initialization:

He initialisation, which represents common practice in U-Net architectures using the ReLu activation function, is kept from the baseline model architecture.[1][77]

4.1.7 Metrics

Metrics choice is guided by the need to ensure comparability with both the baseline study and crop-classification literature, while maintaining consistency the use of the Macro-Dice loss function. John and Zhang (2022) evaluate their model for deforestation classification on IoU, Precision, Recall, and F1-score; IoU is designed for spatial segmentation tasks and thus is not relevant here (the 1D Attention U-Net outputs a single crop label per field rather than a pixel-level mask).[1] Rußwurm *et al.* (2020) employ overall accuracy, average accuracy, weighted f-score, and kappa.[25] Additional metrics used in the literature include multiclass Macro-Average F1 score, which is mathematically aligned with Macro-F1.[76]

To balance these requirements, multiclass Macro-F1 is adopted as the core metric, reflecting both the class-balanced nature of the loss function and the imbalanced distribution of the validation and test data. To allow comparison to both baseline papers, the following additional metrics are also calculated:

1. Per-Class Precision, Recall, and F1-Score to align with the deforestation baseline
2. Overall Accuracy, Average Accuracy, Weighted F-score and Kappa to align with the BreizhCrops baseline. However, these metrics are imbalance-sensitive and thus are not relied on.

4.2 Hyperparameter Tuning

Bayesian optimisation allows efficient and intelligent tuning of batch size, weight decay, and learning.[78] John and Zhang (2022) use a batch size of 1 on their image segmentation task.[1] This is inappropriate for time series data. Batch size choice in relevant crop classification literature varies from 8 to 256.[25][79][80][67] larger batch sizes can speed up training but may generalize less effectively.[81] To balance these considerations, the batch sizes investigated were $32, 64, 128$. John and Zhang (2022) find a learning rate of 0.0005 to be optimal for the Attention U-Net model in deforestation segmentation.[1] The optimal learning rates observed in crop classification literature generally fall within the range of 10-5 to 10-3, which is therefore chosen as the search range.[25][46][79][80] The weight decay search range is defined as 10-8, 10-4 to balance the BreizhCrops Transformer optimal weight decay of 5.52 x 10-8 and common practice.[25][82] Tuning is performed over 15 iterations with patience for ReduceLROnPlateau and early stopping also at 5 and 15 respectively. Performance is evaluated on validation Macro F1 score, for which the highest is Epoch 7 of Iteration 4. The model is rerun for this configuration and model weights are saved; the optimised hyperparameters are displayed below.

Hyperparameter	**Value**
Batch Size	64
Learning Rate	4.6×10^{-5}
Weight Decay	4.0×10^{-8}

Table 4: Hyperparameters that minimise Macro F1 on the validation set, obtained via BayesSearch.

5 Model Evaluation

5.1 Initial Results and Analysis of Failure Cases

Initial results demonstrate poor performance of the Attention U-Net model on this crop classification task. Apart from *Corn*, the second most populous class which exhibits a reasonable Macro F1-score (0.7246) and high precision (0.8488),

Macro F1-Score	Overall Accuracy (OA)	Average Accuracy (Macro Recall)	Weighted F-score	Kappa Statistic
0.2574	0.4446	0.2807	0.4673	0.3051

Table 5: Summary Performance Metrics for Initial Model.

Class ID	Class Name	Support	Precision	Recall	F1-Score
0	Barley	287088	0.2076	0.5378	0.2996
1	Wheat	816432	0.4293	0.3543	0.3882
2	Rapeseed	155712	0.0743	0.2313	0.1124
3	Corn	1505328	0.8488	0.6321	0.7246
4	Sunflower	96	0.0000	0.0000	0.0000
5	Orchards	26496	0.0000	0.0000	0.0000
6	Nuts	528	0.0000	0.0000	0.0000
7	Permanent meadows	1254432	0.3469	0.3963	0.3700
8	Temporary meadows	1843872	0.4817	0.3746	0.4214

Table 6: Per-Class Performance Metrics for Initial Model.

the crop types are poorly distinguished. Oversampling and feature augmentation failed to improve classification of the three classes underrepresented in the original dataset, all of which show F1-scores of zero.

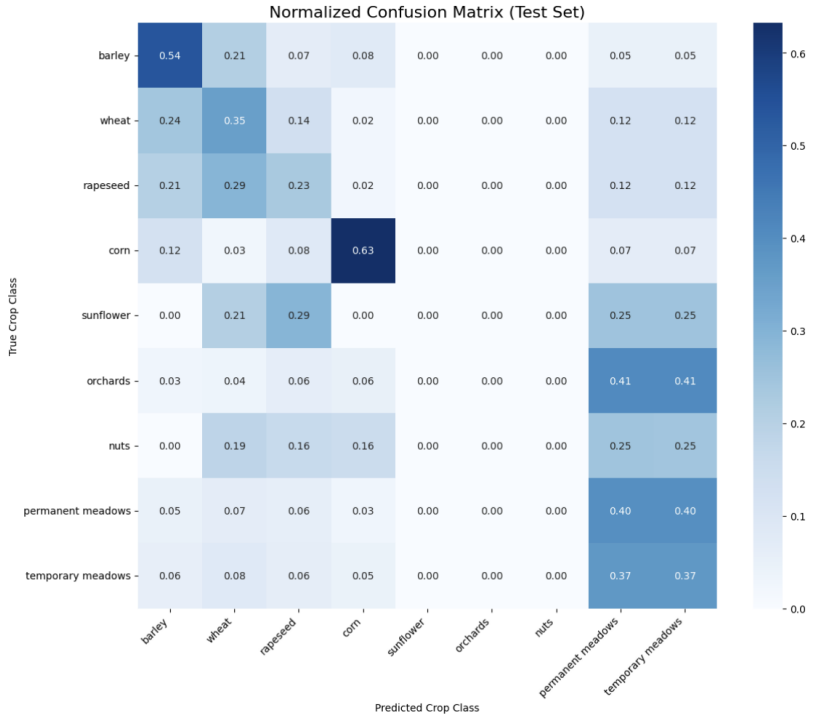

Figure 7: Confusion Matrix for Initial Model.

The confusion matrix allows elucidation of key misclassifications. Wheat, barley, and temporary meadows are the major misclassification sinks. Barley is most often misclassified as wheat (21% of the time), and wheat as barley (24%). Rapeseed suffers most from confusion with both barley (21%) and wheat (29%). Corn is fairly well classified, but is confused with barley 12% of the time. The model completely fails on sunflowers, orchards, and nuts. Sunflowers are approximately uniformly misclassified as wheat (21%), rapeseed (29%), and permanent and temporary meadows (both 25%). Orchards are most frequently misclassified as permanent and temporary meadows (41% of the time for both), as are nuts (25% of the time for both). Meadows are well distinguished from other crop types, but poorly from each other (permanent meadows are misclassified as temporary meadows 40% of the time, and temporary as permanent 37% of the time). There is clear positive correlation between the number of data points for each crop in the test set and classification performance, suggesting issues with weighting and training set balance that are addressed in the following ablation studies.

5.2 Ablation Studies

Four ablation studies, described below, are conducted to refine model design and improve performance. The results are detailed in Table 8, and the related confusion matrices and Per-Class statistics are available in the Appendix.

(i) *Inverse Frequency Class Weighting:* The poor performance of the initial model indicate that the Macro-Dice Loss function is failing to manage the weight disparity. Inverse Frequency Weighting is introduced as a possible means of mitigating this issue and, in turn, improving the currently ineffective classification of minority classes. It is observed that this results in minor improvements to minority classes *Orchards* and *Rapeseed*. However, it has no impact on the complete misclassification of *Sunflower* and *Nuts* previously observed, and is detrimental to the performance of more common classes. For example, the F1-score of corn falls from 0.7246 to 0.4931. Macro F1, OA, Average Accuracy, Weighted F-score, and Kappa also all fall.

(ii) *Analysis of Training Set Imbalance:* An imbalanced version of the training set is preprocessed using random stratified subsampling, replacing the current oversampling and feature augmentation techniques. This is employed to validate whether prediction bias introduced by under- and oversampling is detrimental to the model. Now training, validating, and testing on imbalanced sets necessitates a change in loss function from the Macro Dice-Loss that was specifically employed due to the balanced-imbalance disparity between training and validation/test sets. Weighted Cross entropy loss commonly employed [68] is commonly employed. This additional change should be considered when analysing the results.

(iii) *Feature Set Optimisation:* Mutual information scores coupled with Sequential Forward Selection highlights MTVI2 as the most predictive feature, and demonstrates (contrary to Yuan*et al.* (2023)'s findings, that additional VIs do not offer significant improvement. As such, only MTVI2 is kept in the final model.

(iv) *Vegetation Index Scaling:* Reflectance bands were initially scaled before calculation of VIs. Despite this scaling inevitably diminishing their physical interpretability, it was assumed that scaling was necessary for faster convergence and avoidance of saturation. However, upon testing, calculation of MTVI2 prior to scaling results in large performance improvements, suggesting that model performance benefits

Ablation Study	Macro F1-Score	Overall Accuracy (OA)	Average Accuracy (Macro Recall)	Weighted F-score
(i)	0.2226	0.3681	0.2393	0.3970
(ii)	0.1712	0.3374	0.2409	0.3240
(iii)	0.2626 (Validation)	-	-	-
(iv) (Final Model)	0.3456	0.4014	0.3565	0.4676

Table 7: Summary of results of ablation studies. Kappa is recorded but not reported here, as per Farhadpour et al. (2024)'s discussion of its lack of utility.[68]

5.2.1 Comparison with U-Net and Baseline

Finally, the Attention U-Net model is compared against a U-Net model with the attention gates removed. Interestingly, Test Macro F1 exhibits a large increase from 0.3456 to 0.4361. The per-class metrics are shown in the table below for closer analysis.

	Attention U-Net			**U-Net**		
Class Name	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Barley	0.8115	0.1595	0.2665	0.8169	0.4831	0.6072
Wheat	0.9563	0.6315	0.7607	0.9457	0.7247	0.8206
Rapeseed	0.9929	0.6449	0.7819	0.9848	0.7088	0.8243
Corn	0.9968	0.2874	0.4462	0.9954	0.5144	0.6783
Sunflower	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Orchards	0.0115	0.6333	0.0226	0.0155	0.5274	0.0301
Nuts	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Permanent meadows	0.4049	0.5393	0.4626	0.4320	0.5701	0.4915
Temporary meadows	0.4533	0.3127	0.3701	0.5040	0.4447	0.4725

Table 8: Comparison of Per-Class Performance Metrics for Attention U-Net and U-Net

It is evident that low recall is key to the reduced performance of Attention U-Net. The Attention-based model exhibits similar precision to U-Net, but is far more cautious, resulting in a significantly higher rate of false negatives (drastically so for Barley and Corn). Both models show complete failure on minority class classification, perhaps suggesting that more aggressive oversampling approaches could be valuable.

To further investigate the performance of the attention mechanism, attention weights are plotted (here, two corn samples are displayed). The plots, while qualitative, suggest inconsistent alignment of the attention weights with phenological patterns. It is possible that the TAGs are over-filtering important low-level phenological features, while the additional trainable parameters and more complex loss landscape may outweigh any gains from feature refinement.

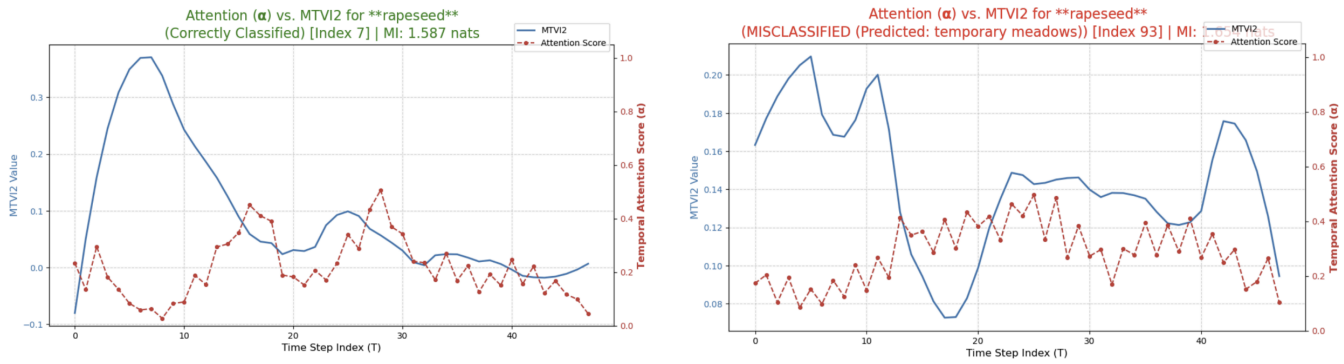

Figure 8: Comparing the relationship between attention weights and MTVI2 for classified and misclassified samples of rapeseed.

More generally, we observe performance is drastically inferior to the baseline (and to performance of LSTMs, Transformers, and RNNs on this dataset (which achieve OAs of 0.8 [83]). The baseline 2D convolutional structure and skip connections clearly excel at retaining spatial localisation and context; it is possible that localised nature of the 1D convolutions prevent the 1D U-Net from successfully capturing the long-range temporal dependencies needed to discriminate between full-season phenological curves.

5.3 Conclusions

It is evident, and perhaps unsurprising, that the U-Net model is a relatively ineffective architecture in the domain of time series crop classification. More interestingly, it is observed that Temporal Attention Gates are of clear detriment to performance, perhaps indicating over-filtering of important phenological features. Methodologically, key takeaways are the detrimental impact of calculating vegetation indices based on scaled reflectances, which appears to destroy their physical interpretability, and the surprisingly high mutual information of MTVI2 with crop class (NDVI and NDRE are far more commonly employed). These pre-processing and feature engineering methods should be further explored with more appropriate models, as they may result in improved performance in other contexts.

A Appendix

A.1 Ablation Study Tables and Confusion Matrices

Class ID	Class Name	Support	Precision	Recall	F1-Score
0	Barley	287088	0.1922	0.4435	0.2682
1	Wheat	816432	0.3333	0.3428	0.3380
2	Rapeseed	155712	0.0842	0.2410	0.1248
3	Corn	1505328	0.8032	0.3558	0.4931
4	Sunflower	96	0.0000	0.0000	0.0000
5	Orchards	26496	0.0022	0.0001	0.0001
6	Nuts	528	0.0000	0.0000	0.0000
7	Permanent meadows	1254432	0.3357	0.3943	0.3626
8	Temporary meadows	1843872	0.4678	0.3760	0.4169

Table 9: Ablation 1: Per-Class Performance Metrics with addition of inverse frequency weighting.

Macro F1-Score	Overall Accuracy (OA)	Average Accuracy (Macro Recall)	Weighted F-score	Kappa Statistic
0.2226	0.3681	0.2393	0.3970	0.2216

Table 10: Ablation 1: Model Performance Metrics with addition of inverse frequency weighting.

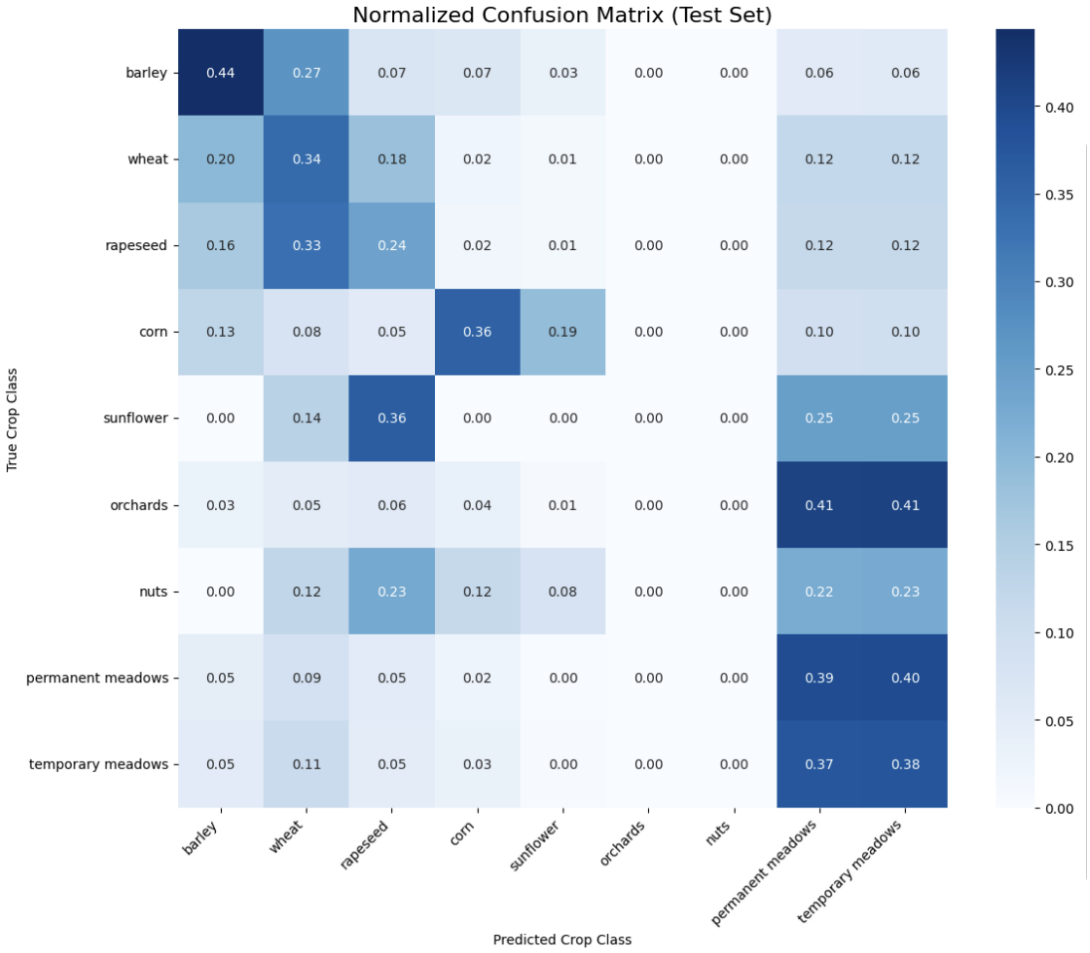

Figure 9: Ablation 1: Confusion Matrix after Inverse Frequency Weighting is applied.

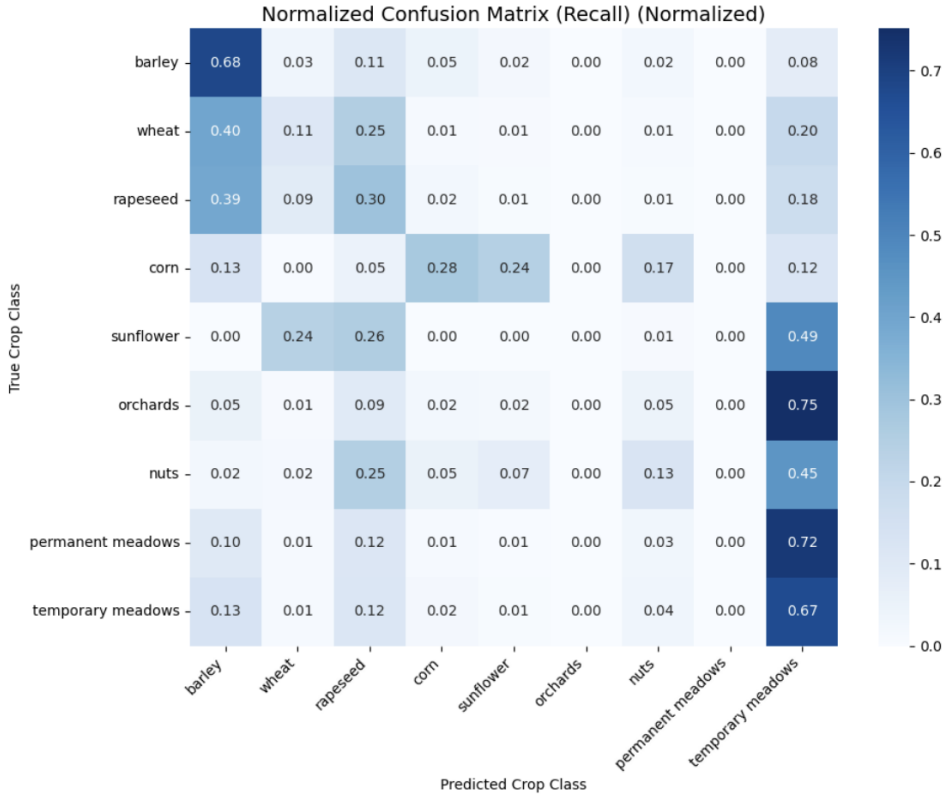

Figure 10: Ablation 2: Stratified Crop Class Distributions.

References

[1] David John and Ce Zhang. An attention-based u-net for detecting deforestation within satellite sensor imagery. *International Journal of Applied Earth Observation and Geoinformation*, 107:102685, 2022. ISSN 1569-8432. doi: https:

Class ID	Class Name	Support	Precision	Recall	F1-Score
0	Barley	287088	0.1692	0.6817	0.2711
1	Wheat	816432	0.5635	0.1114	0.1860
2	Rapeseed	155712	0.0640	0.3001	0.1055
3	Corn	1505328	0.8261	0.2782	0.4162
4	Sunflower	96	0.0000	0.0000	0.0000
5	Orchards	26496	0.0000	0.0000	0.0000
6	Nuts	528	0.0002	0.1269	0.0004
7	Permanent meadows	1254432	0.4367	0.0001	0.0002
8	Temporary meadows	1843872	0.4837	0.6696	0.5617

Table 11: Ablation 2: Per-Class Performance Metrics with Imbalanced Training Set.

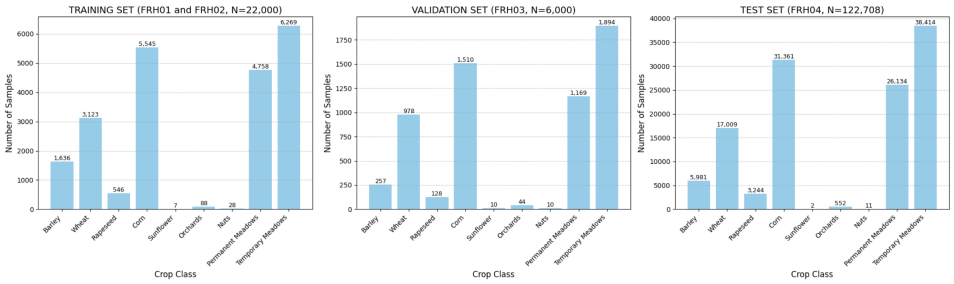

Figure 11: Ablation 2: Confusion Matrix with Imbalanced Training Set

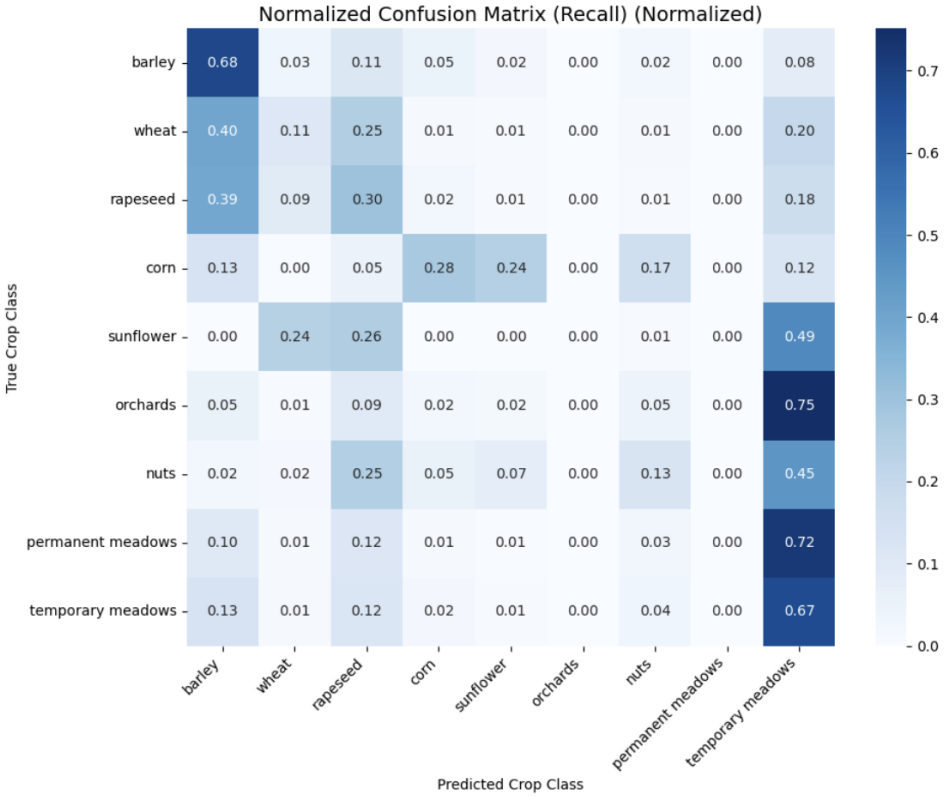

Figure 12: Ablation 4: Unscaling of VIs

//doi.org/10.1016/j.jag.2022.102685. URL https://www.sciencedirect.com/science/article/pii/S0303243422000113.

[2] Lucimara Bragagnolo, Roberto Valmir da Silva, and José Mario Vicensi Grzybowski. Amazon rainforest dataset for semantic segmentation, May 2019. URL https://doi.org/10.5281/zenodo.3233081.

[3] Lucimara Bragagnolo, Roberto Valmir da Silva, and José Mario Vicensi Grzybowski. Amazon and atlantic forest image datasets for semantic segmentation, February 2021. URL https://doi.org/10.5281/zenodo.4498086.

[4] Yacob Abrehe Zereyesus, Lila Cardell, Jarrad Farris, Kayode Ajewole, Michael E. Johnson, Jessie Lin, Constanza Valdes, and Wendy Zeng. Global food assessment, 2025–35, sep 2025. URL https://ers.usda.gov/sites/default/files/_laserfiche/publications/113294/GFA-36.pdf?v=69234.

Vegetation Index	MI Score
MTVI2	0.1300
MSAVI	0.0842
SAVI	0.0663
IRECI	0.0636
GCVI	0.0617
VARI	0.0596
RVI	0.0569
EVI	0.0528
BI	0.0516
NDRE	0.0402
NDWI	0.0372
NDVI2	0.0368
MNDWI	0.0365
GNDVI	0.0356
BCI	0.0312

Table 12: Ablation 3: Mutual Information Ranking of Vegetation Indices

Class ID	Class Name	Support	Precision	Recall	F1-Score
0	Barley	287088	0.8115	0.1595	0.2665
1	Wheat	816432	0.9563	0.6315	0.7607
2	Rapeseed	155712	0.9929	0.6449	0.7819
3	Corn	1505328	0.9968	0.2874	0.4462
4	Sunflower	96	0.0000	0.0000	0.0000
5	Orchards	26496	0.0115	0.6333	0.0226
6	Nuts	528	0.0000	0.0000	0.0000
7	Permanent meadows	1254432	0.4049	0.5393	0.4626
8	Temporary meadows	1843872	0.4533	0.3127	0.3701

Table 13: Ablation 4: Per-Class Performance Metrics Attention U-Net

[5] Food and Agriculture Organization of the United Nations (FAO). Smallholders and family farmers - Sustainability Pathways (factsheet), 2012. URL https://www.fao.org/fileadmin/templates/nr/sustainability_pathways/docs/Factsheet_SMALLHOLDERS.pdf.

[6] Atakilte Beyene. Small farms under stress: Smallholder Agriculture and Emerging Global Challenges. Technical Report 5, The Nordic Africa Institute, 2014. URL https://www.files.ethz.ch/isn/183918/FULLTEXT01small%20farmholders.pdf.

[7] Chenchen Ren, Liyin He, Yuchi Ma, Stefan Reis, Hans Van Grinsven, Shu Kee Lam, and Lorenzo Rosa. Trade-offs in agricultural outcomes across farm sizes. *Earth Critical Zone*, 1(1):100007, 2024.

[8] David Lam, Murray Leibbrandt, and James Allen. The demography of the labor force in sub-saharan africa: Challenges and opportunities, November 2019. URL https://g2lm-lic.iza.org/wp-content/uploads/2019/11/glmlic_sp010.pdf.

[9] Van Touch, Daniel K.Y. Tan, Brian R. Cook, De Li Liu, Rebecca Cross, Thong Anh Tran, Ariane Utomo, Sophea Yous, Clemens Grunbuhel, and Annette Cowie. Smallholder farmers' challenges and opportunities: Implications for agricultural production, environment and food security. *Journal of Environmental Management*, 370:122536, 2024.

[10] Austine Phiri, George T Chipeta, and Winner D Chawinga. Information needs and barriers of rural smallholder farmers in developing countries: A case study of rural smallholder farmers in malawi. *Information Development*, 35(3):421–434, 2019.

[11] Rajveer Dhillon and Qianna Moncur. Small-scale farming: A review of challenges and potential opportunities offered by technological advancements. *Sustainability*, 15(21), 2023.

[12] Food and Agriculture Organization of the United Nations. [Untitled FAO Document]. Technical report, FAO, Rome, 2008. URL https://openknowledge.fao.org/server/api/core/bitstreams/e01f4a96-ebdc-40d6-848c-14e8257da656/content.

[13] V. Ongoma, Youssef Brouziyne, E. H. Bouras, and A. Chehbouni. Closing yield gap for sustainable food security in sub-saharan africa – progress, challenges, and opportunities. *Frontiers in Agronomy*, 7:1572061, 2025. doi: 10.3389/fagro.2025.1572061. URL https://doi.org/10.3389/fagro.2025.1572061.

[14] René Schils, Jørgen E. Olesen, Kurt-Christian Kersebaum, Bert Rijk, Michael Oberforster, Valery Kalyada, Maksim Khitrykau, Anne Gobin, Hristofor Kirchev, Vanya Manolova, Ivan Manolov, Mirek Trnka, Petr Hlavinka, Taru Palosuo, Pirjo Peltonen-Sainio, Lauri Jauhiainen, Josiane Lorgeou, Hélène Marrou, Nikos Danalatos, Sotirios Archontoulis, Nándor Fodor, John Spink, Pier Paolo Roggero, Simona Bassu, Antonio Pulina, Till Seehusen, Anne Kjersti Uhlen, Katarzyna Żyłowska, Anna Nieróbca, Jerzy Kozyra, João Vasco Silva, Benvindo Martins Maçãs, José Coutinho, Viorel Ion, Jozef Takáč, M. Inés Mínguez, Henrik Eckersten, Lilia Levy, Juan Manuel Herrera, Jürg Hiltbrunner, Oleksii Kryvobok, Oleksandr Kryvoshein,

Metric	Value
Macro F1-Score (Class-Balanced)	0.3456
Overall Accuracy (OA)	0.4014
Average Accuracy (Macro Recall)	0.3565
Weighted F-score	0.4676
Kappa Statistic	0.2861

Table 14: Ablation 4: Overall Model Performance Metrics

Metric	Value
Macro F1-Score (Class-Balanced)	0.4361
Overall Accuracy (OA)	0.5372
Average Accuracy (Macro Recall)	0.4415
Weighted F-score	0.5912
Kappa Statistic	0.4234

Table 15: Ablation 5: Overall Model Performance Metrics for U-Net.

Roger Sylvester-Bradley, Daniel Kindred, Cairistiona F.E. Topp, Hendrik Boogaard, Hugo de Groot, Jan Peter Lesschen, Lenny van Bussel, Joost Wolf, Mink Zijlstra, Marloes P. van Loon, and Martin K. van Ittersum. Cereal yield gaps across europe. *European Journal of Agronomy*, 101:109–120, 2018.

[15] Gillian Klucas. Yield gap study highlights potential for higher crop yields in africa. *Nebraska Today*, September 2015. URL https://news.unl.edu/article/yield-gap-study-highlights-potential-for-higher-crop-yields-in-africa.

[16] Global Yield Gap Atlas (GYGA). Global Yield Gap Atlas. https://www.yieldgap.org/, 2025. Coordinated by University of Nebraska-Lincoln and Wageningen University & Research.

[17] SERVIR West Africa. Desert locust risk mapping (p-locust). https://servir.icrisat.org/desert-locust-risk-mapping-p-locust/, 2023. Accessed 9 Dec. 2025.

[18] Fatoumata Balde, Nana Yaw Asabere, Amie Diouf, and Bamba Gueye. Senagripreci: Improving precision agricultural yields through a crop recommender system. In Kohei Arai, editor, *Intelligent Systems and Applications*, pages 726–737, Cham, 2024. Springer Nature Switzerland.

[19] Sherrie Wang, François Waldner, and David B. Lobell. Unlocking large-scale crop field delineation in smallholder farming systems with transfer learning and weak supervision. *Remote Sensing*, 14(22), 2022.

[20] Lily T. Janjigian. Exploring smallholder field delineation. Master of engineering thesis, Massachusetts Institute of Technology, Cambridge, MA, May 2025.

[21] G. Cheng, Y. Jiang, Y. Li, K. Yang, J. Yang, W. Li, and J. Luo. Crop type classification with combined spectral, texture, and radar features of time-series sentinel-1 and sentinel-2 data. *International Journal of Remote Sensing*, 44(4):1215–1237, 2023. doi: 10.1080/01431161.2023.2176723.

[22] Zhonglin Ji, Yaozhong Pan, Xiufang Zhu, Dujuan Zhang, and Jinyun Wang. A generalized model to predict large-scale crop yields integrating satellite-based vegetation index time series and phenology metrics. *Ecological Indicators*, 137:108759, 2022.

[23] Fuyao Zhang, Jielin Yin, Nan Wu, Xinyu Hu, Shikun Sun, and Yubao Wang. A dual-path model merging cnn and rnn with attention mechanism for crop classification. *European Journal of Agronomy*, 159:127273, 2024.

[24] Zhihui Zhu, Yuling Chen, Chengzhuo Lu, Minglong Yang, Yonghua Xia, Dewu Huang, and Jie Lv. Research on crop classification using u-net integrated with multimodal remote sensing temporal features. *Sensors*, 25(16), 2025. ISSN 1424-8220. doi: 10.3390/s25165005. URL https://www.mdpi.com/1424-8220/25/16/5005.

[25] Marc Rußwurm, Charlotte Pelletier, Maximilian Zollner, Sébastien Lefèvre, and Marco Körner. Breizhcrops: A time series dataset for crop type mapping, 2020. URL https://arxiv.org/abs/1905.11893.

[26] Wikipedia contributors. Sustainable Development Goal 10. https://en.wikipedia.org/wiki/Sustainable_Development_Goal_10, 2025. Page last edited 12 November 2025. Accessed on 11 December 2025.

[27] John A. Dixon, Aidan Gulliver, and David Gibbon. Sub-saharan africa. In *Farming Systems and Poverty: Improving farmers' livelihoods in a changing world*, chapter 2. Food and Agriculture Organization of the United Nations and The World Bank, 2001. URL https://www.fao.org/4/Y1860E/y1860e04.htm.

[28] Wikipedia contributors. Sustainable Development Goal 16. https://en.wikipedia.org/wiki/Sustainable_Development_Goal_16, 2025. Accessed: 11 December 2025.

[29] J.E. Dobson. On reading and interpreting black box deep neural networks. *Int J Digit Humanities*, 5:431–449, 2023. doi: 10.1007/s42803-023-00075-w. URL https://doi.org/10.1007/s42803-023-00075-w.

Class ID	Class Name	Support	Precision	Recall	F1-Score
0	Barley	287088	0.8169	0.4831	0.6072
1	Wheat	816432	0.9457	0.7247	0.8206
2	Rapeseed	155712	0.9848	0.7088	0.8243
3	Corn	1505328	0.9954	0.5144	0.6783
4	Sunflower	96	0.0000	0.0000	0.0000
5	Orchards	26496	0.0155	0.5274	0.0301
6	Nuts	528	0.0000	0.0000	0.0000
7	Permanent meadows	1254432	0.4320	0.5701	0.4915
8	Temporary meadows	1843872	0.5040	0.4447	0.4725

Table 16: Ablation 5: Per-Class Performance Metrics for U-Net.

[30] M. Machefer, A. Thomas, M. Meroni, J. Veiga Lopez Pena, M. Ronco, C. Corbane, and F. Rembold. Potential and limitations of machine learning modeling for forecasting acute food insecurity. *Global Food Security*, 45:100859, 2025.

[31] Global Indigenous Data Alliance (GIDA) and Research Data Alliance (RDA) International Indigenous Data Sovereignty Interest Group. CARE Principles for Indigenous Data Governance, october 2019. URL https://gida-global.org/care/. Accessed: November 3, 2025.

[32] Hauke Hoppe, Peter Dietrich, Philip Marzahn, Thomas Weiß, Christian Nitzsche, Uwe Freiherr von Lukas, Thomas Wengerek, and Erik Borg. Transferability of Machine Learning Models for Crop Classification in Remote Sensing Imagery Using a New Test Methodology: A Study on Phenological, Temporal, and Spatial Influences. *Remote Sensing*, 16(9):1493, 2024. doi: 10.3390/rs16091493. URL https://www.mdpi.com/2072-4292/16/9/1493.

[33] Superlinear. VITO transform agricultural monitoring with AI. https://superlinear.eu/impact/vito-partners-with-superlinear-to-transform-agricultural-monitoring-with-ai, 2025. Accessed: 6 December 2025.

[34] Rahat Tufail, Patrizia Tassinari, and Daniele Torreggiani. Deep learning applications for crop mapping using multi-temporal sentinel-2 data and red-edge vegetation indices: Integrating convolutional and recurrent neural networks. *Remote Sensing*, 17(18):3207, 2025. doi: 10.3390/rs17183207. URL https://www.mdpi.com/2072-4292/17/18/3207.

[35] Marc Rußwurm and Charlotte Pelletier and Maximilian Zollner and Sébastien Lefèvre and Marco Körner. Hands-on tutorial on time series at madics maclean. https://colab.research.google.com/drive/1i0M_X5-ytFhF0N0-FjhKiqnraclSEIb0?usp=sharing, 2019. Accessed: 2025-12-03.

[36] Conghe Song, Curtis E. Woodcock, Karen C. Seto, Mary Pax Lenney, and Scott A. Macomber. Classification and change detection using landsat tm data: When and how to correct atmospheric effects? *Remote Sensing of Environment*, 75(2): 230–244, 2001. ISSN 0034-4257. doi: https://doi.org/10.1016/S0034-4257(00)00169-3. URL https://www.sciencedirect.com/science/article/pii/S0034425700001693.

[37] Jie He, Wenzhi Zeng, Chang Ao, Weimin Xing, Thomas Gaiser, and Amit Kumar Srivastava. Cross-regional crop classification based on sentinel-2. *Agronomy*, 14(5), 2024. ISSN 2073-4395. doi: 10.3390/agronomy14051084. URL https://www.mdpi.com/2073-4395/14/5/1084.

[38] Sentinel-2. Sentinel-2 dataset (aws open data registry), 2015–.

[39] Sustainability Directory. What are the ethical considerations of agricultural data?, 2025. URL https://climate.sustainability-directory.com/question/what-are-the-ethical-considerations-of-agricultural-data/. Published 20 April 2025.

[40] Paycompliance. The ethics of ai in compliance: Bias, transparency, and governance, June 2025. URL https://paycompliance.com/2025/06/19/the-ethics-of-ai-in-compliance-bias-transparency-and-governance/. Accessed: December 6, 2025.

[41] CFA Institute Research and Policy Center. Chapter 10: Ethical ai in finance, 2025. URL https://rpc.cfainstitute.org/research/foundation/2025/chapter-10-ethical-ai-in-finance. Accessed: December 6, 2025.

[42] Xiaoguang Yuan, Shiruo Liu, Wei Feng, and Gabriel Dauphin. Feature importance ranking of random forest-based end-to-end learning algorithm. *Remote Sensing*, 15(21), 2023. ISSN 2072-4292. doi: 10.3390/rs15215203. URL https://www.mdpi.com/2072-4292/15/21/5203.

[43] Haitian Zhang, Maofang Gao, and Chao Ren. Feature-ensemble-based crop mapping for multi-temporal sentinel-2 data using oversampling algorithms and gray wolf optimizer support vector machine. *Remote Sensing*, 14(20), 2022.

[44] A. Bandar and A. Coşkunçay. Crop classification with attention based bi-lstm and temporal convolution neural network combination for remote sensing breizhcrop time series data. *Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 29 (1):173–188, 2024. doi: 10.53433/yyufbed.1335866. URL https://doi.org/10.53433/yyufbed.1335866.

[45] Google Developers. Imbalanced datasets. https://developers.google.com/machine-learning/crash-course/overfitting/imbalanced-datasets, 2025. Part of the Machine Learning Crash Course.

[46] Tuanfei Zhu, Cheng Luo, Zhihong Zhang, Jing Li, Siqi Ren, and Yifu Zeng. Minority oversampling for imbalanced time series classification. *Knowledge-Based Systems*, 247:108764, 04 2022. doi: 10.1016/j.knosys.2022.108764.

[47] AI Innovator's Journal. Structure-preserving oversampling for imbalanced multivariate time series data, sep 2023. URL https://medium.com/@ai_innovator/oversampling-for-imbalanced-multivariatetime-series-data-49e77502bf5f.

[48] Tarid Wongvorachan, Surina He, and Okan Bulut. A comparison of undersampling, oversampling, and smote methods for dealing with imbalanced classification in educational data mining. *Information*, 14(1), 2023.

[49] Ibraheem M Alkhawaldeh, Ibrahem Albalkhi, and Abdulqadir J Naswhan. Challenges and limitations of synthetic minority oversampling techniques in machine learning. *World Journal of Methodology*, 13:373–378, dec 2023. doi: 10.5662/wjm.v13.i5.373.

[50] Sarem Seitz. Why smote is not necessarily the answer to your imbalanced dataset. Blog post, Towards Data Science, jun 2022. URL https://towardsdatascience.com/why-smote-is-not-necessarily-the-answer-to-your-imbalanced-dataset-ef19881da57a/. Accessed: 2025-12-05.

[51] Abdallah Ashraf. Oversampling for better machine learning with imbalanced data. Medium, dec 2023. URL https://medium.com/@abdallahashraf90x/oversampling-for-better-machine-learning-with-imbalanced-data-68f9b5ac2696. Accessed: 2025-12-05.

[52] Milvus. How does data augmentation help with class imbalance? https://milvus.io/ai-quick-reference/how-does-data-augmentation-help-with-class-imbalance, 2025. Accessed: 2025-12-05.

[53] A. A. Awan. A Complete Guide to Data Augmentation. DataCamp Tutorial, November 2022. URL https://www.datacamp.com/tutorial/complete-guide-data-augmentation. Accessed: December 5, 2025.

[54] Mustafa Ustuner, Fusun Balik Sanli, Saygin Abdikan, Mustafa Esetlili, and Yusuf Kurucu. Crop type classification using vegetation indices of rapideye imagery, 09 2014.

[55] Marta Pasternak and Kamila Pawluszek-Filipiak. The evaluation of spectral vegetation indexes and redundancy reduction on the accuracy of crop type detection. *Applied Sciences*, 12(10), 2022. ISSN 2076-3417. doi: 10.3390/app12105067. URL https://www.mdpi.com/2076-3417/12/10/5067.

[56] Yupeng Kang, Qingyan Meng, Miao Liu, Youfeng Zou, and Xuemiao Wang. Crop classification based on red edge features analysis of gf-6 wfv data. *Sensors*, 21(13), 2021.

[57] Rahat Tufail, Patrizia Tassinari, and Daniele Torreggiani. Assessing feature extraction, selection, and classification combinations for crop mapping using sentinel-2 time series: A case study in northern italy. *Remote Sensing Applications: Society and Environment*, 38:101525, 2025.

[58] Jordi Inglada, Marcela Arias, Benjamin Tardy, Olivier Hagolle, Silvia Valero, David Morin, Gérard Dedieu, Guadalupe Sepulcre, Sophie Bontemps, Pierre Defourny, and Benjamin Koetz. Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. *Remote Sensing*, 7(9):12356–12379, 2015. ISSN 2072-4292. doi: 10.3390/rs70912356. URL https://www.mdpi.com/2072-4292/7/9/12356.

[59] Fukang Feng, Maofang Gao, Ronghua Liu, Shuihong Yao, and Guijun Yang. A deep learning framework for crop mapping with reconstructed sentinel-2 time series images. *Computers and Electronics in Agriculture*, 213:108227, 2023. ISSN 0168-1699. doi: https://doi.org/10.1016/j.compag.2023.108227. URL https://www.sciencedirect.com/science/article/pii/S0168169923006154.

[60] Hengbin Wang, Zijing Ye, Yu Yao, Wanqiu Chang, Junyi Liu, Yuanyuan Zhao, Shaoming Li, Zhe Liu, and Xiaodong Zhang. Improving cross-regional model transfer performance in crop classification by crop time series correction. *Geospatial Information Science*, 28(4):1581–1596, 2025.

[61] Yady Tatiana Solano-Correa, Francesca Bovolo, Lorenzo Bruzzone, and Diego Fernández-Prieto. A method for the analysis of small crop fields in sentinel-2 dense time series. *IEEE Transactions on Geoscience and Remote Sensing*, 58(3):2150–2164, 2020. doi: 10.1109/TGRS.2019.2953652.

[62] Jie Zhou, Li Jia, Massimo Menenti, and Ben Gorte. On the performance of remote sensing time series reconstruction methods – a spatial comparison. *Remote Sensing of Environment*, 187:367–384, 2016.

[63] EO Research. How to normalize satellite images for deep learning, Sep 2022. URL https://medium.com/sentinel-hub/how-to-normalize-satellite-images-for-deep-learning-d5b668c885af. Published in *Planet Stories* on Medium; Accessed: 2025-12-04.

[64] Flavio Piccoli, Mirko Paolo Barbato, Marco Peracchi, and Paolo Napoletano. Estimation of soil characteristics from multispectral sentinel-3 imagery and dem derivatives using machine learning. *Sensors*, 23(18), 2023.

[65] Manan Thakkar and Rakeshkumar Vanzara. Enhancing crop yield estimation from remote sensing data: a comparative study of the quartile clean image method and vision transformer. *Discov Appl Sci*, 6(11):610, 2024. doi: 10.1007/s42452-024-06329-8. URL https://doi.org/10.1007/s42452-024-06329-8.

[66] Sivaramakrishnan Rajaraman, Ghada Zamzmi, and Sameer K. Antani. Novel loss functions for ensemble-based medical image classification. *PLOS ONE*, 16(12):1–18, 12 2022.

[67] Yu Bian, Linhui Li, and Wenpei Jing. Cacpu-net: Channel attention u-net constrained by point features for crop type mapping. *Frontiers in Plant Science*, 13:1030595, 2023. doi: 10.3389/fpls.2022.1030595. URL https://doi.org/10.3389/fpls.2022.1030595.

[68] Sarah Farhadpour, Timothy A. Warner, and Aaron E. Maxwell. Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices. *Remote Sensing*, 16(3), 2024.

[69] Devanshi Pratiher. Understanding loss functions for deep learning segmentation models. Medium blog post, May 2024. URL https://medium.com/@devanshipratiher/understanding-loss-functions-for-deep-learning-segmentation-models-30187836b30a. Accessed on December 11, 2025.

[70] Seyd Teymoor Seydi, Meisam Amani, and Arsalan Ghorbanian. A dual attention convolutional neural network for crop classification using time-series sentinel-2 imagery. *Remote Sensing*, 14(3), 2022.

[71] Xin Zhou, Jinfei Wang, Bo Shan, and Yongjun He. Early-season crop classification based on local window attention transformer with time-series rcm and sentinel-1. *Remote Sensing*, 16(8), 2024.

[72] Sebastian C. Ibañez and Christopher P. Monterola. A global forecasting approach to large-scale crop production prediction with time series transformers. *Agriculture*, 13(9), 2023.

[73] Hongwei Zhao, Zhongxin Chen, Hao Jiang, Wenlong Jing, Liang Sun, and Min Feng. Evaluation of three deep learning models for early crop classification using sentinel-1a imagery time series—a case study in zhanjiang, china. *Remote Sensing*, 11(22), 2019.

[74] Pan Zhou, Xingyu Xie, Zhouchen Lin, and Shuicheng Yan. Towards understanding convergence and generalization of adamw. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):6486–6493, 2024. doi: 10.1109/TPAMI.2024.3382294.

[75] Ricardo Llugsi, Samira El Yacoubi, Allyx Fontaine, and Pablo Lupera. Comparison between adam, adamax and adam w optimizers to implement a weather forecast based on neural networks for the andean city of quito. In *2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM)*, pages 1–6, 2021. doi: 10.1109/ETCM53643.2021.9590681.

[76] Krishna Karthik Gadiraju, Bharathkumar Ramachandra, Zexi Chen, and Ranga Raju Vatsavai. Multimodal deep learning based crop classification using multispectral and multitemporal satellite imagery. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3234–3242, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403375. URL https://doi.org/10.1145/3394486.3403375.

[77] Piyush Kashyap. Understanding weight initialization techniques in neural networks. *Medium*, Nov 2024.

[78] Vu Nguyen. Bayesian optimization for accelerating hyper-parameter tuning. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 302–305, 2019. doi: 10.1109/AIKE.2019.00060.

[79] Lucas Wittstruck, Thomas Jarmer, and Björn Waske. Multi-stage feature fusion of multispectral and sar satellite images for seasonal crop-type mapping at regional scale using an adapted 3d u-net model. *Remote Sensing*, 16(17), 2024.

[80] Xiangsuo Fan, Chuan Yan, Jinlong Fan, and Nayi Wang. Improved u-net remote sensing classification algorithm fusing attention and multiscale features. *Remote Sensing*, 14(15), 2022.

[81] Andy Wang. The underlying dangers behind large batch training schemes, nov 2022. URL https://towardsdatascience.com/the-underlying-dangers-behind-large-batch-training-schemes-6cdc0e511ef1/.

[82] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[83] Marc Rußwurm, Charlotte Pelletier, Maximilian Zollner, Sébastien Lefèvre, and Marco Körner. Breizhcrops: A time series dataset for crop type mapping. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences ISPRS (2020)*, 2020.