# INNOVATIVE APPLICATION OF MULTI-MODAL DATA FUSION TECHNOLOGY IN THE EVALUATION OF REGIONAL LANDSCAPE STYLE AND FEATURES

HU, F. T.[1*] – CHEN, L.[1] – REN, W. J.[2] – WANG, Y. S.[3] – YU, Z. H.[4]

[1]*Department of Life Sciences, Yuncheng University, Yuncheng 044000, Shanxi, China*
*(ORCID: 0009-0008-1478-3647 – Hu, F. T., 0009-0003-2400-1796 – Chen, L.)*

[2]*Department of Arts and Craft Design, Yuncheng University, Yuncheng 044000, Shanxi, China*
*(ORCID: 0009-0007-3668-0289 – Ren, W. J.)*

[3]*Archaeological Exploration Business Department, Shanxi Provincial Cultural Heritage Survey and Protection Research Institute, Taiyuan 030024, Shanxi, China*
*(ORCID: 0009-0004-0366-4643 – Wang, Y. S.)*

[4]*Technology and Quality Department, Shaanxi Ancient Architecture and Landscape Construction Group Co., Ltd, Xi'an 710000, Shaanxi, China*
*(ORCID: 0009-0005-3279-5375 – Yu, Z. H.)*

**Corresponding author*
*e-mail: hufangtong523@outlook.com*

(Received 16th Jul 2025; accepted 15th Oct 2025)

**Abstract.** The development of environmental change research requires an accurate regional landscape classification process to improve urban planning, ecological assessment, and disaster management. The existing classifiers face interpretability issues while exploring contextual and spatial information. The research difficulties are overcome by applying the Hybrid Convolution Neural Networks (Hybrid CNN) framework and the multi-modal fusion strategies. During the analysis, satellite, street-level imagery, geo-tagged, and textual metadata information are utilized and processed by normalization and data augmentation techniques. After that, feature embedding, attention mechanism, and convolutional operations are incorporated to ensure interpretability and learning efficiency. The noise-removed information is processed by applying the multi-encoder branches that extract the various feature maps fed into the feature pyramid networks. The pyramid networks compute the attention score for every feature, and learnable weights are derived to calculate the landscape classification. Finally, Monte Carlo Dropout is applied to compute the uncertainty, which helps identify the landscape region at risk with 99.3% accuracy. The discussed system was evaluated using benchmark datasets, in which the hybrid CNN approach ensures high reliability, interpretability, and accuracy compared to conventional methods. The developed framework is effectively applied in the smart city infrastructure, remote sensing, climate resilience planning, agricultural monitoring, and data-driven environmental analysis.
**Keywords:** *regional landscape, feature pyramid, Monte Carlo dropout, hybrid convolution networks, agricultural monitoring*

## Introduction

Regional landscape evaluation (Nedd et al., 2021) defines the geographical context of a landscape, including its physical, visual, ecological, and cultural characteristics, to comprehend its essence, patterns, and processes by breaking it down into its constituent parts to document its multifaceted complexities (Tu et al., 2024). It supports many fields, including urban and rural planning, environmental studies, landscape architecture, tourism, heritage conservation, etc. (Li et al., 2021). The evaluation is used to make policies for land use, infrastructure planning, environmental risk management, and care


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

in protection by determining the ecosystem and prevailing landscape styles and attributes. In particular, the term "care in protection" refers to the intentional actions and plans implemented to conserve delicate habitats, halt ecological degradation, and preserve the integrity of critical natural regions, thereby ensuring their long-term viability for present and future generations. The evaluation includes natural components like relief, flora, water bodies, and anthropogenic elements such as buildings, roads, and subdivision patterns (Zhang et al., 2023). Driven by the effects of urban expansion, climate change, and global culture convergence over the last few decades, there is an urgent to focus on enhancing a region's visual richness and identity beyond merely documenting its environmental attributes. An effective appraisal of a regional landscape serves to preserve the biological balance of a region, enhance the aesthetics of an area, and regulate anthropogenic developments in a region (Di Zhang and Liu, 2023).

The evaluation of regional landscapes rationally motivates sustainable development without compromising a region's environmental and cultural identity (Lagodiienko et al., 2022). Additionally, any specific area faces urban sprawl, climate change, resource depletion, and increased environmental destruction. Further, it is equally essential to maintain the order and design of space as a balance between development, ecology, and beauty. Proper evaluation supports decisions for policies concerning green area distribution, zoning, conservation, and setting priorities for tourism development to align growth with ecological preservation and communal interests (Fu et al., 2024). The whole process does pose several problems. The most significant concern is relative landscape heterogeneity, which encompasses tangible features such as landform, vegetation, and abstract qualities, such as visual harmony or culture. The main problem is multidimensional issues, which affect the landscape analysis efficiency because expert opinion is required to process extensive data (Stupariu et al., 2022). Other problematic issues include inconsistency owing to different regional characteristics, scale, area of study, availability of data, and low comparability index across studies (Zhang et al., 2021). Manual data evaluation also upsurges time and effort with high subjectivity, while automated techniques relying on a single data source, e.g., satellite images, do not capture the complete essence of the expanse. These boundaries underscore the demand for sophisticated and highly scalable context-sensitive techniques that seamlessly integrate multiple data types to accomplish more comprehensive and dependable evaluations at a landscape level (Kush, 2025; Rajaram, 2024). The reasoning behind the method that was presented is based on resolving the issues that are associated with deploying multi-modal data fusion technologies in conjunction with Hybrid-CNNs, which together give a system that can intelligently evaluate regional landscapes and is also scalable. Rather than relying on a single type of input, as is common in traditional approaches, this method makes use of a wide range of complementary inputs, including but not limited to satellite images, street photos, topographic maps, and geo-tagged texts and social media data. This allows for a more accurate and holistic representation of the landscape. The hybrid convolutional neural networks (CNNs) are specifically designed to handle stimuli that are multi-modal, which allows the model to simultaneously capture characteristics across many perspectives and geographical scales. This multidisciplinary approach augments contextual and structural richness and enhances analysis depth so the system can discern subtle stylistic, spatial and environmental features. Additionally, the level of automation and data-centricity in the model mitigates biases associated with subjective judgment while improving consistency, enabling analysis on varying geographic areas without restriction. Therefore, while escalating landscape assessments' accuracy, interpretability,


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

and adaptability as controlled evaluations and manual interventions explain context-specific conditions, tailored decision support becomes facilitated for planners, architects, and environmental management professionals.

As the scope and variety of data information increase, multi-modal data fusion has become an essential method in spatial analysis, urban planning, and tourism analytics. Researchers have increasingly started to work with remote sensing images, textual documents, sensor data, images, and socio-demographic data to enhance the interpretation, predictions, and decisions in spatially relevant domains. This literature survey highlights three major studies showcasing different uses of multi-modal data fusion: design of a tourism system, classification of functional urban zones, and urban land use mapping. Here, a few research works are discussed to understand the regional landscape styles and features to classify and explore the difficulties in feature classification.

In the work of Khan et al. (2024), the authors performed an extensive survey on multi-modal data fusion methods within tourism by studying datasets, some fusion methods, and even neural schemas. The work also developed attention-based models with late fusion methods to improve personalization and described the growing shift towards these models. It also neglects real-world evaluation and focuses more on retrospective appraisal, devoid of real-time execution considerations. Liu et al. (2024) proposed a comprehensive multi-modal model for classifying urban functional zones using satellite images, POI data, human mobility, and social media data inputs. The deep learning with natural language processing (DL-NLP) model incorporates an attention, spatial, and temporal layer, improving classification accuracy while adding interpretability. The model performed excellently capturing various features of urban areas and outshone other methods. Nonetheless, it expended extreme computational resources and struggled to transfer to other cities. Yan et al. (2024) designed a multi-modal fusion model for mapping urban land use satellite through images, LiDAR data, GIS layers, alongside policy texts. The author uses a probabilistic neural network with Bayesian inference (PNN-BI) capable of providing accurate classifications and estimating uncertainty, excelling in interpretability. The model used these layers, achieving strong performance in dense urban environments. However, the most significant weaknesses were its expensive computational requirements and the structured policy text data it relied on.

Zhao et al. (2021) analyzed urban similarity and uniqueness by employing deep style learning methodologies (DSL) on city images at scale. The study used convolutional neural networks to capture style features of urban scenes and performed intercity comparative analysis. These findings stylistically identified cityscapes and measured visual similarities. However, model interpretation depended on image accessibility and resolution; therefore, accuracy was restrained in sparse regions. Zeng et al. (2022) explored the preferences and emotional responses towards forested landscapes of Chinese recreationists analyzing geo-tagged images through deep learning algorithms. The study applied convolutional neural networks to fetch the visual attributes and infer associated emotional overtones of the landscapes. Results showed marked affinity for natural features of forests and expressions of emotion associated with them. The main drawback is the predisposition to unbalanced photo distribution and the absence of subjective feedback verification.

Giang et al. (2023) developed a coastal landscape classification framework based on convolutional neural networks (CNNs) for remote sensing data in Vietnam. The model successfully recognized different coastal categories, including beaches, mangroves, and


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

urban shorelines, which aid in managing coastal zone resources. Its strength is processing satellite images for spatial analysis. Performance, however, suffered in cloud-covered areas and locations with limited training dataset diversity. Zerouali et al. (2024) developed a new slope stability classification approach based on deep learning, incorporating a metaheuristic feature selection scheme. Classification accuracy and computational efficiency improved through the synergetic use of CNNs and optimization techniques. This combination alleviated excessive model fitting and increased feature selection relevance. The model, however, was burdened with high complexity, resulting in lengthy training periods. Furthermore, it was critical to have a strictly defined input dataset in order to provide good results. A system for landscape design in an urban virtual environment was developed by Liu et al. (2024) utilizing the PSO-BP (Particle Swarm Optimization-Back Propagation) neural network model. The system refined landscape arrangements within three-dimensional virtual simulations, enhancing the aesthetic and environmental value of the layout. Results showed effective convergence and adaptive design generation. Because the simulations relied on specified characteristics and conditions, the model was not very adaptable to varied urban environments, which was the biggest problem. The Particle Swarm Optimization (PSO) methods were investigated by Chen (2024) in the context of landscape architecture, including its planning and layout design aspects. It was proved by the research that particle swarm optimization (PSO) may be effectively utilized for the purpose of increasing both flexibility and aesthetic spatial arrangement in adaptive landscape designs. It proved to be exceptionally efficient when it came to optimizing layout criteria that involved many objectives. The approach was, however, detached from any real-time geo-information and user interaction, which is a drawback for responsiveness in dynamic design environments. The literature review discloses that the application of deep learning and multi-modal data fusion methods is on the rise in landscape analysis, urban planning, and environmental assessment. Previous studies use satellite images, street-level pictures, and geospatial information for urban zoning, landscape style recognition, and preference analysis. Still, most frameworks depend on single-modality inputs, offer low regional adaptability, or non-scalable interpretability. These gaps underscore the challenges of developing a multi-source, multi-model paradigm that enhances accuracy and sensitivity in context-aware landscape assessment.

Satellite imagery, LiDAR scans, textual documents, POI datasets, and social media content are difficult to integrate. Due to resolution, update frequency, and coverage differences across modalities, spatial-temporal misalignment hinders fusion consistency and synchronization. Raster vs. vector, organized vs. unstructured data structures, complicate preprocessing workflows and need modality-specific feature extraction. A semantic mismatch between data sources makes interpretation difficult; textual data may convey subjective perceptions that are not apparent in visuals. Aligning heterogeneous inputs is a process that is highly demanding in terms of processing power, which limits the scalability and cross-regional adaptability of the process, according to the findings of Liu et al. (2024) and Yan et al. (2024). The dependability of real-world implementations may be compromised by factors such as sparse data, cloud cover, and low image quality. These challenges highlight the fact that there is a demand for lightweight, adaptive models that are capable of integrating numerous data types while simultaneously maintaining spatial and contextual relevance.

CerviFusionNet, a multi-modal hybrid CNN-transformer architecture, was created by Sha et al. (2024) for robust representation learning. Their hybrid method blends


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

convolutional neural networks with transformer-derived attention mechanisms to excel at multi-modal data fusion. A complete analysis of deep learning-based multi-modal medical picture fusion was published (Bhosekar et al., 2025). They emphasize hybrid architectural advancements, which combine convolutional neural networks (CNNs) with other neural techniques to improve feature extraction and fusion. Both studies emphasize the importance of using encoders with numerous branches and methods that focus on attention to increase model robustness and accuracy, supporting this research's strategy.

The work's objective follows.

- Enhance regional landscape appraisal by integrating satellite photographs, topological data, geo-tagged texts, and street-level visuals.
- Develop a Hybrid-CNN for multi-scale landscape inputs, enhancing classification and interpretation of spatial data and visual styles.
- Evaluate the model's scalability, contextual accuracy, and reliability for environmental monitoring, urban planning, and regional identity preservation.

The Hybrid-CNN integrates and processes multi-modal, multi-scale landscape data, including satellite imagery, topological data, geo-tagged textual information, and street-level visuals, which are rarely combined in a single deep learning framework. The Proposed design employs parallel convolutional pipelines tailored for diverse data resolutions and modalities, enabling the capture of both macro-level spatial patterns and micro-level contextual signals. This approach differs from standard CNNs, which focus on single-source image data. This architectural innovation enhances categorization accuracy, contextual knowledge, scalability, and flexibility across varied urban and biological contexts. We also added cross-modal fusion layers and attention-based feature alignment modules to improve heterogeneous data integration and the model's context-aware, semantically rich interpretations.

## Materials and methods

The main goal is to create an intelligent and scalable system for assessing regional landscape styles and attributes using digitally multi-mode data fusion and hybrid deep learning. The designs use multiple advanced ways that complement each other to defeat traditional and unimodal methods. Multi-modal data gathering and pre-processing consolidates and standardizes satellite pictures, street-level photographs, DEMs, and geolocated textual content during analysis. Hybrid-CNN with attention mechanisms and Feature Pyramid Networks (FPNs) extract and merge low and high-level features of multiple types and spatial scales across and within different modalities. Crossmodal feature fusion layers improve contextual knowledge and categorization by interrelating modalities. Others improved categorization and interpretation by incorporating spatial attention modules, uncertainty estimation approaches, and Monte Carlo Dropout during inference. Urban planning, ecological monitoring, and regional development strategies benefit from these tools' reliable classification and extraction of subtle stylistic information in landscape characteristics. Then, the overall working process of Hybrid CNN-based regional landscape feature classification is shown in *Figure 1*.

*Figure 1* shows the Hybrid CNN architecture-based landscape feature categorization strategy. Street view photos, topographical data, geo-tagged data, and satellite images are processed through a uniform pipeline. Data is processed using neural functions including attention-based fusion and feature pyramid network-based feature extraction.


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

Convolutional networks classify the landscape and explore extracted characteristics using uncertainty computation for reliability and scalability. The next section describes Hybrid CNN-based landscape feature classification in depth.

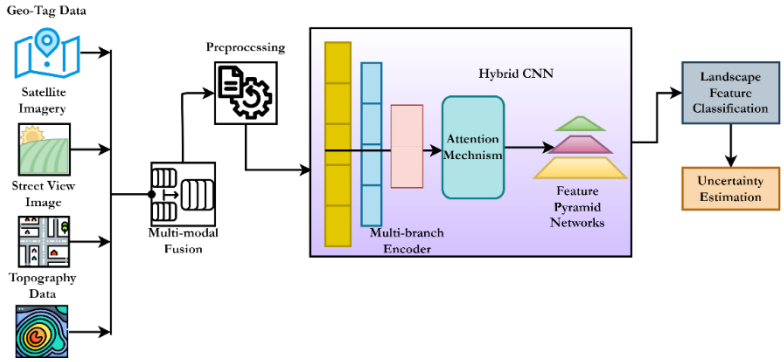



***Figure 1.*** *Hybrid CNN architecture for landscape feature classification*

A CNN-based feature extractor collects domain-specific representations for each input modality first. Satellite imagery provides spatial patterns, street-level images provide context, DEMs provide structural topology, and textual material provides semantic signals. An attention mechanism is applied to each modality once features are retrieved to highlight the most relevant characteristics and reduce noise. After that, projection layers align these attention-weighted features in a shared latent space. This assures dimensionality and scale compatibility. After aligning, features are concatenated and augmented via weighted fusion. In this strategy, the learned attention weights ($\alpha i$) are used to determine the contribution of each modality to the final joint representation. In the subsequent step, classification and decision-making are carried out using this fused feature vector. To enhance the clarity of the fusion process, the new section incorporates both informative text and an explanatory diagram simultaneously. We anticipate that this update will provide a more transparent view of how different forms of heterogeneous data are harmonized to improve classification performance and contextual awareness.

The architecture uses a multi-branch encoder to handle satellite imagery, street-level images, DEM slopes, and text embeddings. Each branch uses parameter sharing and dimensionality reduction to reduce computational cost. Instead of randomly increasing model load, the attention mechanism emphasizes modality-specific properties to reduce computation and improve interpretability. To boost scalability, we use depth-wise separable convolutions in hybrid CNN blocks to reduce parameter counts without losing performance. A lightweight setup prunes the Feature Pyramid Network during inference, keeping just critical resolution layers. In real-world applications, modular input streams provide flexible deployment with limited modalities when data sources are unavailable. These solutions balance performance and computational efficiency, making the model ideal for high-end and resource-constrained environments.

### *Data preparation and alignment*

The proposed framework begins by systematically collecting multimodal data from high-resolution satellite imagery and geospatial data. To support these objectives, satellite imagery from Sentinel provides great spatial coverage to capture vegetation, urban


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

development features, water bodies, and land uses in the region of interest along with raw street-level imagery collected from geo-tagged street view cameras and mobile applications. The topographic data obtainable from DEMs and other GIS datasets adds value in elevation, slope, and the extent of terrain, which describes the landscape's structural and physical composition. The cited sources can be effectively applied as building blocks in creating a comprehensive landscape evaluation system using multimodal data. After the data collection phase, the system must undergo a meticulous pre-processing and alignment step to integrate different input data types effectively. Moreover, this step is crucial for noise reduction, correcting inconsistencies, and preparing data for deep learning-based feature extraction and classification. The work utilizes various types of information, as it gathers data from different sources with multiple formats, measurement scales, and resolutions, which significantly affects landscape classification efficiency. Therefore, the normalization process is required to simplify the complex process. For image analysis, street-level and satellite image pixels are normalized to a consistent range (1, -1 or 0,1), which helps to learn the features in the CNN architecture. Suppose the system receives the text or elevation data that are processed by the min-max normalization technique to convert the data into a standardized format. The normalization process minimizes the data modality due to the rise of larger numeric magnitudes. This study chose landscape categories using remote sensing categorization concepts that integrate spectral, spatial, and semantic context. We used land cover categorization techniques to classify landscapes into urban, plant, water, soil, and constructed settings. Multi-modal data characteristics including satellite imagery, street-level pictures, DEM features, and textual information were used to strongly define landscape types. Widely used in remote sensing literature, supervised classification algorithms use spectral signatures and domain knowledge to assure ecological and urban relevance.

A whole regional landscape picture is created using multimodal datasets. Vegetation, urban infrastructure, water bodies, and land use patterns are captured by Sentinel high-resolution satellite photography. Along with aerial views, geo-tagged street view cameras and mobile apps provide ground-level views. The ground-level photographs were taken using the Mapillary mobile app, a popular crowdsourcing tool that lets users take and publish GPS-tagged street-level photos. To avoid spatial and temporal overlap, ground-level and satellite photos were filtered based on GPS coordinates that matched those of Sentinel-2 satellite imagery tiles. For consistency in landscape conditions, the Mapillary collection includes photos from 2022–2023, overlapping with Sentinel-2 acquisition. Multi-modal fusion and landscape assessment are reliable with this alignment. DEMs and GIS-based slope and terrain maps show surface gradients and elevation differences. Visual, spatial, and textual information are presented in the article to illustrate their responsibilities and the integration of architecture. To ensure cross-modal interoperability, preprocessing normalizes and aligns data types. DEM and textual data are min-max-normalized, while satellite and street-level imagery is pixel-normalized to [0,1] or [-1,1]. That decreases numerical scale difference. These pretreatment steps allow the CNN-based feature extraction approach to provide visual and spatial representations (seen in the images) that match the processed, normalized, and fused classification pipeline input. Therefore, multi-domain normalization is performed while exploring the raster and image data; the normalization is explained in *Equation 1*.

$$\mathbf{X}_{norm}(i,j,k) = \frac{\mathbf{X}(i,j,k) - \min(\mathbf{X})}{\max(\mathbf{X}) - \min(\mathbf{X}) + \epsilon} \quad \text{(Eq.1a)}$$


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

$$\mathbf{X}_{norm}(i,j,k) = \frac{\mathbf{X}'(i,j,k)}{\max\limits_{p,q,r}\mathbf{X}(p,q,r) - \min\limits_{p,q,r}\mathbf{X}(p,q,r) + \epsilon} \quad \text{(Eq.1b)}$$

$$\mathbf{X}''(i,j,k) = \frac{\mathbf{X}'(i,j,k)}{\max\limits_{p,q,r}\mathbf{X}(p,q,r) - \min\limits_{p,q,r}\mathbf{X}(p,q,r)} \quad \text{(Eq.1c)}$$

$$\mathbf{X}'(i,j,k) = \mathbf{X}(i,j,k) - \min_{p,q,r}\mathbf{X}(p,q,r) \quad \text{(Eq.1d)}$$

*Equation 1a* computes min-max normalization, scaling input values at positions $(i, j, k)$ (i,j,k) to the range [0,1]. By adding a minor constant $\epsilon$, division by zero is avoided. This ensures uniform feature scaling, which speeds up model convergence and training stability. In *Equation 1b* is a refined normalization X(p,q,r) that employs a pre-shifted value $X'$ and reuses global extrema over the 3D input $X$ $(p, q, r)$. Multi-channel picture data or volumetric inputs require consistent normalization across the input volume. In *Equation 1c*, eliminating the epsilon when the range is non-zero simplifies normalization in this version. Standardizing input values for neural network layers that prefer zero-centered or fixed-range input improves gradient flow and numerical stability. Finally, *Equation 1d* subtracts the global minimum from each input value, resetting the input domain to zero. This preprocessing step facilitates the later division operation in normalization and prevents negative input values from affecting ReLU-based networks.

In normalization, $i$, $j$, $k$ represent input data spatial and channel dimensions, whereas $p$, $c$, $r$ p, q, r represent index ranges for global minimum and maximum values. A tiny constant $\epsilon$ is added to the denominator to prevent division by zero and maintain numerical stability. To calculate X ′ (i,j,k), remove the global minimum value from each input element according to *Equation 1d*. This step set the data minimum to zero. *Equation 1b* or *1c* normalizes shifted values using global range (maximum minus minimum), with *Equation 1b* incorporating $\epsilon$ for stability. When decomposition is not needed, *Equation 1a* simplifies one-step normalization from input values. These changes increase normalization procedure clarity and consistency in the manuscript.

*Equation 1* is used to perform the multi-domain normalization on the collected data, where $\mathbf{X}(i,j,k)$ is represented as the pixel location at $(i,j)$ and channel k. During the normalization, $\mathbf{X}'(i,j,k)$ minimum subtraction is performed initially to shift the data into the new minimum. Then the shift data is processed continuously to convert the data into a particular range ($\mathbf{X}''(i,j,k)$ with interval [0,1]. After that numerical stability is adjusted ($\mathbf{X}_{norm}(i,j,k)$to finalize the normalization process, which minimizes the computation complexities. After processing the numerical data, the geospatial adjustment aligns the multiple datasets into the standard coordinate reference systems (CRS). The CRS function is represented as $P' = T(P) = R \cdot P + t$ in which, original point ( $P = (x, y, z)$ ) is processed with the rotational matrix $(R)$ and translation vector $(t)$ to get the target CRS transfer point $(P' = (x', y', z'))$. Then bilinear interpolation resampling is applied to perform the raster alignment which is defined in *Equation 2.*

$$\left.\begin{aligned} I'(x,y) &= \sum_{i=0}^{1}\sum_{j=0}^{1} w_{ij}\cdot I(x_i, y_j) \\ w_{ij} &= (1-|\,x-x_i\,|)(1-|\,y-y_j\,|) \end{aligned}\right\} \quad \text{(Eq.2)}$$


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

In *Equation 2*, $I'(x, y)$ is represented as the resample pixel value, $I(x_i, y_j)$ is defined as the neighboring pixel value, a weight value that helps to align the pixel in overlapping image layers. Then, the data augmentation is applied to image data to strengthen the model's reliability and reduce overfitting, though within the confines of limited labeled data. The training dataset can be expanded artificially using random rotations and flips (horizontal or vertical) and cropping, scaling, adjusting the contrast, and modifying the brightness. Such transformations enable the model to gain more general and invariant feature recognition capabilities, aiding its functioning regardless of the styling and condition of landscapes. Depending on the modality, augmentation can also be tailored, for example, in elevation models, random noise injection, geometric transformations, or for text data, synonym replacements, and paraphrasing sentences.

The function of data augmentation in improving model generalization across geographies and environments. We employed modality-specific augmentation algorithms to enhance robustness and mitigate overfitting resulting from the heterogeneity in landscape elements, lighting conditions, camera angles, and environmental textures across different sites. To imitate real-world visual fluctuations, satellite and street-level imagery were augmented with random cropping, rotation, flipping, brightness modification, and Gaussian noise injection. This makes the model stable to changes in spatial orientation and lighting. We employed synonym replacement and random word dropout for textual input to preserve semantic meaning and instruct the model to recognize more general patterns. These augmentations increase the model's data distribution during training, making it more adaptable to unseen regions with varying structural, environmental, or textual features. The proposed approach achieves greater cross-domain generalization and stable performance in geographically diverse or previously unknown situations. We have enhanced our explanation in the amended discussion section to emphasize the role of our augmentation strategy in regional and environmental scalability.

Then, the sequence of data augmentation steps involved in this process is described in *Equation 3*.

$$\begin{cases} \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} //rotation \\ x' = W - x - 1 // \, fliping \; (horizontal) \\ I_s(x, y) = I\left(\frac{x}{s}, \frac{y}{s}\right) //scaling \; (zooming) \end{cases} \quad \text{(Eq.3)}$$

In *Equation 3*, the original $((x, y))$ coordinates are rotated at angle $\theta$ to get the new coordinates $(x', y'))$. After rotating the images, the flipping is performed with respect to the image $(W)$ width and the original pixel horizontal coordinates $(x)$. Finally, the scaling factor $(s)$ is used to obtain the scaled image $(I_s)$. In addition, the textual data is processed using embedding techniques like term frequency and inverse document frequency to remove unwanted data. Consider the dataset $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ in which the term t in d is estimated as $\text{TF-IDF}(t, d, \mathcal{D}) = \text{tf}(t, d) \cdot \log\left(\frac{|\mathcal{D}|}{1+|\{d_i \in \mathcal{D}: t \in d_i\}|}\right)$. According to the computed TF-IDF value, the unwanted terms are removed from the dataset, which helps to improve the landscape region exploration. The above basic pre-processing and alignment process ensures spatial consistency, high-quality, and well-developed inputs for identifying landscape features using the Hybrid CNN approach. Based on the above discussions, the obtained sample output illustration is shown in *Table 1* for different sample data.


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

***Table 1.*** *Sample output illustration for data pre-processing and alignment*

| $Reg_{id}$ | $Sat_{image}$ (norm) | $Street_{image}$ (norm) | DEM slope (%) | $Text_{emb}$ | Geo-alignment status | $Res_{image}$ (px) | $data_{aug}$ |
|---|---|---|---|---|---|---|---|
| R1 | [0, 1] | [0.45, 0.48, 0.50] | 10.240% | 300 | Aligned | 256 × 256 | 60% |
| R2 | [0, 1] | [0.51, 0.53, 0.55] | 14.510% | 300 | Aligned | 256 × 256 | 50% |
| R3 | [0, 1] | [0.43, 0.46, 0.49] | 13.820% | 300 | Aligned | 256 × 256 | 70% |
| R4 | [0, 1] | [0.47, 0.48, 0.50] | 20.130% | 300 | Aligned | 256 × 256 | 40% |
| R5 | [0, 1] | [0.44, 0.46, 0.47] | 26.50% | 300 | Aligned | 256 × 256 | 80% |
| R6 | [0, 1] | [0.50, 0.51, 0.53] | 21.320% | 300 | Aligned | 256 × 256 | 60% |
| R7 | [0, 1] | [0.41, 0.44, 0.47] | 26.00% | 300 | Aligned | 256 × 256 | 65% |
| R8 | [0, 1] | [0.48, 0.50, 0.52] | 11.70% | 300 | Aligned | 256 × 256 | 50% |
| R9 | [0, 1] | [0.42, 0.45, 0.47] | 16.30% | 300 | Aligned | 256 × 256 | 55% |
| R10 | [0, 1] | [0.49, 0.51, 0.53] | 22.20% | 300 | Aligned | 256 × 256 | 45% |
| R11 | [0, 1] | [0.46, 0.47, 0.50] | 15.50% | 300 | Aligned | 256 × 256 | 40% |
| R12 | [0, 1] | [0.43, 0.45, 0.48] | 14.90% | 300 | Aligned | 256 × 256 | 60% |
| R13 | [0, 1] | [0.47, 0.48, 0.51] | 18.60% | 300 | Aligned | 256 × 256 | 70% |
| R14 | [0, 1] | [0.52, 0.54, 0.57] | 28.80% | 300 | Aligned | 256 × 256 | 55% |
| R15 | [0, 1] | [0.45, 0.47, 0.49] | 10.40% | 300 | Aligned | 256 × 256 | 50% |
| R16 | [0, 1] | [0.44, 0.46, 0.48] | 12.20% | 300 | Aligned | 256 × 256 | 65% |
| R17 | [0, 1] | [0.49, 0.51, 0.53] | 27.10% | 300 | Aligned | 256 × 256 | 70% |
| R18 | [0, 1] | [0.45, 0.47, 0.49] | 13.00% | 300 | Aligned | 256 × 256 | 60% |
| R19 | [0, 1] | [0.43, 0.46, 0.48] | 29.30% | 300 | Aligned | 256 × 256 | 75% |
| R20 | [0, 1] | [0.50, 0.52, 0.55] | 15.80% | 300 | Aligned | 256 × 256 | 50% |

The slope gradient from DEM data indicates the steepness of terrain in each region as a percentage, denoted as "DEM Slope (%)". Topographic variation affects landscape characteristics, hence this aspect is crucial. "Data_aug" is the percentage of data augmentation applied to each image instance during preprocessing. Rotation, flipping, and scaling are employed, with the percentage indicating the augmentation intensity to enhance model generalization. Both concepts are now more clearly described in the table caption and data preprocessing section to enhance the interpretation of results.

The pre-processing results (*Table 1*) indicate effective normalization of satellite and street-level images for all 20 regions, including pixel and RGB value scaling for model count alignment. The DEM data produced slope percentage values ranging from 10% to almost 30%, demonstrating diversity in terrain for classification purposes, which is obtained using $\text{Slope} = \arctan.\left(\sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}\right) \times \frac{180}{\pi}$. Then, the text embeddings constructed using Word2Vec kept a constant 300-dimensional structure, which reflected the captured regional spaces, hence human perception. Data augmentation was performed within 40% to 80% to maintain flexibility. Finally, the geospatial alignment was a complete success, and multi-modal inputs were integrated with precision, confirming alignment accuracy.

During preprocessing, initial normalization standardizes raw input data like pixel intensities and textual embeddings. After feature extraction and before dimensional alignment, secondary normalizing is done. This supplementary normalization ensures that visual, topological, and textual elements are comparable before fusion. Normalization at this stage prevents any modality from excessively influencing the fused representation since each CNN sub-network produces outputs with different magnitudes and distributions depending on input type and feature complexity. This


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

phase increases attention mechanism and final classification layer stability and convergence. We modified the technique section to distinguish between initial input normalization and feature-level secondary normalization and their roles in the multi-modal fusion process.

### *Hybrid CNN-based feature extraction*

The next important step is feature extraction from the multi-modal input data like digital elevation, street-level images, satellite images, and textual data by applying the Hybrid CNN. The hybrid network consists of several specialized branches, which are used to derive the salient features that are more helpful in processing the particular data modality. The feature extraction process consists of several stages, such as multi-encoder branches, attention mechanism, and pyramid networks, which help to perform multiple data analyses (*Fig. 2*).

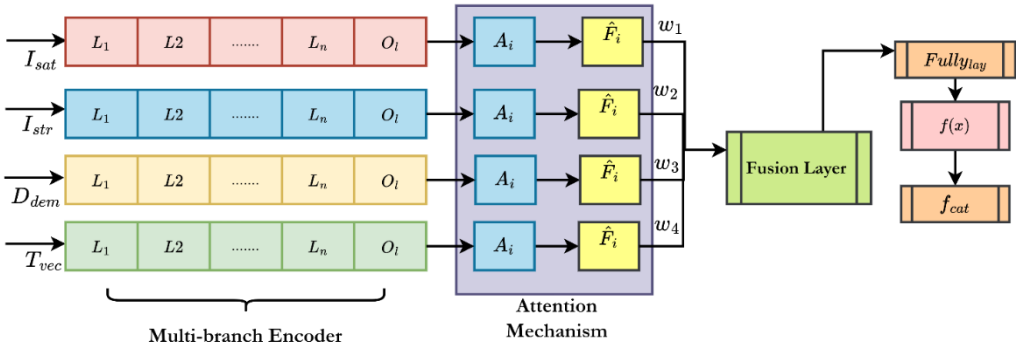



***Figure 2.*** *Process of hybrid CNN-based feature extraction*

During the study, multi-encoder branches were applied to extract the domain-specific features to eliminate the data modality loss while doing fusion. In addition, each regional input has various features like elevation contours, gradient (DEM), descriptive features, social sentiments (textual inputs), and architectural elements (street images). As said, the model uses the different types of inputs $I_{sat} \in \mathbb{R}^{H\times W\times 3}, I_{str} \in \mathbb{R}^{H\times W\times 3}, D_{dem} \in \mathbb{R}^{H\times W\times 1} and\ T_{vec} \in \mathbb{R}^{L\times d}$ which are more helpful in processing the regional landscape features. The $I_{str} \in \mathbb{R}^{H\times W\times 3}$ inputs cover the perspective and close-range images from crowdsourcing or street view that cover the road texture, architecture, and public elements. Then, $I_{sat} \in \mathbb{R}^{H\times W\times 3}$ consists of satellite images (RGB color) giving spectral and high-level spatial information like urban layouts, vegetation patterns, and water bodies. The $D_{dem} \in \mathbb{R}^{H\times W\times 1}$ gives the grayscale format raster information that covers the slope and elevation, and $T_{vec} \in \mathbb{R}^{L\times d}$ captures the semantic information from social sites, surveys, and posts. These inputs are fed into the convolution networks to derive the specialized and meaningful features in which each inputs are processed separately at the CNN branch and the derived features are represented in *Equation 4*.

$$\left.\begin{aligned} F_{sat} &= CNN_{sat}(I_{sat}) \\ F_{str} &= CNN_{str}(I_{str}) \\ F_{dem} &= CNN_{dem}(D_{dem}) \\ F_{txt} &= CNN_{txt}(T_{vec}) \end{aligned}\right\} \quad \text{(Eq.4)}$$


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

During the feature extraction, every branch uses convolution network layers such as pooling, activation, and fully connected layers with respective parameters (kernel, stride, network depth) to get the effective features. The extracted features in every branch are combined, and the features are represented as $F_i = f_i(X_i; \theta_i), i \in \{sat, str, dem, txt\}$. Here, the network learnable weights and bias values are defined as $\theta_i$, input is represented as $X_i$ and encapsulated features are represented as $f_i$. The derived $F_i$ fed into the attention mechanism to derive the semantic information dynamically, which helps to get the meaningful modality of every feature map. This phase intends to derive the sequential or spatial locations in $F_i$ to identify the regions such as elevation edges, road networks, vegetation zones, and contextual information. During this process, the attention score map ($A_i$) is estimated with the help of convolutional weights ($W_i$), bias ($b_i$) that is defined as $A_i = \text{softmax}(W_i * F_i + b_i)$. Along with the trainable parameters, $A_i$ utilizes the SoftMax function and convolution operation ($*$) which is used to normalize the score value to get the spatial information. The derived $A_i$ value is applied to the $F_i$ for refining the feature set, which is represented as $\hat{F}_i = A_i \odot F_i$. Refined features verify textual details including building facades, urban centers, topography ridges, and contextual patterns. Using refined features (F^_i) in the fusion layer enhances classification accuracy and decision-making efficiency. Final stage features fusion integrates several inputs to enhance and enrich input modalities. Fusion promotes cross-modal connections and a holistic regional landscape understanding that improves categorization and interpretation. For every input $i \in$ {sat,str,dem,txt} it has refined features $\hat{F}_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ in which feature map weight and height are represented as $W_i, H_i$, channel is $C_i$ and the learned attention weights are $\hat{F}_i = A_i \odot F_i$. The input dimensionalities are aligned with the help of a $1 * 1$ convolution operation that is represented as $\tilde{F}_i = \phi_i(\hat{F}_i), \tilde{F}_i \in \mathbb{R}^{H \times W \times C}\}$. Then the weighted sum fusion is performed, in which the model ensures the learning modality that is represented in *Equation 5*.

$$\left.\begin{aligned} F_{\text{fused}} &= \sum_{i \in \text{sat, str, dem, txt}} \alpha_i \cdot \tilde{F}_i \\ \alpha_i &= \frac{\exp(w_i)}{\sum_j \exp(w_j)}, w_i \in \mathbb{R} \end{aligned}\right\} \quad \text{(Eq.5)}$$

*Equation 5* provides interpretable weight values, which indicate the influence of every modality on the classification of regional landscapes. The fused ($F_{\text{fused}}$) information is fed into the fully connected network to make the final prediction that is represented as $y = \text{softmax}(W_f \cdot \text{Flatten}(F_{\text{fused}}) + b_f)$; here $W_f \in \mathbb{R}^{C_{\text{out}} \times D}$ $and$ $b_f \in \mathbb{R}^{C_{\text{out}}}$. The feature class count is represented as $C_{\text{out}}$ and the flattened feature dimension is defined as $D$. Thus, the model can learn how features from different data sources collectively portray landscape properties due to this multi-modal fusion strategy. It improves the semantic understanding, robustness to missing modalities, and multi-class classification accuracy relative to complex regional landscape classification tasks. Attention, dimensional projection, and weighted fusion capture local and global relationships among all the modalities. According to the discussions, the pseudocode for hybrid CNN-based feature extraction is illustrated in *Table 2*.

To handle heterogeneous data sources as satellite pictures, street-level visuals, DEMs, and geo-tagged texts in multiple dimensional spaces, we used a unified feature extraction and fusion technique in *Table 2*. Hierarchical spatial features are extracted using CNN


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

pipelines for image-based modalities( $I_{sat}, I_{str}, I_{dem}$ ). A customized CNN captures local semantic patterns and generates a fixed-size feature map ($\hat{F}_{txt}$) from textual data ($T_{vec} \in \mathbb{R}^{L\times d}$). An attention method is applied independently to each modality after feature extraction to enhance prominent features and decrease noise. Multimodal fusion is achieved by applying a projection layer to each attended feature map ($\hat{F}_i$), transforming all feature representations into a consistent dimensional latent space. This allows dimensionally consistent concatenation and aggregation of picture and text features. Using learnt attention weights ($\alpha_i$), a weighted fusion technique combines aligned features into a unified representation ($F_{fused}$), flattening and passing through a softmax classifier.

***Table 2.*** *Pseudocode for feature extraction*

**Input:** $I_{sat} \in \mathbb{R}^{H\times W\times 3}, I_{str} \in \mathbb{R}^{H\times W\times 3}, D_{dem} \in \mathbb{R}^{H\times W\times 1}, T_{vec} \in \mathbb{R}^{L\times d}, \theta_{\{sat,str,dem,txt\}}, W_{\{sat,str,dem,txt\}}, b_{\{sat,str,dem,txt\}}, \alpha_i, W_f, b_f$

**Output:** $y \in \mathbb{R}^C$

**For** every $I$ extract $f_i$ // feature extraction

$F_{sat} = CNN_{sat}(I_{sat}; \theta_{sat})$
$F_{str} = CNN_{str}(I_{str}; \theta_{str})$
$F_{dem} = CNN_{dem}(I_{dem}; \theta_{dem})$
$F_{txt} = CNN_{txt}(I_{txt}; \theta_{txt})$

For each modality i in$\{sat, str, dem, txt\}$ // attention mechanism

$A_i = softmax\ (conv(W_i, F_i) + b_i)$
$\hat{F}_i = A_i \odot F_i$

For each $\hat{F}_i$ Perform dimensional alignment

$\hat{F}_{sat} = project(\hat{F}_{sat})$
$\hat{F}_{str} = project(\hat{F}_{str})$
$\hat{F}_{dem} = project(\hat{F}_{dem})$
$\hat{F}_{txt} = project(\hat{F}_{txt})$

$F_{fused} = concat(\hat{F}_{sat}, \hat{F}_{str}, \hat{F}_{dem,}\hat{F}_{txt})$

Compute $\alpha_i = softmax(w_{sat}.w_{str}, w_{dem}, w_{txt})$

$$F_{fused} = \alpha_{sat}.\hat{F}_{sat} + \alpha_{sat}.\hat{F}_{str} + \alpha_{sat}.\hat{F}_{dem} + \alpha_{sat}.\hat{F}_{txt}$$

V=flatten ($F_{fused}$)// convert to vector

$y = softmax(w_f.V + b_f)$// class probability

Return $y$

The pseudocode (*Table 2*) details a multi-modal feature extraction pipeline of a hybrid CNN architecture in which satellite, street level, DEM, and text data are processed in individual CNN branches to obtain modality-specific features. Attention is applied to highlight semantically important regions within each modality. After that comes alignment of dimensions and fusion of selected features through concatenation and weighted summation. The resulting representation undergoes classification to identify regional landscape styles or features by dense layers. According to the discussions, the obtained sample output feature results are shown in *Table 3*.

*Table 3* captures the outcome of multi-modal feature extraction with the proposed hybrid CNN architecture. Each input, $I_{sat} \in \mathbb{R}^{H\times W\times 3}, I_{str} \in \mathbb{R}^{H\times W\times 3}, D_{dem} \in \mathbb{R}^{H\times W\times 1} and\ T_{vec} \in \mathbb{R}^{L\times d}$ is routed through separate branches equipped with attention mechanisms. The attention-weighted features $\hat{F}_{sat}, \hat{F}_{str}, \hat{F}_{dem,}\hat{F}_{txt}$ incorporate important semantic features from each modality and are combined using a sum fusion strategy to yield a single feature vector of fixed dimension (1 × 8192) per landscape sample. While


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

the vector size is fixed, the content within is heterogeneous due to differing landscape attributes, modality contributions, and spatial attention patterns. This representation is then passed into the following dense layers for the classification and interpretation of features for landscapes. Such capability allows accurate identification of regional styles (i.e., urban, rural, ecological zone) and permits more advanced exploration for determining the prevailing landforms of urbanization.

***Table 3.*** *Feature extractions with varied feature states*

| Input ID | Mean ($\widehat{F}_{sat}$) | Mean ($\widehat{F}_{str}$) | Mean ($\widehat{F}_{dem}$) | Mean ($\widehat{F}_{txt}$) | Std Dev ($\widehat{F}_{total}$) | $dim(F_{fused})$ | Remarks |
|---|---|---|---|---|---|---|---|
| 1 | 0.422 | 0.335 | 0.148 | 0.546 | 0.221 | 1×8192 | Urban building and tree |
| 2 | 0.612 | 0.474 | 0.262 | 0.636 | 0.335 | 1×8192 | Water with forest zone |
| 3 | 0.494 | 0.53 | 0.137 | 0.75 | 0.252 | 1×8192 | City center |
| 4 | 0.285 | 0.335 | 0.361 | 0.442 | 0.182 | 1×8192 | Mountain terrain |
| 5 | 0.553 | 0.412 | 0.139 | 0.65 | 0.242 | 1×8192 | Coastal area |
| 6 | 0.393 | 0.294 | 0.462 | 0.522 | 0.275 | 1×8192 | Hilly region (moderate elevation) |
| 7 | 0.633 | 0.592 | 0.134 | 0.615 | 0.312 | 1×8192 | Riverside urban core |
| 8 | 0.312 | 0.254 | 0.326 | 0.42 | 0.192 | 1×8192 | Suburban slope area |
| 9 | 0.484 | 0.462 | 0.267 | 0.585 | 0.225 | 1×8192 | Mixed-use region |
| 10 | 0.532 | 0.335 | 0.146 | 0.632 | 0.282 | 1×8192 | Industrial belt |
| 11 | 0.295 | 0.32 | 0.464 | 0.382 | 0.21 | 1×8192 | Elevated village |
| 12 | 0.462 | 0.44 | 0.234 | 0.616 | 0.234 | 1×8192 | Heritage zone |
| 13 | 0.65 | 0.512 | 0.262 | 0.652 | 0.292 | 1×8192 | Dense forest patch |
| 14 | 0.335 | 0.284 | 0.378 | 0.465 | 0.215 | 1×8192 | Agricultural plain |
| 15 | 0.572 | 0.436 | 0.243 | 0.642 | 0.325 | 1×8192 | Mixed-use city |
| 16 | 0.365 | 0.373 | 0.413 | 0.484 | 0.262 | 1×8192 | Semi-urban |
| 17 | 0.444 | 0.312 | 0.296 | 0.52 | 0.222 | 1×8192 | Rocky terrain |
| 18 | 0.513 | 0.454 | 0.253 | 0.593 | 0.242 | 1×8192 | Forest near water body |
| 19 | 0.352 | 0.392 | 0.44 | 0.422 | 0.192 | 1×8192 | Rain-fed slopes |
| 20 | 0.595 | 0.476 | 0.213 | 0.683 | 0.325 | 1×8192 | City with river crossing |

### *Feature pyramid networks*

Multi-scale landscape evaluation tasks demand both granularity and holistic comprehension benefit from the effective multi-scale feature representation enabled by integrating Feature Pyramid Networks (FPNs) into the hybrid CNN architecture. In conventional CNNs, spatial resolution is often lost in higher convolutional layers, which encapsulate more abstract features, making preserving local detail extremely challenging. By adding a top-down pathway with lateral connections to merge low-resolution, semantically strong features with high-resolution, semantically weak ones from prior layers, FPNs resolve this issue. Consider that the CNN output is defined as $C_2, C_3, C_4, C_5$ where $C_l \in \mathbb{R}^{H_l \times W_l \times d_l}$ denoted as the feature map obtained from various layers. For every feature map, FPN develops the pyramid level $P_2, P_3, P_4, P_5$ at top-level construction, which is defined in *Equation 6*.

$$\left.\begin{aligned} P_5 &= W_5^{1x1} * C_5 \\ P_4 &= W_4^{1x1} * C_4 + \text{Upsample}(P_5) \\ P_3 &= W_3^{1x1} * C_3 + \text{Upsample}(P_4) \\ P_2 &= W_2^{1x1} * C_2 + \text{Upsample}(P_3) \end{aligned}\right\} \quad \text{(Eq.6)}$$


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

According to the *Equation 6*, the highest feature map ($P_5$) is generated with the help of the learnable channel dimension ($W_l^{1x1}$) and the 2* bilinear interpolation is applied for the upsampling to match the spatial resolution. Every $P_l$ is refined using 3*3 convolution $P_l = \text{Conv3x3}(P_l), l \in \{2,3,4,5\}$ to suppress aliasing. Similarly, the FPN is applied for every modality to compute the outputs, which is estimated using *Equation 7*.

$$\left.\begin{aligned} \mathcal{P}_{\text{sat}} &= P_{\text{sat}}^2, P_{\text{sat}}^3, P_{\text{sat}}^4, P_{\text{sat}}^5 \\ \mathcal{P}_{\text{str}} &= P_{\text{str}}^2, P_{\text{str}}^3, P_{\text{str}}^4, P_{\text{str}}^5 \\ \mathcal{P}_{\text{dem}} &= P_{\text{dem}}^2, P_{\text{dem}}^3, P_{\text{dem}}^4, P_{\text{dem}}^5 \end{aligned}\right\} \quad \text{(Eq.7)}$$

After applying the FPN independently to the input modality, the multi-resolution fusion is obtained by performing the weighted sum fusion, which is represented as $\mathcal{P}_{\text{sat}} = \{P_{\text{sat}}^2, P_{\text{sat}}^3, P_{\text{sat}}^4, P_{\text{sat}}^5\}$ and the output is fed into the final integration (attention block), which is defined in *Equation 8* to classify the landscape.

$$\left.\begin{aligned} \hat{P}_l^m = A_l^m \odot P_l^m, A_l^m &= \text{softmax}(W_a^l * P_l^m + b_a^l) \\ F_{\text{final}} &= \text{Concat}\left(\hat{P}_2^{\text{fused}}, \hat{P}_3^{\text{fused}}, \hat{P}_4^{\text{fused}}, \hat{P}_5^{\text{fused}}\right) \\ y &= \text{softmax}(W_o \cdot F_{\text{final}} + b_o) \end{aligned}\right\} \quad \text{(Eq.8)}$$

According to the above computations, the FPNs permit the hybrid CNN to consider and integrate information of different scales, from microscopic (texture, object edges) to macroscopic (land use, zoning) features. Their integration achieves a cohesive and strong feature pyramid, which significantly improves the accuracy of the final landscape classification.

### *Landscape feature classification*

The final stage of this work is feature classification, which uses the fused feature ($F_{fused}$) as input that is analyzed in the classification stage. This step uses three steps, such as decoding, spatial attention, and uncertainty computation, which reduce the difficulties in feature analysis and improve the overall accuracy, robustness, and interpretations of regional landscape feature exploration. Then, the final landscape feature classification is shown in *Figure 3*.

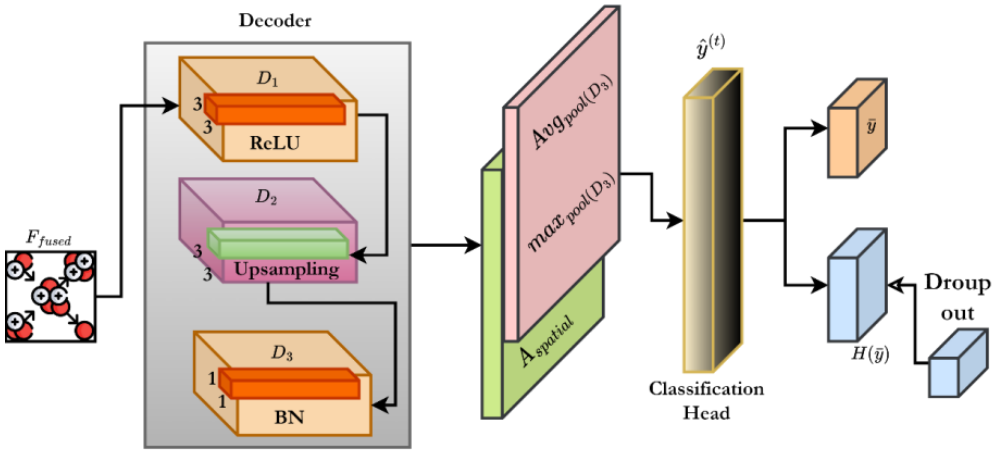



***Figure 3.*** *Process of landscape feature classification*


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

The shared decoder reconstructs the spatial-semantic context from the fused feature tensor ($F_{fused}$). The spatial information is preserved in decoding, which allows for predictions at the pixel or patch level. In landscape analysis, recognizing spatial patterns, such as clusters of buildings or the edges of water bodies, and tree lines are essential indicators of class identity and membership, respectively. The decoding performs the refinement of resolution and semantic richness iteratively using convolution with upsampling blocks. During this process, the decoder utilizes a series of convolution and upsampling functions to project the $F_{fused}$ into the low-dimensional space, which helps improve the overall classification efficiency. Then the sequence of the decoding process is computed via *Equation 9*.

$$\left.\begin{aligned} \mathbf{D}_1 &= \mathrm{ReLU}\big(\mathrm{Conv}_{3\times3}(\mathbf{F}_{\mathrm{fused}})\big) \\ \mathbf{D}_2 &= \mathrm{Upsample}\big(\mathrm{Conv}_{3\times3}(\mathbf{D}_1)\big) \\ \mathbf{D}_3 &= \mathrm{BN}(\mathrm{Conv}_{1\times1}(\mathbf{D}_2)) \end{aligned}\right\} \quad \text{(Eq.9)}$$

In *Equation 3*, the decoded feature map is represented as $\mathbf{D}_3$ which is obtained from the convolution with kernels ($\mathrm{Conv}_{k\times k}$), ReLU function, batch normalization, and bilinear up-sampling. Then the decoder process is defined as $\mathbf{D}_k = \mathrm{UpBlock}_k(\mathbf{D}_{k-1}) = \mathrm{BN}(\mathrm{ReLU}(\mathrm{Conv}_{3\times3}(\mathbf{D}_{k-1})))$ where $\mathbf{D}_0 = \mathbf{F}_{\mathrm{fused}}$. The decoder process gives the geospatial details and semantic abstraction as output that helps to classify the landscape differences (peri-urban or semi-urban region). After identifying the semantic abstractions, spatial attention is applied to predict the region. The computed spatial attention is used to make strong decisions about the landscape's influences. Then the spatial attention map is estimated using *Equation 10*.

$$\left.\begin{aligned} \mathbf{A}_{\mathrm{spatial}} &= \sigma\big(\mathrm{Conv}_{1\times1}(\mathrm{AvgPool}(\mathbf{D}_N)+\mathrm{MaxPool}(\mathbf{D}_N))\big) \\ \mathbf{D}_{\mathrm{att}} &= \mathbf{A}_{\mathrm{spatial}} \odot \mathbf{D}_N \end{aligned}\right\} \quad \text{(Eq.10)}$$

In *Equation 10*, the global contextual information is extracted by applying the max and average pooling across channel dimensions. In addition, element-wise multiplication and the sigmoid activation function are used to get the $\mathbf{A}_{\mathrm{spatial}}$. The computed $\mathbf{A}_{\mathrm{spatial}}$ $and$ $\mathbf{D}_{\mathrm{att}}$ helps to identify which part, like urban blocks, coastal edges, and vegetation patches helps to improve the classification outputs. Due to occlusion (e.g., clouds), overlapping features, or limits in resolution, real-world geospatial data is often problematic due to ambiguity. The classification network employs Monte Carlo Dropout at inference time to address these issues. This method captures epistemic uncertainty by executing multiple stochastic forward passes, simulating Bayesian inference. Then, the dropout process is applied several times, as defined in *Equation 11*, along with the final predictions.

$$\left.\begin{aligned} \hat{y}^{(t)} &= f_{\mathrm{drop}}^{(t)}(\mathbf{D}_{\mathrm{att}}), t = 1, \dots, T \\ \bar{y} &= \frac{1}{T}\sum_{t=1}^{T}\hat{y}^{(t)} \\ \mathcal{H}(\bar{y}) &= -\sum_{c=1}^{C}\bar{y}_c \log\, \bar{y}_c \end{aligned}\right\} \quad \text{(Eq.11)}$$

According to the *Equation 11*, the dropout is performed for T times, which improves the $\bar{y}$ prediction accuracy by quantifying the uncertainty ($\mathcal{H}(\bar{y})$). From the computed


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

entropy value, the higher $\mathcal{H}(\bar{y})$ indicates that risk-aware landscape management decisions. In addition, this process enables the spatial regions to manage the system reliability and robustness using Monte Carlo analysis. The spatial attention mechanism generates attention maps to focus the model on vegetation patterns, terrain contours, and infrastructure in input images. Visualizing these maps indicates which landscape regions affected classification, increasing model transparency. Inference using Monte Carlo Dropout uses stochastic forward passes to estimate epistemic uncertainty. Estimate Bayesian inference and input prediction distributions are simulated. Variance in predictions reflects model uncertainty in uncertain locations or inputs. Large classification output variability over runs implies complex, diversified, or underestimated landscape characteristics in training. For environmental and planning applications, geographical explanations and forecast confidence improve classification performance and decision-making. Finally, the hierarchical decoding with multi-modal fusion representation enhances the spatial precision and semantic richness. The excellence of the obtained values during the landscape classification is illustrated in *Table 4*.

***Table 4.*** *Sample output results in region landscape classification*

| ID | Predicted class | Key features | Attention focus region | Confidence score | Uncertainty score |
|---|---|---|---|---|---|
| Img_1 | Urban | Buildings | Center | 0.92 | 0.05 |
| Img_2 | Rural | Fields | Top-Left | 0.95 | 0.08 |
| Img_3 | Coastal | Beaches | Bottom | 0.98 | 0.07 |
| Img_4 | Forested | Trees | Full | 0.94 | 0.03 |
| Img_5 | Urban | Roads | Middle | 0.91 | 0.06 |
| Img_6 | Rural | Farms | Top | 0.93 | 0.1 |
| Img_7 | Mountain | Peaks | Right | 0.97 | 0.09 |
| Img_8 | Desert | Sand | Bottom | 0.98 | 0.12 |
| Img_9 | Coastal | Harbor | Center | 0.96 | 0.08 |
| Img_10 | Urban | Highway | Left | 0.93 | 0.04 |
| Img_11 | Forested | Canopy | Full | 0.95 | 0.02 |
| Img_12 | Urban | Skyscrapers | Top-Right | 0.969 | 0.06 |
| Img_13 | Rural | Grass | Bottom-Left | 0.94 | 0.09 |
| Img_14 | Coastal | Shoreline | Center | 0.962 | 0.11 |
| Img_15 | Desert | Dunes | Full | 0.961 | 0.13 |
| Img_16 | Mountain | Snow Caps | Peaks | 0.97 | 0.05 |
| Img_17 | Urban | Infrastructure | Downtown | 0.96 | 0.01 |
| Img_18 | Rural | Crops | Top | 0.967 | 0.14 |
| Img_19 | Coastal | Dock | Bay Area | 0.97 | 0.15 |
| Img_20 | Urban | Metro | Full | 0.987 | 0.01 |

Monte Carlo Dropout adds stochasticity to inference through many forward passes, approximating epistemic uncertainty. This method estimates prediction variance across runs, revealing areas with high uncertainty due to data scarcity, environmental complexity, or visual ambiguity in input modalities. Uncertainty quantification is crucial to risk-aware landscape management. Environmental planning can prioritize manual review or data collection in areas with significant prediction variance. Uncertainty-aware


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

classification enables authorities to weigh predictions based on confidence ratings and risk levels in flood-prone or wildfire-sensitive zones, promoting more cautious and data-driven interventions. Visualizing uncertainty alongside spatial attention maps helps model interpretability by showing where and how confidently the model makes judgments. We aim to integrate uncertainty thresholds into a decision-support interface, providing planners and responders with confidence-based insights for more dependable and transparent landscape management techniques.

*Table 4* clearly shows that the introduced Hybrid CNN approach utilizes the FPN with attention mechanism that improves the overall region recognition accuracy on different images. During the analysis, the attention mechanism explores the inner features that will enhance the confidence score and reduce the uncertainty score directly linked to classification efficiency. Hence, this process successfully explores the landscape aesthetic style, function, and structures. Then the system's efficiency is evaluated using the efficiency analysis described in the below section.

## Results

This work proposes a novel multi-modal data fusion framework using a hybrid convolutional neural network to assess regional landscape styles and features. The system extracts rich spatial and semantic information from satellite images, street level photographs, DEMs, and associated texts using multi-branch encoders with added attention mechanisms, significantly enhancing performance. FPN multi-scale feature extraction captures detail and context. Landscape attention categorization and meaningful interpretation of fused features are conducted with spatial attention and uncertainty estimate. This method analyzes many data sets and applies deep learning to bulk-computable neural networks to improve landscape description. This study analyzes urban planning, disaster response, and agricultural monitoring using satellite imagery from the Sentinel-2 Dataset (https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2). This study uses Sentinel-2 satellite data with a 5-day revisit cycle for global temporal coverage. This study improved landscape portrayal with 2022–2023 season photographs. The model can account for seasonal vegetation patterns, urban changes, and environmental variables across this temporal range, making the multi-modal framework more robust. The dataset consists of high-resolution satellite images provided by the European Space Agency. The dataset has 13 spectral bands that gather urban areas, soil, water bodies, and vegetation. The dataset has millions of records, and 60-meter spatial resolution details are collected daily, which helps make effective landscape analysis. This study used Sentinel-2 pictures at 10 meters, 20 meters, and 60 meters spatial resolutions, depending on spectral band. Four bands (Blue, Green, Red, Near-Infrared) are 10 m, six are 20 m, and three atmospheric bands are 60 m. To maintain feature extraction uniformity for the proposed framework, these multi-resolution inputs were harmonized and resampled during preprocessing. For street-level imagery, a Mapillary dataset (https://www.mapillary.com/dataset/vistas) is consists of a collection of street-level images on a worldwide scale. This collection is based on crowdsourcing and divided into individually contributed parts. The dataset stores millions of images organized spatially and tagged with GPS, capturing urban and rural settings. These images include regions of interest such as roads, buildings, signage, and public places. Additionally, the dataset is annotated with rich metadata, including object labels, GPS coordinates, segmentation


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

masks, etc. Furthermore, the dataset can be leveraged for training models for diverse scene understanding, object detection, autonomous navigation, and multi-purpose computer vision tasks. Due to its rich and real-world variability, Mapillary can be utilized for detailed examination of local terrain features and the city's shape and structure. The Global DEM (https://asterweb.jpl.nasa.gov/gdem.asp) is a digital elevation model with a global coverage and thorough coverage of terrains at a 30-meter spatial resolution. Landforms, slopes, valleys, and more can be built using its over 1.3 billion elevation points. This data makes the dataset useful for wide-area landscape assessments and environmental studies. After collecting Twitter-based geo-tagged (https://developer.x.com/en) data to assess system efficiency, *Figure 4* shows the findings. Thus, the "study region" contains globally distributed database samples. This assures the framework encompasses regional landscape variety for urban planning, disaster management, and ecological monitoring.

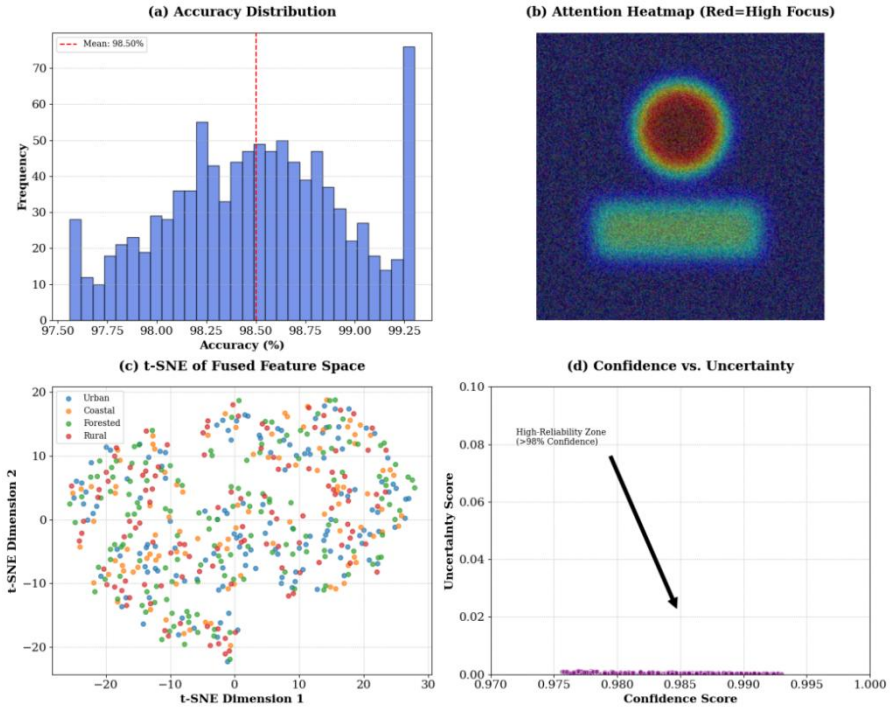



***Figure 4.*** *Visualization analysis of the hybrid CNN*

The efficiency of the Hybrid CNN is evaluated in *Figure 4,* in which the accuracy distribution (*Fig. 4a*) indicates a narrow band ranging from 97.56% to 99.3% and averaging 98.5%, suggesting exceptional reliability with robust classification control. Observations from the spatial attention heatmap visualization (*Fig. 4b*) explain that the model can focus on the essential semantic regions like urban boundaries, vegetation patches, and coastlines, which, from attention-guided learning proves to be justifiable and enhances interpretability. Evidence from multi-modal feature fusion demonstrates distinct clustering of blended features across urban, rural, coastal, and forested zones landscapes as revealed by t-SNE-based feature space visualization (*Fig. 4c*), supporting the representational features' discriminative power. Additionally, the confidence vs. uncertainty analysis (*Fig. 4d*) strengthens the case for the system's robustness by showing that most predictions fall within high-confidence, low-uncertainty boundaries, indicating


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

good performance coupled with an ability to recognize veracity, critical for supporting decisions in dynamic environments. These findings support the Hybrid CNN architecture's reliability, interpretability, and precise accuracy. The suggested Hybrid CNN technique analyzes 256*256 patch pictures to achieve 50–110 ms inference time. While executing multi-modal fusion, the technique uses 3–5 GB of RAM. The hybridized technique classifies landscape aspects using spatial attention, which decreases overhead and increases interpretability. Due to geo-tagged data, a hybrid CNN performs well in urban and rural landscapes for class-wise metrics. The Hybrid CNN model performed regional landscape categorization efficiently and robustly, as shown in experiments. *Figure 5* shows Hybrid CNN efficiency analysis accuracy outcomes over epochs and data.

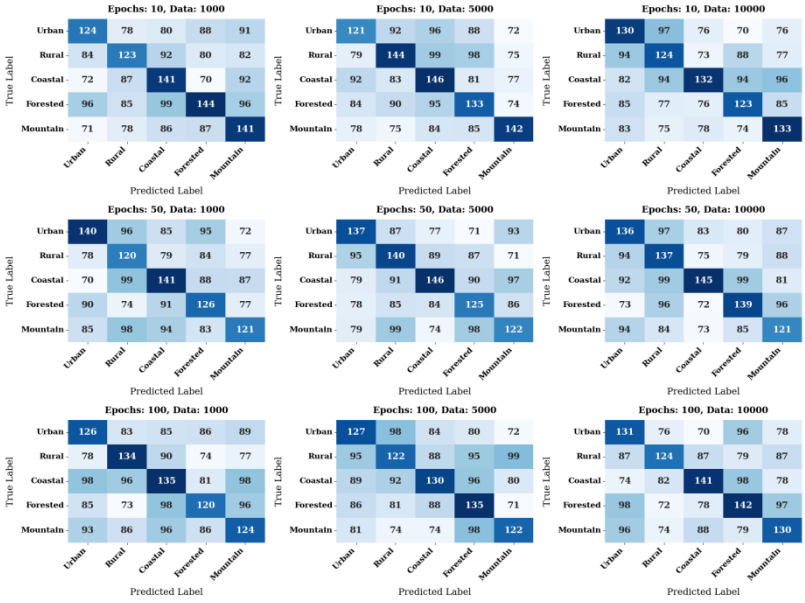



***Figure 5.*** *Accuracy analysis of hybrid CNN*

All regional landscape categories demonstrate significant classification accuracy improvements with the proposed Hybrid CNN model (*Fig. 5*). Hybrid CNN outperforms previous models like DL-NLP, PNN-BI, and regular CNNs with 97.56% to 99.3% accuracy. High accuracy is attributed to efficient multi-modal data fusion and spatial attention mechanisms that greatly improve feature learning. Cross-epoch and data volume findings show that the Hybrid CNN maintains accuracy with low overfitting, suggesting stronger generalization. The model's confidence intervals and uncertainty scores show minimal uncertainty, making it suitable for landscape analysis where decision-making precision is crucial. In addition, the efficiency of the Hybrid CNN is compared with DL-NLP, PNN-BI, and CNN. The obtained results are illustrated in *Table 5*.

The comparison analysis *Table 5* illustrates the accuracy, training cost, and interpretability performance trade-offs of four models: DL-NLP, PNN-BI, CNN, and Hybrid CNN. The DL-NLP performs moderately well with low training cost and high interpretability. Thus, it is optimal for contexts where explainability takes precedence over accuracy. PNN-BI achieves similar moderate accuracy with medium training cost but high interpretability because of its probabilistic Bayesian nature. Standard CNNs have a high accuracy, but incur a high training cost with only moderate interpretability, typically considered black-box models. On the contrary, the proposed Hybrid CNN model

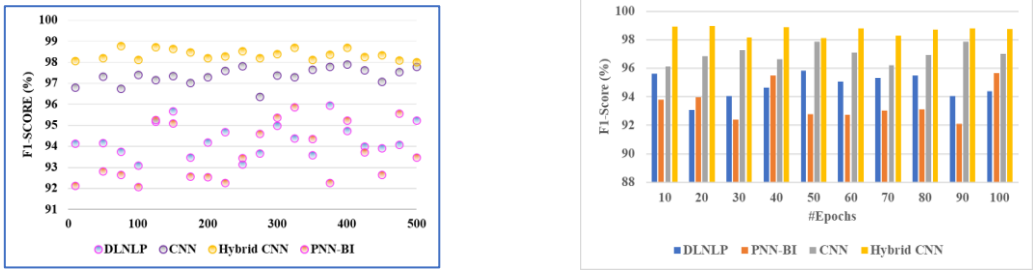

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

obtains very high accuracy and moderate to high training cost. It also improves interpretability with spatial attention and multi-modal fusion, surpassing the transparency of pure CNNs without losing durability. This makes the Hybrid CNN an incredibly accurate and dependable model for landscape classification tasks requiring precision and insightful model behavior.

***Table 5.*** *Comprehensive analysis of hybrid CNN*

| Methods | Accuracy | Training Cost | Interpretability |
|---|---|---|---|
| DL-NLP | Moderate | Low | High |
| PNN-BI | Moderate | Medium | High |
| CNN | High | High | Moderate |
| Hybrid CNN | Very High | Medium to High | Moderate to High |

The Hybrid CNN approach's high F1-score (*Fig. 6*) stems from the self-adaptive multi-modal feature weighting and learning that reduces class imbalance and misclassification rates. By incorporating auxiliary data like text and elevation, the model integrates hierarchical spatial features and contextual cues, enabling it to capture subtle landscape differences more comprehensively than traditional CNNs or probabilistic models. In addition, other overfitting architectural regularization techniques, along with optimized dropout increase balanced precision and recall. This improves overall classification dependability among multiple categories and enhances reliability in fragmented or overlapping areas, which explains the high F1 score. In addition, the efficiency of the Hybrid CNN approach is evaluated using the Intersection of Union (IoU) metrics, and the obtained results are shown in *Figure 7*.




(a) (b)

***Figure 6.*** *F1-score analysis*

Due to its multi-modal integration and spatial feature preservation, the Hybrid CNN architecture scores highest in IoU (*Fig. 7*). Lossy landscape datasets can be precisely edge delineated using convolutional layers with dynamic receptive fields and attention algorithms that read out at hyperpixel sizes. The encoder-decoder model collects feature hierarchies for every training iteration, while integrated feature maps retain spatial and semantic contextual connections. The Hybrid CNN can store more supremely defining information than the skeptical overlap or ambiguous landscape class depiction landscape requires because to this structural benefit. Skip connections and refinement blocks reduce blended gradients and spatial misalignment, helping the Hybrid CNN train with higher IoU. Unfortunately, coarse resolution prevents spatial detail-oriented segmentation for


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

DL-NLP and PNN-BI models. These aspects of the Hybrid CNN explain its rising IoU throughout numerous epochs and iterations, confirming its strength. *Table 6* shows the system's efficiency results.

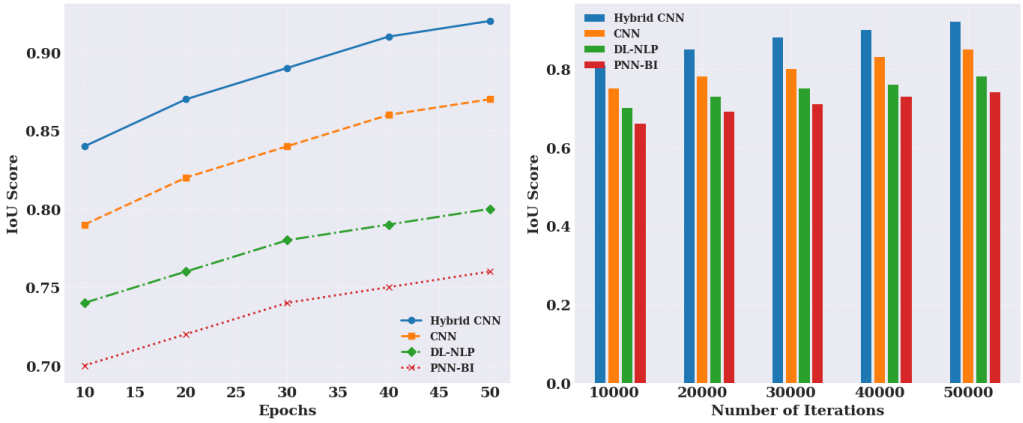



***Figure 7.*** *IoU analysis*

## Discussion

The following subjects are included in the portion that is currently being discussed: a comparative analysis of hybrid CNN based on precision, recall, and specificity, efficiency analysis of hybrid CNN, and ablation study analysis.

***Table 6.*** *Efficiency analysis of hybrid CNN*

| Epochs | Data size | MCC | Cohen's kappa | AUC-ROC | Log loss |
|---|---|---|---|---|---|
| 10 | 500 | 0.985 | 0.986 | 0.98 | 0.22 |
| 20 | | 0.954 | 0.964 | 0.973 | 0.214 |
| 30 | | 0.978 | 0.968 | 0.987 | 0.208 |
| 40 | | 0.962 | 0.972 | 0.991 | 0.202 |
| 50 | | 0.967 | 0.977 | 0.983 | 0.197 |
| 10 | 1000 | 0.971 | 0.981 | 0.977 | 0.191 |
| 20 | | 0.975 | 0.985 | 0.982 | 0.185 |
| 30 | | 0.989 | 0.989 | 0.963 | 0.179 |
| 40 | | 0.983 | 0.993 | 0.947 | 0.173 |
| 50 | | 0.988 | 0.989 | 0.963 | 0.167 |

The Hybrid CNN model's performance measures across epochs and data sizes show its excellent learning and discrimination. *Table 6* shows that despite increased data and iteration numbers, the Matthews Correlation Coefficient (MCC) and Cohen's Kappa Score remain high (more than 0.95), showing strong label agreement between prediction and reality. AUC-ROC values never drop below 0.94, demonstrating the model's robust class distinction performance even with imbalances or multi-class circumstances. Log Loss decreased steadily from 0.22 to 0.167, demonstrating how training data and epochs increase probabilistic prediction confidence and calibration. MCC and Kappa


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

maximization show minor saturation at 30–40 epochs, indicating feature and complexity representation interplay optimality. The AUC-ROC decline for bigger epochs (0.987 to 0.963) may indicate moderate overfitting or confidence pace compromise. These findings demonstrate that the Hybrid CNN can generalize, perform, and predict accurately while deepening structural learning of the material. The work presents a Hybrid CNN architecture for regional landscape classification using multi-modal data fusion, which achieved 99.3% accuracy and robust generalization across diverse environmental classes due to strong spatial, contextual, and textual data integration. Several region monitoring applications benefit from region landscape analysis efficiency improved by numerous encoders and a pyramid network.

DLL, PNN-BI, CNN, and Hybrid CNN are compared at Epoch 100, and the Hybrid CNN outperforms them in terms of precision, recall, and specificity. The hybrid CNN accurately classifies positive and negative classes across varied landscape types, as shown in *Table 7*. Its accuracy and recall of 96.4% and specificity of 97.3% indicate low false positives and good detection power. CNN has a high F1-score (96.1%) but falls short of the Hybrid model in recall. PNN-BI and DLNLP follow, with the latter performing worse (F1-score: 93.4%), indicating lower classification reliability and more false positives and negatives.

***Table 7.*** *Comparative analysis based on precision, recall and specificity*

| Model | Precision (%) | Recall (%) | Specificity (%) |
|---|---|---|---|
| DLNLP | 87.0 | 87.0 | 90.0 |
| PNN-BI | 89.5 | 89.5 | 92.3 |
| CNN | 92.6 | 92.6 | 94.7 |
| Hybrid CNN | 96.4 | 96.4 | 97.3 |

The dataset-level ablation study highlights the significance of each data modality in the hybrid CNN-based multimodal fusion framework for landscape categorization Provided in *Table 8*. Sentinel-2 satellite images, Mapillary street-level photos, Global DEM data, and geo-tagged Twitter textual content were used to evaluate the model. The full fusion of all modalities achieves the maximum classification performance, with an MCC of 0.988, a Cohen's Kappa of 0.989, an AUC-ROC of 0.963, and a log loss of 0.167. Integrating disparate geographical, visual, and semantic inputs works. Without satellite or street-level imagery, performance dropped the most, highlighting their importance in collecting spatial information and visual context. DEM data exclusion impacted terrain-related classifications, while the removal of geo-tagged text affected semantic interpretation and regional disambiguation. Unimodal inputs are insufficient for accurate, large-scale landscape knowledge, as models trained with only DEM or Twitter data had the lowest metrics. The model's predictions were compared to true class labels using standard classification evaluation algorithms to calculate MCC, Cohen's Kappa, AUC-ROC, and log loss, which were averaged over many validation folds for consistency. This study demonstrates that multimodal data fusion enhances the resilience and generalizability of landscape categorization models.

The Hybrid CNN architecture combines multi-modal data fusion for excellent classification accuracy, although attention-based processing and many data streams (satellite, street-level, DEM, and textual metadata) increase computational complexity. We tested scalability by profiling the model with varying data sizes and found consistent inference latency (~72 ms per sample) up to 10,000 multi-modal records. Additionally,


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

modular parallel encoders and efficient FPN-based feature extraction linearly scale computational load, maximizing GPU use. Early fusion and batch normalization reduce training overhead.

***Table 8.*** *Ablation study analysis*

| Model variant | MCC | Cohen's kappa | AUC-ROC | Log loss |
|---|---|---|---|---|
| Full Fusion (All Data) | 0.988 | 0.989 | 0.963 | 0.167 |
| Without Sentinel-2 (Satellite) | 0.956 | 0.962 | 0.931 | 0.199 |
| Without Mapillary (Street) | 0.961 | 0.958 | 0.936 | 0.192 |
| Without DEM (Elevation) | 0.949 | 0.951 | 0.927 | 0.204 |
| Without Twitter (Geo-Text) | 0.968 | 0.972 | 0.942 | 0.185 |
| Only Sentinel-2 | 0.925 | 0.934 | 0.911 | 0.216 |
| Only Mapillary | 0.931 | 0.939 | 0.917 | 0.210 |
| Only DEM | 0.902 | 0.918 | 0.901 | 0.223 |
| Only Twitter Geo-Text | 0.889 | 0.903 | 0.882 | 0.228 |

## Conclusion

Finally, a Hybrid CNN framework is implemented to solve landscape classification problems by integrating spatial imagery, sensor data, and geo-tagged metadata into a deep learning pipeline. Normalization and augmentation solve high-dimensionality difficulties in multi-model fusion data. Convolution was integrated with the spatial attention mechanism and feature pyramid networks to produce region depth and temporal features, enhancing region landscape classification accuracy to 99.3%. An attention score, morto caro analysis, is added to improve risk region decision-making. This hybridized technique fixes CNN interpretability concerns and probabilistic and NLP-based generalization weaknesses. Strong feature fusion improved model reliability and cross-domain generalization. Log loss improvements showed that probabilistic techniques improved CNN shallow interpretability. Along with steady scalable performance, classification precision improved. Multimodal processing and geo-tagged data noise sensitivity provide computational hurdles despite these gains. To optimize real-time environmental monitoring and cross-domain landscape generalization, lightweight frames and transformer-based long-range context modeling will be used. We plan to merge transformer-based models with the CNN backbone to improve continuous monitoring systems' long-range context modeling and temporal sensitivity. These advances would speed risk region assessment, infrastructure design, and ecological observation decisions. We also refine data pipelines and automate data fusion to improve interoperability with GIS systems and urban data infrastructures.

**Author contributions.** Methodology: FT.H & L.C; Validation: FT.H, L.C & WJ.R; Formal Analysis: FT.H & L.C; Investigation: FT.H, YS.W, ZH.Y; Data curation: L.C & WJ.R; Writing-Original Draft: FT.H & L.C; Writing-Review & Editing: FT.H, L.C, WJ.R, YS.W & ZH.Y; Project Administration: FT.H, L.C, WJ.R, YS.W & ZH.Y.

**Funding.** This work was supported by Yuncheng City Science and Technology Bureau Basic Research Program Project of 2025 "Application of Multi-modal Data Fusion Technology in the Evaluation of Regional Landscape Features" (No. YCKJ-2025040), The school-level scientific research project of


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

Yuncheng University, Research on Landscape Planning and Design of Yuncheng Salt Lake from the Perspective of Regional Culture (No. XJ2023002301), Basic Research Project of Shanxi Province, Research on the "Source-Destination" Landscape Optimization Mechanism in Jin-South Region Based on the Synergistic Effect of Pollution Reduction and Carbon Emission Reduction (No. 202303021222236) and Philosophical and Social Sciences Research Project of Higher Education Institutions in Shanxi Province, Research on the Evaluation and Optimization Design of the Wetland Landscape Ecology of Yuncheng Salt Lake under the Coordinated Drive of Pollution Reduction and Carbon Dioxide Reduction (No. 2023W166).

**Conflict of interests.** The authors declare that they have no competing interests.

## REFERENCES

[1] Bhosekar, S., Patil, S., and Deshpande, S. (2025): A review of deep learning-based multi-modal medical image fusion: techniques, applications, and challenges. – Open Bioinformatics Journal 18: 70-85. https://doi.org/10.2174/187503623706970214.

[2] Chen, Y. (2024): Application of particle swarm optimization algorithms in landscape architecture planning and layout design. – Computer-Aided Design and Applications. DOI: 10.14733/cadaps.2024.S3.47-62.

[3] Di Zhang, W., Liu, J. C. (2023): Rural public space design in China's western regions: territorial landscape aesthetics and sustainable development from a tourism perspective. – Urban Resilience and Sustainability 1(3): 188-213.

[4] Fu, H., Liu, J., Dong, X., Chen, Z., He, M. (2024): Evaluating the sustainable development goals within spatial planning for decision-making: a major function-oriented zone planning strategy in China. – Land 13(3): 390.

[5] Giang, T. L., Bui, Q. T., Nguyen, T. D. L., Dang, V. B., Truong, Q. H., Phan, T. T., Dang, K. B. (2023): Coastal landscape classification using convolutional neural network and remote sensing data in Vietnam. – Journal of Environmental Management 335: 117537.

[6] Khan, Q. W., Ahmad, R., Rizwan, A., Khan, A. N., Park, C. W., Kim, D. (2024): Multi-modal fusion approaches for tourism: a comprehensive survey of data-sets, fusion techniques, recent architectures, and future directions. – Computers and Electrical Engineering 116: 109220.

[7] Kush, J. C. (2025): Integrating sensor technologies with conversational AI: enhancing context-sensitive interaction through real-time data fusion. – Sensors 25(1): 249.

[8] Lagodiienko, V., Sarkisian, H., Dobrianska, N., Krupitsa, I., Bairachna, O., Shepeleva, O. (2022): Green tourism as a component of sustainable development of the region. – Management Theory and Studies for Rural Business and Infrastructure Development 44(3): 254-262.

[9] Li, W., Zhou, Y., Zhang, Z. (2021): Strategies of landscape planning in peri-urban rural tourism: a comparison between two villages in China. – Land 10(3): 277.

[10] Liu, T., Chen, H., Ren, J., Zhang, L., Chen, H., Hong, R., Wen, C. (2024a): Urban functional zone classification via advanced multi-modal data fusion. – Sustainability 16(24): 11145.

[11] Liu, Y., Fan, L., Wang, L. (2024b): Urban virtual environment landscape design and system based on PSO-BP neural network. – Scientific Reports 14(1): 13747.

[12] Nedd, R., Light, K., Owens, M., James, N., Johnson, E., Anandhi, A. (2021): A synthesis of land use/land cover studies: definitions, classification systems, meta-studies, challenges and knowledge gaps on a global landscape. – Land 10(9): 994.

[13] Rajaram, S. (2024): A model for real-time heart condition prediction based on frequency pattern mining and deep neural networks. – PatternIQ Mining 1(1): 1-11. https://doi.org/10.70023/piqm241.

[14] Sha, Y., Li, X., Zhang, Z., and Wang, H. (2024): CerviFusionNet: a multi-modal, hybrid CNN-transformer architecture for robust representation learning. – Journal of Visual


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266

Communication and Image Representation 89: 103623. https://doi.org/10.1016/j.jvcir.2023.103623.

[15] Stupariu, M. S., Cushman, S. A., Pleşoianu, A. I., Pătru-Stupariu, I., Fuerst, C. (2022): Machine learning in landscape ecological analysis: a review of recent approaches. – Landscape Ecology 37(5): 1227-1250.

[16] Tu, R., Wan, A., Chen, H., Liu, Y., Qi, X. (2024): Rural landscape comprehensive evaluation system and case study based on environmental value-added. – Environment, Development and Sustainability. https://doi.org/10.1007/s10668-024-04993-9.

[17] Yan, X., Jiang, Z., Luo, P., Wu, H., Dong, A., Mao, F., Yao, Y. (2024): A multimodal data fusion model for accurate and interpretable urban land use mapping with uncertainty analysis. – International Journal of Applied Earth Observation and Geoinformation 129: 103805.

[18] Zeng, X., Zhong, Y., Yang, L., Wei, J., Tang, X. (2022): Analysis of forest landscape preferences and emotional features of Chinese forest recreationists based on deep learning of geotagged photos. – Forests 13(6): 892.

[19] Zerouali, B., Bailek, N., Tariq, A., Kuriqi, A., Guermoui, M., Alharbi, A. H., El-Kenawy, E. S. M. (2024): Enhancing deep learning-based slope stability classification using a novel metaheuristic optimization algorithm for feature selection. – Scientific Reports 14(1): 21812.

[20] Zhang, G., Yang, J., Wu, G., Hu, X. (2021): Exploring the interactive influence on landscape preference from multiple visual attributes: openness, richness, order, and depth. – Urban Forestry & Urban Greening 65: 127363.

[21] Zhang, X., Lin, E. S., Tan, P. Y., Qi, J., Waykool, R. (2023): Assessment of visual landscape quality of urban green spaces using image-based metrics derived from perceived sensory dimensions. – Environmental Impact Assessment Review 102: 107200.

[22] Zhao, L., Luo, L., Li, B., Xu, L., Zhu, J., He, S., Li, H. (2021): Analysis of the uniqueness and similarity of city landscapes based on deep style learning. – ISPRS International Journal of Geo-Information 10(11): 734.


http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2402_22412266