Published: 2023-09-16

Machine Learning Techniques for Environmental Sustainability Score Prediction

Layla Evans, Leah Morris, Leo Stewart

A novel research paper on hybrid multi-modal learning for automated environmental assessment

Abstract

This paper introduces a novel, hybrid machine learning framework for predicting comprehensive Environmental Sustainability Scores (ESS) for urban and industrial entities, a task traditionally reliant on manual, resource-intensive audits. Departing from conventional single-model approaches or purely economic-environmental metrics, our methodology uniquely integrates three disparate data modalities: high-resolution, multi-spectral satellite imagery for land-use and vegetation analysis; unstructured textual data from corporate sustainability reports and regulatory filings, processed via a custom domain-adapted Natural Language Processing (NLP) pipeline; and structured time-series data on resource consumption (energy, water, waste). The core innovation lies in a Hierarchical Attention Fusion Network (HAFN), a bespoke neural architecture that dynamically learns and weights the contribution of each data modality, mimicking a human expert's integrative assessment. We formulate the prediction not as a simple regression but as a structured output learning problem, simultaneously predicting the overall ESS and its constituent sub-scores (e.g., carbon efficiency, biodiversity impact, circular economy adherence). Trained and validated on a newly curated dataset of 5,000 global entities, our model achieves a mean absolute error (MAE) of 4.2 points on a 0-100 ESS scale, significantly outperforming benchmark models like Gradient Boosting (MAE: 7.8) and standard Multi-Layer Perceptrons (MAE: 9.1). More importantly, the HAFN's attention mechanisms provide unprecedented, actionable interpretability, identifying, for instance, that for manufacturing sectors, satellite-derived green space metrics are the dominant predictive feature, whereas for financial services, the sentiment and specificity of disclosure in narrative reports are paramount. This research demonstrates that a consciously designed, multi-modal ML system can transcend traditional analytical silos, offering a scalable, transparent, and highly accurate tool for automated sustainability assessment, with profound implications for investors, regulators, and policymakers seeking to accelerate the transition to a sustainable economy.

Keywords: Environmental Sustainability Score, Multi-modal Machine Learning, Hierarchical Attention Fusion, Satellite Imagery Analysis, Natural Language Processing, Interpretable AI, Structured Output Prediction.

1 Introduction

The imperative for robust, transparent, and scalable assessment of environmental sustainability has never been more critical. Stakeholders ranging from global investors and regulatory bodies to municipal planners and consumers demand reliable metrics to guide capital allocation, policy formulation, and operational decisions. Traditional methods for deriving such metrics, most notably the Environmental Sustainability Score (ESS), are predominantly manual, relying on expert audits, self-reported questionnaires, and the aggregation of disparate quantitative indicators. These processes are not only labor-intensive and costly but also prone to inconsistencies, subjectivity, and significant time lags, rendering them inadequate for the dynamic, data-rich landscape of the 21st century. The central research question this paper addresses is: Can a purpose-built machine learning system, integrating heterogeneous and unconventional data sources, accurately and interpretably predict a holistic ESS, thereby automating and enhancing a process fundamental to sustainable development?

Existing computational approaches have largely operated within narrow silos. Econometric models focus on the relationship between financial performance and a limited set of environmental metrics, such as carbon emissions per revenue unit. Remote sensing applications excel at monitoring specific physical phenomena like deforestation or urban heat islands but rarely integrate this data with socio-economic or narrative context. Recent forays into using machine learning for sustainability have often applied off-the-shelf algorithms like random forests or support vector machines to structured tabular data, missing the rich informational tapestry contained in imagery and text. This constitutes a significant gap: the lack of a holistic, automated framework that mimics the integrative reasoning of a sustainability expert by synthesizing visual evidence of environmental impact, narrative disclosure of policies and performance, and hard numerical data on resource flows.

Our work presents a fundamental departure from these trajectories. We propose that accurate ESS prediction is an inherently multi-modal problem, requiring the fusion of geospatial, linguistic, and numerical data streams. The novelty of our contribution is threefold. First, we introduce a new, curated dataset that aligns high-resolution satellite imagery, the full text of sustainability reports, and time-series operational data for a diverse set of 5,000 entities. Second, we architect and implement the Hierarchical Attention Fusion Network (HAFN), a novel neural model that does not merely concatenate features from different modalities but employs a hierarchical attention mechanism to learn context-dependent importance weights for each data type and even within specific segments of the data (e.g., specific sentences in a report, specific spectral bands in an image). Third, we reformulate the prediction task itself. Instead of a single-target regression, we frame it as structured output learning, where the model predicts a vector comprising the overall ESS and its key sub-component scores. This forces the model to learn the internal structure of the sustainability construct, improving both accuracy and the granularity of interpretability.

The subsequent sections detail this innovative approach. The Methodology section elaborates on the data curation process, the architecture of the HAFN, and our training paradigm. The Results section presents a comprehensive quantitative evaluation against strong benchmarks and, crucially, a qualitative analysis of the model's interpretability outputs, demonstrating how it identifies sector-specific drivers of sustainability performance. The Conclusion discusses the broader implications of this research for the fields of environmental informatics and machine learning, arguing for the necessity of such cross-disciplinary, integrative AI systems in addressing complex societal challenges.

2 Methodology

The methodological core of this research is the design and implementation of an end-to-end machine learning pipeline for multi-modal Environmental Sustainability Score prediction. This process encompasses three major phases: the construction of a novel, aligned multi-modal dataset; the development of specialized feature extractors for each data modality; and the synthesis of these features through the novel Hierarchical Attention Fusion Network (HAFN) for structured output prediction.

2.1 Data Curation and Preprocessing

A significant contribution of this work is the assembly of the Multi-modal Environmental Assessment Dataset (MEAD). MEAD comprises 5,000 entities, including corporations, municipal districts, and industrial facilities, globally distributed. For each entity, we collected three synchronized data streams corresponding to the same fiscal year. The first stream consists of high-resolution (0.5m/pixel) multi-spectral satellite imagery from the QuickBird and IKONOS constellations, covering the entity's primary operational footprint. These images were processed to extract beyond-RGB features, including Normalized Difference Vegetation Index (NDVI) for green cover, Normalized Difference Water Index (NDWI) for water body presence, and texture metrics for land-use classification, resulting in a 512-dimensional feature vector per entity.

The second stream is unstructured text, comprising the complete publicly available sustainability reports, annual reports (environmental sections), and relevant regulatory filings (e.g., EPA TRI reports in the U.S.). A custom NLP pipeline was developed. It begins with domain-specific pre-processing, expanding acronyms common in sustainability discourse (e.g., ESG, CSR, GHG) and lemmatizing using a lexicon enriched with environmental terms. Subsequently, we employed a two-tier feature extraction strategy. A Bag-of-Concepts model, built upon a custom ontology derived from the Global Reporting Initiative (GRI) standards, captures the presence and frequency of key sustainability topics. Concurrently, a fine-tuned DistilBERT model, pre-trained on general corpora and further trained on a corpus of sustainability literature, generates dense document embeddings that capture semantic nuance and narrative tone, yielding a 768-dimensional vector.

The third stream is structured, time-series numerical data. This includes 36 months of monthly records for energy consumption (by source), water withdrawal and discharge, waste generation (by type), and relevant production output metrics. These series were transformed into a fixed-length feature vector by extracting statistical descriptors (mean, trend, volatility) and engineering efficiency ratios (e.g., energy use per unit output), resulting in a 120-dimensional vector. The ground-truth ESS and its eight sub-scores (Air Quality, Water Stewardship, Waste Circularity, Biodiversity, Carbon Management, Environmental Compliance, Innovation, and Transparency) were obtained from the independent auditing consortium SustainAudit International, providing a rigorous, expert-validated target for supervised learning.

2.2 The Hierarchical Attention Fusion Network (HAFN) Architecture

The HAFN is designed to address the key challenge of dynamically and interpretably combining the three heterogeneous feature vectors: image features $\mathbf{v}_i$, text features $\mathbf{v}_t$, and numerical features $\mathbf{v}_n$.

The architecture operates in three stages. In the first, *Modality-Specific Encoding*, each feature vector is passed through a dedicated fully-connected neural network layer with ReLU activation, projecting them into a common latent space of dimension $d = 256$: $\mathbf{h}_i = \text{ReLU}(\mathbf{W}_i\mathbf{v}_i + \mathbf{b}_i)$, and similarly for $\mathbf{h}_t$ and $\mathbf{h}_n$.

The second stage is the core *Hierarchical Attention Mechanism.* It computes attention in two levels. At the first level (Modality Attention), the network learns the importance of each modality for the given entity. A context vector $\mathbf{c}$, initialized as a trainable parameter, is used to score each modality's encoded representation:

$$e_m = \mathbf{c}^T \tanh(\mathbf{W}_a\mathbf{h}_m + \mathbf{b}_a), \quad m \in \{i, t, n\}$$
$$\alpha_m = \frac{\exp(e_m)}{\exp(e_i) + \exp(e_t) + \exp(e_n)}$$

The modality-aware representation is then the weighted sum: $\mathbf{s} = \alpha_i\mathbf{h}_i + \alpha_t\mathbf{h}_t + \alpha_n\mathbf{h}_n$.

At the second level (Intra-Modality Attention), applied specifically to the text modality due to its sequential nature, the model learns to attend to the most relevant sections of the processed text features. The text encoder output is treated as a sequence of concept embeddings. A separate attention layer computes weights β_j for each concept j, based on its contribution to the fused context $\mathbf{s}$, allowing the model to, for example, focus on sentences discussing carbon reduction targets over generic governance statements.

The third stage is *Structured Output Prediction.* The fused and contextually weighted representation $\mathbf{s}$ is fed into a multi-head output layer. One head, a linear layer, predicts the overall ESS, $\hat{y}_{\text{total}} \in \mathbb{R}$. Eight additional heads, also linear layers, predict each of the sub-scores concurrently, $\hat{\mathbf{y}}_{\text{sub}} \in \mathbb{R}^8$. The total loss function $\mathcal{L}$ is a composite of the Mean Squared Error (MSE) for the total score and the sub-scores, encouraging the model to learn the inter-dependencies within the sustainability construct:

$$\mathcal{L} = \text{MSE}(y_{\text{total}}, \hat{y}_{\text{total}}) + \lambda \sum_{k=1}^{8} \text{MSE}(y_{\text{sub}}^k, \hat{y}_{\text{sub}}^k)$$

where λ is a weighting hyperparameter. The model was trained using the Adam optimizer with a learning rate of 0.001 over 500 epochs, with early stopping based on a held-out validation set.

3 Results

The proposed HAFN model was evaluated extensively against several strong benchmark models and subjected to rigorous interpretability analysis to validate its novel approach to ESS prediction.

3.1 Quantitative Performance

We compared the HAFN to four benchmark models: (1) a Gradient Boosting Regressor (GBR) operating on hand-engineered features from all modalities (concatenated); (2) a standard Multi-Layer Perceptron (MLP) with two hidden layers on the same concatenated features; (3) a Late Fusion model, which trains separate MLPs on each modality and averages their predictions; and (4) a Simple Attention Fusion network, a baseline version of HAFN without the hierarchical (intra-modality) attention. The dataset was split into training (60%), validation (20%), and test (20%) sets, ensuring no entity leakage. The primary evaluation metric was Mean Absolute Error (MAE) on the overall ESS (0-100 scale).

The results, summarized in Table 1, demonstrate the clear superiority of the HAFN. It achieves an MAE of 4.23, a 39% improvement over the strong GBR baseline (MAE: 7.82) and a 22% improvement over the Simple Attention Fusion model (MAE: 5.41). The high R^2 score of 0.901 indicates that the model explains over 90% of the variance in the expert-assigned scores. Furthermore, the HAFN also achieved the lowest average MAE across all eight sub-scores (5.87),

Table 1: Model Performance Comparison on ESS Prediction (Test Set)

Model	**MAE (ESS)**	**R^2 Score**
Gradient Boosting Regressor (GBR)	7.82	0.741
Multi-Layer Perceptron (MLP)	9.15	0.682
Late Fusion (Averaging)	6.93	0.768
Simple Attention Fusion	5.41	0.842
Hierarchical Attention Fusion Network (HAFN)	**4.23**	**0.901**

outperforming the next best model (Simple Attention Fusion, 7.12), confirming the benefit of the structured output learning approach.

3.2 Interpretability and Qualitative Analysis

The hierarchical attention mechanism provides a natural window into the model's decision-making process. Analyzing the learned modality attention weights $(\alpha_i, \alpha_t, \alpha_n)$ across different industry sectors reveals compelling, intuitive patterns. For entities in the *Manufacturing and Heavy Industry* sector, the image modality received the highest average attention weight ($\alpha_i = 0.52$), followed by numerical data ($\alpha_n = 0.31$), and then text ($\alpha_t = 0.17$). This aligns with expert intuition that the direct visual evidence of emissions (via thermal bands), site greening, and water body health is paramount for assessing such entities. In contrast, for the *Financial Services and Insurance* sector, the text modality dominated ($\alpha_t = 0.61$), with numerical data secondary ($\alpha_n = 0.28$) and imagery minimal ($\alpha_i = 0.11$). This suggests the model learns that for low-physical-footprint sectors, the quality, specificity, and commitment expressed in sustainability disclosures are the most reliable proxy for environmental performance.

Drilling deeper, the intra-modality text attention for a high-scoring technology company highlighted sentences detailing specific, quantitative targets for renewable energy procurement and e-waste recycling rates. For a poorly scoring utility, the attention focused on sections of its report that were flagged by our NLP pipeline as containing vague, non-committal language and a high density of generic sustainability buzzwords with little operational detail. This capacity to not only predict but also explain *why* a score was assigned, by pointing to influential data sources and even specific content within them, represents a significant advance in transparent, accountable AI for sustainability.

4 Conclusion

This research has presented a novel, integrative machine learning framework for predicting Environmental Sustainability Scores. By moving beyond traditional single-modality or simple ensemble approaches, we have demonstrated that a carefully architected system—the Hierarchical Attention Fusion Network (HAFN)—can effectively synthesize geospatial imagery, unstructured narrative, and structured numerical data to achieve a level of predictive accuracy that begins to approach the consistency of expert human audit teams. The key original contributions are the formulation of ESS prediction as a multi-modal, structured output learning problem; the introduction of the HAFN architecture with its dual-level attention mechanism for dynamic, interpretable fusion; and the creation and public release of the corresponding MEAD dataset to spur further research.

The implications are substantial. For practitioners in finance and investment, this tool offers a scalable, real-time alternative to costly third-party ratings, enabling more dynamic portfolio screening and engagement. For regulators, it provides a mechanism for continuous, evidence-based compliance monitoring across vast jurisdictions. For the entities themselves, the interpretability of the model offers a clear diagnostic: it identifies which aspects of their operational footprint, disclosure quality, or resource efficiency are most critically influencing their perceived sustainability performance.

Future work will focus on several exciting avenues. First, incorporating temporal dynamics more explicitly, treating the prediction as a sequence-to-sequence task to forecast future ESS trajectories. Second, expanding the data modalities to include social media sentiment, supply chain network data, and real-time sensor feeds from IoT devices. Third, exploring the use of this framework for "what-if" scenario analysis, allowing stakeholders to simulate the potential impact of specific interventions (e.g., adding a green roof, switching to renewable energy) on their predicted score. In conclusion, this research underscores the transformative potential of cross-disciplinary, intelligently designed machine learning in addressing one of the most complex and urgent challenges of our time: the accurate, fair, and scalable measurement of our progress toward environmental sustainability.

References

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, 19*(2), 263–311.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Elkington, J. (1997). *Cannibals with forks: The triple bottom line of 21st century business.* Capstone.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232.

Global Reporting Initiative (GRI). (2002). *Sustainability reporting guidelines.* GRI.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

LeCun, Y., Bengio, Y., & Hinton, G. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł, & Polosukhin, I. (2005). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.