Confident Naturalness Explanation (CNE): A Framework to Explain and Assess Patterns Forming Naturalness

Ahmed Emam, Mohamed Farag, and Ribana Roscher, *Member, IEEE*

***Abstract*— Protected natural areas characterized by minimal modern human footprint are often challenging to assess. Machine-learning (ML) models, particularly explainable methods, offer promise in understanding and mapping the naturalness of these environments through the analysis of satellite imagery. However, current approaches encounter challenges in delivering valid and objective explanations and quantifying the contribution of specific patterns to naturalness. These challenges persist due to the reliance on hand-crafted weights assigned to contributing patterns, which can introduce subjectivity and limit the model's ability to capture relationships within the data. We propose the confident naturalness explanation (CNE) framework to address these issues, integrating explainable ML and uncertainty quantification. This framework introduces a new quantitative metric to describe the confident contribution of patterns to the concept of naturalness. Additionally, it generates segmentation masks that depict the uncertainty levels in each pixel, highlighting areas where the model lacks knowledge. To showcase the framework's effectiveness, we apply it to a study site in Fennoscandia, utilizing two open-source satellite datasets. In our proposed metric scale, moors and heathlands register high values of 1 and 0.81, respectively, indicating pronounced naturalness. In contrast, water bodies score lower on the scale, with a metric value of 0.18, placing them at the lower end.**

***Index Terms*— Explainable machine learning (ML), naturalness index, pattern recognition, remote sensing, uncertainty quantification.**

I. Introduction

PROTECTED natural areas are regions of the Earth that have remained largely untouched by significant human intervention, such as urbanization, agriculture, and other human activities [1]. These areas are characterized by their preserved and authentic state, and they boast high levels of biodiversity and provide numerous ecological benefits. They also offer unique opportunities to study natural ecosystem processes, including water and pollination cycles, in their unaltered form. To maintain the authenticity of these areas, it is crucial to conduct careful and comprehensive mapping and monitoring efforts. One obstacle, however, is the vague definition of the land cover class, which prevents a comprehensive mapping. Nevertheless, these efforts help reveal the complex geo-ecological patterns essential for preserving these regions' naturalness. As a result, the monitoring and understanding of natural areas have gained significant attention in both remote sensing and environmental research fields recently [2], [3]. Satellite imagery emerges as an effective technique for consistently observing vast protected natural expanses, which can be challenging for humans to access. This technology enables efficient and cost-effective data collection while minimizing disturbances to delicate ecosystems. By using machine-learning (ML) models, specifically convolutional neural networks (CNNs), it becomes possible to accurately classify natural regions by analyzing satellite imagery datasets.

In previous studies on naturalness analysis, such as [4] and [5], the quantification of naturalness has been done by defining a feature-engineered modern human footprint index within a predefined range of 0–10. These studies use proxies like railways, electric power structures, and population density to assess the impact of human activity. Our approach distinguishes itself by specifically focusing on naturalness as the primary concept of interest and assessing and explaining the underlying patterns that contribute to it. With a similar goal, though a different approach, explanatory frameworks designed by [6], [7] generate attribution maps that effectively highlight patterns indicative of protected natural areas in satellite imagery. However, despite the effectiveness of these methods in identifying natural regions using specific indices or distinct patterns characterizing natural regions, they struggle to provide a quantitative metric that accurately reflects the patterns' importance, while also considering their certainty.

To address these limitations, we propose a framework that extracts patterns that contribute to the concept of naturalness in satellite imagery and assesses the importance and certainty of these patterns with a novel metric called confident naturalness explanation (CNE). This metric combines existing tools for explainability and uncertainty quantification that enable

Manuscript received 1 September 2023; revised 20 October 2023 and 29 January 2024; accepted 8 February 2024. Date of publication 12 February 2024; date of current version 26 February 2024. This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant RO 4839/5-1, Grant SCHM 3322/4-1, Grant 458156377-MapinWild, and Grant RO 4839/6-1-459376902-AID4Crops; and in part by the DFG's Excellence Strategy under Grant EXC-2070-390732324-PhenoRob. *(Corresponding author: Ahmed Emam.)*

Ahmed Emam and Mohamed Farag are with the Institute of Geodesy and Geoinformation, University of Bonn, 53113 Bonn, Germany (e-mail: aemam@uni-bonn.de; mibrahi2@uni-bonn.de).

Ribana Roscher is with the Institute of Bio- and Geosciences, Forschungszentrum Jülich, 52428 Jülich, Germany, and also with the Institute of Geodesy and Geoinformation, University of Bonn, 53113 Bonn, Germany (e-mail: r.roscher@fz-juelich.de).

Digital Object Identifier 10.1109/LGRS.2024.3365196

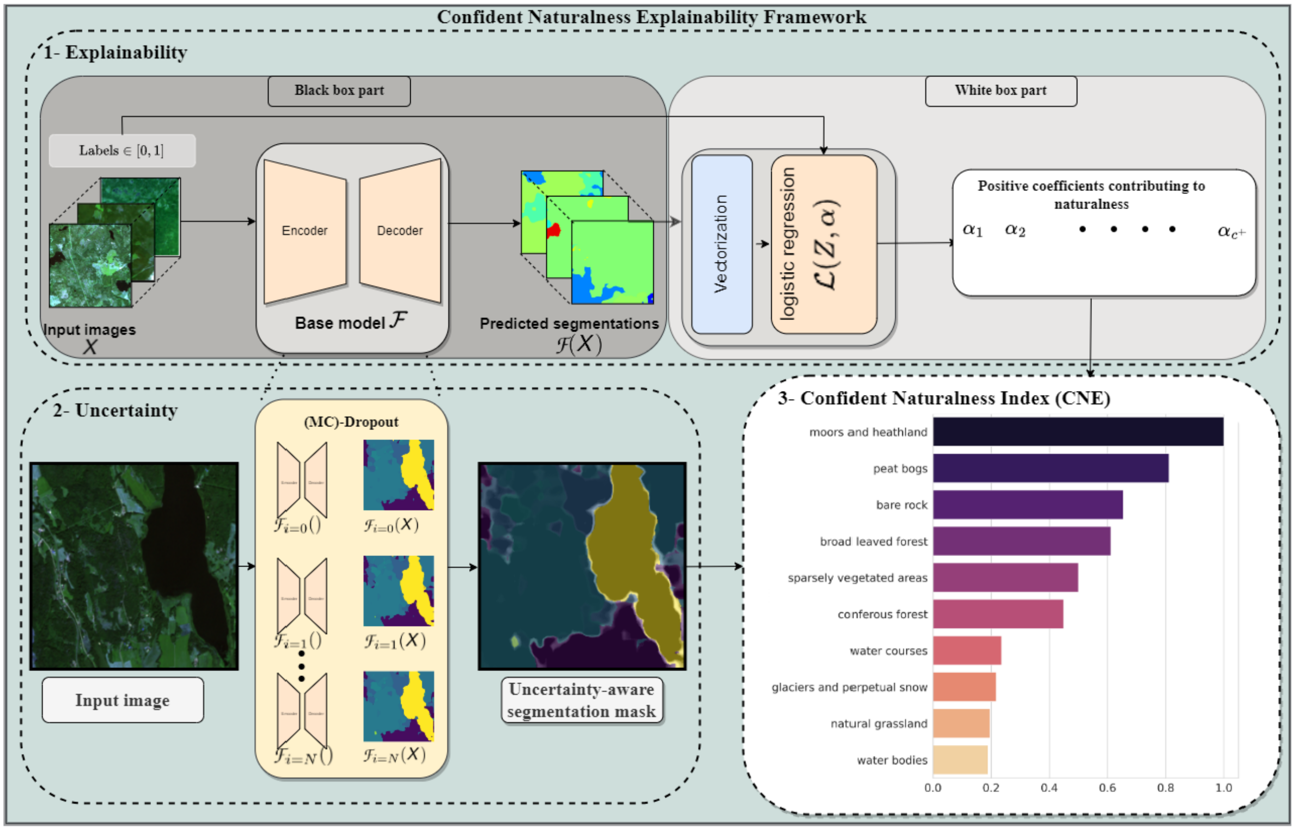
Confident Naturalness Explainability Framework
1- Explainability
Black box part
White box part
Labels ∈ [0, 1]
Encoder
Decoder
Base model $\mathcal{F}$
Input images X
Predicted segmentations $\mathcal{F}(X)$
Vectorization
logistic regression
$\mathcal{L}(Z, \alpha)$
Positive coefficients contributing to naturalness
α_1 α_2 • • • • α_{c^+}
2- Uncertainty
(MC)-Dropout
$\mathcal{F}_{i=0}()$ $\mathcal{F}_{i=0}(X)$
$\mathcal{F}_{i=1}()$ $\mathcal{F}_{i=1}(X)$
$\mathcal{F}_{i=N}()$ $\mathcal{F}_{i=N}(X)$
Input image
Uncertainty-aware segmentation mask
3- Confident Naturalness Index (CNE)
moors and heathland
peat bogs
bare rock
broad leaved forest
sparsely vegetated areas
conferous forest
water courses
glaciers and perpetual snow
natural grassland
water bodies
0.0
0.2
0.4
0.6
0.8
1.0

Fig. 1. **Illustration for the CNE framework.** In the explainability part, the input images are fed to the segmentation model, resulting in predicted segmentation masks; they are fed to logistic regression with ground-truth labels of the input images to vectorize patterns and classify the input into naturalness and anthropogenic areas. After training the logistic regression, we use only the positive coefficients to calculate the CNE metric. In the uncertainty, the MC-Dropout resembles multiple sampled models used to quantify the uncertainty of each pattern in the input image. In the lower right corner, the knowledge gained from parts 1 and 2 is combined to calculate the CNE metric in part 3 and assign a quantifiable metric value to each pattern, reflecting its confident contribution to the concept of naturalness. The uncertainty part is shown in detail in Fig. 2.

prioritization and sorting of the contributing patterns based on their quality. Our main contributions are as follows.

1) We demonstrate our CNE framework using a study site in Fennoscandia. We extract patterns associated with the concept of naturalness and protected areas from the AnthroProtect dataset [6] by using domain knowledge from the comprehensive, well-known, and well-understood CORINE dataset [8].
2) We generate uncertainty-aware segmentation masks to highlight the pixels where the model exhibits the lowest certainty.
3) We compute CNE, merging both explainability and uncertainty to show the significance of each pattern to the concept of naturalness.

Here, we use the term “pattern” for a land cover class. However, our framework is versatile in this regard, and the term can be used more broadly, for example, for a temporal or spatial regularity in a dataset, as already proposed in [4].

II. CNE Framework

In the following, the CNE framework (Fig. 1) is presented in the context of our study site in Fennoscandia.

A. Study Site and Datasets

The goal of the CNE framework is to extract patterns from a dataset and explain them by utilizing domain knowledge from another source of information. Here, we intend to find patterns associated with the concept of naturalness in the coarse AnthroProtect dataset [6] with the help of the more nuanced, well-understood CORINE dataset [8]. The AnthroProtect dataset comprises 24 000 multispectral Sentinel-2 images of the 256 × 256 pixels in the Fennoscandia region. In our study, we focus on the red, green, and blue bands. The reference images are either classified as protected [WDPA categories: “strict nature reserve” (Ia), “wilderness” (Ib), “national park” (II)] or anthropogenic areas. The protected areas are used as a proxy for naturalness; by doing so, we aim to establish a foundation for understanding and explaining naturalness, starting with these well-protected and minimally influenced environments. Anthropogenic areas in the AnthroProtect dataset align closely with “artificial surfaces” and “agricultural areas” in the CORINE land cover classes (patterns), which means that both datasets are consistent. CORINE dataset’s unique composition captures the broad spectrum of naturalness in Fennoscandia. The original ($1 \times H \times W$)-dimensional label masks are converted to a $C \times H \times W$ one-hot-encoded mask, where $H \times W$ is the size of the image, and C is set to 44, creating a channel for each class from the original label. This results in 1 and 0 s in these channels, with the rest as 0 s.

B. Components of the Framework

The CNE framework shown in Fig. 1 consists of three parts, detailed below. In the first part, the explainability part,

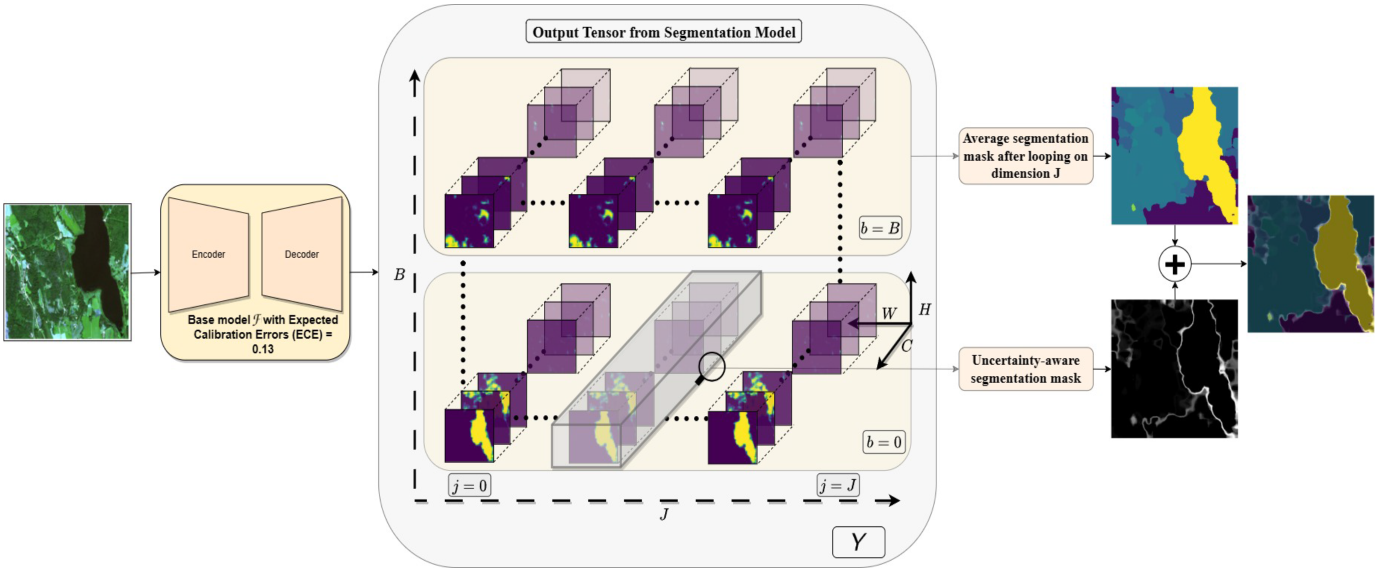
Output Tensor from Segmentation Model
Encoder
Decoder
Base model ℱ with Expected Calibration Errors (ECE) = 0.13
B
b = B
b = 0
W
H
C
j = 0
j = J
J
Y
Average segmentation mask after looping on dimension J
Uncertainty-aware segmentation mask
+

Fig. 2. **Illustrative diagram for the tensor generated by the segmentation model.** The input images are passed to the model, and the middlebox includes the tensor Y where b is the image index at a single batch, and j is the number of sampled models. On the right, we have two outputs. The upper image shows the output after taking the average over dimension J and assigning each pixel to the class with the highest probability, and at the bottom, an uncertainty-aware segmentation mask is generated by getting the standard deviation across MC runs where high-intensity pixels—white—represent high uncertainty and vice versa.

we train a black-box semantic segmentation model with the CORINE land cover segmentation masks. We assign an importance value to predefined patterns by learning a white-box logistic regression model that uses as input the segmentation model output and the AnthroProtect labels. Both together build a gray-box approach that is inspired by [9]. In the second part, the Monte Carlo (MC)-Dropout technique is used to quantify the uncertainty in predicting the patterns contributing to naturalness. The gained knowledge is integrated to create the CNE metric in the third part, which assigns confident importance to the patterns forming naturalness in Fennoscandia.

1) Explainability: The first part of the gray box approach is a black-box semantic segmentation model $\mathcal{F}$. In this work, we utilize the DeepLabV3 based on the ResNet-50 backbone [10]. It efficiently captures multiscale contextual information, enabling precise object boundary delineation. DeepLabV3 uses varying dilation rates to comprehend intricate image details and context through its spatial pyramid pooling module, resulting in highly accurate and fine-grained segmentation outcomes [10]. The architecture has a dropout layer localized after the last fully connected layers and before the output layer. The layer in our architecture serves a dual purpose: It contributes to regularization during training and plays a crucial role in uncertainty quantification. By employing MC-Dropout [11], we use the dropout layer to evaluate prediction uncertainty, offering insights into the model's confidence. The first prime component of the gray box resembles as

$$\mathcal{F}(X) = Y. \tag{1}$$

In our case, the output $Y^{B\times J\times C\times H\times W}$ is a 5-D tensor, with B being the number of images, H and W being the width and height of each image, C being the number of classes, and J the number of MC-Dropout runs as shown in Fig. 2. For simplicity, we omit the indices b for the image j for the MC dropout run for further consideration unless we consider multiple images and runs. Each predicted segmentation mask Y comprises C binary masks, one for each class, indicating which pixels are classified as class c. The segmentation mask is transformed into a $C \times 1$-D vector of patterns z, where each element in the vector represents the abundance of a specific pattern in the predicted segmentation maps. The segmentation mask vectorization is described as follows:

$$z_c = \sum_{h,w} Y_{w,h,c} \tag{2}$$

which means that all pixels that are predicted as class c in each binary segmentation mask are summed.

In the second part of the gray box approach, we employ logistic regression, known for its high interpretability and alignment with algorithmic transparency criteria. It optimizes coefficients to classify predicted segmentation masks as "naturalness" or "anthropogenic." It is worth noting that the gray box approach may sacrifice some spatial information, but this tradeoff serves our primary goal of achieving explainable results.

For this, a logistic regression classification model $\mathcal{L}(z;\boldsymbol{\alpha})$ is used

$$\mathcal{L}(z;\boldsymbol{\alpha}) = \frac{1}{1+e^{-\alpha^{\top}z}} \tag{3}$$

with $\boldsymbol{\alpha}$ being a C-dimensional vector where each coefficient is assigned to one class c.

2) Uncertainty Quantification: Improving models by providing additional information about outputs to increase confidence plays an important role in uncertainty estimation. Models approximate the real world, creating two sources of uncertainty: *Epistemic* uncertainty (model uncertainty) and

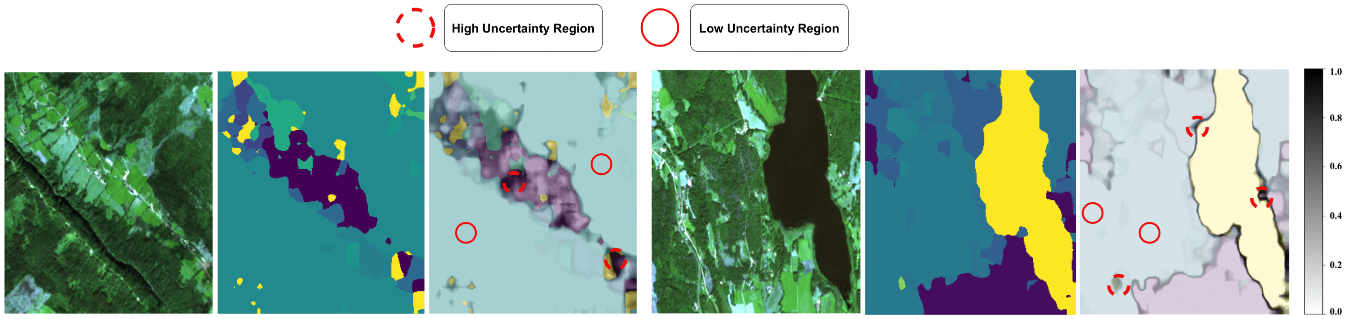

Fig. 3. **Qualitative results.** Demonstrating RGB Sentinel 2 images, predicted segmentation masks, and uncertainty-aware segmentation maps for two examples. The greyscale bar indicates pixel uncertainty in the segmentation maps. Each color in the segmentation masks represents a different pattern.

Aleatoric uncertainty (data uncertainty) [12]. Model uncertainty represents the lack of knowledge about the best model, while data uncertainty relates to the inherent stochastic component in the data-generating process. Predictive uncertainty is the combination of both. A well-known method that provides aleatoric and epistemic uncertainty estimates is Bayesian neural networks (BNNs) [13]. Also Softmax outputs can be used to assess uncertainty in a fast way, but they provide only the aleatoric part. Set-valued predictive techniques like conformal prediction (CP) [14], [15], [16] are used for estimating the overall predictive uncertainty. In deep learning (DL), the common framework is to obtain a point estimate of the model's weights. However, there is a growing need to estimate the model's knowledge, or epistemic uncertainty, which has led to the development of methods for addressing this issue. BNNs are used to generate weight distributions that capture the model's uncertainty. Nevertheless, due to the complexity of obtaining such distributions, approximate techniques such as MC-Dropout and ensemble learning [17] are widely adopted.

Dropout is a regularization approach utilized in DL to reduce overfitting [18]. During training at each forward pass j, dropout randomly deactivates some of the neurons controlled by a hyperparameter p_{drop}, which represents the fraction of neurons to switch off to force the DL model to have distributed representations. During inference, the dropout layer is switched off to avoid getting stochastic outputs.

MC-Dropout exploits the previous idea to estimate the epistemic uncertainty by allowing the model to get a stochastic output during multiple forward runs $j = 1, \ldots, J$ at the inference phase. The mean prediction of each pattern c for different model samples $\mathcal{F}_j$ is A_c for a single image X and obtained as

$$A_c = \frac{1}{J}\sum_{j=1}^{J} Y_{c,j}. \tag{4}$$

Furthermore, the standard deviations S can be analyzed to check the change in pixel-level values generated per class

$$S_c = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(Y_{c,j} - A_c)^2}. \tag{5}$$

3) CNE Metric: We combine the knowledge gained in both the explainability and the uncertainty quantification parts and propose the CNE metric, which assesses the quality of explanations of the patterns contributing to the concept of naturalness in Fennoscandia. The metric is bounded between 0 and 1, where the pattern that has a value of one contributes significantly to the concept of naturalness with high certainty, and a pattern that has a value of zero will either have a significantly low contribution to the concept of naturalness or high uncertainty.

The metric is calculated as follows:

$$\text{CNE}_c = \frac{\boldsymbol{\alpha}_{c^+}}{u_c} \tag{6}$$

with

$$\begin{aligned} \boldsymbol{\alpha}_{c^+} &= \max(\boldsymbol{\alpha}_c, \mathbf{0}) \\ u_c &= \sum_{h,w} S_c. \end{aligned}$$

The term α_{c^+} represents the modified trained logistic regressor's coefficient at which negative values are set to zero. The max function ensures that only positive coefficients are retained to include patterns that positively contribute to the naturalness concepts to ensure valid and objective explanations of naturalness. At the same time, u is the summation of the patterns' pixels across the spatial dimensions after taking the standard deviation over the MC-dropout dimension J.

III. Experiments, Results, and Discussions

A. Experimental Setup

In DeepLabV3 [10], a dropout layer is placed after the last fully connected layers with $p_{\text{drop}} = 0.1$. The model achieved 91.2% IOU on the training set and 80.3% on the test set after 100 epochs using 80% of the AnthroProtect dataset.

B. Results Interpretation

To produce a single number representing the uncertainty for each class, we take the standard deviation of the output Y across the MC-Dropout (25 forward iterations or models) dimension J. The generated output S_c is a 3-D tensor that contains the uncertainty-aware segmentation map for each class for a single test sample as shown in Fig. 3. We further sum over the spatial dimensions H and W to obtain a single value for each class. We used the maximum and minimum CNE values to normalize the metric between 0 and 1, as shown in Section II-A. The standard deviation will be zero if there

TABLE I

CNE Metric Values. Each Pattern Has a Value Representing Its Confident Contribution to the Concept of Naturalness. We Exclude Nondesignating Patterns With Importance Coefficients Below 0.01 to Enhance the Interpretability of the Results

Pattern	Metric	Distribution%
moors and heathland	1.00	13.2
peat bogs	0.81	6.1
bare rock	0.65	6.8
broad-leaved forest	0.61	13.4
sparsely vegetated areas	0.49	24.1
coniferous forest	0.44	18.2
watercourses	0.23	0.2
glaciers and perpetual snow	0.21	1.1
natural grassland	0.19	0.05
water bodies	0.18	5.2

are no discrepancies among predictions of the sampled models, which will generate a small value after summation. Conversely, a nonzero standard deviation will indicate high uncertainty and produce a larger summation. Furthermore, we achieved expected calibration error (ECE) [19] value of 0.1317.

As illustrated in Table I, our investigation unveiled that various wetland patterns possess notably high CNE metric values, ranging from 0.8 to 1. These scores signify the existence of top-tier patterns that significantly boost the concept of naturalness. Wetlands are important ecosystems renowned for their roles in carbon storage, safeguarding biodiversity, regulating water resources, and providing niches for unique plant and animal species finely adapted to their specific surroundings.

In contrast, glaciers, grasslands, and water bodies exhibit relatively low-quality patterns, with an approximate metric value of 0.2. These values indicate patterns with a diminished contribution to the naturalness concept, accompanied by heightened uncertainty.

IV. Conclusion

We utilize ML models to analyze satellite imagery, focusing on understanding naturalness. Our novel approach combines explainable ML and uncertainty quantification to provide comprehensive and valid explanations for intricate natural patterns, addressing the limitations of existing methods. Our framework, CNE, offers a quantitative metric and certainty-aware segmentation masks, transforming the understanding of naturalness in Fennoscandia by delivering objective, valid, and quantifiable explanations. In addition, our proposed approach holds promising potential for enhancing protected area preservation and monitoring efforts. Our results provide a quantifiable index, ranging from 0 to 1, for assessing pattern importance in the context of wilderness. This approach ensures validity by encompassing all distinctive patterns. At the same time, objectivity is maintained through the use of logistic regression coefficients, mitigating hand-engineered features and potential subjectivity and entanglement associated with previous indices and heatmap-based methods.

References

[1] J. S. Sze, L. R. Carrasco, D. Childs, and D. P. Edwards, “Reduced deforestation and degradation in indigenous lands pan-tropically,” *Nature Sustainability*, vol. 5, no. 2, pp. 123–130, Nov. 2021.

[2] R. A. Mittermeier et al., “Wilderness and biodiversity conservation,” *Proc. Nat. Acad. Sci. USA*, vol. 100, pp. 10309–10313, Sep. 2003.

[3] R. J. Smith and A. N. Gray, “Strategic monitoring informs wilderness management and socioecological benefits,” *Conservation Sci. Pract.*, vol. 3, no. 9, p. e482, Sep. 2021, doi: 10.1111/csp2.482.

[4] E. W. Sanderson, M. Jaiteh, M. A. Levy, K. H. Redford, A. V. Wannebo, and G. Woolmer, “The human footprint and the last of the wild,” *BioScience*, vol. 52, pp. 891–904, Oct. 2002.

[5] B. Ekim, Z. Dong, D. Rashkovetsky, and M. Schmitt, “The naturalness index for the identification of natural areas on regional scale,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, Dec. 2021, Art. no. 102622.

[6] T. T. Stomberg, T. Stone, J. Leonhardt, I. Weber, and R. Roscher, “Exploring wilderness characteristics using explainable machine learning in satellite imagery,” 2022, *arXiv:2203.00379*.

[7] A. Emam, T. T. Stomberg, and R. Roscher, “Leveraging activation maximization and generative adversarial training to recognize and explain patterns in natural areas in satellite imagery,” *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024, Art. no. 8500105, doi: 10.1109/LGRS.2023.3335473.

[8] *CORINE Land Cover 2018 Europe 6-Yearly—Version 2020_20u1*, Copenhagen, Denmark, 2019.

[9] A. Bennetot, G. Franchi, J. Del Ser, R. Chatila, and N. Diaz-Rodriguez, “Greybox XAI: A neural-symbolic learning framework to produce interpretable predictions for image classification,” 2022, *arXiv:2209.14974*.

[10] L. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, *arXiv:1706.05587*.

[11] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” 2015, *arXiv:1506.02142*.

[12] E. Hüllermeier and W. Waegeman, “Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods,” *Mach. Learn.*, vol. 110, no. 3, pp. 457–506, Mar. 2021.

[13] H. Wang and D.-Y. Yeung, “A survey on Bayesian deep learning,” *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–37, Sep. 2020.

[14] G. S. V. Vovk and A. Gammerman, *Algorithmic Learning in a Random World*, 1st ed. New York, NY, USA: Springer, Dec. 2005.

[15] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman, “Inductive confidence machines for regression,” in *Machine Learning*, T. Elomaa, H. Mannila, and H. Toivonen, Eds. Berlin, Germany: Springer, 2002, pp. 345–356.

[16] J. Lei and L. Wasserman, “Distribution-free prediction bands for nonparametric regression,” *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 76, no. 1, pp. 71–96, Jan. 2014.

[17] L. Hoffmann and C. Elster, “Deep ensembles from a Bayesian perspective,” 2021 *arXiv:2105.13283*.

[18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015, *arXiv:1512.03385*.

[19] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” 2017, *arXiv:1706.04599*.