Engineering Applications of Artificial Intelligence 178 (2026) 115118

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Research paper

GeoFusion: A hybrid convolutional neural network-transformer model for boreal peatland classification

Fahimeh Farahnakian [a,b,*], Luca Zelioli [a], Farshad Farahnakian [a], Maarit Middleton [b], Javad Sheikh [a], Jukka Heikkonen [a]

[a] Department Of computing, University of Turku, Turku, Finland
[b] Geological Survey of Finland (GTK), Espoo, Finland

ARTICLE INFO

Keywords:
Remote sensing
Peatland classification
Multi-sensor fusion
Deep learning
Vision transformers
Environmental monitoring and management

ABSTRACT

The generation of high-resolution boreal peatland maps through the integration of multi-source remote sensing data provides critical support for greenhouse gas inventories and environmental management. In this study, we propose GeoFusion, a pixel-wise peatland classification architecture that leverages open-access radar imagery from Sentinel-1 and optical imagery from Sentinel-2. GeoFusion combines convolutional neural networks (CNNs) for hierarchical local feature extraction with vision transformers (ViTs) for modeling long-range dependencies, enabling effective fusion of spectral–textural and structural information. The architecture also incorporates data augmentation, multi-head self-attention, and dropout regularization to enhance robustness and mitigate overfitting. Performance was assessed at two classification levels: site type and fertility level, capturing both ecological diversity and functional characteristics of complex peatland ecosystems. Comparative experiments against well-known deep learning models and hybrid Transformer-based models were conducted across three ecologically distinct regions in Finland. The obtained results demonstrate that GeoFusion outperforms other models and achieves average accuracies of 79.25% and 72.68% for fertility-level and site-type predictions, respectively.

1. Introduction

Peatlands are unique ecosystems characterized by the accumulation of organic matter, in the form of peat, which is a dense layer of decayed vegetation characteristic to wetlands. These ecosystems play a crucial role in global carbon storage, acting as significant carbon sinks by sequestering vast amounts of atmospheric carbon dioxide over millennia (Minasny et al., 2024). Although peatlands cover only about 3%–4% of the Earth's land surface, they store approximately one-third of the world's soil carbon, which is more than the carbon stored in all other ecosystems (Intergovernmental Panel on Climate Change (IPCC), 2023). This carbon storage capability makes peatlands vital to mitigating climate change.

In Finland, peatlands are especially important, covering nearly 30% of the country's land area (Turunen and Valpola, 2020), as shown in Fig. 1. These peatlands are a vital part of natural heritage, playing a crucial role in biodiversity, hydrology, and carbon dynamics (Finland, 2022; Ghobadi et al., 2012). However, they have historically been used for agriculture, forestry, and energy production, leading to significant drainage and changes in land use, which pose challenges to their ecological integrity (de Waard et al., 2024). Therefore, the conservation and sustainable management of these ecosystems are essential, as they contribute not only to carbon storage but also to the regulation of water resources and the support of diverse habitats. One important task for protecting peatlands is the development of a classification model that allows us to identify and assess different types of peatlands along with their unique characteristics.

Remote sensing (RS) technology has revolutionized the way we observe and analyze the Earth's surface, providing detailed information about land and water bodies. By utilizing sensors installed on satellites or aircraft, electromagnetic radiation reflected/backscattered or emitted from the Earth's surface (Aplin, 2004) is recorded in RS data. These sensors can operate across various ranges of the electromagnetic spectrum, such as visible, infrared, and microwave, allowing for the identification and analysis of different materials and surfaces (Aplin, 2004). This technology enables the continuous monitoring of vast and often inaccessible areas, offering valuable insights into the Earth's land, including vegetation health (Nay et al., 2018), water bodies (Yigit Avdan et al., 2019), and changes in land use over time (Chughtai et al., 2021).

* Corresponding author at: Geological Survey of Finland (GTK), Espoo, Finland.
E-mail address: fahimeh.farahnakian@gtk.fi (F. Farahnakian).

https://doi.org/10.1016/j.engappai.2026.115118
Received 17 October 2024; Received in revised form 25 March 2026; Accepted 11 May 2026
Available online 20 May 2026

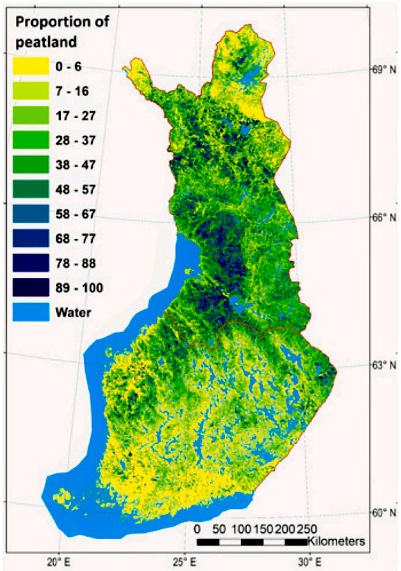

Fig. 1. Distribution map of peatlands in Finland (Haakana et al., 2023).

RS has proven to be a cost-effective and timely tool for land cover classification tasks, including mapping of various vegetation habitats (Phiri and Morgenroth, 2017). It enables the differentiation of various surface types, including grasslands, forests, concrete structures, and bodies of water, based on their unique spectral signatures. RS technology is important for peatland classification as it can provide consistent, large-scale coverage of peatlands that are often located in remote or challenging terrains. Peatlands, characterized by their unique wetland vegetation and waterlogged conditions, require specialized techniques to distinguish them from other types of land cover. RS enables the monitoring of peatland extent, health, and hydrological status by capturing subtle variations in vegetation and moisture content. However, the complexity and diversity of peatlands continue to pose significant challenges for their accurate mapping.

Recent advancements in Deep Learning (DL) have produced promising results in peatland classification by utilizing various neural network architectures to process and learn from multi-scale complex RS data effectively (Hosseiny et al., 2022; Jafarzadeh et al., 2022; DeLancey et al., 2020). Despite improvements in RS for peatland classification, the problems with current DL models remain unsolved. These models struggle to capture important long-range relationships and the complex patterns over time and space found in peatlands (Aleissaee et al., 2023).

To address the challenges of accurately classifying peatland types across spatial variations of fertility and moisture, we developed GeoFusion, a novel hybrid architecture that integrates CNNs and ViTs for pixel-level peatland classification. GeoFusion is, to our knowledge, the first approach in this domain to employ a hybrid (CNN+Transformer) framework. By fusing two open-source RS data sources, including radar imagery from S-1 constellation and multispectral optical data from S-2, GeoFusion could capture local spatial detail and global contextual relationships within peatland ecosystems. We implement a two-level classification scheme, distinguishing both site type and fertility status, thereby offering a refined ecological characterization of peatland conditions.

Through extensive evaluation across three distinct study sites in Finland, GeoFusion consistently outperformed traditional deep learning benchmarks, including both multi-stream CNNs (Farahnakian et al., 2023b) and encoder–decoder fusion models (Zelioli et al., 2024), in terms of classification accuracy for both site type and fertility. Unlike conventional CNNs, which are inherently limited by local receptive fields, GeoFusion leverages its Transformer component to model long-range spatial dependencies—a critical capability for classifying peatland landscapes characterized by extended hydrological and vegetation patterns. The key contributions of our work are summarized as follows:

1. **Development of a multi-modal architecture:** While hybrid CNN-Transformer models have been applied to domains like land cover classification (Li et al., 2023; Yan et al., 2023; Shailaja et al., 2025), GeoFusion represents the first application of a dual-stream CNN + Transformer model for peatland classification using S-1 and S-2 data.
2. **Token-level fusion across modalities:** Unlike previous approaches that rely on straightforward feature concatenation or late fusion, e.g., combining S-1, S-2 with terrain derivatives (Karlson and Bastviken, 2023; Pan et al., 2025), GeoFusion performs token-level fusion, combining class (CLS) and spatial embeddings across all four modality streams. This allows the Transformer to learn richer cross-modal relationships for classification.
3. **Inter-modal attention with multi-head cross-attention:** GeoFusion fuses features by explicitly modeling cross-modal dependencies via multi-head attention. It includes two rounds of cross-modal attention processing, enabling it to capture sophisticated interactions between SAR coherence and optical spectral features.
4. **Extensive evaluation:** The proposed framework is rigorously tested across three distinct peatland sites in Finland, showcasing its effectiveness and adaptability in diverse environments.
5. **Comparison with leading Transformer variants:** To demonstrate GeoFusion strengths, we conduct a benchmark comparison against state-of-the-art ViT architectures, including Swin Transformer (Liu et al., 2021), ViT-UNet (Zhou et al., 2024) and Data-efficient Image Transformers (DeiT) (Touvron et al., 2021).
6. **Public availability for reproducibility:** In contrast to other studies that only shared results and architecture details, our implementation with code and models is publicly available in GitHub at https://github.com/MaaTi-project/GeoFusion

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature on peatland classification and recent deep learning approaches used in land-cover mapping. Section 3 describes the study areas, datasets, preprocessing steps, and the proposed architecture. Section 4 presents the experimental results and discussion. Finally, Section 6 concludes the article with a summary of the findings and directions for future work.

2. Related work

Recent studies on peatland classification have utilized three primary types of sensor data to achieve satisfactory pixel-wise accuracy and address the complexities of peatlands: (1) Optical RS images, (2) Synthetic Aperture Radar (SAR), and (3) LiDAR (Light Detection and Ranging). Optical RS utilizes multispectral and hyperspectral imagery to capture the unique spectral signatures of peatlands, distinguishing them based on vegetation type, moisture content, and surface reflectance (Rezaee et al., 2018a). However, SAR is particularly useful in peatland environments due to its ability to penetrate cloud cover and detect subtle variations in surface roughness and moisture, which are key indicators of peatland health and extent (de Waard et al., 2024). LiDAR provides high-resolution elevation data, allowing for detailed mapping of peatland topography and hydrology, which are crucial for understanding peatland dynamics and their role in carbon sequestration (Guo et al., 2017).

Traditional machine learning (ML) approaches for RS segmentation have commonly employed algorithms such as Support Vector Machines (SVM) (Merchant et al., 2016; Middleton et al., 2012) and Random

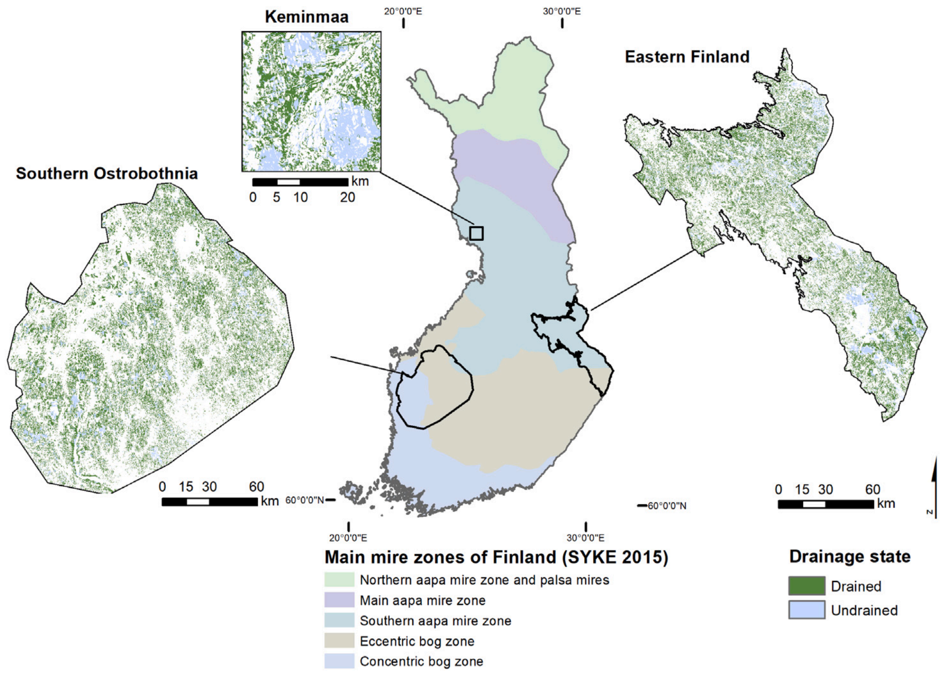

Fig. 2. Study pilot areas including Keminmaa, Southern Ostrobothnia, and Eastern Finland.

Forests (RF) (Millard and Richardson, 2015; Bourgeau-Chavez et al., 2017). These models rely on hand-crafted features derived from the imagery, such as texture, color, and spatial attributes, which can insufficiently capture the complex, hierarchical patterns inherent in RS data. As a result, they are limited in their ability to model data heterogeneity and are sensitive to noise, which reduces their effectiveness across varying peatland types and environmental conditions (Cheng et al., 2020). To address some of these limitations, traditional supervised classifiers have also been applied in the object-based image analysis (OBIA) domain, particularly for northern peatlands, where segment-based classification can enhance spatial context (Mahdianpari, 2019). However, a comprehensive study (Jafarzadeh et al., 2022) comparing multiple ML algorithms for wetland classification found no single traditional method to be consistently superior.

In recent years, DL models have attracted increasing attention for their ability to model complex spatial dependencies in data. For instance, in Hosseiny et al. (2022) introduced WetNet, a spatial–temporal ensemble DL model, which integrates Sentinel-1 (S-1) and Sentinel-2 (S-2) data to improve wetland classification by capturing both spatial and temporal variations. In another study (Jafarzadeh et al., 2022), the authors proposed Wet-GC, a novel multimodel graph convolutional approach that effectively handles limited training samples by exploiting the graph structure inherent in RS imagery. In DeLancey et al. (2020), the researchers compared DL models with traditional shallow learning methods, demonstrating the superiority of DL, particularly CNNs, in capturing the intricate patterns of wetlands across large-scale datasets. In O'Neil et al. (2020), the use of DL with physically informed input data was explored, enhancing the identification of wetland areas by integrating domain knowledge into the modeling process. In Rezaee et al. (2018b), researchers applied a deep convolutional neural network specifically to complex wetland classification tasks, highlighting the potential of DL to discern subtle differences in optical RS imagery. In Pouliot et al. (2019), the effectiveness of CNNs for wetland mapping with Landsat data was assessed, underscoring the ability of the model to generalize across diverse ecological regions. CNNs work best within a small area, so they often miss connections in an image, like peatland patches that are far apart but still related. In addition, CNNs have fixed filter weights that apply the same way to the whole image, which means they do not consider the specific content (Aleissaee et al., 2023). In our previous studies (Farahnakian et al., 2023b; Zelioli et al., 2024; Farahnakian et al., 2023a; Zelioli et al., 2025), we attempted to address some of the limitations of CNNs by incorporating skip connections to maintain feature continuity and utilizing a sliding-window strategy to improve local context modeling. Although these modifications resulted in noticeable improvements in classification accuracy, the inherent structural constraints of CNNs still limit their ability to capture complex spatial arrangements effectively.

To address these limitations, researchers have explored the potential of Transformers, specifically ViTs, for the pixel-wise classification task (Han et al., 2022). ViTs have demonstrated strong performance in various vision tasks by effectively modeling global dependencies and capturing complex spatial hierarchies without relying on local receptive fields, using self-attention layers (Vaswani, 2017). Their capability to handle heterogeneous spatial patterns suggests potential advantages over CNN-based approaches in the context of RS applications. However, Transformer-based models may miss certain local details when extracting features (Li et al., 2023). DeiT (Touvron et al., 2021) follows the same architectural design as ViT but introduces enhanced training strategies to improve data efficiency. In particular, DeiT incorporates knowledge distillation (KD) (Hinton et al., 2015), enabling a student network to learn from both ground-truth labels and a teacher model. Swin Transformer (Jamali and Mahdianpari, 2022) introduces a hierarchical design tailored for high-resolution imagery. Instead of global self-attention, Swin computes attention within local non-overlapping windows and shifts these windows between consecutive layers. ViT-UNet (Zhou et al., 2024) combines a ViT encoder with a UNet-style decoder to support dense prediction tasks. The Transformer encoder captures long-range contextual dependencies, while the decoder progressively upsamples feature maps and integrates skip connections from intermediate encoder layers to recover fine-grained spatial details.

3. Methodology

3.1. Study areas

Fig. 2 shows the three test areas, namely Keminmaa (Kemi), Southern Ostrobothnia (SO), and Eastern Finland (EF). They are distributed to cover the most common main mire vegetation zones and the drainage

Table 1
Description of input data sources used in this study.

Data	Bands	Grid cell size (m)	No. of rasters
S-1 intensity	Microwave band C (5.5 cm) June and September 2019, VV, VH, VV/VH	10	6
S-1 coherence	Microwave band C (5.5 cm) mean and max. of VV, VH, VV/VH	10	6
S-2	B2 (490 nm), B3 (560 nm), B4 (665 nm), B5 (705 nm), B6 (740 nm), B7 (783 nm), B8 (842 nm), B8A (865 nm), B11 (1610 nm), B12 (2190 nm)	10	10

status related to forestry, agriculture, and peat extraction. Kemi is characterized by a wide variation in peatland site types related to southern *aapa* mires ranging from the most nutrient-rich minerotrophic to oligotrophic types. The SO site is dominated by nutrient-poor site types of eccentric raised bogs. It is heavily influenced by past peat extraction and has a high number of active and abandoned agricultural organic soil field plots. EF test area, compared to the two other sites, has more topographic variation, a relatively lower number of training data representing a typical area outside of the densely inventoried peatland regions, and possesses a relatively high portion (73%) of drained peatlands. For a more detailed description of the test areas, see Middleton et al. (2023).

3.2. Data

Table 1 shows the summary of SAR S-1 and optical S-2 datasets that are used in this article. The S-1 data is comprised of intensity images (i.e., ground range detected products) including both VV and VH polarizations (V=vertical, H=horizontal) and their ratio. These data represent the backscattered signals detected, multi-looked, and projected into a single image. Mosaics of two instances in time, namely June and September 2019, were constructed. Secondly, SAR coherence time series were produced from single-look complex images. Data were acquired once a month over the summer season, May-September 2019, and mean and maximum coherence values of both polarizations and their ratio were input into the classification.

A S-2 mosaic was created using Level-1C images from June 2019 to August 2021. The images selected for the mosaic had low cloud coverage (0 – 2.5...10%) and low haze, depending on availability. All ten S-2 bands at 10/20 meters resolution were used and atmospherically corrected to Level-2A. The mosaic was created in two stages: first by merging tiles for smaller areas and then combining these subareas into a country-wide mosaic. A thorough band-wise compensation process was carried out to correct for reflectance variations between adjacent images. The imagery was finally resampled to a uniform 10-meter raster stack, with values scaled to 8-bit integer values, similar to S-1.

Separate training datasets were constructed for the drained and undrained subareas. A drainage mask raster created from the artificial drainage ditches stored in the topographic database by NLS (National Land Survey of Finland (NLS), 2023) was also made to apply the models on drained/undrained areas separately. This approach was chosen after extensive testing to minimize the misclassification of site types between drained and undrained areas. The training data were

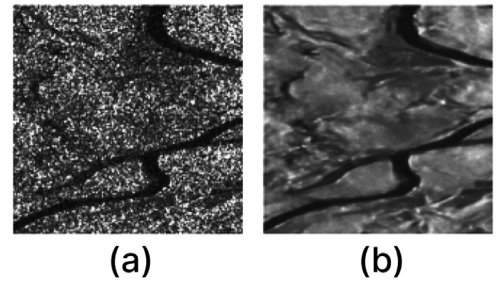

Fig. 3. (a) Original SAR images and (b) SAR images after denoising (Dalsasso et al., 2021).

sourced from the Geological Survey of Finland (GTK) peatland inventory (Geological Survey of Finland (GTK), 2022) and Luke national forest inventory datasets (Mäkisara et al., 2016). Each observation site was field-verified, and peatland site types were determined by qualified personnel. These site-type data are stored as point-wise observations in GIS databases, following the Finnish peatland site type classification system (Laine et al., 2018), which is grounded in the forest site type classification theory by A.K (1913). The classification uses site fertility and wetness as the key gradients. Of these, fertility is a more stable property, while wetness can be altered by artificial drainage.

For peatland classification, we adopted a two-level classification approach that includes both site type and fertility level. The site type classification is focused on the ecological and physical characteristics of the peatland, such as vegetation type, moisture conditions, and surface structure. By classifying peatlands at the site type level, we can identify specific ecological zones within peatlands, which is crucial for targeted conservation and management efforts.

Fertility level (FL) classification, on the other hand, provides insight into the productivity and nutrient status of the peatland. By including fertility level in the classification, we can assess the potential for peatlands to be used sustainably or identify areas that may require restoration to enhance their ecological functions. This dual-level classification approach is particularly important for creating detailed and actionable peatland maps. Table 2 shows the class specific numbers of the field reference locations.

3.3. Image preprocessing

Satellite image preprocessing is essential for land cover classification because it corrects distortions, noise, or inconsistencies in the data, leading to faster model convergence during training and improving generalization by reducing overfitting (Sowmya et al., 2017; Asokan et al., 2020). In this work, all S-1 and S-2 images are initially resized to fixed-size image patches into 128×128 pixels, preserving immediate neighborhood context without relying on larger-scale tiling.

The next step in this work for the image preprocessing pipeline is data normalization, another important step in preparing data for training ML-based models. Normalization improves training efficiency and model stability by rescaling all pixel values to a standardized range between 0 and 1. With this, we ensure that features with larger numerical values do not overshadow the learning process, allowing for a more balanced contribution from all features during training. After that, each input stream: S-1 (coherence), S-1 (intensity), and S-2, is processed through its augmentation pipeline. Augmentation protocols typically include random horizontal and vertical flips (each with ≈50% probability) and rotations by 0°, 90°, 180°, or 270°, applied independently to each 128×128 patch.

While SAR imagery obtained from S-1 satellites can generate radar images in all weather conditions and at any time of day, these images are prone to a type of multiplicative noise called speckle, which can degrade image quality and make further analysis more challenging. To

Table 2
Number of field reference locations used as training data in the study areas for each site type and corresponding fertility class fertility level (FL) (Laine et al., 2018).

Site type	Abbrev.	Keminmaa		South Ostrobothnia		Eastern Finland		FL***
		Undr.	Dr.	Undr.	Dr.	Undr.	Dr.	
Eutrophic paludified hardwood-spruce forest	LbK	3	19	4	1	1	2	1
Herb-rich hardwood-spruce swamp	RhK	22	9	16	20	34	29	1
Herb-rich sedge hardwood-spruce fen	RhSK	2	16	1	2	4	5	1
Herb-rich type I* heathy peatland	Rhtkg I	24	18	224	187	52	87	1
Eutrophic hardwood-spruce fen	VLK	0	6	1	0	0	1	1
Eutrophic birch fen	KoLK	7	25	0	0	1	2	1
Herb-rich type II* heathy peatland	Rhtkg II	0	5	0	1	0	0	1
Paludified *Vacc. myrtillus* spruce forest	KgK	57	46	46	41	42	29	2
Vacc. myrtillus spruce swamp	MK	42	2	54	27	67	50	2
Vacc. myrtillus type I* heathy peatland	Mtkg I	7	0	275	419	82	304	2
Herb-rich sedge fen	RhSN	28	24	23	12	14	16	2
Eutrophic fen	VL	18	40	1	1	0	1	2
Eutrophic flark fen	RiL	45	0	0	0	1	0	2
Herb-rich flark fen	RhRiN	42	9	4	2	5	17	2
Herb-rich sedge fen	RhSN	20	80	20	69	13	146	2
Eutrophic pine fen	VLR	22	2	4	4	4	3	2
Tall-sedge hardwood-spruce fen	VSK	29	1	4	0	3	10	2
Vacc. myrtillus type II* heathy peatland	Mtkg II	1	19	166	231	13	112	2
Vacc. vitis-idaea spruce swamp	PK	3	71	24	14	10	10	3
Spruce-pine swamp	KR	4	40	37	49	27	161	3
Paludified pine forest	KgR	32	58	105	123	46	41	3
C. globularis pine fen	PsR	18	22	29	96	35	63	3
C. globularis spruce swamp	PsK	6	44	4	1	0	0	3
Vacc. vitis-idaea type I* heathy peatland	Ptkg I	9	31	360	1265	62	902	3
Flark fen	VRiN	16	3	7	10	4	16	3
Tall-sedge pine fen	VSR	13	14	52	433	54	458	3
Cottongrass-sedge pine fen	TSR	1	44	8	3	21	3	3
Tall-sedge fen	VSN	64	40	56	20	54	40	3
Vacc. vitis-idaea type II* heathy peatland	Ptkg II	15	21	174	555	7	159	3
Dwarf scrub pine bog	IR	4	0	192	912	72	96	4
Dwarf shrub type I* heathy peatland	Yatkg I	1	17	91	864	28	476	4
Low-sedge *S. papillosum* fen	LkKaN	17	14	31	7	103	22	4
Cottongrass pine bog	TR	15	2	143	544	129	322	4
Low-sedge pine fen	LkR	0	20	95	74	59	160	4
Dwarf shrub type II* heathy peatland	Yatkg II	0	0	0	9	0	0	4
S. fuscum pine bog	RaR	18	311	464	1014	41	60	5
C. adina type I* heathy peatland	Jätkg I	0	65	4	70	0	10	5
S. fuscum bog	RaN	55	16	138	24	13	2	5
Ridge-hollow pine bog	KeR	1	68	150	42	24	2	5
Low-sedge bog	LkN	25	106	140	33	144	36	5
C. adina type II* heathy peatland	Jätkg II	0	31	0	1	0	0	5
Peatland agricultural field	–	0	0	1	157	0	7	NA
Abandoned field on peat soil	–	20	0	31	97	0	13	NA
Mineral soil**	–	0	0	198	293	274	453	NA

Notes: * Type I indicates heathy peatland that was naturally wooded in its original state (before draining), whereas Type II indicates heathy peatland that was originally open or hybrid (open/wooded) mire.
** Mineral soils observed in the field but mapped as peatland in the topographic database.
*** Fertility class (FL): 1 = eutrophic, 2 = mesotrophic, 3 = oligotrophic tall sedge type, 4 = oligotrophic low sedge type, and 5 = ombrotrophic.

mitigate the noise effect, especially in SO and EF study areas, this work first used all the processed S-1 coherence image pairs to calculate pixel-wise maximum and average values of both VV and VH polarizations, as well as their ratio. Then, we applied "deepdespeckling" Python package, which offers a powerful solution through two DL-based models: MERLIN and SAR2SAR, tailored to different S-1 data products (Dalsasso et al., 2022, 2021). Fig. 3 shows an example of the results of applying the despeckling method.

Beyond preprocessing, GeoFusion itself enhances robustness. The attention-based modules effectively capture long-range dependencies across images, enabling the model to understand noisy or spectrally variable pixels within a broader spatial context (Perera et al., 2022). This capability helps diminish the impact of random fluctuations that lack spatial coherence. However, it is important to note that spectral variability stemming from phenological shifts, fluctuations in soil moisture, or atmospheric effects remains a significant challenge. While GeoFusion addresses some aspects of this issue through multimodal fusion of SAR and optical data, it does not specifically model spectral variability by incorporating variability terms directly into the representation.

3.4. GeoFusion

This study aims to develop a standard model to handle multi-modal feature fusion for peatland classification. GeoFusion architecture is a

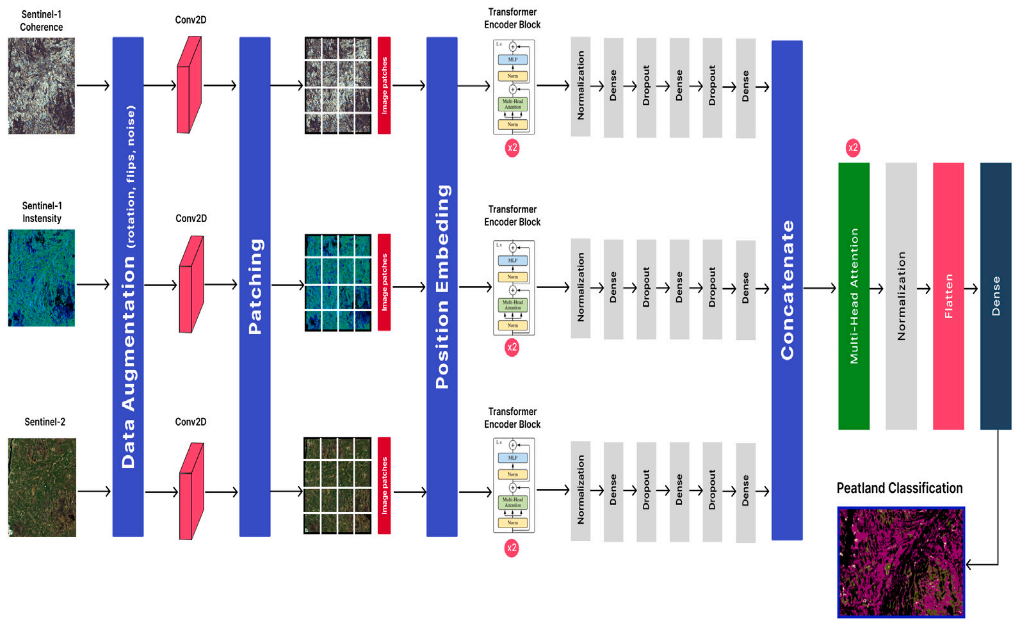

Fig. 4. Overview of Geofusion architecture.

three-stream hybrid CNN-Transformer framework (see Fig. 4) designed for pixel-wise classification of peatland site types and fertility levels, leveraging both S-1 and S-2 imagery. While we present the architecture with three branches in this instance, GeoFusion is flexible and can support as many modules as necessary, depending on the number of available data sources. These branches process S-1 coherence, S-1 intensity, and S-2 reflectance rasters. Each branch processes a distinct input modality or feature stream, with each input patch covering a 5 × 5 spatial window centered on the target pixel. To enhance generalization, a data augmentation layer $\mathcal{A}(\cdot)$ is applied independently to each input stream.

$$\tilde{\mathbf{X}}_i = \mathcal{A}(\mathbf{X}_i), \quad \mathbf{X}_i \in \mathbb{R}^{C_i \times 5 \times 5}$$

where $\mathbf{X}_i$ denotes the ith modality input patch with C_i channels, $\tilde{\mathbf{X}}_i$ is the augmented patch, and $\mathcal{A}()$ represents the stochastic data augmentation function.

Following augmentation, a Conv2D layer extracts fine-grained local spatial patterns from each branch, producing shallow feature maps. These are then passed through a patch embedding module, which flattens the feature maps and projects them into a d-dimensional latent space via a learnable projection matrix $\mathbf{E}_p$. A learnable classification token $\mathbf{x}_{\mathrm{CLS}}$ is prepended to the sequence.

$$\mathbf{Z}_i^{(0)} = [\mathbf{x}_{\mathrm{CLS}}; \mathbf{E}_p \cdot \mathrm{vec}(\tilde{\mathbf{X}}_i)]$$

where $\mathrm{vec}(\cdot)$ denotes vectorization, $\mathbf{E}_p \in \mathbb{R}^{d \times (C_i \cdot 5 \cdot 5)}$ is the patch embedding projection matrix, and $[\cdot;\cdot]$ indicates concatenation along the sequence dimension.

Each embedded sequence is processed by a modality-specific ViT encoder consisting of two Transformer layers. Each layer begins with Layer Normalization, followed by a Multi-Head Self-Attention (MHSA) operation with h heads to capture long-range spatial dependencies.

$$\mathbf{Z}_i^{(\ell)'} = \mathrm{MHSA}(\mathrm{LN}(\mathbf{Z}_i^{(\ell-1)})) + \mathbf{Z}_i^{(\ell-1)}$$

where ℓ denotes the Transformer layer index, LN() is layer normalization, and the residual connection adds the MHSA output to the layer input for stable gradient flow.

The MHSA output is followed by a second Layer Normalization and a feed-forward MLP block, with residual connections, dropout, and stochastic depth for regularization.

$$\mathbf{H}_i = \mathrm{MLP}(\mathrm{LN}(\mathbf{Z}_i^{(\ell)'})) + \mathbf{Z}_i^{(\ell)'}$$

where MLP() consists of two fully connected layers with non-linear activations, and $\mathbf{H}_i$ is the final encoded representation for modality i.

From each branch, the output CLS token is concatenated with spatially pooled patch embeddings to create a compact token representation. These tokens from all four branches form the input to a cross-modal attention block, which aligns features across S-1, S-2, and other complementary modalities.

$$\mathcal{T} = \{[\mathbf{z}_i^{\mathrm{CLS}}; \mathbf{z}_i^{\mathrm{spatial}}]\}_{i=1}^{4}, \quad \mathbf{f} = \mathrm{CrossModalAttn}(\mathcal{T})$$

where $\mathbf{z}_i^{\mathrm{CLS}}$ is the CLS token of branch i, $\mathbf{z}_i^{\mathrm{spatial}}$ is the pooled spatial token for branch i, $\mathcal{T}$ is the set of fused tokens from all branches, and $\mathrm{CrossModalAttn}()$ is the cross-modal multi-head attention operator.

The fused representation $\mathbf{f}$ is passed through a fully connected layer with eight output neurons, corresponding to the number of peatland classes, followed by a Softmax activation to produce the final probability distribution $\hat{\mathbf{y}}$. This classification block, serving as the final stage of GeoFusion, assigns each neuron's output as the predicted probability of its respective class. The class with the highest probability is selected as the final prediction of the model for the input patch, enabling multi-class peatland classification.

$$\hat{\mathbf{y}} = \mathrm{Softmax}(\mathbf{W}_c \mathbf{f} + \mathbf{b}_c)$$

where $\mathbf{W}_c \in \mathbb{R}^{K \times d}$ and $\mathbf{b}_c \in \mathbb{R}^{K}$ are the weights and biases of the classifier, $K = 8$ is the number of classes, and $\hat{\mathbf{y}} \in \mathbb{R}^{K}$ contains the predicted class probabilities.

This mechanism, meaning fusing CNN-based local feature extraction with Transformer-based global context modeling, GeoFusion enables capturing both fine-scale peatland textures and long-range ecological dependencies. In addition, this multi-branch fusion approach ensures that complementary spatial spectral, backscatter and coherence information are jointly exploited, resulting in improved classification robustness across diverse peatlands.

3.5. Implementation details

We compared the proposed GeoFusion models with our previous CNN and Encoder-Decoder (ED) models as presented in Farahnakian et al. (2023b) and Zelioli et al. (2024), respectively. Additionally, this work also compared GeoFusion to three more famous DL-based architectures: ViT-UNet (Zhou et al., 2024), Swin (Jamali and Mahdianpari, 2022), and DeiT. These Transformer-based models are adjusted based on peatland classes in this work. For this purpose, the input data were pre-processed and divided into patches of 5 × 5 pixels for all models. In addition, the final classification layers of all models were modified to perform pixel-wise classification, matching the number of peatland classes for both site type and fertility level prediction tasks.

To ensure fair comparison, all models were implemented under identical experimental settings. Each model was trained using the same training and validation datasets. We also used the same computational setup included an Intel i9-14900KS processor, 192 GB (4 × 48 GB) of RAM, and a NVIDIA RTX 5090 GPU with 24 GB of memory. We implemented the framework using Python 3.9, along with TensorFlow 2.9.1. Both training and inference of all models were carried out on the same machine.

In our implementation of the GeoFusion architecture, we utilized eight transformer layers for each data stream, with each layer employing eight attention heads and a model dimensionality of 64. The input data encoded using a learnable patch embedding of 64 dimensions. The concatenated outputs from the three streams were further processed through two dense layers with 16 and 8 neurons and ReLU activations, followed by dropout regularization with a rate of 0.1 to mitigate overfitting.

A categorical cross-entropy loss function was used for classification as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log \hat{p}_{i,c} \tag{1}$$

where N is the number of samples, C is the number of classes. $y_{i,c}$ is the binary indicator (0 or 1) if class label c is the correct classification for sample i, $\hat{p}_{i,c}$ is the predicted probability of sample i being class c.

The compared models were implemented following their original configurations as closely as possible. For example, all models are optimized with Adam. Learning rates, batch sizes, epochs, and patch sizes are provided in Tables 3. We also apply data augmentations suited to multi-sensor stacks like random flips/rotations. More details on the shared hyperparameters and the model-specific settings are provided in Tables 3 and 4.

Table 3
Shared training hyperparameters across comparison models.

Setting	Value
Optimizer	AdamW (β_1=0.9, β_2=0.999, weight decay = 1×10^{-4})
Learning Rate (LR)	10^{-3}
Batch size (tiles)	16
Epochs	200
Patch size P	5 pixels
Loss	cross entropy
Augmentations	Flip/rotate

4. Experiments

In this section, we present the results of our GeoFusion architecture for the classification of boreal peatlands in three study areas. We report the classification outcomes at two hierarchical levels: site type and fertility level.

4.1. Performance metrics

We selected three metrics commonly used in classification tasks, especially in the context of RS and land cover mapping.

1. **Overall Accuracy (OA):** Overall Accuracy quantifies the proportion of correctly classified samples relative to the total number of samples across all classes. It serves as a global performance indicator of the classification models. OA is computed as:

$$\mathrm{OA} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FP_i + FN_i)}$$

where n is the total number of classes, TP_i is the number of true positives for class i, FP_i is the number of false positives, and FN_i is the number of false negatives.

2. **Average Producer's Accuracy (PA):** provides a single metric summarizing the classification performance from the producer's perspective, indicating the overall reliability of the model in correctly classifying each class. The Average PA is the mean of the Producer's accuracies for all classes:

$$\mathrm{PA} = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FN_i}$$

where n is the number of classes, TP_i is the number of true positives for class i, and FN_i is the number of false negatives for class i.

3. **Average User's Accuracy (UA):** also known as precision, measures the reliability of the classification for all classes from the user's perspective:

$$\mathrm{UA} = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FP_i}$$

where FP_i is the number of false positives for class i.

4.2. Performance comparison

Accuracy. The experimental results, as shown in Fig. 5, indicate that the proposed GeoFusion architecture consistently achieves the highest performance in classifying site types across all three peatland study areas, with accuracies ranging from approximately 68% to 76%. This marks a significant improvement compared to retrained baseline models (CNN and encoder–decoder models, both ≤31%) and state-of-the-art transformer variants (ViT-UNet, DeiT, Swin). For instance, GeoFusion outperforms Swin by about 10% in the EF-Drained class (68.6% vs. 58.9%) and surpasses DeiT by approximately 20%. In SO, the advantage spans 9%–17% across transformer comparisons.

The accuracy of the fertility level classification (see Fig. 6) demonstrates a consistent performance difference between GeoFusion and competing models. This increase highlights significant trends in the advantages of multi-modal fusion for ecological mapping. GeoFusion shows remarkable stability across diverse peatland conditions, maintaining high overall accuracies above 73% in all classes, whereas other architectures exhibit greater variability, with some fertility categories dropping below 50% accuracy. The advantages of this architecture are clear in areas with subtle differences in fertility, such as EF-Undrained. By integrating features from S-1 and S-2, the model can pick up small details about vegetation and soil moisture that other models, like single-modality models or pure transformers, often miss.

PA and UA. Tables 5 and 6 summarize PA and UA metrics for both the site type and the fertility level classification, respectively. The results reveal that GeoFusion consistently achieves the highest PA and UA across the majority of classes, regions, and classification tasks, outperforming both CNN-based (baseline model) and transformer-based models. For PA (Table 5), GeoFusion scores range from 0.48 (EF-Site Type Drained) to 0.81 (Kemi–Fertility Level Drained), with a mean PA of 0.631, substantially higher than Swin (0.57), DeiT (0.47), ViT-UNet (0.450), CNN (0.32), and ED (0.13).

Similarly, in UA (Table 6), GeoFusion achieves a mean UA of 0.72, markedly higher than Swin (0.66), ED (0.42), DeiT (0.45), ViT-UNet (0.45), and CNN (0.31). In specific cases, such as Kemi–Fertility Level Drained (0.83) and Kemi–Site Type Drained (0.81), GeoFusion demonstrates exceptional reliability, with minimal commission errors. Competing transformer models, while occasionally competitive in isolated classes, exhibit greater variability across regions, suggesting weaker generalization. CNN and encoder–decoder baselines show substantial degradation, particularly for undrained fertility levels, reflecting a higher false-positive rate.

Table 4
Model-specific architectural and training settings.

Model	Specifics
GeoFusion	Dim = 64; heads = 8; two dense layers with 16 and 8 neurons.
CNN (Farahnakian et al., 2023b)	Three-stream late fusion; each stream consists of two convolutional layers and two pooling layers; max-pooling with 2 × 2 kernel and stride = 1 after each convolution.
ED (Zelioli et al., 2024)	Encoder: two convolutional layers (10 and 20 filters) with ReLU and pooling (×2); decoder: transposed convolution for upsampling.
DeiT	Same input size as ViT; knowledge distillation with a CNN/strong teacher; distillation token enabled.
Swin (Jamali and Mahdianpari, 2022)	Window size = 5 or = 1, depth = 8; shifted-window attention; multi-scale outputs.
ViT-UNet (Zhou et al., 2024)	ViT encoder (dim = 256, depth = 128); UNet decoder with channels {512, 256}.

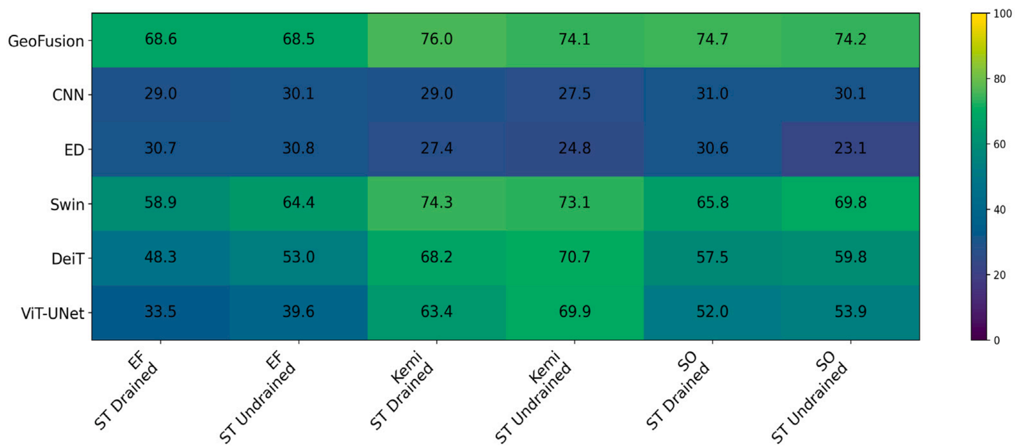

Fig. 5. Accuracy heatmap based on site types — ST (drained and undrained) by architectures and targeted research areas.

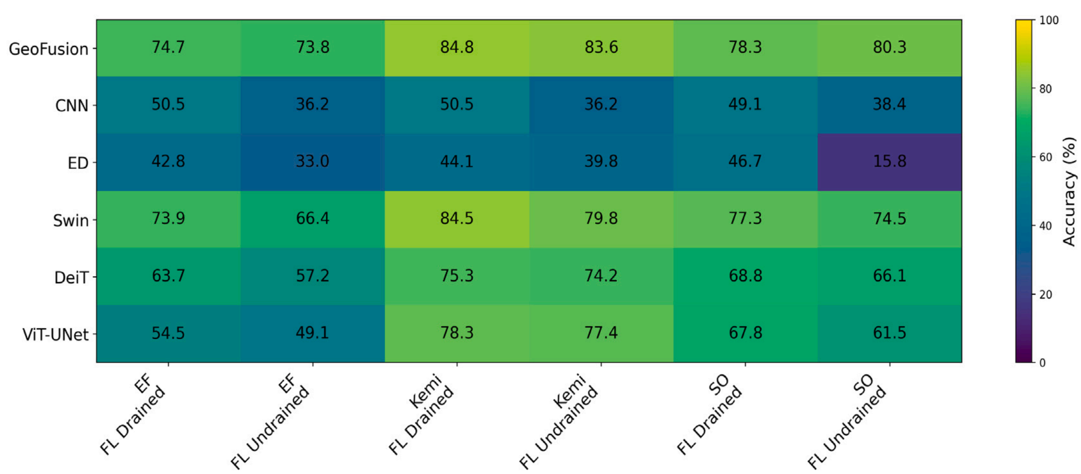

Fig. 6. Accuracy heatmap based on fertility level — FL (drained and undrained) by architectures and targeted research areas.

In addition, we visualized the confusion matrix in Fig. 7 that shows the classification results for peatland fertility levels across the three study areas. For Kemi (Fig. 7(a) and 7(b)), the highest correct prediction rates are observed for FL2 in the undrained dataset (0.87) and OSAF in the drained dataset (0.92), both benefiting from a larger number of training samples and distinctive spectral–textural signatures. In SO (Fig. 7(c) and 7(d)), FL5 in the undrained dataset (0.98) and FL4 in the drained dataset (0.83) achieve the highest accuracy, reflecting strong separability in these categories. For EF (Fig. 7(e) and 7(f)), FL5 (0.96) in the undrained dataset and FL3 (0.84) in the drained dataset lead in performance.

Across all regions, misclassification patterns reveal common challenges. In EF drained (Fig. 7(f)), notable confusion occurs between FL4 and FL5, consistent with the high similarity of these classes in S-1 and S-2 data during the acquisition period. Similarly, in SO undrained (Fig. 7(c)), FL1 and FL2 are occasionally confused, which may be attributed to transitional vegetation zones where canopy composition and soil moisture properties overlap.

4.3. Model complexity

Tables 7 and 8 summarize the model complexity and computational efficiency for peatland classes across the three study areas. The results reveal several notable trends. First, the largest transformer model, ViT-UNet, consistently exhibits the highest parameter count (16–17 M, ∼64 MB) and the slowest inference times, exceeding 950 ms in all regions and surpassing 2 s in some undrained cases. In contrast, GeoFusion maintains a moderate footprint (6–7 M parameters, 23–28 MB)

Table 5
The PA evaluation metrics for each architecture across three regions for fertility level and site types classification.

Area	Architecture	Fertility level		Site types	
		Drained	Undrained	Drained	Undrained
SO	GeoFusion	**0.68**	**0.65**	**0.52**	**0.62**
	CNN (Farahnakian et al., 2023b)	0.45	0.28	0.11	0.14
	ED (Zelioli et al., 2024)	0.25	0.005	0.04	0.01
	Swin (Jamali and Mahdianpari, 2022)	0.66	0.60	0.47	0.56
	DeiT (Hinton et al., 2015)	0.49	0.70	0.25	0.51
	ViT-UNet (Zhou et al., 2024)	0.63	0.53	0.25	0.31
Kemi	GeoFusion	**0.81**	**0.70**	**0.71**	**0.62**
	CNN (Farahnakian et al., 2023b)	0.74	0.64	0.35	0.28
	ED (Zelioli et al., 2024)	0.27	0.21	0.10	0.23
	Swin (Jamali and Mahdianpari, 2022)	0.66	0.62	0.62	0.59
	DeiT (Hinton et al., 2015)	0.55	0.47	0.48	0.49
	ViT-UNet (Zhou et al., 2024)	0.81	0.76	0.47	0.51
EF	GeoFusion	**0.57**	**0.64**	**0.48**	**0.57**
	CNN (Farahnakian et al., 2023b)	0.40	0.34	0.10	0.11
	ED (Zelioli et al., 2024)	0.12	0.15	0.08	0.13
	Swin (Jamali and Mahdianpari, 2022)	0.54	0.62	0.45	0.58
	DeiT (Hinton et al., 2015)	0.56	0.43	0.43	0.33
	ViT-UNet (Zhou et al., 2024)	0.40	0.41	0.10	0.22

Table 6
The UA evaluation metric for each architecture across three regions for fertility level and site types classification.

Area	Architecture	Fertility Level		Site Types	
		Drained	Undrained	Drained	Undrained
SO	GeoFusion	**0.80**	**0.73**	**0.67**	**0.72**
	CNN (Farahnakian et al., 2023b)	0.35	0.23	0.13	0.16
	ED (Zelioli et al., 2024)	0.42	0.008	0.67	0.72
	Swin (Jamali and Mahdianpari, 2022)	0.74	0.69	0.59	0.65
	DeiT (Hinton et al., 2015)	0.54	0.68	0.22	0.50
	ViT-UNet (Zhou et al., 2024)	0.52	0.42	0.34	0.40
Kemi	GeoFusion	**0.83**	**0.74**	**0.81**	**0.69**
	CNN (Farahnakian et al., 2023b)	0.54	0.42	0.50	0.67
	ED (Zelioli et al., 2024)	0.47	0.34	0.81	0.69
	Swin (Jamali and Mahdianpari, 2022)	0.79	0.69	0.57	0.64
	DeiT (Hinton et al., 2015)	0.58	0.44	0.43	0.41
	ViT-UNet (Zhou et al., 2024)	0.70	0.64	0.59	0.61
EF	GeoFusion	**0.76**	**0.69**	**0.64**	**0.67**
	CNN (Farahnakian et al., 2023b)	0.26	0.21	0.16	0.16
	ED (Zelioli et al., 2024)	0.05	0.17	0.64	0.61
	Swin (Jamali and Mahdianpari, 2022)	0.73	0.64	0.66	0.64
	DeiT (Hinton et al., 2015)	0.48	0.45	0.41	0.44
	ViT-UNet (Zhou et al., 2024)	0.29	0.34	0.29	0.29

while delivering inference speeds roughly three times faster than ViT-UNet (e.g., 293.01 ms vs. 976.74 ms in Kemi, drained class), yet achieving superior accuracy in challenging categories such as EF–Soil Drained and Kemi–Fertility Undrained.

Secondly, although GeoFusion's training time varies by region, ranging from approximately 21 s in Kemi undrained to 224 s in SO drained, it remains significantly lower than ViT-UNet in most cases and offers richer multi-modal representation capacity than the lightweight CNN. Among Transformer-based models, Swin achieves the most efficient computation (1.3–2.4 M parameters, ~5–9 MB) and inference speeds comparable to GeoFusion, but without matching its classification performance.

4.4. Feature maps

Fig. 8 provides a domain-informed view of how GeoFusion extracts and integrates evidence in peatland scenes. Starting from the raw inputs (Fig. 8-a), the first convolutional layer (Fig. 8-b) consistently amplifies linear drainage features, ecotonal boundaries, and micro-textures associated with hummock–hollow structure patterns that field studies identify as key discriminants of site type and fertility (Nungesser, 2003). The tokenization stage (Fig. 8-c, d, and e) transitions from pixels to position-aware patch tokens: the content-only embedding (Fig. 8-e) preserves local spectral and texture signatures, while the positional–content delta (Fig. 8-d) is highest along ditch margins, string–flark edges, and stand transitions, indicating that spatial priors are actively exploited where class meaning depends on location and arrangement, not appearance alone. The attention maps then clarify where the model gathers evidence: local attention (Fig. 8-f) concentrates on boundary neighborhoods and moisture-sensitive textures typical of undrained fens, whereas global attention (Fig. 8-g) links distant, co-oriented structures (e.g., parallel drainage lines or string patterns) that ecologists recognize as landscape-scale hydrological organization.

4.5. Qualitative results

The prediction maps generated by the GeoFusion model for site type and fertility level classifications demonstrate the ability of the model to delineate different classes across study areas accurately. These maps provide visual confirmation of the model's performance and highlight its effectiveness in capturing spatial variations in peatland characteristics. The site type prediction maps, illustrated in Fig. 9, show the spatial distribution of various site types across the three study areas (Kemi, SO, and EF). Fig. 10 presents the prediction map for fertility level classification. This map illustrates the spatial distribution of different fertility levels, showcasing the detailed and accurate classification. For

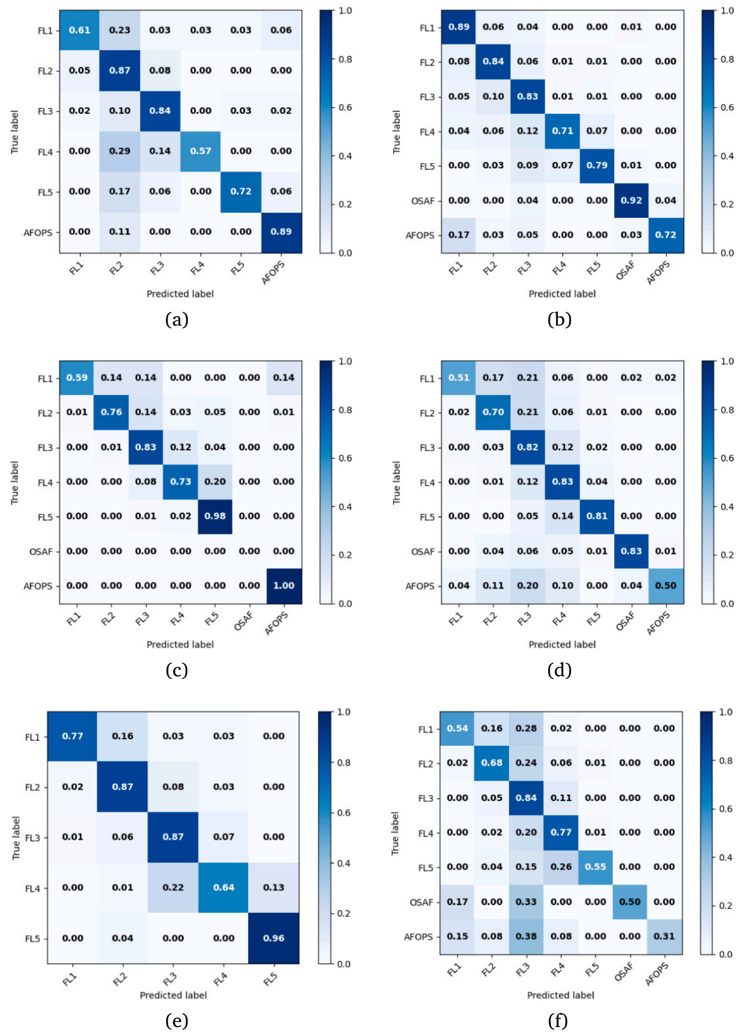

Fig. 7. The confusion matrix of GeoFusion for fertility level classification includes (a) Kemi undrained, (b) Kemi drained, (c) SO undrained, (d) SO drained, (e) EF undrained, and (f) EF drained.

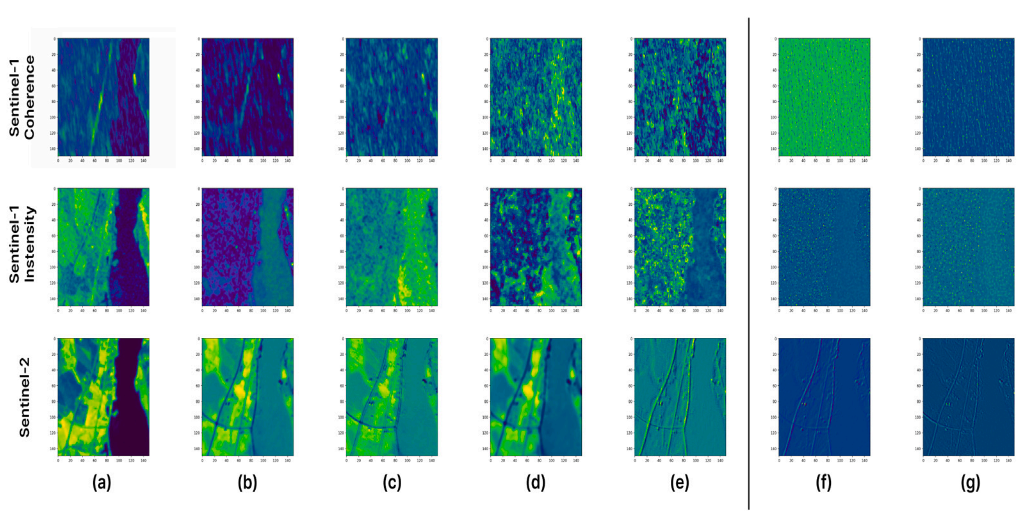

Fig. 8. Feature-map progression for S-1 coherence, S-1 intensity, and S-2 streams: (a) input; (b) post-conv edge/texture maps; (c) token+position embeddings; (d) positional–content delta (spatial prior gain); (e) content-only tokens; (f) local attention (windowed heads); (g) global attention (rollout across layers).

Table 7
Model complexity across areas for the drained class.

Area	Model	Total parameters	Estimated total size (MB)	Training (s)	Inference (ms)
SO	GeoFusion	7,260,701	27.70	224.0	474.74
	CNN (Farahnakian et al., 2023b)	273,280	0.92	97.34	180.61
	ED (Zelioli et al., 2024)	12,108,112	46.19	106.06	235.00
	Swin (Jamali and Mahdianpari, 2022)	2,357,121	8.99	**150.87**	511.77
	DeiT	8,433,275	32.17	300	1068.98
	ViT-UNet (Zhou et al., 2024)	16,715,555	63.76	454.49	**1859.51**
Kemi	GeoFusion	6,000,889	22.89	34.41	293.01
	CNN (Farahnakian et al., 2023b)	1,622,496	6.19	32.87	141.08
	ED (Zelioli et al., 2024)	13,866,055	52.89	19.75	147.91
	Swin (Jamali and Mahdianpari, 2022)	1,349,151	5.15	23.09	303.34
	DeiT	4,825,399	18.41	44.34	577.64
	ViT-UNet (Zhou et al., 2024)	9,558,259	36.46	61.40	976.74
EF	GeoFusion	7,268,925	27.73	89.42	293.28
	CNN (Farahnakian et al., 2023b)	237,280	0.92	33.87	95.85
	ED (Zelioli et al., 2024)	1,220,528	4.66	36.35	119.28
	Swin (Jamali and Mahdianpari, 2022)	2,385,825	9.10	59.53	305.57
	DeiT	8,434,331	32.17	108.28	580.03
	ViT-UNet (Zhou et al., 2024)	**16,716,611**	**63.77**	151.22	956.98

Table 8
Model complexity across areas for the undrained class.

Area	Model	Total parameters	Estimated total size (MB)	Training (s)	Inference (ms)
SO	GeoFusion	7,263,271	27.71	105.76	475.76
	CNN (Farahnakian et al., 2023b)	273,280	0.92	47.26	163.25
	ED (Zelioli et al., 2024)	1,498,544	49.19	5.76	233.82
	Swin (Jamali and Mahdianpari, 2022)	2,366,091	9,03	71.74	512.70
	DeiT	8,433,605	32.17	146.13	1066.55
	ViT-UNet (Zhou et al., 2024)	16,715,885	**63.77**	**222.93**	**2043.43**
Kemi	GeoFusion	6,002,945	22.90	20.93	297.36
	CNN (Farahnakian et al., 2023b)	1,622,496	6.19	13.85	141.89
	ED (Zelioli et al., 2024)	**13,866,127**	52.90	12.66	159.89
	Swin (Jamali and Mahdianpari, 2022)	1,353,255	5.16	14.89	303.64
	DeiT	4,825,663	18.41	29.39	580.66
	ViT-UNet (Zhou et al., 2024)	9,558,559	36.46	42.18	972.87
EF	GeoFusion	5,908,317	22.54	20.93	294.78
	CNN (Farahnakian et al., 2023b)	237,280	0.92	14.00	96.88
	ED (Zelioli et al., 2024)	1,220,591	4.66	15.90	129.20
	Swin (Jamali and Mahdianpari, 2022)	1,347,843	5.14	25.40	304.99
	DeiT	4,815,003	18,37	47.31	574.77
	ViT-UNet (Zhou et al., 2024)	9,547,899	36.42	68,01	976.11

both maps, we illustrate the maps for three selected subsections of each study area.

5. Discussion

5.1. Observations and findings

The integration of multi-modal RS data with advanced DL-based architectures offers a promising pathway for mapping complex ecosystems such as boreal peatlands. In this study, we proposed GeoFusion, a dual-branch CNN-Transformer fusion network that jointly processes S-1 (SAR) and S-2 optical imagery to enhance feature extraction for pixel-wise classification of peatland at two levels: fertility level and site type. The experimental design of this work followed the Finnish peatland classification framework (Laine et al., 2018), with evaluations conducted across three heterogeneous study areas in Finland: SO, Kemi, and EF, each presenting distinct hydrological, peatland vegetation, land use, and training data availability. The key findings from this research are as follows:

- The results revealed that Geofusion could significantly improve both accuracy and class-specific accuracies (PA and UA) for all three sites. For fertility-level classification, GeoFusion achieved accuracies of 85.2%, 82.7%, and 78.9% for EF, Kemi, and SO, respectively, representing gains of 8%–20% over Swin and DeiT, and up to 30% over the retrained CNN. In the site types classes, GeoFusion's advantage was similarly clear, with performance improvements of 9%–17% over Swin and 15%–22% over DeiT in the categories, such as EF–Site type Drained and SO–Site type Undrained.
- GeoFusion demonstrated strong cross-site stability, maintaining accuracy variations within 6% across the three sites, suggesting robust domain generalization despite ecological variability of peatlands. However, the gains were not uniform; performance in EF's site type categories was comparatively lower (e.g., 68.6% for Drained, 70.3% for Undrained) due to high intra-class similarity in the RS data, seasonal phenological overlap, and class imbalance.
- Across nearly all sites, GeoFusion achieved higher PA and UA than other DL and Transformer-based models. For instance, in Kemi–Fertility Drained, GeoFusion's PA reached 0.81 compared to Swin's 0.66 and DeiT's 0.55. This indicates that GeoFusion's attention layers implicitly learn to down-weight noisy or redundant modality-specific tokens and focus on discriminative patterns. It means that the model is performing an internal form of adaptive feature selection that could reduce the need for manual pre-filtering of input features, a valuable capability for large-scale automated monitoring.
- Performance variations relate to class imbalance and spectral similarity. Despite overall gains, GeoFusion's PA drops in certain

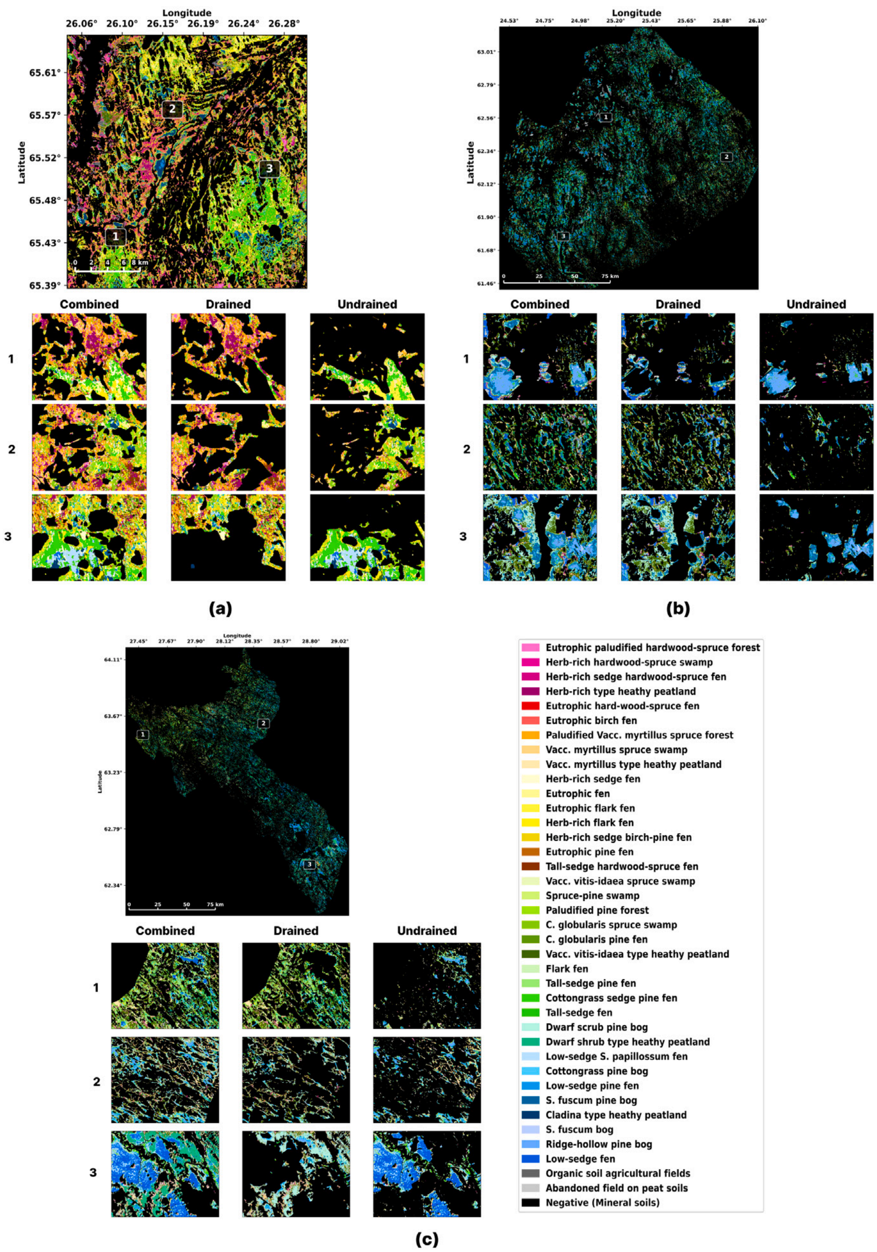

Fig. 9. (a) Keminmaa, (b) Southern Ostrobothnia, and (c) Eastern Finland. Details of the prediction map for three subregions (1–3) can be compared between drained and undrained.

EF site type classifications (Drained: 0.48, Undrained: 0.57). This is likely due to strong similarity between some site types, where vegetation and surface moisture conditions are similar, reducing the contrast in both SAR backscatter/coherence and reflectance. Class imbalance in these subsets, evident in Table 2, may also bias the model towards dominant classes by lowering omission performance for underrepresented categories. Furthermore, fine-scale spatial heterogeneity in EF may limit the discriminative benefit of patch-level token fusion, increasing confusion between ecologically similar classes.

- Model complexity interacts with training sample availability. Our complexity analysis shows that GeoFusion maintains a parameter count (6–7M) that balances representational capacity and computational efficiency, requiring far fewer resources than ViT (16M+) while outperforming smaller CNN baselines. However, model capacity must be interpreted in light of available training data: large models such as ViT struggled in low-sample categories due to insufficient data-to-parameter ratios, leading to unstable PA/UA scores. In contrast, GeoFusion's intermediate complexity allows it to generalize better in small-to-moderate sample regimes while still capturing fine-grained patterns in large-sample classes. This relationship underscores the need to align model complexity with dataset size and distribution to optimize mapping performance.
- The feature maps in Fig. 8 indicate that GeoFusion's performance gains arise from the deliberate coupling of CNN-derived local detail with Transformer-based global context, precisely over peatland structures that practitioners rely on for mapping and validation. Panels (a–g) show a clear progression of abstraction: early convolutions emphasize edges and micro-texture (Fig. 8-b); tokenization introduces spatial priors and yields position-aware representations (Fig. 8 c–e); and the attention modules aggregate evidence across scales—focusing locally near boundaries and coherently linking distant, semantically related regions (Fig. 8 f–g).

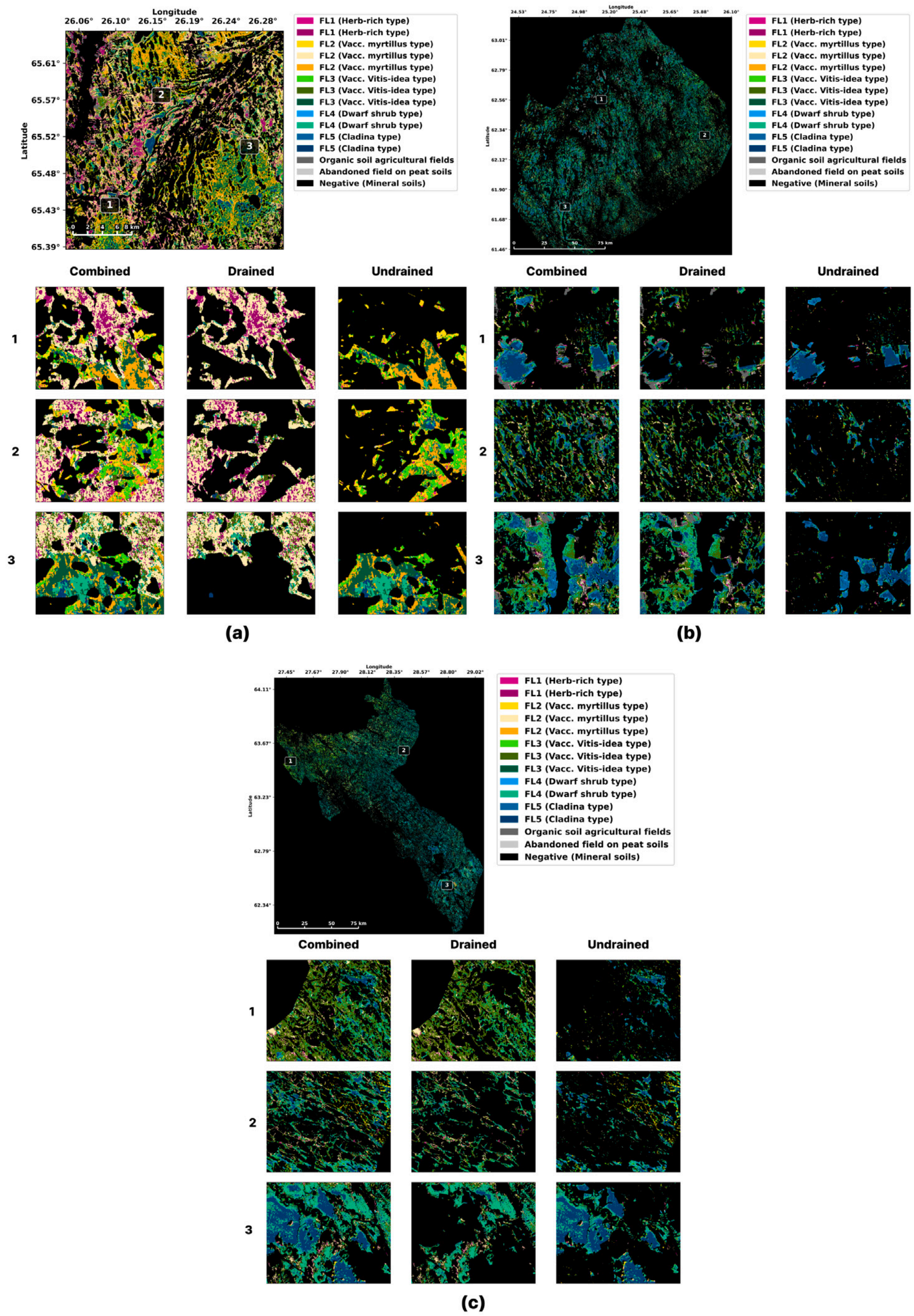

Fig. 10. Prediction map of fertility level for (a) Keminmaa, (b) Southern Ostrobothnia, and (c) Eastern Finland. Details of the prediction map for three subregions (1–3) can be compared between drained and undrained.

- The consistent performance of GeoFusion across various ecologically distinct boreal peatland regions indicates its potential for broader application in national or circumpolar peatland monitoring. However, further research should focus on addressing sensitivity to class imbalance, possibly through methods such as focal loss, class-aware sampling, or synthetic minority oversampling. Furthermore, incorporating multi-seasonal optical and SAR data could help reduce the spectral-textural ambiguity in classes that are sensitive to phenological changes.
- GeoFusion's ability to generalize across ecologically distinct sites without major performance degradation shows its cross-modal representations are transferable. This makes it a strong candidate for foundational model pretraining in boreal peatland contexts. Additionally, we plan to fine-tune models trained on Finnish peatlands to classify peatlands in other regions with minimal retraining.

Overall, our findings show that accurate peatland pixel-level classification requires combining different types of satellite RS data with models that capture both local and global features by combination of CNN and ViT.

5.2. Limitations

GeoFusion is constrained by several limitations that highlight directions for future improvement. First, the limited availability of labeled training samples, especially for undrained peatland site types, restricts the model's ability to learn distinctive spectral–textural patterns. Severe class imbalance further amplifies this issue, lowering accuracy in minority categories. Addressing these challenges will require advanced data augmentation strategies and the generation of synthetic minority samples through diffusion-based generative models (Liu et al., 2024; Sebaq and ElHelw, 2024), alongside large-scale self-supervised pretraining on unlabeled RS data to enhance feature robustness.

Second, peatland ecosystems are inherently dynamic, with seasonal vegetation cycles, hydrological fluctuations, and human interventions driving spatio-temporal variability. The current model does not explicitly capture these dynamics, and its reliance on C-band SAR (S-1) and optical imagery (S-2) introduces sensor-specific limitations: SAR backscatter struggles under dense canopies, while optical data is sensitive to cloud cover and cannot penetrate vegetation layers. Expanding GeoFusion to integrate multi-temporal observations, L-band SAR, LiDAR, or hyperspectral imagery could mitigate these constraints.

Finally, GeoFusion does not explicitly model spectral variability (e.g., endmember shifts caused by illumination or soil moisture changes), which may introduce additional uncertainty in classification. Future work could draw on spectral-variability-aware formulations such as Augmented Linear Mixing Model (ALMM) (Hong et al., 2018) to better capture intra-class spectral diversity when applied to multimodal fusion. Moreover, Transformer-based fusion introduces high computational costs for large-scale, high-resolution datasets, which may hinder operational scalability without further optimization, model pruning, or compression techniques.

6. Conclusion

Peatland ecosystems present an intricate mapping challenge due to their high spatial heterogeneity, similarity between classes, and dynamic seasonal variability. Accurately predicting both ecological site types and fertility levels at a national scale requires models capable of capturing fine-grained local features while also integrating broader spatial–contextual patterns. In this study, we introduced GeoFusion, a hybrid deep learning architecture that unites the localized feature extraction strengths of CNNs with the global contextual modeling of Transformers, enabling effective fusion of multi-source geospatial datasets, including optical and SAR imagery. This design was particularly effective for pixel-level classification of 40 site type classes and 7 fertility levels in boreal peatland, showing consistent improvements. To evaluate the effectiveness of the proposed architecture, we compared it with well-known DL and Transformer-based models: CNN, ED, Swin, and DeiT. Experiments conducted across three study areas in Finland demonstrated that GeoFusion outperforms the other models.

Despite inherent challenges such as class imbalance and the gradational nature of peatland classes, which can be captured with RS data only partially, GeoFusion achieved high accuracy across diverse ecological conditions, with notably high performance in fertility-level classification and competitive results for site-type mapping in both drained and undrained peatlands. The intermediate complexity of the model allowed it to balance representational capacity and computational efficiency, supporting scalability to large-area mapping. These results confirm that integrating optical and SAR modalities through a precisely designed fusion architecture can significantly advance operational peatland monitoring.

CRediT authorship contribution statement

Fahimeh Farahnakian: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Luca Zelioli:** Validation, Software, Methodology, Data curation. **Farshad Farahnakian:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Maarit Middleton:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Javad Sheikh:** Writing – original draft, Formal analysis. **Jukka Heikkonen:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Formal analysis.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jukka Heikkonen reports financial support was provided by Ministry of Agriculture and Forestry of Finland. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is part of the Advances in Soil Information- MaaTi project funded by the Ministry of Agriculture and Forestry of Finland (2021–2022, funding decision VN/27416/2020-MMM-2). The authors would also like to thank Timo.P. Pitkänen, from the Natural Resources Institute Finland (Luke), for his valuable support in preprocessing Sentinel data.

Data availability

The data that has been used is confidential.

References

A.K, C., 1913. Studies about finnish peatlands. Silva Fennica 2 (3).

Aleissaee, 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. Sens. 15 (7), 1860.

Aplin, P., 2004. Remote sensing: Land cover. Prog. Phys. Geogr. 28, 283–293.

Asokan, A., Anitha, J., Ciobanu, M., Gabor, A., Naaji, A., Hemanth, D.J., 2020. Image processing techniques for analysis of satellite images for historical maps classification—An overview. Appl. Sci. 10 (12), 4207.

Bourgeau-Chavez, L.L., Endres, S., Powell, R., Battaglia, M.J., Benscoter, B., Turetsky, M., Kasischke, E.S., Banda, E., 2017. Mapping boreal peatland ecosystem types from multitemporal radar and optical satellite imagery. Can. J. Forest Res. 47 (4), 545–559.

Cheng, G., Xie, X., Han, J., Guo, L., Xia, G.-S., 2020. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 13, 3735–3756.

Chughtai, A.H., Abbasi, H., Karas, I.R., 2021. A review on change detection method and accuracy assessment for land use land cover. Remote. Sens. Appl.: Soc. Environ. 22, 100482.

Dalsasso, E., Denis, L., Tupin, F., 2021. SAR2SAR: A semi-supervised despeckling algorithm for SAR images. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 14, 4321–4329.

Dalsasso, E., Denis, L., Tupin, F., 2022. As if by magic: Self-supervised training of deep despeckling networks with MERLIN. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

de Waard, F., Connolly, J., Barthelmes, A., Joosten, H., van der Linden, S., 2024. Remote sensing of peatland degradation in temperate and boreal climate zones – a review of the potentials, gaps, and challenges. Ecol. Indic. 166, 112437, URL https://www.sciencedirect.com/science/article/pii/S1470160X2400894X.

DeLancey, E.R., Simms, J.F., Mahdianpari, M., Brisco, B., Mahoney, C., Kariyeva, J., 2020. Comparing deep learning and shallow learning for large-scale wetland classification in alberta, Canada. Remote. Sens. 12 (1), URL https://www.mdpi.com/2072-4292/12/1/2.

Farahnakian, F., Zelioli, L., Middleton, M., Seppä, I., Pitkänen, T.P., Heikkonen, J., 2023a. Cnn-based boreal peatland fertility classification from sentinel-1 and sentinel-2 imagery. In: 2023 IEEE International Symposium on Robotic and Sensors Environments. ROSE, IEEE, pp. 1–7.

Farahnakian, F., Zelioli, L., Pitkänen, T., Pohjankukka, J., Middleton, M., Tuominen, S., Nevalainen, P., Heikkonen, J., 2023b. Multistream convolutional neural network fusion for pixel-wise classification of peatland. In: 2023 26th International Conference on Information Fusion. FUSION, IEEE, pp. 1–8.

Finland, S., 2022. Greenhouse gas emissions in Finland 1990 to 2020. Natl. Inven. Rep. under UNFCCC Kyoto Protoc..

Geological Survey of Finland (GTK), 2022. Suot ja turvemaat.

Ghobadi, Y., Pradhan, B., Kabiri, K., Pirasteh, S., Shafri, H.Z.M., Sayyad, G.A., 2012. Use of multi-temporal remote sensing data and GIS for wetland change monitoring and degradation. In: 2012 IEEE Colloquium on Humanities, Science and Engineering. CHUSER, pp. 103–108.
Guo, M., Li, J., Sheng, C., Xu, J., Wu, L., 2017. A review of wetland remote sensing. Sensors 17 (4), 777.
Haakana, M., Tuominen, S., Heikkinen, J., Peltoniemi, M., Lehtonen, A., 2023. Spatial patterns of biomass change across Finland in 2009–2015. ISPRS Open J. Photogramm. Remote. Sens. 8, 100036, URL https://www.sciencedirect.com/science/article/pii/S2667393223000078.
Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al., 2022. A survey on vision transformer. IEEE Trans. Pattern Anal. Mach. Intell. 45 (1), 87–110.
Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv:1503.02531. URL https://arxiv.org/abs/1503.02531.
Hong, D., Yokoya, N., Chanussot, J., Zhu, X.X., 2018. An augmented linear mixing model to address spectral variability for hyperspectral unmixing. IEEE Trans. Image Process. 28 (4), 1923–1938.
Hosseiny, B., Mahdianpari, M., Brisco, B., Mohammadimanesh, F., Salehi, B., 2022. WetNet: A spatial–temporal ensemble deep learning model for wetland classification using sentinel-1 and sentinel-2. IEEE Trans. Geosci. Remote Sens. 60, 1–14.
Intergovernmental Panel on Climate Change (IPCC), 2023. Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.
Jafarzadeh, H., Mahdianpari, M., Gill, E.W., 2022. Wet-GC: A novel multimodel graph convolutional approach for wetland classification using sentinel-1 and 2 imagery with limited training samples. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 15, 5303–5316.
Jamali, A., Mahdianpari, M., 2022. Swin transformer and deep convolutional neural networks for coastal wetland classification using sentinel-1, sentinel-2, and LiDAR data. Remote. Sens. 14 (2), 359.
Karlson, M., Bastviken, D., 2023. Multi-source mapping of peatland types using sentinel-1, sentinel-2, and terrain derivatives—A comparison between five high-latitude landscapes. J. Geophys. Res.: Biogeosciences 128 (4), e2022JG007195.
Laine, J., Vasander, H., Hotanen, J.-P., Nousiainen, H., Saarinen, M., Penttilä, T., 2018. Key to finnish mire types. p. 1602,
Li, Z., Huang, W., Wang, L., Xin, Z., Meng, Q., 2023. CNN and transformer interaction network for hyperspectral image classification. Int. J. Remote Sens. 44 (18), 5548–5573.
Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.
Liu, Y., Yue, J., Xia, S., Ghamisi, P., Xie, W., Fang, L., 2024. Diffusion models meet remote sensing: Principles, methods, and perspectives. IEEE Trans. Geosci. Remote Sens..
Mahdianpari, M., 2019. Advanced machine learning algorithms for Canadian wetland mapping using polarimetric synthetic aperture radar (PolSAR) and optical imagery (Ph.D. thesis). Memorial University of Newfoundland.
Mäkisara, K., Katila, M., Peräsaari, J., Tomppo, E., 2016. The multi-source national forest inventory of Finland – methods and results 2013. Natural resources institute Finland (luke).
Merchant, M.A., Adams, J.R., Berg, A.A., Baltzer, J.L., Quinton, W.L., Chasmer, L.E., 2016. Contributions of C-band SAR data and polarimetric decompositions to subarctic boreal peatland mapping. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 10 (4), 1467–1482.
Middleton, M., Laatikainen, M., Kivilompolo, J., Harju, A., Lerssi, J., Valkama, M., Pitkänen, T., Pohjankukka, J., Balazs, A., Tuominen, S., et al., 2023. Technical description for the peatland site type data of Finland.
Middleton, M., Närhi, P., Arkimaa, H., Hyvönen, E., Kuosmanen, V., Treitz, P., Sutinen, R., 2012. Ordination and hyperspectral remote sensing approach to classify peatland biotopes along soil moisture and fertility gradients. Remote Sens. Environ. 124, 596–609.
Millard, K., Richardson, M., 2015. On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping. Remote. Sens. 7 (7), 8489–8515.
Minasny, B., Adetsu, D.V., Aitkenhead, M., Artz, R.R., Baggaley, N., Barthelmes, A., Beucher, A., Caron, J., Conchedda, G., Connolly, J., et al., 2024. Mapping and monitoring peatland conditions from global to field scale. Biogeochemistry 167 (4), 383–425.
National Land Survey of Finland (NLS), 2023. Elevation model 2 m.
Nay, J., Burchfield, E., Gilligan, J., 2018. A machine-learning approach to forecasting remotely sensed vegetation health. Int. J. Remote Sens. 39 (6), 1800–1816.
Nungesser, M.K., 2003. Modelling microtopography in boreal peatlands: hummocks and hollows. Ecol. Model. 165 (2–3), 175–207.
O'Neil, G.L., Goodall, J.L., Behl, M., Saby, L., 2020. Deep learning using physically-informed input data for wetland identification. Environ. Model. Softw. 126, 104665, URL https://www.sciencedirect.com/science/article/pii/S1364815219311417.
Pan, Z., Xiang, H., Shi, X., Wang, M., Song, K., Mao, D., Huang, C., 2025. Detecting and mapping peatlands in the Tibetan Plateau Region using the random forest algorithm and sentinel imagery.. Remote. Sens. 17 (2).
Perera, M.V., Bandara, W.G.C., Valanarasu, J.M.J., Patel, V.M., 2022. Transformer-based SAR image despeckling. In: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 751–754.
Phiri, D., Morgenroth, J., 2017. Developments in landsat land cover classification methods: A review. Remote. Sens. 9 (9), 967.
Pouliot, D., Latifovic, R., Pasher, J., Duffe, J., 2019. Assessment of convolution neural networks for wetland mapping with landsat in the central Canadian Boreal Forest Region. Remote. Sens. 11 (7), URL https://www.mdpi.com/2072-4292/11/7/772.
Rezaee, M., Mahdianpari, M., Zhang, Y., Salehi, B., 2018a. Deep convolutional neural network for complex wetland classification using optical remote sensing imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 11 (9), 3030–3039.
Rezaee, M., Mahdianpari, M., Zhang, Y., Salehi, B., 2018b. Deep convolutional neural network for complex wetland classification using optical remote sensing imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 11 (9), 3030–3039.
Sebaq, A., ElHelw, M., 2024. Rsdiff: Remote sensing image generation from text using diffusion model. Neural Comput. Appl. 36 (36), 23103–23111.
Shailaja, P., Kumar, P.M., Nikhitha, N., Reddy, K.N.K., Reddy, E.M., Reddy, G.G., Indu, V., 2025. LCC-net: Swin transformer-CNN hybrid for enhanced land cover classification in natural disaster monitoring. Syst. Soft Comput. 200303.
Sowmya, D., Shenoy, P.D., Venugopal, K., 2017. Remote sensing satellite image processing techniques for image classification: a comprehensive survey. Int. J. Comput. Appl. 161 (11), 24–37.
Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. PMLR, pp. 10347–10357.
Turunen, J., Valpola, S., 2020. The influence of anthropogenic land use on finnish peatland area and carbon stores 1950–2015..
Vaswani, A., 2017. Attention is all you need. arXiv preprint arXiv:1706.03762.
Yan, C., Fan, X., Fan, J., Yu, L., Wang, N., Chen, L., Li, X., 2023. Hyformer: Hybrid transformer and cnn for pixel-level multispectral image land cover classification. Int. J. Environ. Res. Public Health 20 (4), 3059.
Yigit Avdan, Z., Kaplan, G., Goncu, S., Avdan, U., 2019. Monitoring the water quality of small water bodies using high-resolution remote sensing data. ISPRS Int. J. Geo-Information 8 (12), 553.
Zelioli, L., Farahnakian, F., Farahnakian, F., Middleton, M., Heikkonen, J., 2024. Enhancing peatland classification using sentinel-1 and sentinel-2 fusion with encoder-decoder architecture. In: 2024 27th International Conference on Information Fusion (FUSION). pp. 1–7.
Zelioli, L., Farahnakian, F., Middleton, M., Pitkänen, T.P., Tuominen, S., Nevalainen, P., Pohjankukka, J., Heikkonen, J., 2025. Peatland pixel-level classification via multispectral, multiresolution and multisensor data using convolutional neural network. Ecol. Informatics 90, 103233, URL https://www.sciencedirect.com/science/article/pii/S1574954125002420.
Zhou, N., Xu, M., Shen, B., Hou, K., Liu, S., Sheng, H., Liu, Y., Wan, J., 2024. ViT-UNet: A vision transformer based unet model for coastal wetland classification based on high spatial resolution imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens..