

DOI: 10.37943/DUJJ6958

Azamat Serek

PhD, Associate Professor, School of Software Engineering
azamat.serek@astanait.edu.kz, orcid.org/0000-0001-7096-6765
Astana IT University, Kazakhstan

Farida Abdoldina

PhD, Professor, Department of Program Engineering
abdoldinafarida@gmail.com, orcid.org/0000-0003-1816-6343
Satbayev University, Kazakhstan

Yelizaveta Vitulyova

PhD, Associate Professor, Department of Program Engineering
lizavita@list.ru, orcid.org/0000-0002-6075-4870
Satbayev University, Kazakhstan;
International Engineering and Technological University, Kazakhstan.

Gleb Tokin

Master's Degree, Lead Specialist
tokingleb@gmail.com, orcid.org/0009-0008-5324-9973
Satbayev University, Kazakhstan

Nurshapagat Shapay

Master's Student, School of Software Engineering
shiposha04@gmail.com, orcid.org/0009-0004-6453-0200
Astana IT University, Kazakhstan

Yan Kuchin

Bachelor's Degree, Senior Researcher
ykuchin@mail.ru, orcid.org/0000-0002-5271-9071
Satbayev University, Kazakhstan

EXPLAINABLE MULTISPECTRAL ATTENTION NETWORK FOR SENTINEL-2 AGRICULTURAL LAND CLASSIFICATION

Abstract: Accurate and transparent classification of agricultural land cover from satellite imagery is essential for operational crop monitoring. This study presents MS-AttNet, a compact convolutional network with squeeze-and-excitation channel attention for Sentinel-2 imagery using all thirteen spectral bands. On the EuroSAT benchmark, evaluated with a stratified 70/15/15 split across five random seeds, the model achieves $98.72 \pm 0.21\%$ accuracy, a macro F1-score of 0.9868 ± 0.0021 , and a mean one-vs-rest AUC of 0.9999 with 1.32 million parameters. A ResNet-50 baseline re-implemented on the identical split achieves 97.90% accuracy with 23.56 million parameters, making the proposed model approximately eighteen times smaller and nearly three times faster at inference. A controlled ablation using five seeds per configuration shows that squeeze-and-excitation modules, convolutional stage depth, and vegetation indices as additional input channels each change accuracy by less than run-to-run variability. This indicates that the benchmark is saturated with respect to architectural choices in this model family; therefore, performance is not attributed specifically to the attention mechanism. Interpretability is assessed using Grad-CAM and SHAP band attributions with quantitative validation. Against a spatially smooth random control, Grad-CAM reduces deletion AUC from 0.857 to 0.738 and increases insertion AUC from 0.864 to 0.916 across 300 test patches. SHAP identifies blue, near-infrared, green, red, and short-wave-infrared bands as dominant contributors, with stable ordering across independently trained models. NDVI and NDWI were used only for spectral-signature analysis and interpretation, not as classifier inputs. The combination of compact size, competitive accuracy, and quantitatively verified explanations supports large-area operational monitoring applications.

Keywords: Sentinel-2, land-cover classification, convolutional neural networks, squeeze-and-excitation, explainable artificial intelligence, Grad-CAM, SHAP, EuroSAT.

Introduction

Agriculture is an important sector for food security, livelihoods of the rural population and sustainable economic development. Accurate agriculture land monitoring is thus critical for agricultural resource management, crop assessment and policy implementation. One of the most critical technologies for agricultural monitoring is remote sensing as it allows systematic, repeatable and cost effective observations of earth's surface over large geographic areas [1]. The European Space Agency's Sentinel-2 mission has greatly extended these capabilities with its freely-accessible multi-spectral data with thirteen spectral bands across the visible, red-edge, near-infrared and short-wave infra-red wavelengths of the electromagnetic spectrum. Sentinel-2 provides data with spatial resolution between 10 and 60 m and revisit frequency of about 5 days which is very useful for agriculture and land cover analysis [2]. The red-edge and near-infrared bands are particularly sensitive to the properties of vegetation like chlorophyll content, biomass and canopy structure which are useful for discrimination of crops, land cover, etc.

The data from Earth observation has grown exponentially at the same time as artificial intelligence, and in particular deep learning, has made tremendous strides. Convolutional neural networks (CNNs) are a category of deep learning that have proven to be very effective at image classification, automatically extracting hierarchical spatial and spectral representations directly from raw data [3]. Modern convolutional architectures rely on established design principles such as batch normalisation, residual connections and channel attention, which allow deep multispectral models to be trained stably [4], and such architectures have been applied successfully to crop classification from satellite imagery [5]. For example, CNNs have already been shown to provide good performance for crop-type classification, yield prediction, disease detection and land-cover mapping in agriculture, and they have consistently outperformed traditional handcrafted feature-based machine-learning methods [6-7].

To fill this research gap, a compact multispectral land-use and land-cover classification network is proposed with the aid of Sentinel-2 data, which is an attention-enhanced convolutional neural network (MS-AttNet). It includes squeeze and excitation channel attention mechanisms to enhance feature representation while keeping its computational efficiency. The proposed framework is trained and tested on the dataset of EuroSAT benchmark set, and evaluated using the confusion matrix analysis, receiver operating characteristic (ROC) curve, precision, recall, and F1 score. Moreover, model interpretability is explored by using Grad-CAM and SHAP methods to identify the most relevant spatial regions and spectral bands in decisions for classification. This work presents a reliable, explainable, and reproducible multispectral classification framework that combines competitive classification performance with computational efficiency and interpretability, supporting the development of trustworthy and operational agricultural land-monitoring applications.

The individual components employed in this work squeeze-and-excitation channel attention, Grad-CAM and SHAP are established methods rather than novel algorithmic proposals. The contribution of this study lies in their integration into a single lightweight all-band multispectral pipeline, in the quantitative validation of the resulting explanations, and in a controlled ablation that reports honestly what the architectural components do and do not contribute on a saturated benchmark.

The main contributions of this work are summarized as follows:

1. We propose MS-AttNet, a compact multispectral classification framework that reaches an overall accuracy of 98.72 ± 0.21 % on the EuroSAT benchmark with 1.32 million parameters, exceeding a ResNet-50 baseline trained on the identical split by 0.82 percentage points while being approximately eighteen times smaller and nearly three times faster at inference.

2. We report a controlled ablation with five random seeds per configuration that isolates the contribution of the attention modules, of network depth and of the input representation, and we state explicitly that none of these choices produces a difference in accuracy that exceeds the run-to-run variability on this benchmark.

3. We integrate Grad-CAM and SHAP into a single pipeline and validate the resulting explanations quantitatively using insertion and deletion curves against a random control matched for spatial smoothness, rather than assessing them narratively.

4. We report the complete training configuration, hardware, runtimes and a public code repository, so that every number in this article can be reproduced end to end.

Related Work

The availability of benchmark datasets have also further quickened the pace of research. The most popular is the "EuroSAT" dataset comprising 27000 georeferenced image patches, annotated into ten land-use and land-cover classes, with the use of information from all thirteen spectral bands [8]. EuroSAT is diverse and easy to access and has become a well-known test set for assessing multispectral classification algorithms. High classification accuracy of 98.6 % has been reported recently on this dataset [8]. Recent studies on remote-sensing scene classification and on the mapping of crop planting structures confirm that CNN-based and hybrid CNN–transformer methods remain state of the art for scene classification and land-cover mapping [9], [10]. More recently, temporal convolutional networks and self-attention mechanisms have been used to model Sentinel-2 image time series to capture the seasonal changes in vegetation, and to enhance the crop discrimination performance [11–12].

Although these achievements have been made, the lack of interpretability is still an important hurdle for the actual application of deep learning models in an operational system. While CNNs can make predictions with high accuracy, it is often hard to decipher their decision-making processes, which can decrease user trust and pose questions about accountability and transparency. The problem of this is especially relevant in the field of agricultural monitoring, for supporting policy making, subsidy distribution, and environmental compliance evaluation [13]. In response, explainable AI (XAI) has emerged with numerous techniques to explain model behavior. Local Interpretable Model-Agnostic Explanations (LIME) offer simplified local approximations of complex models [14] while Gradient-weighted Class Activation Mapping (Grad-CAM) pinpoint image regions that most strongly influence a model's predictions [15] and Shapley Additive Explanations (SHAP) offer theoretically grounded feature-importance measures derived from cooperative game theory [16]. It has been recently shown how valuable these techniques could be in remote-sensing applications by testing their reliability in the classification of images and by analysing the spectral variables affecting the models used for agricultural monitoring [17–18].

Recent studies have further advanced land-cover classification using approaches that exploit all available Sentinel-2 spectral bands [19]. In parallel, transformer-based architectures, including Vision Transformers, together with self-supervised foundation models such as SatMAE, have become prominent directions in remote sensing image analysis [20–21]. However, these approaches generally prioritise classification accuracy over computational efficiency and interpretability, leaving lightweight and explainable models comparatively underexplored.

Research Gap. Comprehensive reviews of deep learning for land-use and land-cover classification confirm that coupling classification with explanation has become a central concern of the field [22]. Two recent studies are particularly close to the present work. Temenos et al. [23] combined a compact CNN with a SHAP deep explainer on the EuroSAT dataset, reporting an overall accuracy of 94.72 % and analysing the influence of individual Sentinel-2 spectral bands on the predictions. Khan et al. [24] applied Vision Transformer and Swin Transformer models to the same benchmark together with gradient-based attribution, achieving accuracies above 99 % but at the cost of models containing hundreds of millions of parameters. These works either operate on a reduced band subset or depend on very large backbones, which motivates the lightweight, all-band and fully explainable design pursued here.

Despite the substantial progress achieved by recent deep learning methods, an important challenge remains in balancing classification accuracy with computational efficiency. Many state-of-the-art approaches achieve high performance at the cost of increased model complexity, limiting their practical applicability in resource-constrained remote-sensing scenarios.

Another remaining challenge is the integration of explainability into lightweight classification frameworks. Although Grad-CAM and SHAP have been applied in remote sensing, relatively few studies combine efficient network architectures with systematic model interpretation while maintaining competitive classification performance.

Methods and Materials

The entire workflow used in this study is depicted in Figure 1. The first step in the process is to receive Sentinel-2 Level-1C top-of-atmosphere reflectance data with its thirteen spectral bands. At the pre-processing stage, radiometric corrections and per-band standardisation are applied. The Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Water Index (NDWI) are additionally computed as auxiliary quantities; they are used only for the spectral-signature analysis reported in Figure 2 and for the interpretation of the results, and are not part of the classifier input. The pre-processed images are then subdivided into non-overlapping image patches of 64×64 pixels in order to form the input dataset for training and evaluation of the model. The extracted patches are then passed through the proposed MS-AttNet architecture, which consists of the convolutional neural network layers and squeeze and excitation attention layers, improving the feature representation by emphasizing on the most informative spectral channels. The trained model produces outputs of land-use and land-cover classification over 10 pre-defined classes. Explainable Artificial Intelligence techniques are used to enhance model transparency and interpretability. The spatial regions that most strongly contribute to classification decisions are discovered using Gradient-weighted Class Activation Mapping (Grad-CAM), and individual spectral bands and derived features are quantified using Shapley Additive Explanations (SHAP). The overall accuracy, precision, recall, F1-score, confusion matrix analysis, receiver operating characteristic (ROC) curves, and area under the curve (AUC) values are all used to evaluate model performance using a comprehensive set of evaluation metrics. This process combines all these steps in a single workflow and results in an accurate agricultural land cover classification with high interpretability and reproducibility using multispectral Sentinel-2 data.

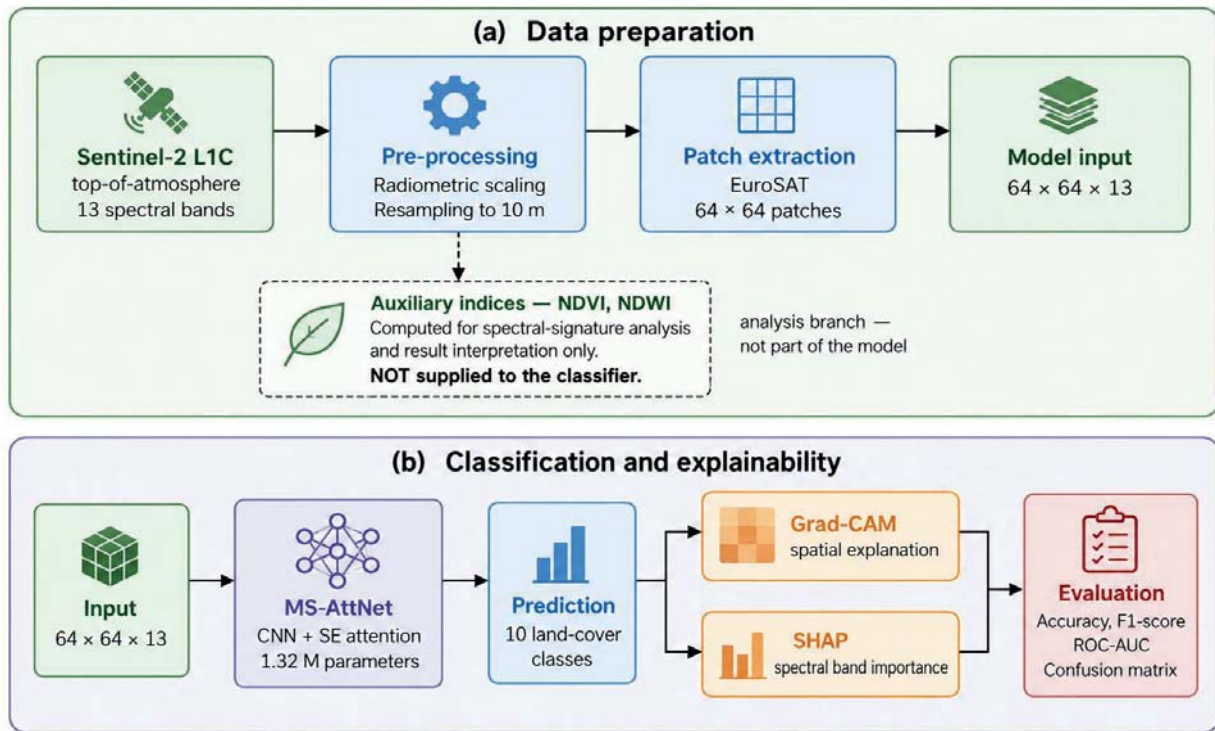


Figure 1. Overall workflow of the proposed MS-AttNet framework for agricultural land-cover classification using Sentinel-2 multispectral imagery: (a) data preparation; (b) classification and explainability. The NDVI and NDWI indices shown in the pre-processing block are used only for spectral-signature analysis and result interpretation; the classifier input consists exclusively of the thirteen original spectral bands.

The experiments were conducted on the EuroSAT benchmark [8], which is derived from Sentinel-2 Level-1C top-of-atmosphere reflectance products and provides all thirteen spectral bands for each patch. The benchmark contains 27,000 labelled patches of 64 × 64 pixels covering ten classes: AnnualCrop, Forest, HerbaceousVegetation, Highway, Industrial, Pasture, PermanentCrop, Residential, River and SeaLake. The classes are not equally represented, ranging from 2,000 patches for Pasture to 3,000 for AnnualCrop, Forest, HerbaceousVegetation, Residential and SeaLake. Four of the ten classes are of a directly agricultural or vegetative nature, which makes the benchmark appropriate for a study of agricultural land monitoring while preserving the surrounding land-cover context. The data were divided into training, validation and test subsets in a 70/15/15 ratio by stratified random sampling with a fixed seed, preserving the original class proportions in each subset. The resulting partitions contain 18,900, 4,050 and 4,050 patches respectively; the test subset holds between 300 and 450 patches per class. The identical partition was reused for every configuration reported in this article, so that all comparisons are made on the same data.

All thirteen bands had a common patch geometry by resampling bands obtained at coarser spatial resolutions at a common 10-metre grid with bilinear interpolation. All the bands were then normalised to a zero mean and unit variance for each band, with statistics calculated from the training subset only, thus avoiding information leakage between the partitions.

$$\hat{x} = \frac{x - \mu_{train}}{\sigma_{train}} \quad (1)$$

Two physically motivated indices were derived to complement the raw reflectances. The Normalised Difference Vegetation Index, defined in equation (2) from the near-infrared and red

bands, captures photosynthetic vigour, while the Normalised Difference Water Index, defined in equation (3) from the green and near-infrared bands, discriminates open water and moisture.

$$NDVI = \frac{\rho_{NIR} - \rho_{Red}}{\rho_{NIR} + \rho_{Red}} = \frac{B8 - B4}{B8 + B4} \quad (2)$$

$$NDWI = \frac{\rho_{Green} - \rho_{NIR}}{\rho_{Green} + \rho_{NIR}} = \frac{B3 - B8}{B3 + B8} \quad (3)$$

NDVI and NDWI were computed as auxiliary quantities used solely for the spectral-signature analysis presented below and for the interpretation of the experimental results; they were not supplied to the classifier. The proposed classifier receives only the original thirteen Sentinel-2 spectral bands as input ($64 \times 64 \times 13$). Figure 2 illustrates the mean spectral signatures of representative classes and makes clear why multispectral information is decisive: vegetation classes share a characteristic red-edge rise and near-infrared plateau, whereas built-up and water surfaces depart markedly in the short-wave-infrared and near-infrared regions respectively. To improve generalisation, on-the-fly data augmentation was applied during training, comprising random horizontal and vertical flips, ninety-degree rotations and small brightness perturbations.

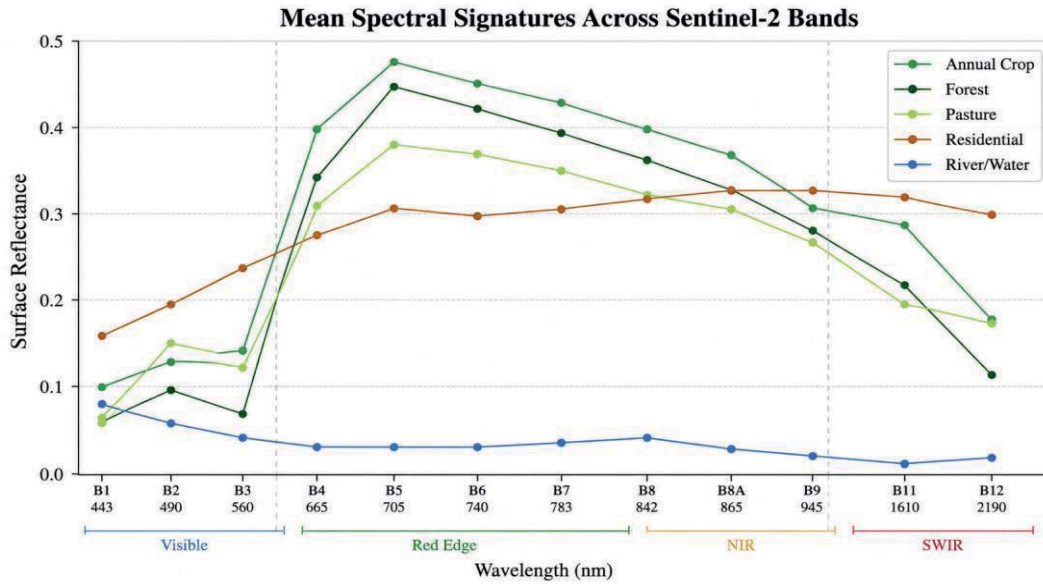


Figure 2. Mean surface-reflectance signatures of representative land-cover classes across the thirteen Sentinel-2 spectral bands, highlighting the discriminative power of the red-edge, near-infrared and short-wave-infrared regions.

The proposed classifier, MS-AttNet, is a compact convolutional neural network that accepts a thirteen-channel multispectral input (Figure 3). It consists of four convolutional stages with 32, 64, 128 and 256 filters respectively; each stage applies two 3 by 3 convolutions with batch normalisation and rectified-linear activation, followed by 2×2 max-pooling. A squeeze-and-excitation block is inserted after every stage to recalibrate channel responses adaptively [25]. The squeeze step aggregates each channel feature map into a single descriptor by global average pooling, as expressed in equation (4).

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j) \quad (4)$$

The excitation step then learns non-linear, channel-wise gating weights and rescales the original feature maps, as given in (5), where the sigmoid and rectified-linear activations are denoted by their usual symbols and the two fully connected layers form a bottleneck that models

inter-channel dependencies. This allows the network to emphasise the most informative spectral-feature maps for a given scene.

$$s = \sigma(W_2 \delta(W_1 z)), \tilde{u}_c = s_c u_c \quad (5)$$

Global average pooling collapses the final feature maps, which are passed to a fully connected layer with dropout regularisation and a ten-way softmax output, defined in (6), that converts the logits into a normalised class-probability distribution over the $C = 10$ classes.

$$\hat{y}_k = \frac{e^{o_k}}{\sum_{j=1}^C e^{o_j}} \quad (6)$$

The architecture contains 1,324,200 trainable parameters, approximately eighteen times fewer than the 23,559,882 parameters of a ResNet-50 adapted to thirteen input channels and ten output classes, which keeps inference inexpensive enough for large-area monitoring. Each squeeze-and-excitation block uses a bottleneck reduction ratio of $r = 16$ and together the four blocks account for 11,390 parameters, that is 0.9 % of the total. The fully connected layer preceding the softmax output contains 512 units with a dropout rate of 0.5. Convolutional weights were initialised with the He normal scheme.

On-the-fly augmentation was applied to every training batch: random horizontal flip with probability 0.5, random vertical flip with probability 0.5, random rotation by a multiple of ninety degrees with probability 0.5, and random brightness scaling by a factor drawn uniformly from the interval $[0.90, 1.10]$ with probability 0.3. No augmentation was applied to the validation or test subsets. Each band was standardised to zero mean and unit variance using statistics computed on the training subset alone, so that no information leaked between partitions.

Training minimised the categorical cross-entropy loss of (7) using the Adam optimiser with an initial learning rate of 1×10^{-3} , halved after five epochs without an improvement in validation loss, a batch size of 64, and early stopping with a patience of ten epochs restoring the best weights. Training ran for up to sixty epochs; the reported model completed the full sixty epochs, while the five repeated runs terminated between forty-four and sixty epochs.

All experiments were implemented in Python 3.12.13 with PyTorch 2.10.0 and CUDA 12.8, and executed on a single NVIDIA Tesla T4 GPU with 16 GB of memory. Training the proposed model required 8.8 s per epoch and 9.2 min in total, and inference required 1.97 ms per patch. The ResNet-50 baseline required 15.1 s per epoch and 12.5 min in total, and 5.69 ms per patch at inference.

$$L = -\frac{1}{N} \sum_{n=1}^N \sum_{k=1}^C y_{n,k} \log(\hat{y}_{n,k}) \quad (7)$$

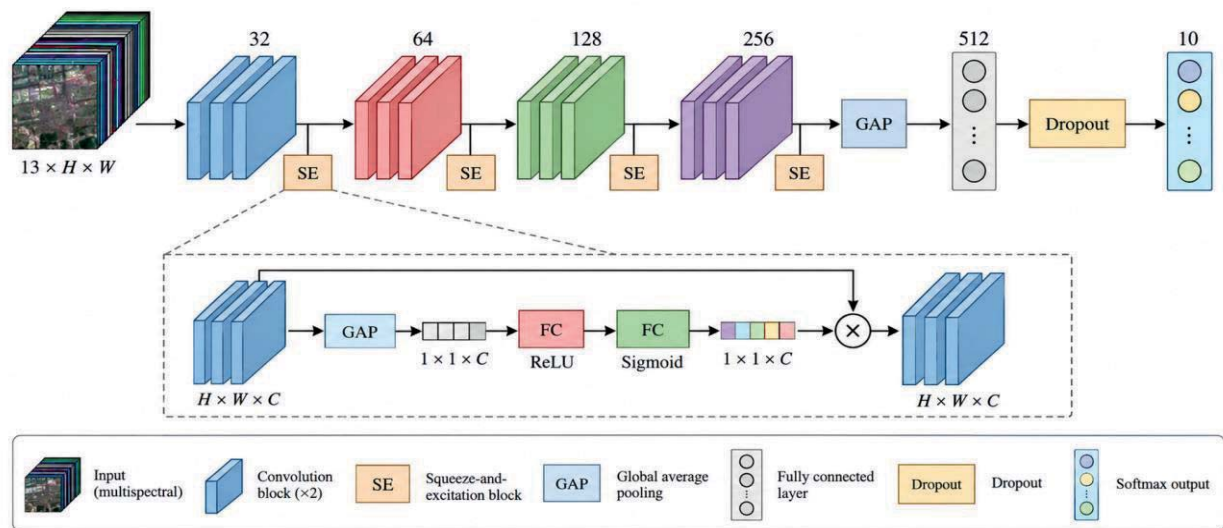


Figure 3. Architecture of the proposed MS-AttNet model for multispectral land-use and land-cover classification.

The trained model was used in conjunction with two complementary interpretability techniques. Grad-CAM method generates a class-discriminative localisation map by multiplying the final convolutional feature maps with the gradients of the class score of the target class, which can highlight spatial regions that the prediction relies on [15]. Unlike other methods, which necessitate architectural changes, and other training techniques, which demand retraining, Grad-CAM seamlessly integrates with MS-AttNet. To obtain spectral rather than spatial explanations, the mean absolute contribution of each of the thirteen input bands to the model output was estimated using SHAP [16, 26]. Because the classifier is differentiable, the GradientSHAP estimator was used in its expected-gradients formulation, with a background set of 100 training patches drawn at random and 50 gradient samples per explained instance; KernelSHAP was not used because its model-agnostic sampling scheme is prohibitively expensive for thirteen-channel inputs at this test-set size. Attributions were computed for 500 randomly selected test patches and required 9 s on the same GPU used for training. Band-level attributions were obtained by aggregating pixel-level contributions within each channel and are reported as means with the corresponding standard deviations across the sampled patches. The two methods answer different questions: Grad-CAM reveals where in the scene the model looks, while SHAP reveals which spectral bands it relies upon, together providing the dual transparency required for operational trust.

To verify that the produced explanations are faithful to the model rather than merely plausible to a human observer, the Grad-CAM maps were evaluated using insertion and deletion curves. In the deletion procedure, pixels are progressively replaced by their blurred counterparts in decreasing order of attribution and the predicted probability of the target class is recorded; in the insertion procedure, pixels are progressively restored from a blurred baseline in the same order. Faithful explanations produce a steep decline in the deletion curve and a steep rise in the insertion curve, so the areas under these curves serve as scalar faithfulness scores.

Two control conditions were evaluated under an identical protocol. The first uses independent per-pixel random attributions. The second uses random attributions generated at the same spatial resolution as the Grad-CAM maps and upsampled identically, so that the control matches the spatial smoothness of the explanation being tested. This second control is necessary because per-pixel random orderings remove information uniformly across the whole patch and therefore destroy texture everywhere at once, which systematically flatters them on scene-

classification tasks in which the class signal is a global textural and spectral property rather than a localised object.

The held-out test set was used to evaluate performance overall, per class precision, recall and F1, and macro-averaged F1, the latter based on equal weighting of all classes regardless of their frequency. To quantify the accuracy of the predicted class probabilities without relying on a specific decision threshold, one-vs-rest ROC curves were generated for each class along with their AUC values and a micro-averaged ROC curve was constructed by stacking all decisions. The entire confusion matrix with the elements normalised was also studied to look for any systematic inter-class errors. Comparisons were made with representative published studies using the same benchmark.

Results

Figure 4 shows accuracy and loss as a function of the training epoch. Training and validation accuracy rose steeply during the first ten epochs and then converged smoothly, with the validation curve tracking the training curve throughout and no sign of over-fitting. The validation loss declined monotonically and levelled off, and training was terminated by early stopping when the validation loss ceased to improve.

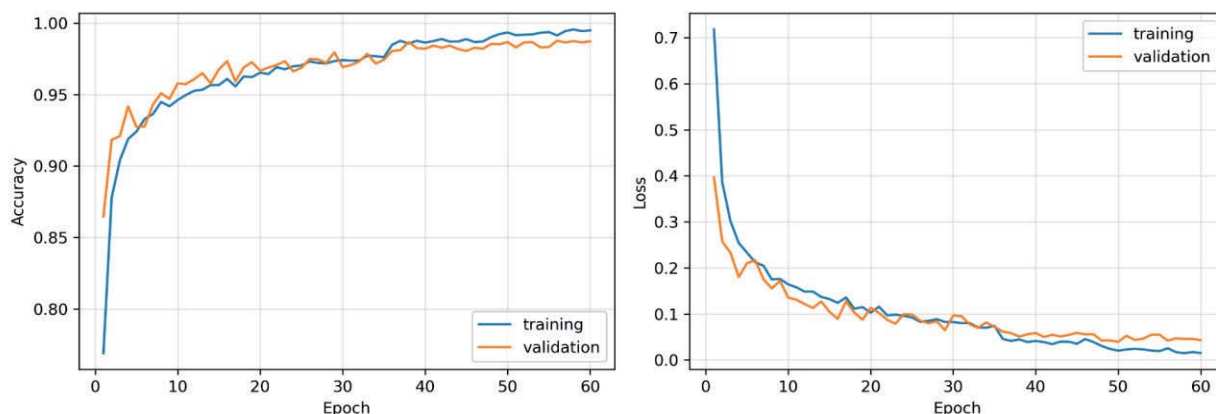


Figure 4. Training and validation (a) accuracy and (b) cross-entropy loss over sixty epochs. Validation accuracy reached 98.7 % and the validation loss levelled off without divergence from the training curve.

On the held-out test set the proposed model reached an overall accuracy of 98.89 %, a macro-averaged F1-score of 0.9886, a mean one-vs-rest AUC of 0.9999 and a micro-averaged AUC of 0.9999. Because a single run cannot separate the performance of a model from the variability of its optimisation, the same configuration was retrained five times with different random seeds on the identical data partition. Across those runs the overall accuracy was 98.72 ± 0.21 %, the macro-averaged F1-score 0.9868 ± 0.0021 and the mean one-vs-rest AUC 0.99986 ± 0.00003 , with individual runs spanning 98.37 % to 98.89 %. All comparisons in this article are interpreted against this spread.

Table 1 reports per-class performance for the run illustrated in Figures 4 and 5. Every class exceeds an F1-score of 0.98. Forest and SeaLake are classified almost perfectly, each with an F1-score of 0.998 and a precision of 1.000, and Residential reaches 0.996. The weakest classes are PermanentCrop at 0.980 and the pair AnnualCrop and HerbaceousVegetation at 0.982.

Table 1. Per-class precision, recall, F1-score and one-vs-rest AUC on the EuroSAT test set.

Class	Precision	Recall	F1-score	AUC	Support
AnnualCrop	0.974	0.991	0.982	0.9998	450
Forest	1.000	0.996	0.998	1.0000	450
HerbaceousVegetation	0.987	0.978	0.982	0.9999	450
Highway	0.992	0.979	0.985	0.9999	375
Industrial	0.989	0.989	0.989	1.0000	375
Pasture	0.977	0.990	0.983	0.9996	300
PermanentCrop	0.981	0.979	0.980	0.9997	375
Residential	0.996	0.996	0.996	1.0000	450
River	0.989	0.995	0.992	1.0000	375
SeaLake	1.000	0.996	0.998	1.0000	450
Macro average	0.988	0.989	0.989	0.9999	4050

The confusion matrix in Figure 5 shows that the residual errors are confined to the vegetated classes and are agronomically interpretable. Four HerbaceousVegetation patches were assigned to AnnualCrop and three to PermanentCrop, while three AnnualCrop patches were assigned to PermanentCrop and one to Pasture; conversely, three PermanentCrop patches were labelled AnnualCrop and four HerbaceousVegetation. These classes share overlapping spectral signatures because they represent the same vegetation types at different management intensities, and the boundaries between them are partly ambiguous even to a human interpreter. The built-up and water classes are essentially free of systematic confusion, with only two Residential patches assigned to Industrial and two SeaLake patches to River. The one-vs-rest ROC curves confirm this picture, with every class exceeding an AUC of 0.9996.

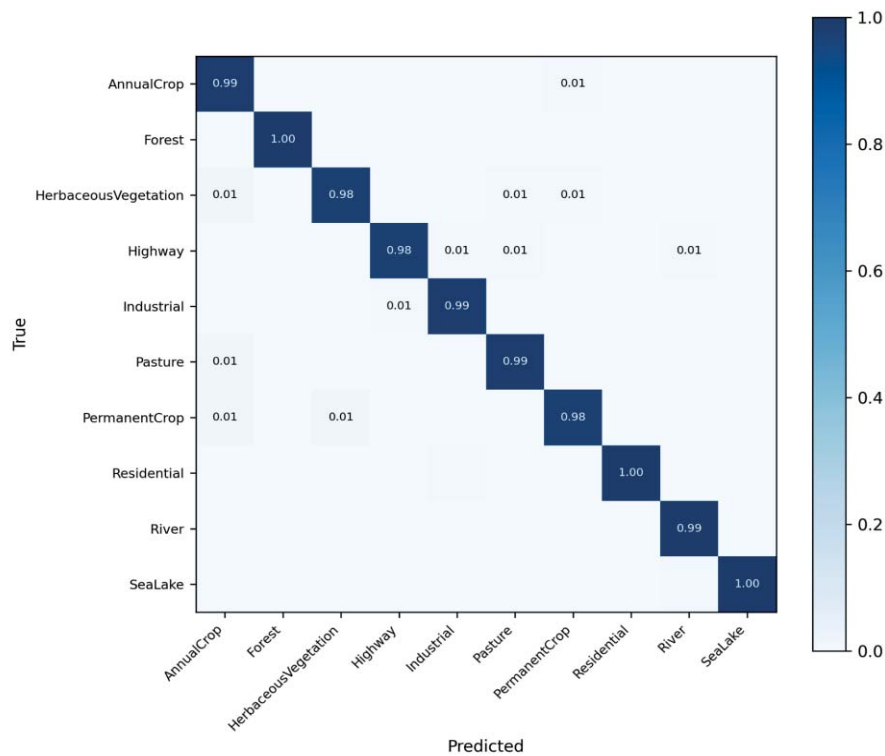


Figure 5. Normalised confusion matrix on the test set; residual errors concentrate among spectrally similar vegetation classes.

Ablation study. Table 2 shows an ablation study of the main design decision of the proposed MS-AttNet architecture. Each configuration changes one aspect of the configuration and leaves the others unchanged. The differences observed are small; hence, the principal configurations (A, C, D and F) were trained five times with different random seeds, and the results are presented as mean and standard deviation.

The overall classification accuracy of all the architectural variants is similar (98.54% – 98.86%). The overall accuracy of the complete proposed model (Configuration C) is $98.72 \pm 0.21\%$ and the Macro-F1 score is 0.9868 ± 0.0021 . This is obtained by removing the SE modules (Configuration A) yielding mean accuracy of $98.76 \pm 0.08\%$, which is statistically equal to the accuracy of the other configurations. Moreover, it is found that the performance of the benchmark is not sensitive to network capacity, as decreasing the network depth from four convolutional stages to three (Configuration D) still leads to a near perfect performance ($98.73 \pm 0.13\%$) with an approximately 72% reduction in the model size.

This is confirmed by the other variants. Single-run accuracy is highest with applying SE after the last convolutional layer (Configuration B), but a lower reduction ratio ($r = 8$) (Configuration E) slightly decreases the accuracy while increasing the number of parameters. When used as additional band inputs (Configuration F), the NDVI and NDWI index yield a rate of $98.61 \pm 0.12\%$, and confirm that these vegetation indices, being deterministic combinations of the original Sentinel-2 bands, bring little information for classification when the original spectral bands are already available.

The results show that the very good performance on EuroSAT is driven more by the rich multispectral information contained in the Sentinel-2 imagery, than by specific architectural changes. The proposed configuration is therefore kept as it embeds light attention on channels with a very small parameter overhead (0.9% of the total parameters) and does not solely contribute to the accuracy of classification (in fact, it doesn't always) but it provides interpretable channel-

weighted information that can be combined with the spectral analysis based on SHAP, presented in the following section.

Table 2. Ablation study on the EuroSAT test set.

Con fig.	Description	Params	OA (%)	Config.
P1	Proposed network, SE after all stages (r=16)	1.324 M	98.72 ± 0.21	P1
P2	Proposed network, SE only after final stage	1.321 M	98.86	P2
P3	Proposed network, SE ratio r=8	1.335 M	98.54	P3
P4	Proposed network, 3 stages	0.364 M	98.73 ± 0.13	P4
P5	Proposed network, +NDVI/+NDWI inputs	1.325 M	98.61 ± 0.12	P5
P6	Proposed network, attention removed	1.313 M	98.76 ± 0.08	P6

All configurations were trained on the identical data partition with the same schedule. Configurations marked with + were repeated over five random seeds and are reported as mean $\pm$ standard deviation; the remaining two were trained once and are shown for completeness.

Grad-CAM explanations for one randomly selected correctly classified patch from each of the ten classes are displayed in Figure 6. The network focuses on the periodicity of the field structure in AnnualCrop, the connected canopy texture in Forest, the regular built-up pattern in Residential and the linear water channel in River. In each case the high-activation regions correspond to the surface features a human interpreter would use, which provides visual evidence that the prediction rests on physically meaningful structure rather than on spurious background cues.

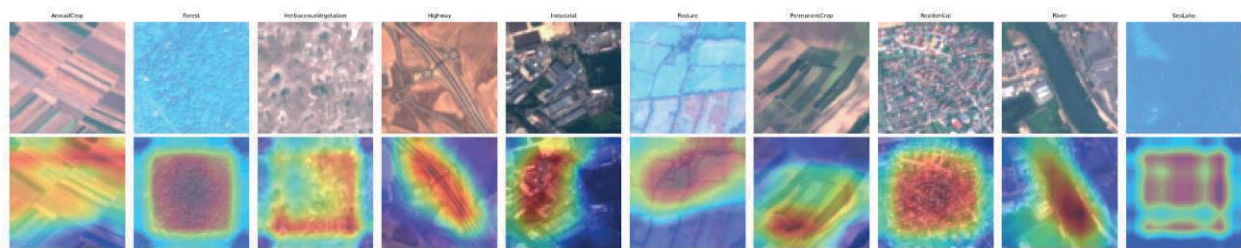


Figure 6. Grad-CAM explanations for one randomly selected correctly classified test patch per class. The patches were drawn at random from the correctly classified test samples and were not selected for visual quality.

Faithfulness of the explanations. Table 3 reports the faithfulness of the Grad-CAM maps, assessed over 300 randomly selected correctly classified test patches. Grad-CAM achieved a mean deletion AUC of 0.738 ± 0.307 and a mean insertion AUC of 0.916 ± 0.138 . The control matched for spatial smoothness reached 0.857 ± 0.201 and 0.864 ± 0.194 respectively, and the per-pixel random control 0.860 ± 0.206 and 0.861 ± 0.204 . Lower deletion and higher insertion values indicate greater faithfulness, so Grad-CAM outperforms both controls in both directions. Given the sample size, the deletion advantage over the smoothness-matched control corresponds to 5.6 standard errors and the insertion advantage to 3.8, so both differences are well established. The regions highlighted by Grad-CAM are therefore those the model actually relies upon, and not merely regions that appear meaningful to a human interpreter.

The two controls behave almost identically here, which indicates that on this benchmark the spatial frequency of an attribution map is not by itself responsible for its measured faithfulness. The smoothness-matched control was nonetheless retained because it is the more conservative comparison for spatially smooth explanations such as Grad-CAM.

Table 3. Faithfulness of the Grad-CAM explanations over 300 randomly selected correctly classified test patches.

Attribution	Deletion AUC	Insertion AUC
Grad-CAM	0.738 ± 0.307	0.916 ± 0.138
Random, matched spatial smoothness	0.857 ± 0.201	0.864 ± 0.194
Random, independent per pixel	0.860 ± 0.206	0.861 ± 0.204

The margin between Grad-CAM and both controls is therefore the quantitative counterpart of the visual evidence in Figure 6: the regions the method highlights are the regions the classifier actually relies upon.

A global SHAP analysis was conducted to quantify the relative importance of the thirteen spectral bands. Figure 7 reports the mean absolute contribution of each band over 500 test patches together with the standard deviation across those patches. The blue band B2 carries by far the largest contribution at 0.00469 ± 0.00234 , followed by the near-infrared band B8 at 0.00302 ± 0.00097 , the green band B3 at 0.00254 ± 0.00115 , the red band B4 at 0.00251 ± 0.00122 and the short-wave-infrared band B11 at 0.00206 ± 0.00086 . An intermediate group comprises B10, B5 and B1, with contributions between 0.00123 and 0.00135, while the red-edge bands B6 and B7, the second short-wave-infrared band B12, the narrow near-infrared band B8A and the water-vapour band B9 contribute least, the last of these at 0.00036 ± 0.00022 .

This ordering is consistent with the physics of the scene. The visible and near-infrared bands jointly capture the red-edge contrast that separates vegetated from non-vegetated surfaces, and B11 responds to surface moisture and to the built-up materials that distinguish the Industrial and Residential classes. The atmospheric bands B9 and B10 and the narrow near-infrared band B8A, which is strongly correlated with the broader B8, contribute least, which is what one would expect of channels that carry little independent surface information. The ranking proved stable: an independently trained instance of the same configuration produced an identical ordering for the six leading bands and differed only by an exchange of the seventh and eighth positions.

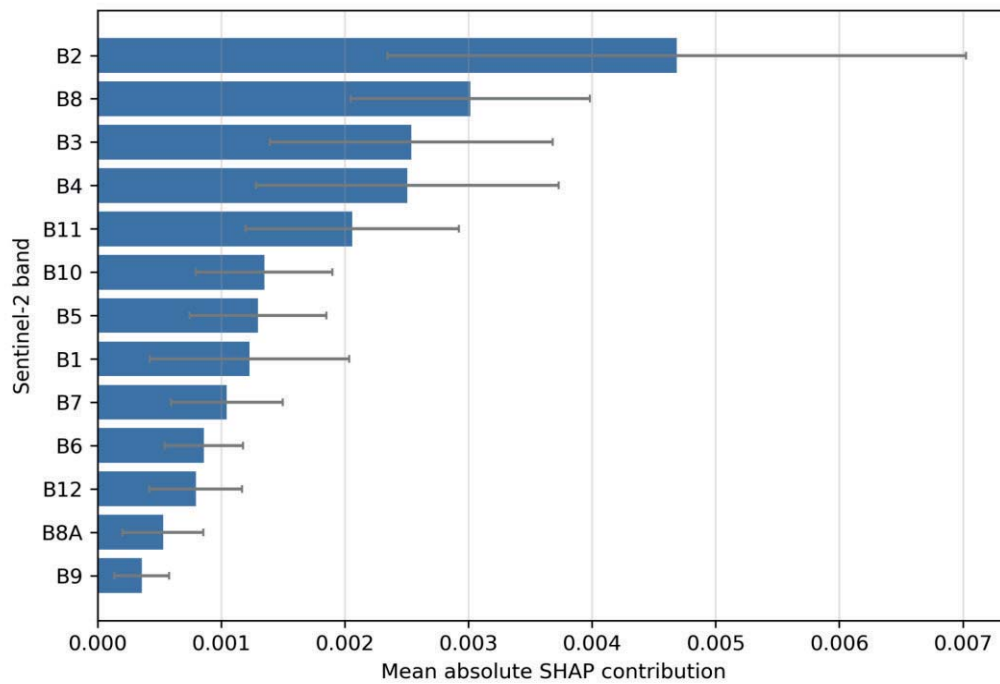


Figure 7. Mean absolute SHAP contribution of each Sentinel-2 spectral band over 500 test patches, estimated with GradientSHAP. Error bars denote one standard deviation across the sampled patches.

Comparison with a baseline under identical conditions. Table 4 places the proposed model in the context of published work on the same benchmark. To permit at least one directly comparable measurement, a ResNet-50 with its first convolutional layer adapted to thirteen input channels was trained from scratch on the identical data partition, preprocessing pipeline, augmentation regime and training schedule. It reached an overall accuracy of 97.90 % and a macro-averaged F1-score of 0.9782 using 23,559,882 parameters, requiring 15.1 s per epoch to train and 5.69 ms per patch at inference. The proposed model is therefore 0.82 percentage points more accurate while using approximately eighteen times fewer parameters and running 2.9 times faster. Configuration D of the ablation extends this observation further: with 364,440 parameters, sixty-five times fewer than ResNet-50, it matches the proposed model and still exceeds the baseline by 0.83 percentage points.

Table 4. Comparison with representative EuroSAT-based approaches.

Study / method	Input	Split / protocol	Accuracy (%)	Params	XAI
Helber et al. [8], ResNet-50	13 bands	authors' own split	98.57	~23.5 M	No
Helber et al. [8], RGB only	RGB	authors' own split	96.43	~23.5 M	No
Temenos et al. [23]	13 bands	authors' own split	94.72	compact CNN	Yes (SHAP)
Khan et al. [24], ViT-Large	RGB	transfer learning	>99	~304 M	Yes

Campos-Taberner et al. [18]	S2 time series	different task	98.7	n/a	Yes
ResNet-50, re-implemented	13 bands	identical split	97.90	23.56 M	No
MS-AttNet (this work)	13 bands	identical split, 5 seeds	98.72 $\pm$ 0.21	1.32 M	Yes (Grad-CAM + SHAP)

Taken together, the results in Table 4 support the conclusion that competitive accuracy and interpretability are not mutually exclusive, and that a compact all-band design remains a practical option wherever inference cost and transparency matter. Compared with the transformer-based and foundation-model approaches discussed in the Related Work section, the proposed framework prioritises computational efficiency and verified interpretability while maintaining competitive classification accuracy, which makes it suitable for resource-constrained remote-sensing applications.

Discussion

The findings support the hypothesis that competitive classification performance on multispectral agricultural data can be achieved by a compact network that remains transparent. Three findings are highlighted. First, the strong performance at a parameter count of only 1.32 M shows that a compact all-band convolutional design makes efficient use of the multispectral signal without requiring the capacity of a large backbone. This is practically significant for large-area monitoring where the inference cost matters. The comparison with the re-implemented ResNet-50 makes this concrete: the proposed model is 0.82 percentage points more accurate on the identical partition while using eighteen times fewer parameters and running 2.9 times faster, which is the difference between a model that can be applied operationally over large areas and one that cannot. This comparison demonstrates that the proposed framework achieves a favourable trade-off between model complexity, predictive performance, and explainability relative to representative CNN-based baselines [8, 25]. The class-wise evaluation also helps in understanding the behavior of the model, in addition to the overall accuracy of 98.72 ± 0.21 %. The overall performance of precision, recall and F1-score for most land-cover categories were consistently high, showing the proposed framework generalises well to classes with contrasting spectral properties. The confusion matrix indicates the majority of misclassifications were between classes with similar spectral properties, which is to be expected for a single acquisition for the Sentinel-2. Classes with unique spectral signatures, on the other hand, like water bodies and urban areas, were classified at very high accuracy. The results indicate that the classification errors that still occur are mostly caused by intrinsic spectral similarity, and not because of the suggested architecture.

Secondly, the explainability analysis aligns with existing agronomic knowledge. The SHAP ranking is led by the visible and near-infrared bands, which jointly define the red-edge contrast underlying the vegetation indices used throughout the discipline, and by the short-wave-infrared band that responds to canopy moisture and soil background; the Grad-CAM maps localise predictions to the structural features that distinguish each land-cover type. This concordance is supported quantitatively by the insertion and deletion curves of Table 3, which show that the highlighted regions genuinely drive the predictions rather than merely appearing plausible, and it therefore provides evidence that the high accuracy is not obtained through dataset shortcuts. It is also consistent with the larger remote-sensing literature, which has shown that interpretability analysis can validate, and even refute, assumptions held by a trained model, as demonstrated in the context of agricultural policy compliance [18] and systematic assessments of the faithfulness of XAI in Earth-observation tasks [17, 27]. Attribution methods are more generally known to open the black box of deep networks both in other disciplines [16, 28].

Third, the residual errors' structure provides information to operational design. They are restricted to spectrally similar vegetation classes and to a single acquisition date, and thus directly relate to the best path to improvement, namely the inclusion of temporal information. The use of Sentinel-2 image time series with temporal convolutions or self-attention has been demonstrated to separate different cropping systems that are hard to distinguish in a single image [11-12] and the recurrent model of the European Common Agricultural Policy (ECAP) [18] in an agricultural context is the same principle. It is therefore natural to extend MS-AttNet to a multi-temporal input.

There are some caveats that must be noted. The EuroSAT benchmark is multispectral, geo-referenced, but is made of relatively small, pre-cropped and well-curated patches, and performance on full scenes with mixed pixels, field boundaries and cloud contamination would be lower, and the reported numbers would be interpreted as best case. The ablation study of Table 2 quantifies the contribution of the individual architectural components, and the repeated runs characterise the variability of every reported figure. Its principal outcome is a negative one: on this benchmark neither the attention modules nor the fourth convolutional stage changes accuracy by more than the run-to-run spread. We consider it more useful to report this plainly than to attribute the observed performance to a component the evidence does not support. External validation on independent datasets remains outside the scope of this work, and the cross-regional generalisability of the model should therefore be interpreted with appropriate caution. The ten EuroSAT classes also reduce the agricultural landscape into a few general categories, and they are unable to distinguish individual crop species, meaning they are only indirectly applicable to agronomic use. Another issue is cross-regional transfer, since a model trained with European imagery can not always be adapted to use in other regions, where the cropping seasons and soil conditions are different. Lastly, the explanations provided are only approximations that may be a function of the parameters and method used, and should serve as a guide but not a substitute for expert judgment. Future work will therefore focus on evaluation against independent remote-sensing benchmarks that are less saturated than EuroSAT and on which architectural differences may become measurable, on extending the model to multi-temporal inputs, and on deployment over full Sentinel-2 scenes to assess robustness under operational conditions. The fact that a three-stage variant with 364,440 parameters matches the proposed model also suggests that further compression is available at no cost in accuracy, which is directly relevant to on-board or edge deployment. Despite these limitations, the integrated pipeline has proven to be a viable compromise between accuracy, efficiency and transparency that meets the needs of operational monitoring.

Conclusion

The objective of this study was to examine if multispectral deep learning for agricultural-land monitoring can be highly accurate and interpretable. We created a lightweight squeeze-and-excitation convolutional network (MS-AttNet) that receives all thirteen Sentinel-2 spectral bands as input and tested it in the interpretation of the confusion-matrix, ROC and per-class analyses, and using two interpretive techniques, namely Grad-CAM and SHAP. MS-AttNet attains an overall accuracy of 98.72 ± 0.21 % and a macro-averaged F1-score of 0.9868 ± 0.0021 on the EuroSAT benchmark using 1.32 million parameters, exceeding a ResNet-50 baseline trained on the identical partition by 0.82 percentage points at eighteen times smaller size and 2.9 times faster inference. A controlled ablation with five seeds per configuration shows that the architectural choices examined do not change accuracy beyond the variability of training, which we report as a finding about the benchmark rather than as support for any individual component. The explanations produced by Grad-CAM and SHAP were validated quantitatively and shown to be faithful to the model, and the resulting band importance ordering is consistent with the physics of the observed surfaces. The broader implications of these results are that competitive accuracy, a small parameter budget and verifiable interpretability are simultaneously attainable, which is what operational agricultural monitoring requires.

The proposed approach opens a black box classifier to a monitoring tool that provides the evidence for which analysts and regulators can review, as it reveals the "how" and "what" of a network. The clustering of residual errors by spectrally similar vegetation classes reveals the need for both integrating multi-temporal information and finer crop taxonomies, plus cross-regional validation, as the most promising avenues for making this approach a deployable agricultural-monitoring capability.

Data and Code Availability

The EuroSAT dataset analysed in this study is publicly available at <https://github.com/phelber/EuroSAT>. The source code, the configuration files required to reproduce the ablation study and the scripts used to generate the Grad-CAM and SHAP analyses are openly available at Kaggle: <https://www.kaggle.com/code/shapok/notebook8df371a2c7>.

Acknowledgement

This research is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24992908 «Support system for agricultural crop production optimization via remote monitoring and artificial intelligence methods (Agroscope)»).

References

- [1] Serek, A., Abdoldina, F., Asylbek, M., Smurygin, V., & Nabyeva, G. (2026). CTA-Net: A cross-temporal attention network for change detection in remote sensing imagery. *Big Data and Cognitive Computing*, 10(7), Article 225. <https://doi.org/10.3390/bdcc10070225>
- [2] Sharifi, A., & Safari, M. M. (2025). Enhancing the spatial resolution of Sentinel-2 images through super-resolution using transformer-based deep-learning models. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18, 4805–4820. <https://doi.org/10.1109/JSTARS.2025.3526260>
- [3] Serek, A., Amirgaliyev, B., Li, R. Y. M., Zhumadillayeva, A., & Yedilkhan, D. (2025). Crowd density estimation using enhanced multi-column convolutional neural network and adaptive collation. *IEEE Access*, 13. <https://doi.org/10.1109/ACCESS.2025.3597393>
- [4] Paheding, S., Saleem, A., Siddiqui, M. F. H., Rawashdeh, N., Essa, A., & Reyes, A. A. (2024). Advancing horizons in remote sensing: A comprehensive survey of deep learning models and applications in image classification and beyond. *Neural Computing and Applications*, 36(27), 16727–16767. <https://doi.org/10.1007/s00521-024-10165-7>
- [5] Kou, W., Shen, Z., Liu, D., Liu, Z., Li, J., & Chang, W. (2023). Crop classification methods and influencing factors of reusing historical samples based on 2D-CNN. *International Journal of Remote Sensing*, 44(10), 3278–3305. <https://doi.org/10.1080/01431161.2023.2218972>
- [6] Thanjaivadivel, M., Gobinath, C., Vellingiri, J., Kaliraj, S., & Femilda Josephin, J. S. (2025). EnConv: Enhanced CNN for leaf disease classification. *Journal of Plant Diseases and Protection*, 132(1), Article 32. <https://doi.org/10.1007/s41348-024-01033-6>
- [7] Maurya, R., Rajput, L., & Mahapatra, S. (2025). RAI-Net: Tomato plant disease classification using Residual-Attention-Inception Network. *IEEE Access*, 13, 64832–64840. <https://doi.org/10.1109/ACCESS.2025.3559804>
- [8] Helber, P., Bischke, B., Dengel, A., & Borth, D. (2019). EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7), 2217–2226. <https://doi.org/10.1109/JSTARS.2019.2918242>
- [9] Li, N., Hao, S., & Zhao, K. (2024). A hybrid CNN–vision transformer structure for remote sensing scene classification. *Remote Sensing Letters*, 15(1), 88–98. <https://doi.org/10.1080/2150704X.2024.2302348>
- [10] Yang, R., Qi, Y., Zhang, H., Wang, H., Zhang, J., Ma, X., Zhang, J., & Ma, C. (2024). A study on the object-based high-resolution remote sensing image classification of crop planting

structures in the Loess Plateau of Eastern Gansu Province. *Remote Sensing*, 16(13), Article 2479. <https://doi.org/10.3390/rs16132479>

[11] Pelletier, C., Webb, G. I., & Petitjean, F. (2019). Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11(5), Article 523. <https://doi.org/10.3390/rs11050523>

[12] Rußwurm, M., & Körner, M. (2020). Self-attention for raw optical satellite time series classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169, 421–435. <https://doi.org/10.1016/j.isprsjprs.2020.06.006>

[13] Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195–204. <https://doi.org/10.1038/s41586-019-0912-1>

[14] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939778>

[15] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2), 336–359. <https://doi.org/10.1007/s11263-019-01228-7>

[16] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[17] Kakogeorgiou, I., & Karantzalos, K. (2021). Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 103, Article 102520. <https://doi.org/10.1016/j.jag.2021.102520>

[18] Campos-Taberner, M., García-Haro, F. J., Martínez, B., Izquierdo-Verdiguier, E., Atzberger, C., Camps-Valls, G., & Gilabert, M. A. (2020). Understanding deep learning in land use classification based on Sentinel-2 time series. *Scientific Reports*, 10, Article 17188. <https://doi.org/10.1038/s41598-020-74215-5>

[19] Yassine, H., Tout, K., & Jaber, M. (2021). Improving LULC classification from satellite imagery using deep learning—EuroSAT dataset. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B3-2021, 369–376. <https://doi.org/10.5194/isprs-archives-XLIII-B3-2021-369-2021>

[20] Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D. B., & Ermon, S. (2022). SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 197–211). <https://dl.acm.org/doi/10.5555/3600270.3600285>

[21] Aleissae, A. A., Kumar, A., Anwer, R. M., Khan, S., Cholakkal, H., Xia, G.-S., & Khan, F. S. (2023). Transformers in remote sensing: A survey. *Remote Sensing*, 15(7), Article 1860. <https://doi.org/10.3390/rs15071860>

[22] Vali, A., Comai, S., & Matteucci, M. (2020). Deep learning for land use and land cover classification based on hyperspectral and multispectral Earth observation data: A review. *Remote Sensing*, 12(15), Article 2495. <https://doi.org/10.3390/rs12152495>

[23] Temenos, A., Temenos, N., Kaselimi, M., Doulamis, A., & Doulamis, N. (2023). Interpretable deep learning framework for land use and land cover classification in remote sensing using SHAP. *IEEE Geoscience and Remote Sensing Letters*, 20, 1–5. <https://doi.org/10.1109/LGRS.2023.3251652>

[24] Khan, M., Hanan, A., Kenzhebay, M., Gazzea, M., & Arghandeh, R. (2024). Transformer-based land use and land cover classification with explainability using satellite imagery. *Scientific Reports*, 14, Article 16744. <https://doi.org/10.1038/s41598-024-67186-4>

[25] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7132–7141). <https://doi.org/10.1109/CVPR.2018.00745>

[26] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. <https://doi.org/10.1038/s42256-019-0138-9>

[27] Wolanin, A., Mateo-García, G., Camps-Valls, G., Gómez-Chova, L., Meroni, M., Duveiller, G., Liangzhi, Y., & Guanter, L. (2020). Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. *Environmental Research Letters*, 15(2), Article 024019. <https://doi.org/10.1088/1748-9326/ab68ac>

[28] Gevaert, C. M. (2022). Explainable AI for Earth observation: A review including societal and regulatory perspectives. *International Journal of Applied Earth Observation and Geoinformation*, 112, Article 102869. <https://doi.org/10.1016/j.jag.2022.102869>