Deep Learning-Based Grassland Classification Using Multi-Modal Sentinel-1, Sentinel-2, and Street-Level Imagery

Konstantinos Christofi[a,b], Charalambos Chrysostomou[a], Iason Tsardanidis[c], Michalis Mavrovouniotis[a], Charalampos Kontoes[c], Diofantos Hadjimitsis[a,b]

[a]*Dept. of Big Earth Data Analytics, Eratosthenes Centre of Excellence, Limassol, Cyprus*
[b]*Dept. of Civil Engineering and Geomatics, Cyprus University of Technology, Limassol, Cyprus*
[c]*Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens, Athens, Greece*

Abstract

Accurate grassland classification is important for sustainable land management and ecological monitoring, particularly in policy contexts such as the EU's Common Agricultural Policy (CAP). This study explores the application of deep learning models to classify grasslands using multi-modal remote sensing data comprising Sentinel-1 SAR, Sentinel-2 multispectral, and street-level images. Each of these data sources provides complementary information: Sentinel-2 provides rich spectral information, Sentinel-1 provides structural and moisture-related information regardless of the weather conditions, and street-level images provide ground-level views with high detail. Our results show that Sentinel-2 alone provides strong classification accuracy. Adding Sentinel-1 provides small but consistent gains, and even when Sentinel-2 is unavailable, Sentinel-1 alone still provides quite accurate results. While the inclusion of street-level imagery does not surpass the performance of satellite-only models, it offers complementary value in certain cases, such as visually complex or ambiguous parcels. The results show the value of fusing satellite and ground-level data for improved grassland classification, demonstrating the potential of deep learning for setting up scalable and high-accuracy environmental monitoring systems.

Keywords:
Grassland Classification, Remote Sensing, Deep Learning, Sentinel-1,

Sentinel-2, Street-Level Imagery, Multi-Modal Data Fusion, Land Cover Mapping, Synthetic Aperture Radar (SAR), Multispectral Imagery

1. Introduction

Grasslands are one of the world's most important ecosystems contributing heavily to biodiversity conservation, carbon sequestration, and land stabilization ([1, 2, 3]). The environmental and agronomic importance of grasslands has placed them at the center of sustainability policies, including the European Union's Common Agricultural Policy (CAP), which promotes the conservation of permanent grasslands to improve ecosystem services such as water management and erosion regulation ([4, 5, 1]).

Accurate classification of grasslands is fundamental to allow informed decision-making on land-use, conservation, and agricultural management. Remote sensing offers scalable and reproducible methods for measuring the health, composition, and temporal trends of grasslands. With the growing availability of open-access satellite images, especially from the European Space Agency's Sentinel missions, there is now the possibility of monitoring grasslands at high spatial, temporal, and spectral resolution.

Sentinel-1 provides Synthetic Aperture Radar (SAR) images (Fig. 1a) that are not impacted by cloud cover or variations in lighting, making it particularly valuable in regions of frequent atmospheric disruption. Its capability in capturing surface roughness and vegetation structure augments the multispectral observations from Sentinel-2, which provides detailed spectral information on vegetation health, biomass, and land cover conditions. The synergy of these two sensors offers complementary viewpoints, and the concurrent usage of both has shown potential in enhancing the accuracy of land cover classification.

While Sentinel-2 optical imagery (Fig. 1b) often leads to better classification performance since it covers broad spectral capabilities, its access might be hindered by cloud or acquisition gaps. Under those circumstances, Sentinel-1 may deliver helpful information independently, as its structural and moisture-sensitive signals offer an alternative view of vegetation conditions.

Building on recent developments in Deep Learning (DL) and multi-modal data fusion, this research explores the integration of satellite imagery (Sentinel-1 and Sentinel-2) with Street-Level imagery (Fig. 1c) for grassland classification. Street-level images provide a ground-level view of vegetation cover,

color, and condition-factors that are not always clearly visible from satellite observation. By integrating satellite data with street-level images, we aim to enhance the classification performance and gain a deeper understanding of grassland ecosystems.

This study explores the impact of various combinations of data modalities-Sentinel-1, Sentinel-2 and street-level imagery-on the classification accuracy in a unifies deep learning framework. The results demonstrate the value of multi-modal fusion and also show that, in the absence of Sentinel-2 imagery, Sentinel-1 by itself is also capable of providing relatively good results. This work contributes to the increasing number of studies on remote sensing monitoring of grasslands and illustrates the value of deep learning for achieving higher accuracy, scalability, and robustness in land-cover classification systems.

(a) Sentinel-1: Synthetic Aperture Radar (SAR) image

(b) Sentinel-2: Multi-spectral optical image

(c) Street-level image

Figure 1: Sentinel-1 and Sentinel-2 satellite images and Street-Level images from the Netherlands region of interest (ROI), as used in our dataset.

2. Related Work

The rapid advancement of remote sensing, alongside the evolution of Deep Learning (DL) techniques, has significantly improved our ability to classify complex land-cover types such as grasslands. Sentinel-1 and Sentinel-2, which are both part of the European Space Agency's Copernicus program [6], are among the most useful data sources in this context. While Sentinel-2 provides rich multispectral imagery which is ideal for vegetation analysis, Sentinel-1 provides Synthetic Aperture Radar (SAR) data, which captures surface

structure and moisture regardless of weather or lighting conditions. These complementary modalities offer unique and synergetic perspectives.

Several studies have proven the effectiveness of deep learning architectures, and especially Convolutional Neural Network (CNNs), in exploiting Sentinel-2's spectral and spatial richness for vegetation monitoring and land-cover classification, For example [7] introduced several DL applications in remote sensing and emphasized CNNs' outstanding performance in spectral-spatial feature extraction. Following these observations, [8] released the EuroSAT dataset from Sentinel-2 imagery that allowed one to train CNN-based models with classification accuracies over 98% for land-use applications. Similarly, [9] demonstrated that deep neural networks significantly outperformed traditional models trained on Sentinel-2 imagery, especially in modeling complex vegetation patterns.

It has also been demonstrated that adding vegetation indices, including the Red Edge band and the Normalized Difference Vegetation Index (NDVI), increases the classification accuracy for grassland monitoring [10]. The ability of neural networks to distinguish between states of healthy and degraded vegetation is improved by these spectral features.

Sentinel-2 is frequently impacted by cloud cover, seasonal gaps, or a low revisit frequency, despite its remarkable performance in the majority of situations. In such cases, Sentinel-1 becomes particularly valuable. As a radar-based system, Sentinel-1 records backscatter information reflecting the moisture content, roughness, and the structure of the vegetation. When processing SAR time-series data, deep learning algorithms have demonstrated a lot of potential. For instance, [11] used Temporal CNNs (TempCNNs) to take advantage of seasonal fluctuations in SAR signals, demonstrating how temporal patterns can significantly improve vegetation classification performance.

Current studies also suggested the benefits of multi-modal data fusion, particularly the fusion of Sentinel-1 and Sentinel-2 to improve model robustness. Both early (feature-level) and late (decision-level) fusion techniques allow networks to learn textural and spectral features and result in more informative representations and better classification performance. Our results are consistent with this literature: Sentinel-2 alone yields strong performance, but Sentinel-1 and Sentinel-2 together are even better, especially in edge cases or cloudy conditions. In addition, when Sentinel-2 data are not available, Sentinel-1 alone still delivers reasonable predictive performance.

Another recent and emerging trend in remote sensing is the combina-

tion of street-view images with satellite imagery. Ground-level imagery provides contextual data-vegetation density, height, and perceived greenness-that might not be apparent from top-down imagery. Recent work by [12] proposed a dual-branch neural network to handle high-resolution satellite and street-view imagery with gated fusion to combine learned features. This approach improved the accuracy of urban village classification by 2% compared to satellite-only models. Similarly, [13] merged Sentinel-2 imagery with Google Street View images [14] for the classification of informal settlements with up to 5% better F1 scores for difficult classes.

Building upon prior studies highlighting the robustness of Random Forest (RF) for remote sensing tasks involving radar data [15, 16, 17], our recent work [18] further confirmed RF's effectiveness in classifying grassland areas using Sentinel-1 Synthetic Aperture Radar (SAR) data. In that study, we conducted a comparative evaluation of several machine learning models, including RF, SVM, XGBoost, Logistic Regression, and Deep Neural Networks, and found that RF achieved the highest accuracy and MCC scores when applied to SAR-based inputs. These findings reinforce the model's suitability for handling SAR-specific challenges such as speckle noise and polarization variability, and extend the evidence supporting its strong performance in vegetation and land-cover classification.

Recent advancements demonstrate that multi-modal deep learning, particularly those capable of integrating heterogeneous data sources like satellite and street-level imagery, offer powerful tools for improving classification performance in complex environments. In this study, we combine Sentinel-1, Sentinel-2, and ground-level imagery to address the challenge of grassland classification. While the addition of street-level images did not consistently improve overall accuracy, our results suggest that they can provide complementary context in specific settings where satellite observations are limited or ambiguous. This reinforces the potential value of ground-level data, particularly when carefully aligned and integrated with overhead imagery.

3. Data

3.1. Satellite-Based Datasets

The dataset utilized in this study is a multi-sensor resource provided by [19], designed to support grassland classification for agricultural monitoring. It comprises Earth Observation (EO) data from the Sentinel-1 and

Sentinel-2 satellites, part of the Copernicus program [6], managed by the European Space Agency (ESA). The dataset is pre-annotated, with Sentinel-1 and Sentinel-2 data geo-referenced and labeled using crop-type information from the Dutch Land Parcel Identification System (LPIS). Sentinel-1 captures Synthetic Aperture Radar (SAR) data, whereas Sentinel-2 provides multi-spectral optical imagery, offering complementary information crucial for vegetation classification. The dataset spans a period of seven months in 2017.

3.1.1. Sentinel-1 SAR Data

Sentinel-1 operates in the microwave spectrum, enabling data acquisition regardless of weather conditions or daylight availability. The dataset used in this study consists of SAR data collected monthly from April to October 2017. It includes two polarization modes: Vertical-Vertical (VV) and Vertical-Horizontal (VH), as well as coherence values from Horizontal-Horizontal (HH) and Horizontal-Vertical (HV) polarization. Each observation is linked to a unique parcel identifier (*parcel_id*), which corresponds to an agricultural parcel, and is labeled as either grassland or non-grassland.

The dataset includes radar backscatter coefficients for each polarization and month, denoted in columns such as $2017 - 04\_VH\_SAR$ and $2017 - 04\_VV\_SAR$. These backscatter values represent the intensity of the radar signal reflected from the surface, influenced by vegetation structure, moisture content. and surface roughness.

3.1.2. Sentinel-2 Optical Data

Sentinel-2 provides 13 spectral bands covering the visible, near-infrared, and short-wave infrared regions (see Table 1. The dataset consists of multi-spectral optical data collected monthly from March to October 2017. Each observation is associated with a *parcel_id* and labeled to indicate whether the area of interest belongs to the *Grassland* class.

The dataset includes Sentinel-2 reflectance values from all spectral bands across multiple time points. However, bands $B1$, $B9$ and $B10$, primarily designed for coastal and atmospheric applications [20], are excluded from analysis. The temporal sequences of spectral values allow dynamic monitoring, which is essential for land cover classification based on spectral properties.

Band	Function
B1	Coastal Aerosol
B2	Blue
B3	Green
B4	Red
B5	Red-edge
B6	Red-edge
B7	Red-edge
B8	NIR
B8a	Red-edge
B9	Water vapour
B10	SWIR
B11	SWIR
B12	SWIR

Table 1: Sentinel-2's 13 spectral bands

3.2. Ground-level dataset

3.2.1. Street-level Images

The third modality dataset we are using for this particular project was again provided from [19]. It consists of street-level images captured from February to October 2017, acquired from Mapillary after appropriate processing [19]. These images capture ground-level perspectives of various locations in the Netherlands, focusing specifically on agricultural parcels. Each image is annotated with a unique *parcel_id*, allowing for precise matching with the corresponding data entries in the Sentinel-1 and Sentinel-2 datasets. By establishing a connection between high-resolution ground-truth visual observations and satellite-based remote sensing data, it enables a comprehensive analysis for our use case.

A localized view of the parcels—classified according to whether or not they represent grassland—is provided by the street-level images. A Comma-Separated Values (CSV) file containing this classification is used to match each image to its corresponding parcel and land cover type.

3.3. Data Integration

The dataset construction process involved several integration steps across the three data modalities. First, Sentinel-1 and Sentinel-2 datasets were

merged using a common `parcel_id` and label field to ensure that only parcels present in both satellite sources were retained. This inner merge resulted in a consolidated dataset of remote sensing observations, which was subsequently aligned with a set of annotated agricultural parcels via a second inner join on `parcel_id`. This step ensured that the retained records had ground-truth labels derived from the Dutch Land Parcel Identification System (LPIS).

Street-level imagery data was handled separately. Each original image was split vertically into two halves to increase the granularity and effective sample size of the dataset. This preprocessing step yielded two distinct samples per image—labeled as `image_id_left.jpg` and `image_id_right.jpg` as shown in figure 2. Each half-image is associated with a `direction` field and mapped to a corresponding `parcel_id`. Since multiple images may correspond to the same parcel, this results in a many-to-one relationship between image records and parcel-level features.

Figure 2: Example of street-level image splitting. The image 689913158387442.jpg is split into 689913158387442_left.jpg and 689913158387442_right.jpg, each treated as an individual input sample for model training.

After all integration steps, over 36,000 records were available for training and evaluation. This is due to the fact that parcels with matching Sentinel-1, Sentinel-2, and annotation entries were retained in full, and street-level im-

agery was expanded through vertical image splitting. Since multiple images can correspond to the same parcel and each original image was split into two halves, the final dataset benefits from increased granularity and diversity of perspective without discarding any data. This design supports robust multi-modal learning, although the geographic extent is still limited to the $\tilde{5}$,000 parcels with complete cross-modal coverage.

3.4. Handling Class Imbalance

The dataset used in this study exhibits a significant class imbalance, with 31,438 records labeled as $Grassland$ and only 4,936 labeled as $Non-Grassland$. This imbalance poses a challenge for machine learning models, as they tend to be biased toward the majority class, potentially leading to poor generalization for the minority class [21, 22].

To address this issue and ensure balanced representation during training, we computed the inverse frequency of each class in the training data to derive appropriate class weights. These weights were then assigned to individual samples using PyTorch's $WeightedRandomSampler$ [23], which ensures that both classes contribute proportionately during the training process. This approach mitigates the effect of class imbalance and promotes more effective model learning.

4. Methodology

4.1. Data Preprocessing

This study utilizes three distinct data modalities—-Sentinel-1, Sentinel-2, and street-level imagery from Mapillary-—to perform grassland classification over Utrecht in the Netherlands. All datasets were georeferenced and spatially aligned at the parcel level to ensure consistency in representation across modalities. No missing values were observed in the preprocessed datasets, eliminating the need for imputation. Input features are standardized using the `StandardScaler` from `scikit-learn` [24] to normalize feature scales and improve training convergence.

All datasets are split into training, validation, and test sets using an 80:10:10 ratio. A stratified 10-fold cross-validation scheme is also applied to ensure balanced class representation and robust generalization.

To ensure consistency of format in input representations across the different modalities, all image inputs–including street-level images–are flattened

into one-dimensional arrays before they are processed by the model. Although this allows for a consistent tabular structure and simplifies the architecture of the fusion models, this design decision may limit the model's ability to capture local spatial features. Future work can explore convolutional or patch-based representations to preserve spatial context and potentially further improve performance results.

4.2. Feature Extraction and Input Representations

For Sentinel-1 data, input features include VV and VH polarizations from monthly acquisitions between April and October 2017, resulting in a total of 28 features (2 polarizations $\times$ 7 months). Coherence was not considered in the feature extraction process. In the case of Sentinel-2 data, 290 features were created by extracting values from 29 different acquisition dates between March and October in 2017, ranging from 10 spectral bands per image. No spatial aggregation, or temporal smoothing, or vegetation indices were calculated. Every band value at each time step was considered as an individual feature. This results to a total of $29 \times 10 = 290$ features per parcel, which corresponds to multi-temporal spectral information at raw band level. No cloud masking or atmospheric correction was performed during preprocessing.

Mapillary street-level imagery was preprocessed prior to feature extraction. Each image was resized to 224×224 pixels and normalized using standard ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]). The RGB values were first scaled to the $[0, 1]$ range and then normalized channel-wise. This process is summarized in the following function:

```
img_array = (img_array / 255.0 - mean) / std
```

After preprocessing, each image was flattened into a one-dimensional array and appended as a new feature column to the tabular Sentinel-1 and Sentinel-2 feature matrices. This resulted in an extended feature vector per sample, combining satellite data with corresponding street-level visual content. These composite vectors are then used as input for downstream neural networks, enabling models to learn from both remote sensing and street-level perspectives simultaneously.

4.3. Model Architectures and Training Configuration

To explore the predictive potential of various modalities in the grassland classification task, we developed both traditional Machine Learning (ML)

[25] baselines and a suite of neural network architectures. The former were applied to Sentinel-1, Sentinel-2, and their combination to establish a performance benchmark, while the latter were designed to handle richer multi-modal data inputs including street-level imagery, which cannot be used directly with traditional algorithms.

4.4. Traditional Machine Learning Baseline

To provide a baseline for vegetation classification before incorporating street-level imagery, several traditional machine learning models were trained on Sentinel-1, Sentinel-2, and their fused feature sets. These models included Logistic Regression [26], Random Forests (RF) [27], XGBoost (XGB) [28], and Support Vector Machines (SVMs) [29], each applied to pixel-level or patch-aggregated features from the remote sensing data.

These algorithms were selected because of their high performance on tabular data and popularity in remote sensing applications. Unlike deep learning models, they are not suited for raw image inputs, therefore, street-level imagery was excluded from these experiments.

For Sentinel-1 and Sentinel-2 data, the pre-processed input features were used as input to the model. When combining the two sources, features from both modalities were concatenated. All models were trained with 10-fold cross-validation, and hyperparameters (e.g., number of trees, max depth, learning rate for XGBoost, etc.) were optimized through grid search.

4.5. Neural Network Architectures

Building upon the insights obtained from traditional machine learning baselines on Sentinel-1 and Sentinel-2 data, we developed a suite of neural network architectures to leverage the richer and more complex multi-modal data available. These models are designed to handle not only satellite-derived features but also high-dimensional street-level imagery, which cannot be directly processed by classical ML algorithms. Implemented in PyTorch [23], the neural network architectures are structured to systematically isolate and assess the predictive contribution of each data source, both individually and in combination

Each architecture in our study falls into one of the main three categories: single-source models, dual-source models, and a multi-source fusion model, with all models sharing a set of common design principles. Data from a single modality is processed by the single-source models. *"CustomNN_S1"* is a multi-layer perceptron (MLP) model designed to handle features exclusively

from Sentinel-1 data, while *"CustomNN_S2"* is a deeper MLP architecture to handle the higher dimensionality of Sentinel-2 data. In contrast, *"CustomNN_SL"* is a Convolutional Neural Network (CNN) designed specifically to process street-level imagery.

The dual-source models combine information from two different data modalities. *"CustomNN_S1_S2"* integrates Sentinel-1 and Sentinel-2 features using separate MLP branches for each source, with their resulting embeddings concatenating and passing through additional fully connected layers for classification. Similarly *"CustomNN_S1_SL"* merges Sentinel-1 features (processed by a MLP) with street-level imagery (processed by a CNN), combining their outputs, passing through additional fully connected layers for joint prediction. *"CustomNN_S2_SL"* follows the same fusion approach as the aforementioned one, substituting Sentinel-1 with Sentinel-2.

Furthermore, we developed a multi-modal architecture, *"CustomNN_S1_S2_SL"*, which combines all three modalities - Sentinel-1, Sentinel-2, and Street-Level imagery. Each modality is processed through its own branch (MLPs for Sentinel data, and CNN for street-level images), and the resulting embeddings are concatenated and passed through subsequent fully connected layers to produce the final classification output.

Across all models, the MLP components consist of stacked fully connected layers that use *ReLU* activations and dropout regularization to prevent overfitting. The CNN components are composed of five convolutional blocks with progressively increasing filter depths, interspersed with max-pooling layers. Dense layers are applied after the final feature maps have been flattened. In fusion models, outputs from each modality-specific branch are concatenated before proceeding with further training into some more layers that produce the final classification. These modular structures support flexible multi-modal experimentation to help isolate the predictive contributions of each data source. A schematic overview of the model designs is provided in figure 3.

4.5.1. Evaluation Setup and Metrics

All models were trained using the *RAdam* optimizer, selected for its improved convergence stability compared to *Adam*. Training was performed for 500 epochs with a batch size of 64. *Cross-entropy* loss was used as the objective function, with class weights adjusted to compensate for class imbalance.

Hyperparameters such as learning rate, dropout rate, and architecture depth were tuned through grid search on the validation folds. The best-

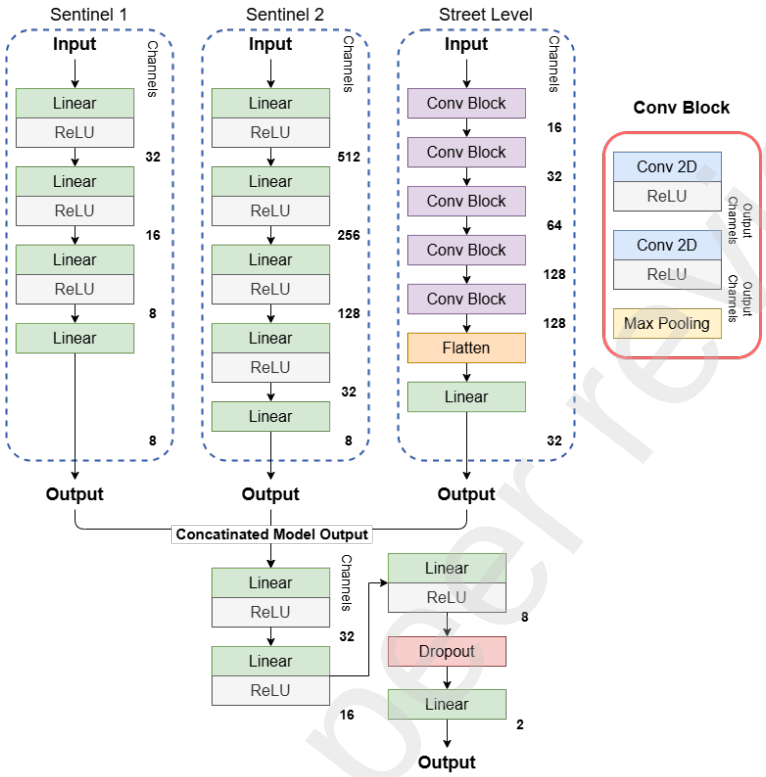

Figure 3: Multi-modal Deep Learning Architecture integrating Sentinel-1, Sentinel-2 and Street-Level image branches using parallel feature extractions, followed by concatenation and fully connected layers for final predictions.

performing model from each fold was selected based on the highest Matthews Correlation Coefficient (MCC) on the validation set.

All experiments were conducted on a high-performance computing server featuring dual *AMD EPYC 7452 32-core processors*, four *NVIDIA A40 GPUs*, and *512 GB of RAM*. It is important to note that not all of these resources were used concurrently during each training session. The models were developed using *PyTorch*, with GPU acceleration employed whenever possible.

5. Results

5.1. Comparison between Traditional Machine Learning models and Neural Networks

To assess the effectiveness of different classification methods for grassland mapping, we compared four traditional Machine Learning (ML) models, Support Vector Machines (SVM), XGBoost (XGB), Random Forest (RF), and Logistic Regression, with a neural network model. The models were trained and tested on three input data: Sentinel-1 (S1), Sentinel-2 (S2), and their fusion (S1_S2). Table 2 presents the average test performance in terms of accuracy and Matthews Correlation Coefficient (MCC), reported as mean ± standard deviation across 10 folds.

Across all datasets, the neural network performed competitively, especially with the fused S1_S2 input, achieving an accuracy of 0.951 ± 0.010 and an MCC of 0.784 ± 0.049. This performance was comparable to, and in some instances exceeded, by ensemble-based traditional models. Notably, the Random Forest classifier achieved the highest overall performance on the S1_S2 dataset, with an accuracy of 0.963 ± 0.013 and an MCC of 0.834 ± 0.060, outperforming the neural network on both metrics.

For the S1 input, traditional models consistently outperformed the neural network. XGBoost and Random Forest achieved considerably better MCC scores (0.662 and 0.634, respectively) than the 0.456 achieved by the neural network. This suggests that the neural network was not able to generalize effectively using only SAR data as input.

In Contrast, the S2 input alone yielded strong performance for all models. The neural network achieved an MCC of 0.765 and an accuracy of 0.946, but Random Forest again surpassed it with an MCC of 0.824 and an accuracy of 0.961. SVM and XGBoost also showed strong performance with MCCs over 0.80.

In general, the results highlight the power of ensemble learning methods like Random Forest and XGBoost in handling high-dimensional remote sensing data, especially when spectral data from Sentinel-2 is available. The neural network approach, while competitive is more sensitive to the input data, benefiting most from multi-modal fusion.

Despite the strong performance of traditional machine learning models, especially ensemble methods like Random Forest and XGBoost, our research extended further to explore the potential of neural networks with the addition of street-level imagery and its fusion with satellite data. Neural Networks

Average Test Results			
Dataset	**Model**	**Accuracy**	**MCC**
S1	SVM	0.893 ± 0.014	0.573 ± 0.063
	XGBoost	**0.929 ± 0.013**	**0.662 ± 0.070**
	Random Forest	0.924 ± 0.010	0.634 ± 0.055
	Logistic Regression	0.878 ± 0.027	0.578 ± 0.082
	Neural Network	0.846 ± 0.020	0.456 ± 0.064
S2	SVM	0.957 ± 0.010	0.807 ± 0.045
	XGBoost	0.957 ± 0.016	0.809 ± 0.072
	Random Forest	**0.961 ± 0.014**	**0.824 ± 0.067**
	Logistic Regression	0.906 ± 0.015	0.649 ± 0.060
	Neural Network	0.946 ± 0.013	0.765 ± 0.059
S1_S2	SVM	0.958 ± 0.007	0.815 ± 0.035
	XGBoost	0.955 ± 0.016	0.800 ± 0.071
	Random Forest	**0.963 ± 0.013**	**0.834 ± 0.060**
	Logistic Regression	0.898 ± 0.016	0.625 ± 0.056
	Neural Network	0.951 ± 0.010	0.784 ± 0.049

Table 2: Comparison of traditional machine learning models and neural network on grassland classification grouped by dataset.

are uniquely suited for handling unstructured data such as images, making them a natural choice for multi-modal learning. Given their capacity to learn complex spatial and visual patterns that cannot be captured by traditional models, we examine if the fusion of ground-level and top-down views can reveal new discriminative features for the grassland classification task.

5.2. Multi-Modal Learning with Satellite and Street-Level Imagery

Table 3 shows the average test results across different input modality combinations. Each setup was evaluated based on Accuracy, Matthews Correlation Coefficient (MCC), and Loss, reported as the mean ± standard deviation over multiple runs.

With an accuracy of 0.951 ± 0.010 and MCC 0.784 ± 0.049, the S1_S2 model performed the best overall, indicating that the most informative representation is provided by fusing Sentinel-1 and Sentinel-2 data. Though it showed a little greater variance and loss, the S1_S2_SL model also performed competitively, suggesting that adding street-level images did not always enhance generalization performance in this setting.

Among the single modality models, S2 outperformed both S1 and SL in all metrics, achieving an accuracy of 0.946 ± 0.013. This highlights the strength of Sentinel-2 imagery in this task. In contrast, the SL branch (street-level only) demonstrated the lowest accuracy and MCC, indicating its limited standalone effectiveness in the grassland classification task.

Fusion models (S1_SL, S2_SL) that combined street-level inputs with remote sensing data, showed intermediate performance. Notably, S2_SL maintained high accuracy (0.945 ± 0.025), but suffered from higher loss and variability. This suggests that although street-level data might provide complementary features, it also presents challenges in consistency or overfitting.

Average Test Results - Neural Networks			
Dataset	**Accuracy**	**MCC**	**Loss**
S1	0.846 ± 0.020	0.456 ± 0.064	0.525 ± 0.060
S2	0.946 ± 0.013	0.765 ± 0.059	1.221 ± 0.327
SL	0.604 ± 0.157	0.116 ± 0.072	0.892 ± 0.269
S1_S2	0.951 ± 0.010	0.784 ± 0.049	1.250 ± 0.471
S1_SL	0.865 ± 0.042	0.529 ± 0.093	0.593 ± 0.215
S2_SL	0.945 ± 0.025	0.770 ± 0.080	1.939 ± 1.127
S1_S2_SL	0.949 ± 0.010	0.778 ± 0.047	1.548 ± 0.642

Table 3: Average test performance (Accuracy, MCC, and Loss) across different input modality combinations for grassland classification. Results are reported as mean ± standard deviation.

6. Discussion

6.1. Key Takeaways from Results

The results highlight a notable difference between the performance of traditional machine learning models and neural network methods, particularly in unimodal settings. Traditional models, especially Random Forest and Support Vector Machines, demonstrated strong performance when trained on Sentinel-2 or the fusion of Sentinel-1 with Sentinel-2 data. For example, the Random Forest model achieved an MCC of 0.834 and accuracy of 0.963 using S1_S2 data, outperforming the neural networks on the same inputs.

However, despite this, neural networks remain competitive, particularly when integrating complex or unstructured modalities like street-level imagery. However, in purely satellite-based settings, traditional models remain

the top performers. The fusion of Sentinel-1 and Sentinel-2 in the neural network (S1_S2) yielded an accuracy of 0.951 and MCC of 0.784, closely trailing the best traditional models. This reinforces the notion that neural networks can effectively learn representations from complex spatial data, though with marginally reduced performance compared to highly optimized traditional models.

The street-level imagery (SL) modality, when used in isolation, proved to be the weakest performer. The neural network that handled this data achieved an accuracy of 0.604 and a notably low MCC of 0.116, indicating poor agreement with ground truth labels and limited standalone utility. However, its contribution became more meaningful in fusion settings. Notably, combining Sentinel-2 street-level imagery (S2_SL) resulted in an MCC of 0.770, suggesting that ground-level images complement overhead views if appropriately integrated. This level of performance was not replicated with traditional models on the same multi-modal data, highlighting the strength of neural networks in handling unstructured inputs like imagery.

The best-performing neural network configuration was the S1_S2 model, which provided a strong balance of accuracy and MCC. Interestingly, it also exhibited one of the highest losses (1.250), suggesting potential miscalibration of predicted probabilities, which warrants further research. These results underscore that while traditional models are strong baselines for satellite-based classification tasks, neural networks provide a flexible framework for integrating unstructured or heterogeneous data sources such as imagery.

6.2. Interpretation of Fusion Models

6.2.1. Advantage of Sentinel data fusion

The strong performance of the models using the S1_S2 fused dataset, observed across both traditional machine learning models and neural networks, is consistent with well-established findings in remote sensing research. Numerous studies have demonstrated that the complementary spectral and structural information provided by Sentinel-1 (SAR) and Sentinel-2 (multispectral) data, enhances land cover and vegetation classification tasks. Combining Sentinel-2's spectral richness with SAR's sensitivity to surface roughness and moisture, improves model robustness across a variety of ecological conditions. This aligns with earlier work suggesting that even in the absence of street-level imagery (SL), the fusion of multiple satellite modalities can yield powerful classification frameworks for ecological monitoring ([30, 31]).

6.2.2. Traditional Models Still Remain Strong Baselines

It is important to note that traditional machine learning models, particularly Random Forest and SVM, achieved comparable or even better performance in certain cases. For instance, the Random Forest model trained on S1_S2 data achieved the highest MCC (0.834) and accuracy (0.963) among all tested configurations, surpassing even the best-performing neural networks. This reinforces the strength of conventional classifiers when feature extraction is well-aligned with the input data, and suggests that such methods remain strong, efficient baselines in remote sensing applications. Their lower computational cost and better interpretability further strengthen their appeal in practical applications.

6.2.3. Street-Level Imagery Limitations

The inclusion of street-level (SL) imagery in fusion models did not consistently improve performance, and in some cases, appeared to reduce it. Notably the S1_S2_SL model underperformed compared to the S1_S2 model, despite having access to an additional data modality. This can likely be attributed to several practical issues. The quality of street-level images often varies due to occlusions (e.g. vehicles, buildings), inconsistent lighting, or gaps in geographic coverage. These factors can increase variance in the training data and degrade overall model robustness, especially when SL inputs are not filtered or weighted by confidence. These findings suggest that, while SL imagery can provide rich ground-level detail, careful integration techniques and data alignment are necessary for its use in large-scale grassland classification.

6.2.4. Loss and Calibration Issues

An interesting observation across the fused neural networks, involving Sentinel-2 (S1_S2, S2_SL, S1_S2_SL) is the relatively high cross-entropy loss, even while accuracy and Matthews Correlation Coefficient (MCC) performance is good. This discrepancy indicates potential model calibration issues, specifically, that class probabilities output by these models are not well-calibrated to the true probability of correct classification. Model calibration ensures that a model's predicted confidence scores accurately reflect the true probability of correctness. For example, a well-calibrated 0.8 confidence should be correct 80% of the time. Miscalibration occurs when this alignment fails, leading to over-confidence (predicted probabilities higher than

actual accuracy) or under-confidence (probabilities lower than actual accuracy). This mismatch undermines trust in model outputs and inflates loss metrics despite superficially strong classification performance ([32]).

In our scenario, the high loss values suggest that the fusion models-especially those with Sentinel-2-can have over-confidence in incorrect predictions or inconsistency in probability outputs. This can be attributed to the spectral ambiguity characteristic of Sentinel-2 data, which often struggles to cleanly separate grasslands from spectrally similar classes such as crops or shrublands ([33, 34]). Furthermore, temporal gaps in the data collection can lead to exclusion of critical phenological cues, hence lowering the capacity of the model to estimate confidence accurately based on the input ([35]).

Model miscalibration has severe consequences for real-world use, particularly in applications involving confidence-aware decision-making, like policy compliance monitoring or alert systems. An accurate yet poorly calibrated model may mislead users about the certainty of its predictions, thus resulting in inappropriate interventions or missed opportunities for verification. Addressing this in future studies could involve applying post-hoc calibration techniques such as modifying the training objective to encourage more calibrated outputs. Incorporating uncertainty estimation or confidence-aware loss functions would also make multi-modal classification models more interpretable and reliable. Interestingly, traditional models, such as Logistic Regression and Random Forest, did not suffer from the same level of miscalibration in our tests. While they may lack the expressive power of deep neural networks, their probabilistic outputs are often more reliable, or at least easier to calibrate post-hoc. This makes them advantageous in settings where output confidence is critical.

6.2.5. Conditional Benefits of SL

While traditional models were not evaluated with SL inputs due to architectural incompatibility, future work could explore engineered feature extraction from SL data to enable such comparisons. In some situations, SL can provide useful context even when the standalone SL network underperforms significantly (MCC = 0.116). Prior work has shown that SL imagery is particularly useful for identifying ecological indicators that are difficult to observe from above. Although this study did not explicitly isolate such contributions, future work could investigate specific visual cues in street-level imagery that contribute to improved classification performance ([13]). In models such as S2_SL, SL images may help marginally when satellite data

lacks sufficient textural or phenological information, especially during seasons when grassland conditions vary quickly. These advantages, however, seem to be very conditional since in order to be beneficial, SL images must be of high quality, correctly aligned, and relevant to the classification label.

7. Conclusion

This study evaluated the effectiveness of combining Sentinel-1, Sentinel-2. and Street-Level (SL) imagery for grassland classification using a Deep Learning architecture. Sentinel-1 and Sentinel-2 data fusion produced the best overall performance demonstrating that their complementary spectral and structural properties provide a strong foundation for vegetation mapping. Importantly, Sentinel-2 alone yielded high accuracy and MCC, making it a reliable standalone input. In cases where only Sentinel-2 data is available, our architecture performs well. When Sentinel-1 data is also accessible, integrating it can further enhance classification, but with modest gains.

Conversely, in cases where Sentinel-2 is unavailable (e.g. due to persistent cloud cover or data gaps), Sentinel-1 can still provide reasonably good results on its own. While performance is lower than Sentinel-2's, SAR data remains a viable alternative when optical imagery is unavailable. The use of SL imagery proved to be highly conditional. Although it offers fine-scale detail, its integration often introduced variance and elevated loss, likely due to scale mismatch, occlusion, or inconsistent image quality. These findings suggest that SL data should be used selectively and after careful preprocessing to align with satellite-scale labels.

8. Acknowledgments

This work was partially supported by the European Union's HORIZON Research and Innovation Programme by the AI-OBSERVER project funded from the European Union's Horizon Europe Framework Programme HORIZON WIDERA-2021-ACCESS-03 (Twinning) under the Grant Agreement No 101079468, and the 'EXCELSIOR': ERATOSTHENES: Excellence Research Centre for Earth Surveillance and Space-Based Monitoring of the Environment H2020 Widespread Teaming project (www.excelsior2020.eu). The 'EXCELSIOR' project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 857510, from the Government of the Republic of Cyprus through the

Directorate General for the European Programmes, Coordination and Development and the Cyprus University of Technology.

References

[1] R. d'Andrimont, G. Lemoine, M. Van der Velde, Targeted grassland monitoring at parcel level using sentinels, street-level images and field observations, Remote Sensing 10 (8) (2018) 1300.

[2] D. S. Ojima, B. O. Dirks, E. P. Glenn, C. E. Owensby, J. O. Scurlock, Assessment of c budget for grasslands and drylands of the world, Water, Air, and Soil Pollution 70 (1993) 95–109.

[3] J. Bengtsson, J. Bullock, B. Egoh, C. Everson, T. Everson, T. O'connor, P. O'farrell, H. Smith, R. Lindborg, Grasslands—more important for ecosystem services than you might think, Ecosphere 10 (2) (2019) e02582.

[4] M. Boval, R. Dixon, The importance of grasslands for animal production and other functions: a review on management and methodological progress in the tropics, Animal 6 (5) (2012) 748–762.

[5] L. Carlier, I. Rotar, M. Vlahova, R. Vidican, Importance and functions of grasslands, Notulae Botanicae Horti Agrobotanici Cluj-Napoca 37 (1) (2009) 25–30.

[6] S. Jutz, M. Milagro-Perez, Copernicus: the european earth observation programme, Revista de Teledetección (56) (2020) V–XI.

[7] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, F. Fraundorfer, Deep learning in remote sensing: A comprehensive review and list of resources, IEEE geoscience and remote sensing magazine 5 (4) (2017) 8–36.

[8] P. Helber, B. Bischke, A. Dengel, D. Borth, 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) (2019) 2217–2226.

[9] J. Nalepa, L. Tulczyjew, M. Myller, M. Kawulok, Segmenting hyperspectral images using spectral-spatial convolutional neural networks with training-time data augmentation, arXiv preprint arXiv:1907.11935 (2019).

[10] E. S. Weeks, A.-G. E. Ausseil, J. D. Shepherd, J. R. Dymond, Remote sensing methods to detect land-use/cover changes in new zealand's 'indigenous' grasslands, New Zealand Geographer 69 (2013) 1–13.
URL https://api.semanticscholar.org/CorpusID:128710565

[11] C. Pelletier, G. I. Webb, F. Petitjean, Temporal convolutional neural network for the classification of satellite image time series, Remote Sensing 11 (5) (2019) 523.

[12] B. Chen, Q. Feng, B. Niu, F. Yan, B. Gao, J. Yang, J. Gong, J. Liu, Multi-modal fusion of satellite and street-view images for urban village classification based on a dual-branch deep neural network, International Journal of Applied Earth Observation and Geoinformation 109 (2022) 102794.

[13] P. Wang, Deep learning for integration of satellite images and google streetview for mapping informal settlements (slums), Master's thesis, University of Twente (2024).

[14] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, J. Weaver, Google street view: Capturing the world at street level, Computer 43 (6) (2010) 32–38.

[15] M. Belgiu, L. Drăguţ, Random forest in remote sensing: A review of applications and future directions, ISPRS journal of photogrammetry and remote sensing 114 (2016) 24–31.

[16] A. A. Alharbi, Classification performance analysis of decision tree-based algorithms with noisy class variable, Discrete Dynamics in Nature and Society 2024 (1) (2024) 6671395.

[17] K. Christofi, C. Chrysostomou, I. Tsardanidis, M. Mavrovouniotis, C. Kontoes, D. G. Hadjimitsis, Deep-learning-based grassland mapping with sentinel-2: prioritizing key spectral bands and time periods, in: Eleventh International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2025), Vol. 13816, SPIE, 2025, pp. 344–353.

[18] K. Christofi, C. Chrysostomou, I. Tsardanidis, M. Mavrovouniotis, G. Guerrisi, C. Kontoes, D. G. Hadjimitsis, Remote sensing of grasslands: Performance comparison of radar and optical data in machine

learning classification, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 48 (2025) 295–300.

[19] G. Choumos, A. Koukos, V. Sitokonstantinou, C. Kontoes, Towards space-to-ground data availability for agriculture monitoring, in: 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2022, pp. 1–5. doi:10.1109/IVMSP54334.2022.9816335.

[20] K. Ali, B. A. Johnson, Land-use and land-cover classification in semi-arid areas from medium-resolution remote-sensing imagery: A deep learning approach, Sensors 22 (22) (2022) 8750.

[21] J. Chakraborty, S. Majumder, T. Menzies, Bias in machine learning software: Why? how? what to do?, in: Proceedings of the 29th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering, 2021, pp. 429–440.

[22] R. Wang, P. Chaudhari, C. Davatzikos, Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies, Proceedings of the National Academy of Sciences 120 (6) (2023) e2211613120.

[23] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems 32 (2019).

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830.

[25] M. I. Jordan, T. M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science 349 (6245) (2015) 255–260.

[26] M. P. LaValley, Logistic regression, Circulation 117 (18) (2008) 2395–2399.

[27] Y. Liu, Y. Wang, J. Zhang, New machine learning algorithm: Random forest, in: International conference on information computing and applications, Springer, 2012, pp. 246–252.

[28] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[29] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intelligent Systems and their applications 13 (4) (1998) 18–28.

[30] Y. Chabalala, E. Adam, K. A. Ali, Machine learning classification of fused sentinel-1 and sentinel-2 image data towards mapping fruit plantations in highly heterogenous landscapes, Remote Sensing 14 (11) (2022) 2621.

[31] F. Razzano, M. R. Iandolo, C. Zarro, G. Yogesh, S. L. Ullo, Integration of sentinel-1 and sentinel-2 data for earth surface classification using machine learning algorithms implemented on google earth engine, in: 2023 IEEE India Geoscience and Remote Sensing Symposium (InGARSS), IEEE, 2023, pp. 1–4.

[32] C. Y. Wijaya, Maximizing machine learning: How calibration can enhance performance (Jan 2023).
URL https://www.nb-data.com/p/maximizing-machine-learning-how-calibration

[33] V. Komisarenko, K. Voormansik, R. Elshawi, S. Sakr, Exploiting time series of sentinel-1 and sentinel-2 to detect grassland mowing events using deep learning with reject region, Scientific Reports 12 (1) (2022) 983.

[34] X. Fan, G. He, W. Zhang, T. Long, X. Zhang, G. Wang, G. Sun, H. Zhou, Z. Shang, D. Tian, et al., Sentinel-2 images based modeling of grassland above-ground biomass using random forest algorithm: A case study on the tibetan plateau, Remote Sensing 14 (21) (2022) 5321.

[35] S. Pérez-Carabaza, V. Syrris, P. Kempeneers, P. Soille, Crop classification from sentinel-2 time series with temporal convolutional neural networks, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, IEEE, 2021, pp. 6500–6503.