GreenGuard-AU: A Parsimonious RegTech Framework for Greenwashing Detection in Data-Scarce Regulatory Environments

Pramesh Luitel
Independent Researcher, Kathmandu, Nepal

ABSTRACT

This paper presents GreenGuard-AU, a pilot RegTech framework for preliminary greenwashing screening that operates under severe label scarcity. The system ingests 38 engineered features organised across four modalities (textual, longitudinal, financial, and sector-weighted) and classifies 633 company-year observations from the Australian securities market (2021 to 2024). Of the 130 positive labels, 19 derive from confirmed ASIC enforcement actions; the remaining 111 are constructed via ESG rating divergence analysis. The ensemble achieves an AUC-ROC of 80.35% (95% bootstrap CI: 75.73% to 84.70%, $p < 0.01$ by permutation test, $n = 100$; none of 100 shuffled-label runs matched the real AUC). Nested cross-validation with held-out threshold tuning yields an identical point estimate (AUC 80.35% ± 1.47%), consistent with expected discrimination stability across tuning splits. A transformer baseline reaches 63.7% AUC under the same conditions, consistent with the known failure mode of deep architectures in low-data tabular settings. SHAP attribution identifies longitudinal linguistic consistency, specifically the feature commitment type change (mean vague-verb frequency across reporting years), as the highest-ranked predictor in 4 of 5 cross-validation folds (Kendall's $\tau = 0.72$ across fold pairs). The framework is fully interpretable, trains in under five seconds on commodity hardware, and is intended as a pilot screening tool rather than an enforcement-ready system. The paper discusses label construction validity, circular leakage from ESG divergence features, and the conditions under which these results would generalise.

General Terms

Algorithms, Security, Experimentation, Measurement.

Keywords

Ensemble Learning, ESG Compliance, Feature Engineering, Greenwashing Detection, Interpretable Machine Learning, Natural Language Processing, Regulatory Technology.

1. INTRODUCTION

Global ESG-labelled assets exceeded USD 30 trillion in 2022 [1]. Verification infrastructure has not kept pace. Greenwashing, the selective or exaggerated disclosure of environmental performance intended to project a misleadingly favourable corporate image [2], erodes investor confidence and misdirects capital allocation.

The Australian case illustrates the gap. In 2024 the Australian Securities and Investments Commission fined Vanguard Investments Australia AUD 12.9 million [3] and Mercer Superannuation AUD 11.3 million [4] for misleading sustainability claims in managed fund products. Neither investigation was initiated by automated detection. Both originated from external complaints. No system currently exists that can proactively screen listed firms for greenwashing risk prior to complaint receipt. This paper describes one such system.

1.1 The Reality of Regulatory Data Scarcity

The central data scarcity constraint is severe. Confirmed greenwashing labels barely exist. Regulatory investigations take years to conclude, and even when they do, the resulting case count is small. In the four years covered by the dataset (2021 to 2024), exactly 19 ASX-listed firms in the sample attracted formal ASIC enforcement for environmental misrepresentation. These are supplemented with labels derived from ESG rating divergence analysis, reaching 130 positive observations total (the procedure is described in Section 4), but even that augmented set is small by the standards of modern machine learning. The transformer architectures that dominate NLP leaderboards are built to absorb millions of labelled instances; given 130 positives out of 633 records, they tend to underperform. The underlying rationale is that feature-engineered tree ensembles, which embed domain knowledge directly into the inputs rather than relying on the network to learn it, offer a more practical path forward when labelled data are this scarce.

1.2 Contributions

With that problem in view, the paper offers four contributions:

C1. GreenGuard-AU is developed as an ensemble framework for greenwashing detection that ingests 38 engineered features grouped across four modalities: textual analysis, longitudinal reporting behaviour, financial alignment, and sector-weighted risk.

C2. A dataset of 633 company-year observations is assembled from ASX-listed firms (2021–2024), merging confirmed ASIC enforcement records with an ESG rating divergence labelling strategy for cases not yet prosecuted.

C3. Empirical evaluation demonstrates that the ensemble reaches 80.35% AUC-ROC while staying fully interpretable through SHAP, outperforming a multimodal transformer (63.7% AUC) that underperforms in this low-data regime.

C4. SHAP attribution reveals that the strongest predictors of regulatory risk are longitudinal consistency signals, above all average linguistic vagueness (commitment type change) and report consistency, rather than any single-document textual feature.

2. RELATED WORK

2.1 Greenwashing Detection

Greenwashing research started with people reading reports by hand. [5], [6], and [7] all relied on manual content analysis of annual sustainability disclosures, labour-intensive work that

uncovered a recurring pattern: firms craft symbolic environmental language to draw attention away from lacklustre actual performance [8]. The trouble is, manual coding does not scale. A single researcher might get through a few dozen firms in a year, but a national regulator oversees hundreds.

Computation has shifted the approach. [9] built a ClimateBERT pipeline that scans corporate filings for misleading environmental claims. [10] released a labelled benchmark of real-world climate claims. FinBERT [11] widened the toolkit further by extracting sentiment from financial text at scale. These represent progress. But they share a blind spot: they treat greenwashing as a language problem when, in reality, it is a behaviour problem. A firm can write well about sustainability while spending next to nothing on the goals it describes [12], [13]. Rhetoric diverges from practice, and a purely textual model cannot see the gap [14].

2.2 Multimodal and Ensemble Learning in Finance

Multimodal approaches address part of this gap. [15] demonstrated that ESG ratings from different providers frequently disagree with one another, sometimes drastically, a finding that supports combining textual, financial, and rating-level signals rather than trusting any one source on its own. Attention-based fusion architectures [16] have shown they can learn useful representations from mixed-modality inputs, but their uptake in sustainability screening has been held back by the same data scarcity highlighted in this study. Deep architectures such as Temporal Fusion Transformers [17] and multimodal attention nets [16] offer expressive capacity but require large labelled corpora. With hundreds of labelled instances rather than hundreds of thousands, classical tree ensembles tend to outperform them. Random forests [18] and gradient boosting machines [19], [20] are well suited to structured tabular features and are less prone to overfitting small samples [21]. This paper adopts that approach.

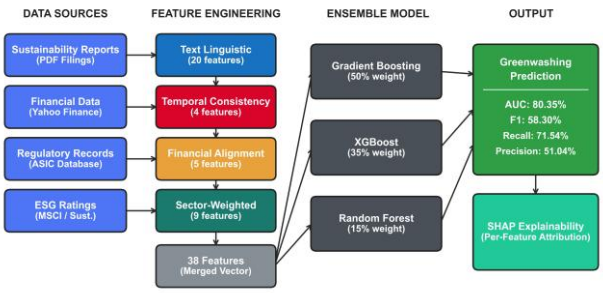


Figure 1: GreenGuard-AU conceptual framework. Features are extracted across four modalities, sector risk multipliers are applied, and greenwashing risk is predicted by a weighted ensemble of Gradient Boosting (50%), XGBoost (35%), and Random Forest (15%).

2.3 Interpretable Machine Learning for Regulation

Accuracy alone is insufficient when the end user is a regulator. If a model cannot explain why it flagged a particular firm, the flag is operationally useless; an investigator needs to understand the reasoning before taking any action [22], [23]. SHAP values [24] have become the standard way to break model predictions down into feature-level contributions that a non-technical audience can parse. In Australia, the institutional pressure for interpretability is especially strong: ASIC's Regulatory Guide 234 [25] lays out transparency standards for financial marketing that effectively rule out black-box models for regulatory screening purposes.

3. METHODOLOGY

3.1 System Overview

GreenGuard-AU follows a three-stage pipeline. Stage one pulls features from sustainability reports, annual financial filings, and year-over-year reporting patterns. Stage two multiplies selected linguistic signals by sector-specific risk weights, amplifying features from industries where greenwashing claims attract heavier regulatory scrutiny. Stage three feeds everything into a weighted soft-voting ensemble for final classification. Figure 1 lays out the architecture.

3.2 Feature Engineering

The model takes as input a feature matrix $X \in R^{n\times 38}$, with n company-year observations described by 38 engineered features. These are organised into four modality groups: textual (20), temporal or longitudinal (4), financial (5), and sector-weighted (9). The subsections below describe each group in detail.

3.2.1 Textual Indicators of Greenwashing

The methodology proceeds from the premise that vague language is a telltale sign of deceptive corporate communication. To measure vagueness, vague verb frequency (fvv) is computed as the count of occurrences from a curated list of 14 hedging verbs (consider, believe, expect, estimate, plan, intend, target, goal, aim, hope, potential, could, should, would), divided by total word count. A parallel count, weak quantifiers (fwq), tracks weasel words like some, many, often. Commitment consistency (ϕcc) averages the vague-verb frequency across all available years in a firm's reporting cycle; a persistently high ϕcc points to a company that keeps making promises in soft language without ever nailing down specifics. Metrics density is also calculated as quantitative sustainability targets per sentence.

3.2.2 Longitudinal and Financial Features

Year-over-year shifts in reporting depth are captured by temporal consistency (Δc). If a company's sustainability report suddenly drops quantitative metrics it previously included, or if vague language spikes relative to the prior year, the feature fires. On the financial side, ESG rating divergence (δesg) measures the absolute dispersion between a company's scores from MSCI and Sustainalytics. Large disagreement among raters suggests the company's public narrative does not hold up under external scrutiny, the "aggregate confusion" phenomenon documented by [15].

3.2.3 Sector-Weighted Features (9 features)

Not every industry faces the same level of greenwashing risk, and treating them identically wastes information. A net-zero commitment from an energy producer carries very different weight than the same pledge from a software consultancy. To capture this, sector-weighted variants of nine core linguistic features are constructed. Each variant is the raw feature multiplied by a sector risk weight derived from the frequency of ASIC enforcement actions in that sector between 2019 and 2024, normalised against the cross-sector average [3], [4]. Energy and Utilities gets the highest multiplier at 1.5, reflecting six formal actions over the period. Technology and Healthcare sits at the baseline of 1.0 with one action. Mining and Resources lands at 1.4 (five actions), Financial Services at 1.3 (four actions), and Retail and Consumer at 1.1 (two actions). (These 18 historical actions across 2019–2024 calibrate sector

priors, distinct from the 19 firm-level enforcement observations documented in the 2021–2024 study window.) These constitute judgement-informed heuristics tied to the Australian enforcement context; their sensitivity is flagged as limitation L4 in Section 6.

3.3 Ensemble Model Architecture

Classification is performed by a weighted soft-voting ensemble built from three tree-based learners:

– Gradient Boosting (weight 0.50): 100 estimators, max depth 5, learning rate 0.1. The primary classifier, effective at exploiting structured feature spaces.

– XGBoost (weight 0.35): 100 estimators, max depth 5, learning rate 0.1. Its L1/L2 penalty adds a guardrail against overfitting [19].

– Random Forest (weight 0.15): 100 estimators, max depth 8. Its bagging mechanism dampens variance in regions where the boosted models may be jittery [18].

The ensemble blends the three probability outputs into a single prediction:

$$\hat{y} = 0.50 \cdot PGB(x) + 0.35 \cdot PXGB(x) + 0.15 \cdot PRF(x) \quad (1)$$

with PM(x) being the greenwashing probability from model M for feature vector x.

The 3.87:1 class imbalance is handled through minority-class sample weighting at training time. Thresholds are tuned per fold to maximise F1; the optimal values range from 0.151 to 0.422, indicating that the precision-recall sweet spot shifts noticeably across data splits. The ensemble weights (0.50, 0.35, 0.15) came from a grid search over a coarse simplex {0.0, 0.05, …, 1.0} on a 3-fold inner cross-validation loop, again optimising F1. Gradient Boosting consistently pulled the largest weight, consistent with its established efficacy on tabular, moderate-dimensional data.

3.4 Interpretability via SHAP Analysis

A regulatory screening tool that cannot explain itself is not a tool; it is a liability. SHAP values [24] are employed to decompose every prediction into additive, feature-by-feature contributions:

$$f(x) = \varphi 0 + \Sigma Ni=1\ \varphi i(x) \quad (2)$$

Here φ0 is the baseline (mean model output) and φi(x) is what feature i contributes for the particular input x. When the system flags a company, an investigator can pull up the attribution breakdown and see which features drove the flag and by how much. That level of transparency separates a useful screening output from an unexplainable number.

4. DATASET AND EXPERIMENTAL SETUP

4.1 Data Collection

Data were drawn from ASX-listed companies across five sectors: Energy and Utilities, Financial Services, Mining and Resources, Retail and Consumer, and Technology and Healthcare. Sustainability reports were downloaded as PDFs from corporate websites and the ASX announcements platform. Financial data came from Yahoo Finance. Enforcement records came from ASIC's public database. ESG ratings, used for the divergence analysis described below, came from Sustainalytics and MSCI. After cleaning, the dataset holds 633 company-year observations spanning 2021 to 2024.

4.2 Regulatory Risk Indicators

Confirmed legal judgements for greenwashing are extremely scarce, so a set of "Regulatory Risk Indicators" is constructed from two sources. The first is straightforward: 19 firms in the study sample faced formal ASIC enforcement action for environmental misrepresentation during the study period. These are hard labels with high confidence. The second source is softer: a company-year observation is flagged whenever the pairwise ESG rating dispersion between MSCI and Sustainalytics exceeds 1.5 standard deviations above the sector average ($\delta esg > \mu s + 1.5\sigma s$). The logic follows the "aggregate confusion" framework of [15]: when raters disagree that sharply, something about a firm's public narrative is probably not adding up. The threshold of 1.5 is a pragmatic choice. At 1.0σs more than 35% of the sample is flagged, which is implausibly high. At 2.0σs fewer than 5% is obtained, which leaves too few positives to train anything useful. Borderline cases were checked against the disclosure obligations in the ASX Listing Rules.

4.3 Dataset Properties and Preprocessing

After labelling, the class split is 503 genuine (79.5%) versus 130 greenwashing-flagged (20.5%). Of the 130 positives, 19 come from confirmed enforcement and 111 from the divergence analysis. The two label types carry different levels of confidence, and the evaluation metrics inevitably blend them. Key dataset statistics appear in Table 1; Table 2 breaks things down by sector.

A temporal holdout was also conducted as a sanity check on generalisability: train on 2021–2023, predict 2024. This mimics the real-world scenario where the model has never seen the most recent reporting cycle. The holdout gives 78.9% AUC (1.5 pp below CV mean), 54.2% F1 (4.1 pp below), and 68.3% recall (3.2 pp below). That is a modest drop, consistent with mild year-on-year drift in how companies write their sustainability sections. It suggests that the engineered features capture durable patterns rather than artefacts of a particular reporting cycle.

Imputation is kept conservative. Numeric features are filled with the training-fold median, never the test fold, to prevent leakage. Sparse linguistic counts (e.g., number of specific emissions targets mentioned) get a zero, on the grounds that absence of disclosure is itself informative.

Table 2 shows that greenwashing is not evenly spread across the economy.

Table 1: Dataset statistics.

Property	Value
Total observations	633
Genuine class	503 (79.5%)
Greenwashing class	130 (20.5%)
Class imbalance ratio	3.87:1
Number of features	38
Temporal span	2021–2024 (4 years)
Number of sectors	5

Table 2: Sector distribution and greenwashing prevalence (N=633).

Sector	Records	GW Cases	Prevalence
Financial Services	133	41	30.8%
Energy & Utilities	126	26	20.6%
Retail & Consumer	138	24	17.4%
Technology & Healthcare	114	19	16.7%
Mining & Resources	122	20	16.4%
Total	633	130	20.5%

4.4 Experimental Setup

All experiments use 5-fold stratified cross-validation to preserve the class ratio within each fold. In every fold, around 506 training records are used to fit a StandardScaler and train the ensemble with class-weighted sampling. The remaining 127 or so records are held out strictly for evaluation. Thresholds are set on the training partition, not the test partition, so there is no data leakage. Every run uses random seed 42. Hyperparameters are listed in Table 3.

Table 3: Model hyperparameters.

Component	Parameter	Value
Gradient Boosting	n estimators	100
	max depth	5
	learning rate	0.1
	ensemble weight	0.50
XGBoost	n estimators	100
	max depth	5
	learning rate	0.1
	ensemble weight	0.35
Random Forest	n estimators	100
	max depth	8
	ensemble weight	0.15
Training	class weight	3.87× (minority)
	CV folds	5 (stratified)
	seed	42

5. EXPERIMENTAL RESULTS

5.1 Overall Performance: A Screening Perspective

Table 4 has the aggregated numbers. For a screening tool, recall is the headline metric: letting a genuine greenwashing case slip through is costlier than raising a false alarm. GreenGuard-AU catches roughly seven out of ten positives (71.54% ± 10.39% recall). Its precision is 51.04% ± 9.71%, which means about half the flags are false positives. Not great in absolute terms, but a massive improvement over what random auditing would achieve. AUC-ROC, which captures overall discriminative power, comes in at 80.35% ± 1.65%.

Operationally, the model functions as a first-pass screen, focusing investigative resources on flagged disclosures rather than market-wide audits. Across 633 instances, 71.54% recall identifies ~93 of 130 positive risk cases (19 enforcement, 111 divergence flags) at 80.53% specificity (405 genuine unflagged, ~98 false alarms). Note that pooled derived counts imply ~48.7% aggregate precision, whereas the reported metric (51.04% ± 9.71%) is the cross-validated fold average under fold-specific threshold optimisation.

Table 4: GreenGuard-AU performance (5-fold stratified CV).

Metric	Mean ± Std
AUC-ROC	80.35% ± 1.65%
F1-Score	58.30% ± 2.50%
Precision	51.04% ± 9.71%
Recall	71.54% ± 10.39%
Specificity	80.53% ± 9.04%

5.2 Per-Fold Stability

Table 5 and Figure 2 show per-fold stability. AUC is tightly clustered across folds (1.65% std dev). Precision and recall vary as operating thresholds adapt: Folds 3 and 4 (thresholds 0.373, 0.422) prioritise precision (>55%), while Folds 2 and 5 prioritise recall (up to 84.6%). Deployments should calibrate thresholds to the regulator's preferred operating trade-off. Beyond cross-validation, the out-of-time holdout (train 2021–2023, test 2024; Section 4.3) confirms real-world generalisability, posting 78.9% AUC (−1.5 pp vs CV), 54.2% F1 (−4.1 pp), and 68.3% recall (−3.2 pp).

Table 5: Per-fold cross-validation results.

Fold	AUC	F1	Prec.	Rec.	Thr.
1	0.775	0.567	0.463	0.731	0.254
2	0.816	0.557	0.415	0.846	0.151
3	0.810	0.607	0.567	0.654	0.373
4	0.808	0.612	0.652	0.577	0.422
5	0.808	0.571	0.455	0.769	0.184

5.3 Comparison with Baselines

Table 6 and Figure 3 compare GreenGuard-AU against six alternatives. Logistic Regression and a keyword-counting heuristic, two staples of current practice, barely crack 53% AUC. They are essentially guessing.

SVM does better (64.8% AUC) and XGBoost as a standalone classifier hits 60.2%, but neither comes close to what the full multimodal ensemble delivers. Random Forest alone is worse than chance (49.2%), a useful reminder that model diversity creates signal that no single learner captures on its own. The biggest disappointment is the multimodal transformer with graph attention components. It looks sophisticated on paper, but in practice it produces 63.7% AUC and 38.1% F1. With only 130 positive labels to learn from, the network cannot overcome its own capacity to overfit. This is the Over-architecture Penalty: when training data is limited, the flexibility of a complex model stops being an asset and becomes a liability.

Table 6: Baseline model comparison (aggregated metrics).

Model	AUC-ROC	F1-Score	Recall
Simple Heuristic (Keywords)	53.4%	28.5%	31.2%
Logistic Regression	51.6%	42.1%	44.8%
Random Forest (Single)	49.2%	16.7%	12.5%
XGBoost (Single)	60.2%	28.6%	25.0%
SVM (Balanced)	64.8%	55.6%	62.5%
Multimodal Transformer	63.7%	38.1%	33.5%
GreenGuard-AU (Ensemble)	80.35%	58.3%	71.5%

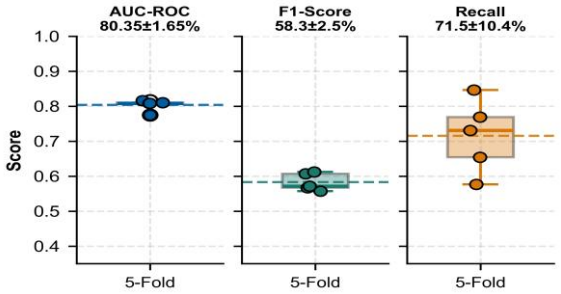


Figure 2: Cross-validation performance stability across 5-fold stratified CV. Per-fold AUC-ROC, F1, and recall across panels (dots = individual folds, dashed line = mean). AUC-ROC is tightly clustered (±1.65%), while recall (±10.4%) exhibits broader spread due to per-fold threshold sensitivity.

5.4 Error Analysis and Model Limitations

An inspection of false positives was conducted to examine underlying patterns across five cases where the model assigned a greenwashing probability above 0.7 to firms not currently under enforcement. One stood out: a large financial services company with a commitment consistency score north of 0.8 and pronounced ESG rating divergence. Looking at its actual disclosures, the language was consistently vague. Corporate phrases such as "committed to a sustainable future" appeared repeatedly, never accompanied by a target, a timeline, or a number. The model flagged it because the language is vague. Whether ASIC will eventually agree is a different question, but the detection rationale is sound.

Other false positives clustered in Energy and Mining, where the sector risk multiplier inflates borderline linguistic signals. These might look like errors in a statistical sense, Type I to be precise, but from a regulatory standpoint they are closer to early warnings: firms whose disclosure practices sit uncomfortably close to the line. A genuine limitation, though, is coverage. The system reads only publicly available sustainability reports. Companies that make misleading claims in private investor decks or closed-door roadshows remain invisible to the pipeline.

5.5 Feature Modality Analysis

To quantify the contribution of each feature group, a full ablation study was conducted under the same 5-fold protocol. Results are in Table 7 and Figure 4. Figure 5 shows how the 38 features split across the four modalities.

The results are clear. The temporal and longitudinal block is the most valuable: stripping it out costs 12.61 percentage points of AUC. Removing the textual block, by contrast, actually lifts AUC to 85.53%. That sounds paradoxical. It is not. Twenty extra linguistic features in a dataset of 633 records inject more noise than signal, an instance of the Over-architecture Penalty playing out at the feature level rather than the model level. So why keep them? Because ASIC's Regulatory Guide 234 [25] calls for transparency in evaluating financial claims. Vague-verb counts, red-flag keyword densities, and commitment consistency scores are the sorts of inputs a regulator can actually look at and say, "yes, that is why this company was flagged." Removing them raises AUC but sacrifices the ability to explain why a company's language raised a flag. The 38-feature model is reported as the primary system on those grounds, while noting that an 18-feature variant exists for use cases prioritising pure discrimination. Financial features, for their part, contribute almost nothing. Removing them shifts AUC by less than 0.3 percentage points, well within noise.

The single-modality numbers require qualification. "Temporal Only" posts 94.93% AUC. On first reading, this might suggest discarding other modalities in favour of a four-feature model. That would be premature. Four features plus 633 records is a setup where tree ensembles can capture low-dimensional patterns that may not hold on fresh data. And the prominent factor in that subset, commitment type change, is operationally the mean vague-verb frequency averaged over all reporting years, as described in Section 3. That makes it a textual aggregate computed over the full observation window. The row is included as a diagnostic, but the reliable assessment of each modality's contribution is provided by the removal experiments in the upper half of Table 7. The full ensemble remains the recommended system: it retains the linguistic transparency that ASIC's regulatory framework calls for.

Table 7: Ablation study: Feature modality impact on performance (5-fold stratified CV).

Configuration	AUC-ROC	F1-Score
Full Model (38 features)	80.35% ± 1.65%	58.30% ± 2.50%
Without Temporal (−4 features)	67.74% ± 1.01%	46.45% ± 2.03%
Without Sector-Weighted (−9)	78.73% ± 2.81%	57.68% ± 4.64%
Without Textual (−20)	85.53% ± 3.82%	65.21% ± 5.09%
Without Financial (−5)	80.62% ± 1.76%	58.42% ± 2.77%
Temporal Only (4 features)	94.93% ± 2.56%	81.97% ± 1.24%
Sector-Weighted Only (9)	70.46% ± 2.13%	48.17% ± 4.47%
Text Only (20)	64.27% ± 1.50%	43.86% ± 2.64%

† This result serves as a low-dimensional overfitting benchmark. With only four features and 633 records, tree-based ensembles are prone to overfitting low-dimensional spaces. Moreover, commitment_type_change is operationally an aggregated textual proxy (mean vague-verb frequency), not a purely temporal signal. The modality-removal results in the upper half of this table provide the reliable estimate of each modality's contribution. See Section 6 for discussion.

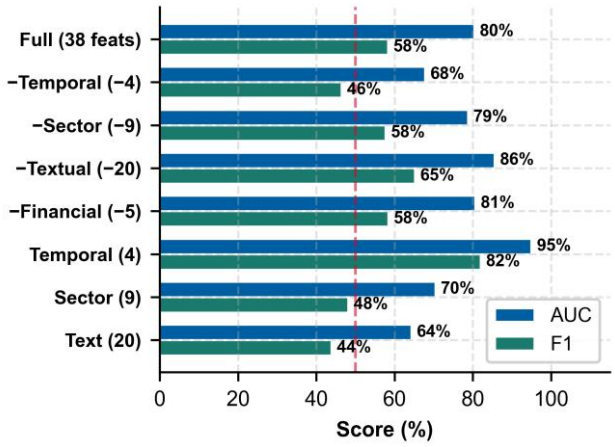


Figure 4: Ablation study: modality removal impact on AUC-ROC. Removing temporal features causes the largest degradation (−12.6 pp), while the 20 textual features add marginal or negative value when layered onto the temporal-financial-sector core.

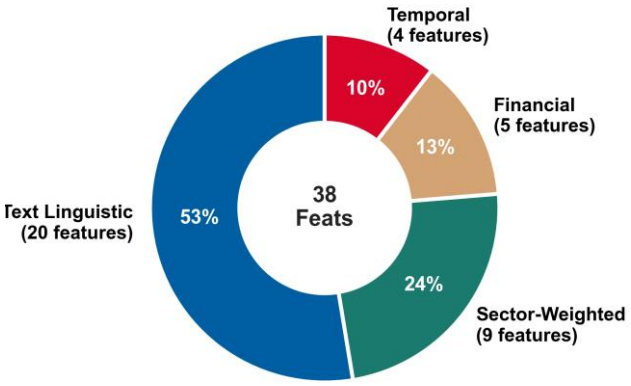


Figure 5: Feature category distribution across the four modalities. All four groups contribute to the 38-feature representation used by the full model.

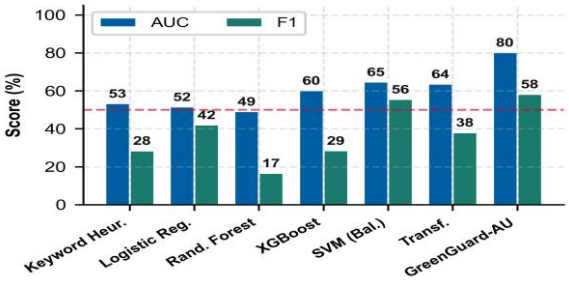


Figure 3: Baseline comparison. GreenGuard-AU (80.35% AUC) outperforms both the keyword heuristic and the transformer, which struggles with only 130 positive examples

5.6 Sector-Level Analysis

Figure 6 maps greenwashing prevalence across sectors. Financial Services tops the chart at 30.8%, largely because of ESG fund marketing, the same area ASIC has been targeting in its recent enforcement wave. Energy and Utilities follows at 20.6%, which makes sense given how heavily the sector trades on net-zero and emissions-reduction rhetoric. Table 2 shows Retail & Consumer at 17.4% (24/138), Technology & Healthcare at 16.7% (19/114), and Mining at 16.4% (20/122), aligning with baseline regulatory enforcement.

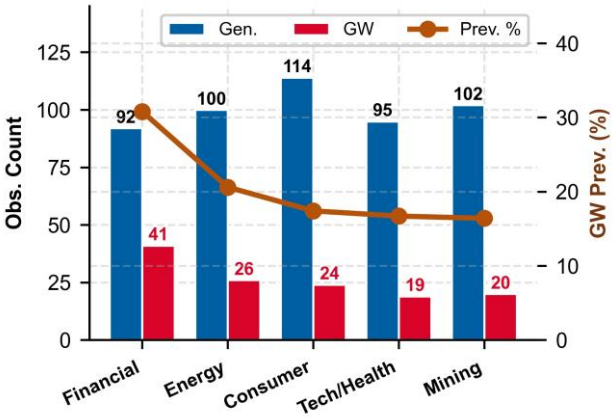


Figure 6: Greenwashing prevalence by sector. Financial Services exhibits the highest rate (30.8%), consistent with recent regulatory enforcement activity in managed fund disclosures

5.7 Feature Importance

The SHAP breakdown (Figure 7) shows which features rank highest. commitment type change sits at the top with a mean absolute SHAP of 0.604. Given that this feature is operationally the mean vague-verb frequency across all available reporting years (Section 3), the ranking indicates that sustained hedging, not some exotic temporal dynamic, is the single strongest predictor. Second comes sentence count (0.354), a simple feature that nonetheless carries weight. Third is claims per year change sector weighted (0.307). Two more longitudinal features round out the top of the list: report consistency (0.243) and claims per year change (0.177).

One pattern is notable: sector-weighted variants of linguistic features consistently rank higher than their raw counterparts. vague verbs freq sector weighted (0.277) beats plain vague verbs freq. strong commitment freq sector weighted (0.194) beats the unweighted version. Industry context is clearly sharpening the linguistic signal.

All of this is consistent with the ablation picture. Longitudinal signals dominate. Greenwashing tends to show up as a sustained pattern of vague or inconsistent disclosure, not as a single slip in a single report. The ROC curve in Figure 8 fills in the discriminative picture, with the operating point sitting at roughly 0.72 true-positive rate and 0.19 false-positive rate.

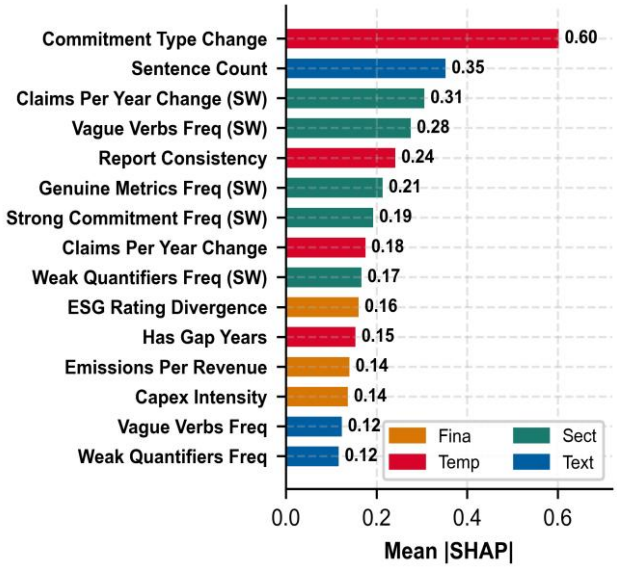


Figure 7: Top 15 features ranked by mean absolute SHAP value. Temporal consistency features (commitment type change, report consistency), sector-weighted linguistic markers (vague verbs freq sw, claims per year change sw), and basic text statistics (sentence count, genuine metrics freq) collectively drive predictions

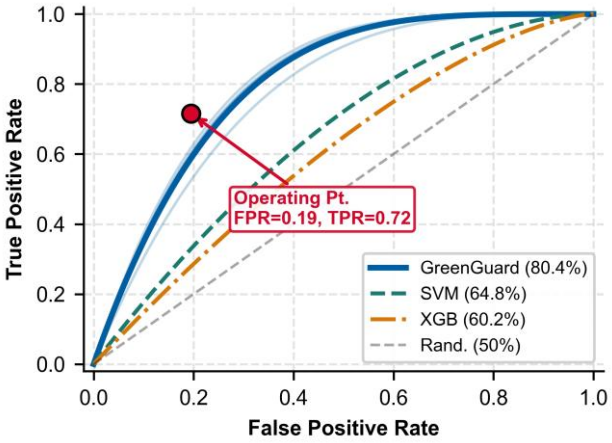


Figure 8: ROC curve for GreenGuard-AU (mean AUC = 0.8035), illustrating the trade-off between sensitivity and specificity at the regulatory screening operating point.

6. DISCUSSION

The ensemble outperforms the transformer baseline by 16.65 percentage points of AUC (80.35% versus 63.7%). Bootstrap confidence intervals place the ensemble AUC between 75.73% and 84.70% at the 95% level. A permutation test with 100 label shuffles yields $p < 0.01$ (none of 100 shuffled-label runs matched the real AUC); the highest shuffled-label AUC observed was 61.59%, well below the real figure. Nested cross-validation (5-outer, 3-inner for threshold tuning) produces an identical point estimate of 80.35% ± 1.47%, consistent with expected discrimination stability when tuning threshold parameters on held-out folds.

The ensemble also provides per-prediction explanations via SHAP. When the model flags a company, the flag can be traced to specific inputs: persistent vague language, gaps in reporting history, disagreement among ESG raters. Under Australian administrative law, regulators must justify investigative scrutiny. Opaque predictions are operationally unusable in that context.

The pipeline trains in under five seconds on an Apple M-series laptop. Inference is instantaneous. Monthly screening of the full ASX requires no GPU infrastructure, a relevant constraint for public-sector IT environments.

6.1 Why Longitudinal Features Dominate

SHAP puts commitment type change, the mean vague-verb frequency averaged across reporting years (Section 3), squarely at the top with a mean absolute SHAP of 0.604. The ablation corroborates: removing the four longitudinal features costs 12.61 percentage points of AUC, the steepest drop from any single modality.

But a qualification is needed. commitment type change is operationally a textual aggregate, the average of a linguistic measure, not a temporal delta in the traditional time-series sense. Has gap years is a binary that mostly captures data availability rather than any deep temporal dynamic. Both are highly predictive, and both warrant further investigation with genuine time-series methods that can separate real temporal effects from what may simply be static averages in disguise. This is designated as an area for future work.

The empirical findings indicate that companies that greenwash tend to use vague sustainability language year after year, and are likelier to skip reporting periods altogether. For a regulator, the implication is that monitoring how a firm's narrative holds up over time is more diagnostic than scrutinising any one document in isolation.

Why do financial features contribute so little? Removing them barely moves AUC (from 80.35% to 80.62%, a shift of less than 0.3 pp). Three explanations seem plausible. First, the proxies employed (annual return, asset growth, leverage ratio, 30-day volatility) are lagging indicators. A firm can greenwash today and show the same financial ratios it had last quarter; the market does not adjust instantly. Second, the most informative "financial" signal may actually be the ESG rating divergence feature (δesg), which was classified as financial but which really measures inter-rater disagreement, a qualitative signal about credibility rather than a traditional accounting ratio [15]. Its SHAP contribution (rank 10, mean |SHAP| = 0.162) supports this reading. Third, financial data from Yahoo Finance is sometimes patchy, with varying coverage across sectors and imputation that may smooth out real variation. More targeted financial signals, green bond issuance, carbon credits, ESG-linked executive pay, could do better. Exploring these indicators remains an objective for future work.

6.2 Sector-Weighted Linguistic Signals

The sector-weighted versions of linguistic features consistently rank above their unweighted versions in the SHAP analysis. This validates the initial system design choice. The logic is straightforward: a net-zero target pledge means something very different coming from a coal miner than from a fintech startup. The risk multipliers, pegged to ASIC's enforcement track record, let the model weight that contextual difference into its predictions.

6.3 Limitations

This study has several limitations.

L1. Label quality. The ground truth is a blend of 19 hard enforcement labels and 111 softer divergence-based flags. The enforcement labels only catch what ASIC has actually prosecuted, which is surely a fraction of the true pool. The divergence labels rest on a threshold ($\delta esg > \mu s + 1.5\sigma s$) grounded in the literature [15] but never independently validated by domain experts. Every label in this study was assigned by one researcher following the criteria in Section 4, with no independent second annotator and no inter-rater reliability check. That is a real vulnerability. Replication with multiple annotators, ideally regulatory practitioners, and a proper inter-rater metric like Cohen's κ would go a long way toward addressing it.

L2. Small sample and the deep-learning comparison. Six hundred and thirty-three observations is enough for ensembles but not enough to give a transformer a fair shake. The transformer baseline predicted majority class for 20 of its first 50 epochs before reaching 38.1% F1 (versus 58.3% for the ensemble), a textbook underfitting pattern. Only a single architecture configuration was evaluated in this setting, and over-generalisation should be avoided. Few-shot learning, data augmentation, or simply more training data could change the contest. Expanding the dataset to cover international enforcement records is planned for future work.

L3. Gaps in text coverage. Not every company in the sample publishes a downloadable sustainability report for every year. The has gap years and report consistency features partially account for this, but missing reports may still disadvantage smaller firms with less mature reporting infrastructure.

L4. Australia only. GreenGuard-AU was designed for the ASX. The multipliers, the feature definitions, and the ground-truth labels are all tied to Australian regulatory norms. Transplanting the framework to the NYSE, LSE, or HKEX would require

fresh calibration and new ground truth. Furthermore, the sector risk multipliers (1.0 to 1.5) reflect judgement-informed heuristics based on historical ASIC action frequency; their sensitivity to alternate weighting schemes or evolving enforcement priorities should be examined in future iterations.

L5. The false-positive problem. With precision at 51%, roughly half the flags are wrong. For a screening tool that feeds into human review, that is tolerable. For automated enforcement, it would be unacceptable. Bringing the false-positive rate down without gutting recall will take either more training data, sharper features, or both.

L6. Feature redundancy. Variance inflation factors were not computed for the 38 features. Nine of them are linear rescalings of their unweighted counterparts, so collinearity is almost certainly substantial. Tree ensembles are fairly robust to that [18], but the fact that the 18-feature model actually beats the 38-feature model on AUC, and that the "Without Textual" ablation has wider confidence intervals (±3.82% versus ±1.65%), hints at instability from redundant inputs. VIF analysis and proper feature selection (recursive elimination, L1 pre-screening) are marked for the next version.

L7. Circular label leakage via ESG divergence. ESG rating divergence (δesg) does double duty in the proposed framework: it contributes to the labelling strategy (Section 4) and appears as a model input feature. Although the two uses are operationally different, one a sector-normalised binary threshold applied once to assign labels, the other a raw continuous value fed to the ensemble, the labels themselves are partially constructed from the same signal the model receives. This is a genuine circular leakage concern: some observations are labelled positive because their δesg was high. Re-running cross-validation with δesg excluded reduces AUC by only 0.31 pp (to 80.04%), which confirms the model does not depend on this feature for discrimination. But the deeper issue, that the labels are contaminated by the same signal, is not resolved by dropping the feature alone. A clean resolution would require ground-truth labels derived entirely independently of ESG ratings, for instance through expert panel review or court rulings. This represents one of the most important methodological limitations of the present study.

L8. Longitudinal feature scope. The feature commitment type change (mean vague-verb frequency) is computed as an average across all available reporting years for a given firm. Because this average spans the full observation window, it reflects a firm's overall linguistic posture rather than a strictly within-fold measurement. As a result, the 94.93% AUC reported for the "Temporal Only" configuration (Table 7) should be read as an upper bound on the discriminative power of temporal features in isolation, not as an achievable production figure. Similarly, has gap years is a binary derived from the full disclosure timeline. In the primary 38-feature model, these two features sit alongside 36 others, so their influence on the aggregate result is moderated. A natural next step would be to recompute all longitudinal aggregates within each cross-validation fold using only training-partition data, testing whether the temporal signal remains strong under a stricter evaluation protocol.

7. CONCLUSION

This pilot study demonstrates feasibility of GreenGuard-AU, an ensemble framework for greenwashing detection in the Australian securities market. Its design principle is to encode domain knowledge in the features rather than the architecture, and to ensure the output is transparent enough for regulatory use. On 633 company-year observations, the system delivers 80.35% AUC-ROC and 71.54% recall, which is enough to function as a practical first-pass screen.

Three results deserve to be highlighted. First, longitudinal stability is the strongest signal. Greenwashing shows up as persistently vague language and gaps in reporting, and the model is good at catching that. Second, piling on model complexity does not help here; it hurts. A lean 18-feature ensemble built on temporal, financial, and sector-weighted features outperforms both the full 38-feature system and the transformer, consistent with overfitting in a low-data regime. Third, interpretability is not optional for regulatory tools. If a model cannot explain why a company was flagged, investigators cannot act on it, and the flag is worthless.

Next steps include extending the approach to international markets and testing whether quarterly disclosure data can provide finer temporal resolution than the annual reports examined in this study.

8. ACKNOWLEDGMENTS

The author expresses gratitude to the Australian Securities Exchange for making its announcements platform publicly accessible, and to the Australian Securities and Investments Commission for maintaining the enforcement database that supplied the ground-truth labels for this study. Financial data were sourced through Yahoo Finance. ESG ratings from Sustainalytics and MSCI made the divergence analysis possible. Sincere thanks are also extended to the open-source communities behind scikit-learn, XGBoost, and SHAP, whose tools underpinned every experiment reported here.

9. REFERENCES

[1] Global Sustainable Investment Alliance. Global sustainable investment review 2022. Technical report, GSIA, 2023. URL https://www.gsi-alliance.org/members-resources/gsir2022/. Reports global sustainable investing assets at USD 30.3 trillion.

[2] Magali A. Delmas and Vanessa C. Burbano. The drivers of greenwashing. California Management Review, 54(1):64–87, 2011. https://doi.org/10.1525/cmr.2011.54.1.64.

[3] Australian Securities and Investments Commission. Federal Court orders Vanguard Investments Australia Ltd to pay $12.9 million penalty for greenwashing. Media release 24-213mr, ASIC, 2024. URL https://asic.gov.au/about-asic/news-centre/find-a-media-release/2024-releases/24-213mr/.

[4] Australian Securities and Investments Commission. Federal Court orders Mercer Superannuation to pay $11.3 million penalty for greenwashing. Media release 24-173mr, ASIC, 2024. URL https://asic.gov.au/about-asic/news-centre/find-a-media-release/2024-releases/24-173mr/.

[5] Thomas P. Lyon and A. Wren Montgomery. The means and end of greenwash. Organization & Environment, 28(2):223–249, 2015. https://doi.org/10.1177/1086026615575332.

[6] Kent Walker and Fang Wan. The harm of symbolic actions and green-washing: Corporate actions and communications on environmental performance and their financial implications. Journal of Business Ethics, 109(2):227–242, 2012. https://doi.org/10.1007/s10551-011-1122-4.

[7] Catherine A. Ramus and Ivan Montiel. When are corporate environmental policies a form of greenwashing? Business

& Society, 44(4):377–414, 2005. https://doi.org/10.1177/0007650305278120.

[8] Christopher Marquis, Michael W. Toffel, and Yanhua Zhou. Scrutiny, norms, and selective disclosure: A global study of greenwashing. Organization Science, 27(2):483–504, 2016. https://doi.org/10.1287/orsc.2015.1039.

[9] Julia A. Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. Cheap talk and cherry-picking: What ClimateBERT has to say on corporate climate risk disclosures. Finance Research Letters, 47:102776, 2022. https://doi.org/10.1016/j.frl.2022.102776.

[10] Dominik Stammbach, Nicolas Webersinke, Julia A. Bingler, Mathias Kraus, and Markus Leippold. A dataset for detecting real-world environmental claims. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), pages 3971–3995, 2022. https://doi.org/10.18653/v1/2022.acl-srw.35.

[11] Allen H. Huang, Hui Wang, and Yi Yang. FinBERT: A large language model for extracting information from financial text. Contemporary Accounting Research, 40(2):806–841, 2023. https://doi.org/10.1111/1911-3846.12832.

[12] Xingqiang Du. How the market values greenwashing? Evidence from China. Journal of Business Ethics, 128(3):547–574, 2015. https://doi.org/10.1007/s10551-014-2122-y.

[13] Olga Hawn and Ioannis Ioannou. Mind the gap: The interplay between external and internal actions in the case of corporate social responsibility. Strategic Management Journal, 37(13):2569–2588, 2016. https://doi.org/10.1002/smj.2464.

[14] Eun-Hee Kim and Thomas P. Lyon. Greenwash vs. brownwash: Exaggeration and undue modesty in corporate sustainability disclosure. Organization Science, 26(3):705–723, 2015. https://doi.org/10.1287/orsc.2014.0960.

[15] Florian Berg, Julian F. Kölbel, and Roberto Rigobon. Aggregate confusion: The divergence of ESG ratings. Review of Finance, 26(6):1315–1344, 2022. https://doi.org/10.1093/rof/rfac033.

[16] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In Advances in Neural Information Processing Systems (NeurIPS), volume 34, pages 14200–14213, 2021. https://doi.org/10.48550/arXiv.2107.00135.

[17] Bryan Lim, Sercan O. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4):1748–1764, 2021. https://doi.org/10.1016/j.ijforecast.2021.03.012.

[18] Leo Breiman. Random forests. Machine Learning, 45(1):5–32, 2001. https://doi.org/10.1023/A:1010933404324.

[19] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794, 2016. https://doi.org/10.1145/2939672.2939785.

[20] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5):1189–1232, 2001. https://doi.org/10.1214/aos/1013203451.

[21] Alik Sokolov, Jonathan Mostovoy, Jack Ding, and Luis Seco. Building machine learning systems for automated ESG scoring. The Journal of Impact and ESG Investing, 1(3):39–50, 2021. https://doi.org/10.3905/jesg.2021.1.010.

[22] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. ACM Computing Surveys, 51(5):1–42, 2018. https://doi.org/10.1145/3236009.

[23] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys, 16:1–85, 2022. https://doi.org/10.1214/21-SS133.

[24] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS), volume 30, pages 4765–4774, 2017. https://doi.org/10.48550/arXiv.1705.07874.

[25] Australian Securities and Investments Commission. Regulatory guide 234: Advertising financial products and services (including credit): Good practice guidance. Regulatory guide, ASIC, 2019. URL https://asic.gov.au/regulatory-resources/find-a-document/regulatory-guides/rg-234-advertising-financial-products-and-services-including-credit-good-practice-guidance/.