

Multimodal Temporal Deep Learning for ESG Assessment and Sustainability Forecasting: Implications for Responsible Business in Emerging Markets

Thinh Tran Thanh

University of Economics and Law

Dang Le Huu

University of Economics and Law

Huong Pham Thai Huynh

University of Economics and Law

Tuyen Pham Thi Ngoc

University of Economics and Law

Trang Nguyen Thi Thuy

University of Economics and Law

Duy Thanh Tran

thanhtd@uel.edu.vn

University of Economics and Law

Research Article

Keywords: ESG measurement and disclosure, Explainable AI in sustainable investment, Corporate sustainability in emerging markets

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10605735/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Multimodal Temporal Deep Learning for ESG Assessment and Sustainability Forecasting: Implications for Responsible Business in Emerging Markets

Thinh Tran Thanh^{1,2}, Dang Le Huu^{1,2}, Huong Pham Thai Huynh^{1,2}, Tuyen Pham Thi Ngoc^{1,2}, Trang Nguyen Thi Thuy^{1,2},
Duy Thanh Tran^{1,2,*}

¹ University of Economics and Law, Ho Chi Minh City, Vietnam

² Vietnam National University, Ho Chi Minh City, Vietnam

* Corresponding Author Email: thanhtd@uel.edu.vn

Abstract: ESG investment has accelerated sharply in Vietnam, yet investors lack quantitative tools to identify companies on a trajectory toward the Vietnam Sustainability Index (VNSI) the country's only officially verified ESG benchmark before membership decisions are announced. This gap persists because two technical barriers have not been resolved simultaneously: corporate ESG information is disclosed predominantly in unstructured Vietnamese text that international NLP models do not process effectively, and existing studies discard the longitudinal ESG trajectory that most reliably signals future index membership. We propose the Multi-Tier Deep Learning Framework (MTDLF), combining PhoBERT based Vietnamese ESG text encoding with Bidirectional GRU temporal classifier. Tier 1 integrates PhoBERT with Hierarchical Chunking and Global Attention Pooling, fuses the resulting representation with 11 quantitative financial indicators, and produces a 128 dimensional ESG Vector per firm year observation. Tier 2 learns the three year ESG trajectory of each company and performs binary VNSI classification. Trained and evaluated on 984 observations from 123 HOSE listed companies across 2017–2024, MTDLF achieves ROC-AUC of 0.8698, PR-AUC of 0.5332 - 2.84 times the random baseline and Recall of 0.89 on the VNSI - positive class across 149 test sequences from companies entirely excluded during training. An integrated What-If Analysis engine further translates predictions into firm specific improvement priorities. These results confirm that Vietnamese ESG disclosures carry genuine predictive content, and language specific NLP combined with multimodal temporal modeling constitutes viable approach to ESG measurement in emerging markets.

Keywords: ESG measurement and disclosure; Explainable AI in sustainable investment; Corporate sustainability in emerging markets

1 Introduction

Over the past two decades, something has shifted silently in the manner in which capital markets distribute value. ESG - Environmental, Social, and Governance - was never meant to be a compliance exercise. Rather, it was born out of a United Nations initiative in 2004 as a result of a recognition by financiers that the risks most likely to destroy long - term value were those which balance sheets were traditionally not accounting for. Investment capital registered under the Principles for Responsible Investment stood at a staggering 123 trillion USD by 2023, compared to 21 trillion USD in 2010 [26]. What one sees here is not a shift in ideology but a shift in risk management: environmental liabilities, labor issues, and governance are now being factored into the cost of capital.

For Vietnam, this global reorientation has come at a particularly timely moment. The country entered the COP26 process in 2021 with commitments to reach net zero by 2050, end coal use by 2040, and end deforestation by 2030 - promises judged to be “among the most ambitious in the ASEAN region”. Domestically, policy has kept pace with this external agenda too: the Circular 155/2015/TT - BTC introduced disclosure obligations for public companies in Vietnam, and subsequent SSC - World Bank guidelines have expanded these to environmental and social reporting too. This context provides the backdrop for the launch of the Vietnam Sustainability Index by the HOSE in July 2017, with the SSC and the German GIZ. Every year, the VNSI selects the 20 companies from the VN100 basket with the best ESG performance according to the GRI criteria and awards the label to the companies - making it the only ESG index for Vietnamese listed companies with regulatory label status [23]. From the perspective of the research design, the annual publication by a regulatory body provides a classification label that sharply distinguishes it from the proprietary ratings of a for-profit ratings company.

Despite institutional progress, a significant gap remains. Currently, there are no quantitative tools that can predict which companies are likely to become members of VNSI. This shortcoming reflects three unresolved methodological issues in ESG studies, particularly in the Vietnamese context.

The first issue concerns the reliability of the measurement. Berg, Kölbel and Rigobon (2022) [6] compared ESG ratings of the same company provided by six reputable organizations and found that the average correlation between the rating agencies was only 0.54. This is quite low compared to the correlation above 0.90 commonly seen in credit rating organizations. This difference stems from three main reasons: the evaluating organizations do not measure the same aspect, they use different indicators to assess the same factor, and even when using the same measurement, the weighting varies significantly.

A company rated an ESG leader by one provider may simultaneously qualify as a laggard by another, and both assessments can be technically defensible. For investors operating in Vietnam, this fragmentation is compounded by near-total absence of international agency coverage. VNSI's annual publication partially addresses this void, but its value depends entirely on whether analytical models can be built to learn from it - and none currently exist.

The second problem is the persistent separation between quantitative and textual approaches in ESG research. One such stream, based on the meta-analysis of over 2,200 studies by Friede, Busch, and Bassen (2015) [13], panel regression results by Le et al. (2024) [18] on six ASEAN markets, and evidence from Vietnam's own stock market by Nguyen & Nguyen (2025) [16], links ESG ratings to financial performance but ignores the process by which ESG ratings are developed. A parallel stream of studies - built on the transformer architecture introduced by Devlin et al. (2019) [10] - including ESG-BERT by Mehra et al. (2022) [21], ESG2PreEM by Lee et al. (2024a) [20], and E-BERT by Zhang et al. (2024) [32], shows NLP models can classify ESG text with high accuracy but uses textual data in isolation, abstracting from the quantitative financial context in which revealed commitment affects measured outcomes. No existing study brings both modalities together within a single architecture. This separation leads to several limitations in research practice: VNSI membership does not solely reflect financial health or the quality of information disclosure, but is the result of simultaneously evaluating both factors according to GRI criteria. However, unipolar models can only approximate this comprehensive assessment.

The third issue concerns the handling of the time factor. Both current research approaches view each company's data year by year as independent observations, as if a company's ESG profile in 2022

were unrelated to its own data in 2021 or 2020. However, this assumption is inappropriate when predicting the likelihood of becoming a VNSI member.

Inclusion in the VNSI is not the result of a single year, but reflects a cumulative improvement in ESG over multiple reporting periods and is assessed in relation to the entire VN100 group. Therefore, to distinguish between a company that truly improves its ESG in the long term and one that merely enhances its disclosures in the short term to attract responsible investment flows, a time-based approach is necessary. This is precisely the perspective that current methods overlook, even though it is a particularly important factor for long-term investors.

All three of these barriers stem from limitations related to the specific linguistic characteristics of Vietnam. More than 90% of ESG - relevant information published by Vietnamese listed companies appears exclusively in Vietnamese, embedded across annual reports and sustainability documents routinely extending to several hundred pages. Vietnamese is a tonal, isolating language whose syntactic and lexical properties differ fundamentally from English, and multilingual models such as XLM - R underperform language - specific architectures on Vietnamese NLP tasks by margins of 2.3 to 4.1 percentage points in F1 score [22]. Despite the availability of PhoBERT - pretrained on a 20GB Vietnamese corpus - no existing study has applied it to ESG text analysis, leaving eight years of accumulated Vietnamese ESG disclosure effectively inaccessible to quantitative research.

This paper proposes the Multi - Tier Deep Learning Framework (MTDLF) as a direct response to these four interconnected problems. The architecture operates in two sequential tiers. Tier 1, PhoESG - DeepFusion, encodes Vietnamese ESG text of arbitrary length through Hierarchical Chunking combined with Global Attention Pooling, fuses the resulting representation with a branch processing 11 standardized financial indicators, and produces a 128 - dimensional ESG Vector for each firm - year observation. Tier 2, a Bidirectional GRU Temporal Classifier, receives three consecutive ESG Vectors and learns the temporal trajectory of ESG development to perform binary VNSI classification. A What - If Analysis engine built on feature perturbation further translates predicted probabilities into firm - specific improvement priorities, advancing the analytical function from prediction to prescription.

Our primary contribution is to offer a new approach to ESG measurement by combining multimodal data within a unified deep learning architecture, contributing to the understanding of ESG disclosure as a substantive, verifiable phenomenon rather than a ceremonial exercise. Empirically, we construct a dataset of 984 firm - year observations from 123 HOSE - listed companies across 2017 - 2024 , with VNSI labels sourced from official HOSE announcements [17], and demonstrate that MTDLF achieves ROC - AUC of 0.8698, PR - AUC of 0.5332 - 2.84 times the random baseline - and Recall of 0.89 on the positive class across 149 test sequences drawn entirely from companies excluded during training. That a model trained on publicly available Vietnamese - language disclosures and financial data can predict VNSI membership with this accuracy suggests that ESG disclosures in Vietnam carry genuine informational content - and that unlocking it requires, as a necessary condition, the language - specific analytical tools this paper develops.

2 Related Researches

2.1 Theoretical Background

From Corporate Social Responsibility to Corporate Sincerity and Public Value Creation: The evolution of business ethics thought within academia has been profoundly shaped by foundational debates concerning the ultimate purpose of the corporate entity. The neoclassical economic school appeared in the 1950s and became popular in the 1970s. It strongly believed that a business's only responsibility is to maximize shareholder profits [14]. From this point of view, when managers spent company money on social or environmental causes, it was seen as agency manipulation. In other words, managers were thought to be using shareholders' money to boost their own reputations or gain moral approval they did not deserve. Due to such a line of thought, CSR activities remained largely superficial charitable marketing efforts over the years. Such attempts were not an inherent aspect of the business, but rather a PR tool to soothe the population and save the reputation of the company.

The implementation of Environmental, Social and Governance (ESG) practices has triggered a paradigm shift in structure towards the old paradigm that is profit-oriented. On the one hand, the modern model of evaluation demands that ethical factors be organized, quantified, and completely incorporated into both financial risk evaluation and value creation strategies in the organization. To ensure that the shift to voluntary, intermittent social engagement to a higher level of rigorous sustainable reporting requirements can take place, the management boards would need to show a willingness to commit to these principles. Reporting of sustainable development must not be a superficial reporting to satisfy regulatory requirements. Rather, they ought to reflect a substantive benefit to the society and demonstrate the value that the enterprise will contribute to society. Such genuineness would be especially vital in new markets where regulatory and legal systems are not fully developed or in a state of flux. In these situations, firms are supposedly supposed to take the initiative to adopt ethical policies that are beyond the minimum legal requirements, making them producers of public goods, such as environmental stewardship, good labor relations, and social equity.

The interconnections of sustainable corporate behavior and financial performance are inherent in the philosophical background of Stakeholder Theory, and are the foundation of the Stakeholder Theory. The theory opposes the shareholder maximization theory and views the corporate structure as a multi-dimensional system of relationships [12]. Past profitability on a balance sheet alone cannot just tell the worth of an organization and its ethics. Rather, it is the capacity and the willingness to meet the potentially conflicting needs of a high density network of organisations that encompasses employees, customers, local communities, governmental agencies and the natural environment.

In this complex system, one of the main challenges to establishing ethical trust is the fact of information asymmetry between the management board and outside parties. Annual Reports and Sustainability Reports have thus left the confines of regular administration or compliance reports to become essential strategic communication instruments [11]. The social responsibility of enterprises is being manifested in these reports, building up an important dialogic space so that social responsibility becomes transparently visible; on the other hand, the governance philosophy of enterprises is put to the test and open to public opinion. Private companies' ethical commitment is directly proportional to the degree of detail, specificity and transparency of the public reporting. These reports thus offer the necessary linguistic information for stakeholders to judge if the enterprise is truly bringing value to the community or whether it is simply enacting formal and symbolic coping acts to justify its ongoing existence.

The academic theoretical system needs to be studied from the twofold and complementary perspective of Legitimacy Theory and Institutional Theory to understand the underlying strategic motives behind businesses' desire to participate in prestigious national indexes like the Vietnam

Sustainability Index in emerging markets. According to Legitimacy Theory, enterprises live on and are constantly operating under the invisible social contract. Breaches of fundamental ecological norms or abuses of human labour may result in harsh and immediate sanctions from the community, boycotting by consumers or strong legal sanctions from the government. Thus, it is an optimal and proactive way to secure the approval of the society, to preserve the legitimacy of the organization and to establish an unblemished ethical reputation, which is difficult to imitate, as a competitive advantage, by striving to become part of a highly visible national sustainability index basket [3].

Within the emerging markets context this legitimacy-seeking is exacerbated by structural, macro-economic pressures that are also grounded in Institutional Theory. At the national level, there are powerful transnational integration commitments in the current economic landscape, as can be seen in the ambitious governmental goals, including the aim of becoming "Net Zero" carbon emissions by 2050. Listed companies are at the same time subject to isomorphic institutional pressures from three different directions. First, coercive pressure is the pressure from state regulators and securities commissions for unprecedented transparency of information. Second, normative pressure comes from international institutional investors and global funds with strict ethical screening filters in their portfolios. Third, mimetic pressure comes when market-leading companies constantly push the limits on sustainable practice standards and smaller companies are forced to follow suit to stay alive. Inclusion in the sustainability index basket is therefore a strong and validating sign to the global capital market, clearly stating that the enterprise has overcome these institutional obstacles and entered the transnational moral value chain.

Measurement Gaps and the Ethical Challenge of Reputation Washing: Despite the general theoretical agreement on the value of ecological and social criteria, there are significant measurement challenges in empirical work to accurately determine the veracity of ethical commitments. The biggest difficult to face is cosmetic greenwashing and reputation washing. This is due to opportunist management, in which the lack of real-world changes in the way the business operates is disguised with "big talk" and campaigns of extensive marketing. These acts are unethical in business and pose great informational risk for capital allocation decisions in financial markets.

Current measurement methodologies show significant structural weaknesses with the problem of information distortion in complexity cases. Notably, there are huge differences between global rating agencies in their ratings, due gaps in definition and weighting of concepts as well as evaluation methods [6] which have resulted in loss of investor confidence in quantitative assessments available. In addition, the current empirical studies are often based on single-modality methods, resulting in large measurement gaps. Focused solely on the financial numbers, it fails to account for the strategic course of the organisation, the adjustment to the new culture and the fundamental values of the organisation's ethics. On the other hand, if all the text is in natural language, the evaluation process is vulnerable to manipulation by employing advanced linguistic strategies by the corporate public relations team. There is no multimodal analytical approach which correlates textual declarations with actual financial practices, making it a one-sided evaluation of ethics and opening it up to profit oriented manipulation. Furthermore, the majority of research in the academic literature has focused on the economy of the developed western world, and there is a significant lack of knowledge regarding economies in transition, in which the semantic structure and reporting standards in the local language largely differ ethnoculturally and within the regulatory context.

Business Ethics as a Recursive Time-Series Trajectory: One of the major conceptual flaws of existing ethical evaluation approaches is to view corporate behaviour as a set of one-off, cross-

sectional observations occurring at discrete year-end financial reports. But, sustainable development should be seen as a recursive process that must continuously be addressed in a long-term perspective strategy. The pay-off for investments in advanced emission reduction technologies, moving towards renewable energy and putting in place comprehensive social welfare systems is usually characterized by highly time-lagged impacts in terms of monetary return or reputation.

This requires then a methodological shift from the static spatial analysis to long time-series monitoring of the ethics of corporate behavior. Only a thoughtful temporal line of sight can help analysts to differentiate between those businesses that have a history of consistent gradual ethical investments and those that are “social opportunists” and only spend money on social campaigns when there is some unexpected goodwill, just before the capital campaigns. Authentic business ethics isn't just a point to reach and a tick to complete: it's an ongoing, dynamic combination of long-term execution capacity and an attitude of transparency that is honed and fostered over a period of time.

2.2 Hypothesis development

The academic community has integrated organizational legitimacy theory and stakeholder expectation theory, and has verified the urgency of transparent information disclosure. A methodological gap exists in emerging markets; this study intends to develop tools to address information distortion and reputation whitewashing, and build a quantitative foundation for predicting enterprises' selection into the national sustainability index basket.

This study proposes a combined impact hypothesis regarding the quality of information disclosure and corporate financial health. Derived from stakeholder theory, this hypothesis holds that only when an enterprise has sufficient and stable material resources can it fulfill its ethical commitments to create social value. Financially weak firms may even violate the baseline compliance requirements for occupational safety and environmental protection. This paper proposes a signaling logic for corporate governance from a reverse perspective: a sound capital structure and high-transparency information disclosure work in tandem to form a dual signal of a firm's governance sincerity. The soft information provided by the sustainability story tells investors about an organization's moral purpose. Strong financial metrics provide evidence that an organization can actually realize its vision. Both are needed to satisfy the comprehensive criteria of regulators and indexes. Weaknesses in sustainability disclosure and/or financial strength will critically impact the organization's credibility with investors. Therefore, we hypothesize:

H1: Better sustainability information disclosure quality and sound financial position will significantly and positively influence firms' probability of being included in Vietnam Sustainability Index.

Our main research hypothesis on predictive strength of unstructured language data builds on limitations existing in conventional reporting standards which include temporal delays, stringent standardized protocols, which could result in missing vital value drivers that contribute to an entity's long-term sustainability. While buried deep within paragraphs of annual reports, subtle cues can be found that indicate management's attitudes towards environment, equality, and governance [4]. When analyzed through intelligent language tools - sifting through fluff and unverifiable statements – this text can transform from corporate speak to a validated measure of integrity and ethical significance [2]. Free from the confines of the balance sheet, this type of forward looking narrative can help

identify what is left behind by numbers; a company's culture and future actions [1, 28]. Based on these assumptions our second hypothesis is as follows:

H2: Predictive power and completeness: Business ethics cues identified using scientific text analysis techniques from unstructured corporate text are much stronger predictors and complements to traditional financial measures for determining how committed and how index-worthy a company truly is.

Hypothesis on the Superiority of the Time-Series Multimodal Deep Learning Framework. Considering the above challenges for measurement and also the fact that corporate greenwashing is dynamic in nature, one cannot possibly assess corporate ethics by using only a single snapshot of data or modality of information. To measure how truthful a company is about its commitments, one needs an algorithmic system combining both natural language processing of managerial text with numerical financial statements. Additionally, longitudinal information allows us to determine whether a company is following through on its promises. Use of a multimodal deep learning model that employs a time-series RNN to analyze historical patterns can help us understand the continual actions of companies [5]. This dynamic approach enables the structural identification of substantive sustainable efforts and systematically filters out short-term opportunistic signals. As a result, this methodology provides empirical support for the theoretical assertion that business ethics is a complex, multidimensional construct requiring an equally multidimensional analytical approach. Accordingly, the third hypothesis is as follows:

H3: An assessment framework utilizing time-series multimodal deep learning demonstrates substantially improved classification accuracy for corporate ethics profiles and more effectively captures authentic corporate sincerity than traditional static single-modality methods.

3 Research Methodology

The Multi-Tier Deep Learning Framework (MTDLF) is a novel multimodal temporal architecture designed to predict the ability of companies to be listed as sustainable in Vietnam. This framework addresses four barriers: (1) extracting ESG signals from lengthy Vietnamese-language corporate reports containing a wealth of non-ESG information, (2) integrating heterogeneous quantitative financial indicators with qualitative textual data, (3) modeling ESG dynamics over time across consecutive fiscal years, and (4) explaining how indicator fluctuations will impact a company's sustainability ability.

3.1 Problem Formulation

The Vietnam Sustainability Index (VNSI), published annually by the Ho Chi Minh Stock Exchange (HOSE), assesses listed companies based on Environmental, Social, and Governance (ESG) criteria. Predicting whether a company will be included in the sustainable list presents several methodological challenges. First, ESG-related data exists in two different forms: financial indicators and natural language text in Annual Reports (AR) and Sustainability Reports (SR). Second, ESG strategies exhibit a lagged effect, as investments in governance, environment, or social are only recognized after one to two reporting cycles [19]. Third, being a sustainable enterprise constitutes a binary data discrepancy problem, with approximately 25% of observations belonging to the positive class. Fourth, in sustainability reports, ESG information spans thousands of words; however,

PhoBERT [22], the most modern Vietnamese language NLP model, is limited to an input sequence of 256 words.

Formally, given a firm c at target year T , the prediction task is defined as estimating the probability $P(\text{VNSI}_{c,T} = 1 \mid X_{c,T-2:T})$, where $X_{c,T-2:T}$ encompasses the multimodal ESG-financial profile of firm c over three consecutive years ($T-2, T-1, T$).

3.2 Overall Architecture

As illustrated in Fig. 1, the proposed MTDLF framework integrates data acquisition, multimodal feature extraction, temporal context modeling, model training and evaluation, and What-If XAI-based recommendation into a unified ESG assessment pipeline.

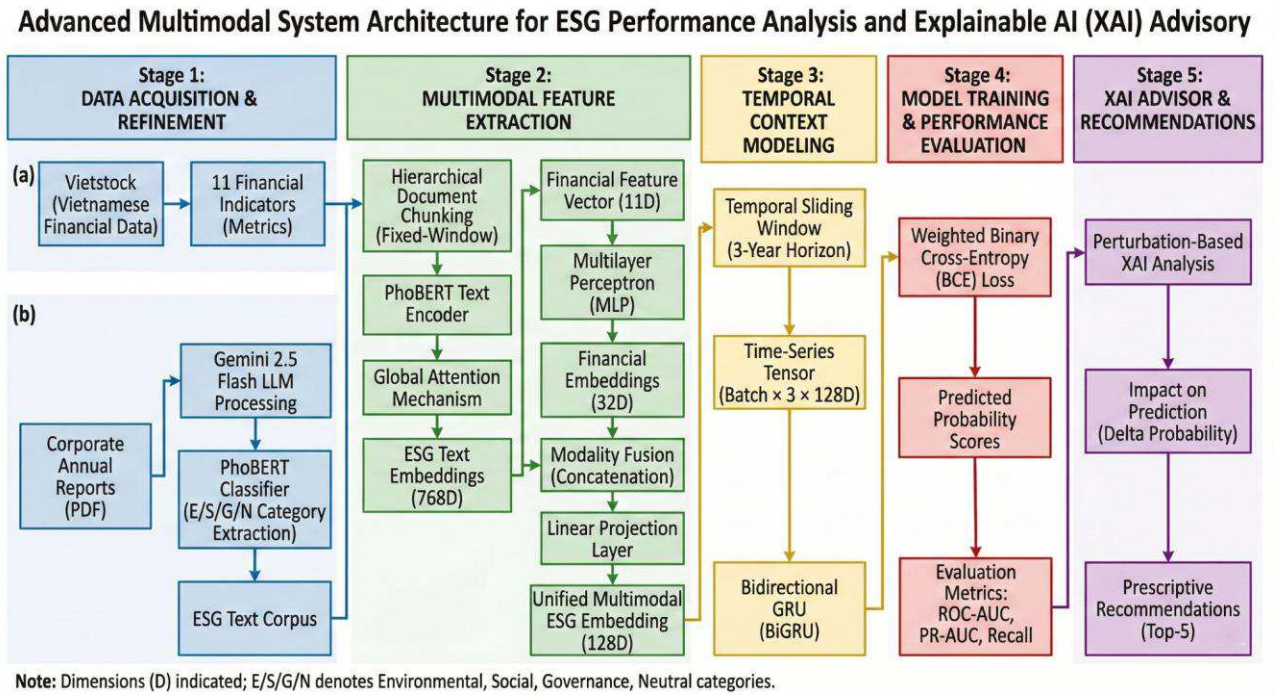


Fig. 1 Overall architecture of MTDLF (Source: Authors)

MTDLF has two architectural layers as follows: Layer 1 (PhoESG-DeepFusion) acts as a static multimodal encoder, transforming each observation (company, year) into a 128-dimensional ESG vector using PhoBERT with financial indicators processed using MLP; Layer 2 (ESG-TimeMachine) uses a Bidirectional Gated Recurrent Unit (Bi-GRU) network [8, 27] to learn ESG potential over time from three-year series of ESG vectors and output binary probabilities of classifying sustainable businesses. In addition, this framework adds a What-If XAI Advisor layer to provide perturbation analysis results to develop strategies for improving the sustainable development capabilities of businesses.

3.3 Data Collection and Preprocessing

3.3.1 Sample and data sources

The dataset comprises 984 annual observations of 123 listed companies on the HOSE and HNX stock exchanges from 2017 to 2024. Two data streams were collected: (a) quantitative financial indicators from Vietstock and (b) qualitative textual data from 1,070 Annual Reports and Sustainability Reports of all companies. The sustainable business label was verified based on the Corporate Sustainability Index (CSI) developed by the Vietnam Chamber of Commerce and Industry (VCCI) and the Vietnam Business Council for Sustainable Development (VBCSD). Table 1 presents the industry distribution of the sample.

Table 1 Sample distribution by industry sector (Source: Authors)

Industry Sector	No. of Firms	Proportion (%)
Industrial	41	33.3
Consumer	24	19.5
Finance	24	19.5
Energy	13	10.6
Real Estate	10	8.1
Healthcare	8	6.5
Technology	3	2.4
Total	123	100.0

Note. The study sample includes companies that have Annual Reports and Sustainability Reports during the study period. Whether or not the Sustainability Report is published will depend on the company.

3.3.2 Financial indicators

The model uses 11 financial indicators belonging to 5 different groups: valuation (EPS, BVPS, P/B, Return on Equity), profitability (ROEA, ROAA), growth (Net Profit Growth, Total Asset Growth, Equity Growth), capital structure (Equity-to-Asset Ratio), and market risk (Beta). Specifically, the P/E ratio is replaced by its inverse, Return on Equity ($E/P = 1/P/E$), to address the extreme right-leaning skew caused by companies with near-zero positive earnings. This transformation reduces the extreme values towards zero while maintaining the direction of the economic signal. Table 2 summarizes the 11 financial indicators used in the quantitative branch of the model, including their variable names, symbols, and corresponding financial categories.

Table 2 Financial indicators used in the quantitative branch (Source: Authors)

Variable	Symbol	Category
Earnings Per Share	<i>eps</i>	Valuation
Book Value Per Share	<i>bvps</i>	Valuation
Price-to-Book	<i>pb</i>	Valuation
Earnings Yield (replaces P/E)	<i>earnings_yield</i>	Valuation

Beta	<i>beta</i>	Market Risk
Return on Equity	<i>roea</i>	Profitability
Return on Assets	<i>roaa</i>	Profitability
Net Profit Growth (%)	<i>net_profit_growth</i>	Growth
Total Asset Growth (%)	<i>total_asset_growth</i>	Growth
Equity Growth (%)	<i>equity_growth</i>	Growth
Equity-to-Asset (%)	<i>equity_to_asset</i>	Capital Structure

3.3.3 Quantitative data preprocessing

The data preprocessing process was implemented in three sequential steps. Missing values were addressed by filling in the median from the company's own available data; if insufficient, the market-wide median was applied. Missing Beta values after both steps were assigned 1.0 to reflect the assumption of market volatility. To handle outliers, we used a two-tier approach. Hard Clipping limited Net Profit Growth to the range $[-100\%, +500\%]$ to help preserve economically significant signals. Smart Clipping adaptively handled the remaining 10 variables: when the first percentile is positive – indicating positive-leaning values – the lowest level is -100 to preserve significant negative outliers (e.g., negative equity); otherwise, standard first percentile clipping was used instead of smart clipping. For data normalization, we apply the Z-score normalization method using estimated parameters from the training dataset, which are then uniformly applied to the test dataset to prevent data leakage.

3.3.4 Qualitative text preprocessing

Business reports contain a lot of content unrelated to ESG (board of directors introductions, company history, marketing materials) which distorts the quality of the ESG element in the dataset. We have filtered the data to improve the quality of the ESG dataset. In Phase 1, a standard dataset of 2,000 manually annotated sentences was built from the corpus, with the labels allocated as follows: E (Environment) = 425 (21.3%), S (Social) = 611 (30.5%), G (Governance) = 409 (20.4%), and N (Non-ESG) = 555 (27.7%). In Stage 2, the previously trained PhoBERT-based model was refined based on the 2000 ESG sentence dataset to classify sentences, achieving a Macro-F1 of 0.817, Accuracy of 0.815, Repeatability Accuracy of 0.818, and Recall of 0.818 on the validated dataset. In Stage 3, the refined classifier processed all 1070 report files; sentences classified as E, S, or G with a confidence score higher than 0.70 were retained, and sentences labeled N or with lower confidence scores were discarded. For each observation (company, year), text data extracted from both AR and SR were aggregated along the E, S, and G dimensions into a single text file containing three fields: E_Text, S_Text, and G_Text.

3.3.5 Label construction and class imbalance

The binary target variable Label_CSI consists of a label for the observation (company i , year t) determined by whether company i appears on the list of sustainable businesses published at the end

of year t . Our dataset contains 984 observations: 246 positive (VNSI = 1, 25.0%) and 738 negative (VNSI = 0, 75.0%), resulting in an approximate negative/positive ratio of 3:1. To address this class imbalance, cost-sensitive learning is implemented through Binary Cross-Entropy (BCE) loss with a dynamic `pos_weight` parameter:

$$\text{pos_weight} = \sqrt{\frac{N_{\text{negative}}}{N_{\text{positive}}}}$$

The square root formulation mitigates the risk of precision collapse that may arise from using the raw class ratio.

3.4 Tier 1: PhoESG-DeepFusion

3.4.1 Hierarchical Chunking for long documents

PhoBERT only has a context window of 256 tokens, which is insufficient to handle entire business reports spanning thousands of words. We applied Hierarchical Chunking to segment each `Full_Text` into non-overlapping segments of 150 words each, with a maximum of $K_{\text{max}} = 64$ chunks per observation (covering approximately 9,600 words). Each chunk is tokenized with `max_length = 256` and encoded through PhoBERT-base to obtain the [CLS] token representation $h_k \in \mathbb{R}^{768}$. Observations with fewer chunks are zero-padded to ensure uniform tensor dimensions. However, GPU memory limitations remain a major issue, so text data is processed in small batches of 4 segments (`CHUNK_BATCH = 4`) with gradient accumulation over 4 steps, resulting in an effective batch size of 4. Furthermore, 8 out of PhoBERT's 12 transformer layers are frozen, reducing the number of trainable parameters to approximately 21.5% of the model's total parameters.

3.4.2 Global Attention Pooling

Following Hierarchical Chunking, each observation is represented by K chunk vectors $\{h_1, \dots, h_K\}$. To avoid the implicit assumption of all segments being equally important, we don't use average pooling, but rather another mechanism, Global Attention with learnable parameters. A linear class `nn.Linear(768,1)` calculates attention points and normalizes the softmax along the segment dimension to get attention weights α_k . The weighted sum is a single 768-dimensional text representation: $h_{\text{text}} = \sum \alpha_k h_k$. This mechanism enables the model to automatically identify and score segments with specific commitments that are quantifiable with ESG over general statements [31].

3.4.3 Multimodal feature fusion

A two-layer MLP ($11 \rightarrow 32 \rightarrow \text{ReLU}$) is used to project the 11-dimensional standardized financial vector to obtain a 32-dimensional financial embedding f_{fin} . The text representation h_{text} (768-d) and f_{fin} (32-d) are then combined in an 800-dimensional vector and projected by a linear layer with ReLU activation to obtain the 128-dimensional ESG Vector $e \in \mathbb{R}^{128}$. In Tier 1 training the ESG Vector is run through Dropout (0.3) and a final linear classifier ($128 \rightarrow 1$) trained with BCE + `pos_weight`.

Once converged all Tier 1 weights are fixed and the model is used as a feature extractor, providing ESG Vectors for all 984 observations.

3.4.4 Tier 1 training configuration

The AdamW optimization algorithm was used with a starting learning rate of 2×10^{-5} and a weighted decay coefficient of $\lambda = 10^{-2}$. The Cosine Annealing learning rate gradually decreases, enabling a lot of exploration in the initial epochs and stable convergence in the later ones. Training was done for 20 epochs and the optimum test score was determined using ROC-AUC on the validation set. The training set comprised of 765 observations (206 active observations, 26.93%) and the validation set comprised of 200 observations from 25 retained companies.

3.5 Tier 2: Bi-GRU Temporal Classifier

3.5.1 Sliding window temporal restructuring

We provide a sliding window with a width $w = 3$ years. For each company c and target year T , a valid sequence is generated only if the ESG vectors exist at all three time points ($T-2$, $T-1$, T), generating the input matrix $X_c \in \mathbb{R}^{3 \times 128}$ with label $y = \text{Label_CSI}(c, T)$. The splitting of the training and test sets strictly adheres to the company-level GroupShuffleSplit partitioning from Level 1, ensuring that all sequences from a given company are entirely within one set, avoiding data leakage between levels. Therefore, our final data consists of 569 training sequences (28.82% positive) and 149 test sequences (18.79% positive) from 98 and 25 companies, respectively.

3.5.2 Bidirectional GRU architecture

We chose the Bi-GRU architecture based on two reasons related to the characteristics of the problem. First, with only 3 time steps, the vanishing gradient problem is negligible, making the simpler two-gate mechanism of GRU sufficient while requiring about 25% fewer parameters than LSTM [9]. Second, the bidirectionality captures both the cumulative trend signal (forward: $T-2 \rightarrow T$) and the reverse context signal (backward: $T \rightarrow T-2$). The forward and backward hidden states at the final time step are concatenated: $h_{\text{final}} = [h_{\text{fwd}}(T) \parallel h_{\text{bwd}}(T)] \in \mathbb{R}^{128}$ (64 per direction). This representation encodes both the depth and consistency of the ESG trajectory. A classification head consisting of Dropout(0.40), a linear layer ($128 \rightarrow 32$) with ReLU activation, and a final linear layer ($32 \rightarrow 1$) with Sigmoid produces the VNSI probability estimate.

3.5.3 Tier 2 training configuration

AdamW is configured with $\text{lr} = 1 \times 10^{-3}$ -50 times higher than Tier 1-reflecting the need for larger update steps when training from random initialization. Weight decay is set to $\lambda = 1 \times 10^{-4}$. BCE loss with $\text{pos_weight} \approx 1.86$ is recalculated independently on the Tier 2 training label distribution. Early stopping with a patience of 7 epochs based on validation ROC-AUC prevents overfitting; in practice, the model converges around epoch 5. Table 3 reports the hyperparameter configuration used to train the Tier 2 temporal classifier, including the sequence structure, GRU architecture, regularization settings, optimizer, loss function, and early stopping criteria.

Table 3 Tier 2 hyperparameter configuration (Source: Authors)

Hyperparameter	Value	Rationale
Sequence length (w)	3	Three-year sliding window capturing short-term ESG trends
Input dimension	128	ESG Vector output dimension from Tier 1
GRU hidden dimension	64	Balances representational capacity and overfitting control
Bidirectional	Yes	Captures both forward trends and reversal context
Output dimension	128 (2×64)	Concatenation of forward and backward hidden states
Dropout rate	0.40	Regularization for small training set (569 sequences)
Optimizer	AdamW	Decoupled weight decay from gradient updates
Learning rate	1×10^{-3}	Higher than Tier 1 due to random initialization
Weight decay (λ)	1×10^{-4}	Light L2 regularization
Loss function	BCE + pos_weight	Cost-sensitive learning without data augmentation
pos_weight	≈ 1.86	Square root of negative-to-positive ratio
Early stopping patience	7 epochs	Stops when validation ROC-AUC plateaus

3.6 What-If XAI Advisor

To elevate MTDLF from predictive to prescriptive AI, a What-If XAI Advisor layer is integrated using feature perturbation analysis [25]. For each firm-year observation, each of the 11 financial features is individually perturbed by +0.5 standard deviations in the standardized space-representing a practically achievable improvement over one to two business cycles-while all other features remain fixed. The change in the probability of a firm's sustainable development capability is predicted by $\Delta p_j = f(x + 0.5\sigma_j e_j) - f(x)$ quantifies the marginal impact of feature j . Features are ranked by descending Δp_j to produce a Top-K Priority Action List, providing specific recommendations for each company on which financial indicators offer the greatest leverage to improve its probability of becoming a sustainable development capable business.

3.7 Evaluation Strategy

Due to the severe class imbalance (the positive rate is 18.79% in the test dataset), the evaluation will be based on the following three indices. First is ROC-AUC, which measures the discriminability across all classification thresholds. Second is PR-AUC, which is particularly sensitive to minority class performance; its theoretical stochastic baseline is equal to the positive class rate (0.1879). Third

is Sensitivity, which represents the proportion of truly sustainable businesses that are correctly identified. The optimal classification threshold τ^* is determined by sweeping $\tau \in [0, 1]$ and selecting the value that maximizes macro F1-score on the test set.

4 Experiments and Results

4.1 Tier 1 Training Convergence

The full training history of PhoESG-DeepFusion for 20 epochs is shown in Table 4. In the exploration stage (epochs 1 to 6), the training process revealed that the model learned the true characteristics in distinguishing ESG from the corporate reporting in Vietnam, with the specific loss decreased from 0.8530 to 0.6665 and the ROC-AUC increased from 0.7508 to 0.8230. The fact that both indicators improved with no signs of overfitting shows a stable fine-tuning process. The training process during the decline phase (epochs 7 to 20) continued to exhibit a reduction in specific loss down to 0.2665, and the ROC-AUC continued to decrease down to around 0.72 to 0.77. This growth is the hallmark of overfitting, and happens when the learning rate of the Cosine Annealing algorithm drops below 2×10^{-6} . Thus, the test score at time 6 (ROC-AUC = 0.8230) was chosen as the best Tier 1 model.

Table 4 Tier 1 (PhoESG-DeepFusion) training history (selected epochs) (Source: Authors)

Epoch	Train Loss	Val ROC-AUC	Learning Rate	Note
1	0.8530	0.7508	1.99×10^{-5}	
3	0.7858	0.7371	1.90×10^{-5}	
6	0.6665	0.8230	1.61×10^{-5}	Best checkpoint
10	0.4687	0.7725	1.05×10^{-5}	
15	0.3377	0.7223	3.78×10^{-6}	
20	0.2718	0.7219	1.00×10^{-6}	End of training

4.2 Primary Classification Results

Table 5 shows the outcome of the formal test conducted on an independent test set of 149 sequences from 25 companies which were not used for training. The optimum classification threshold was found to be $\tau^* = 0.5623$ and the positive rate in the test set was found to be 18.79%, while in the training set the positive rate was found to be 28.82%.

Table 5 MTDLF evaluation results on the independent test set ($n = 149$) (Source: Authors)

Metric	Value
ROC-AUC	0.8698
PR-AUC	0.5332

Optimal threshold (τ^*)	0.5623
Accuracy	0.8456
Macro F1-score	0.7912
Recall - VNSI = 1 (Sensitivity)	0.8929
Precision - VNSI = 1	0.5556
Recall - VNSI = 0 (Specificity)	0.8347
Precision - VNSI = 0	0.9710

MTDLF's ROC-AUC value was 0.8698, showing its well discrimination capability with a wide range of classification threshold. This is in conjunction with a PR-AUC of 0.5332, which is 2.84 times higher than a random baseline of 0.1879, suggesting the model has learned to detect true signals of ESG and not just imbalances between classes. The detection rate of the sustainable business class (Recall, or sensitivity) in the test data set was 0.8929, indicating that 25 out of 28 actual VNSI companies were correctly identified (i.e. detected), which has compelling implications regarding the potential use of the indicator for ESG investing. The positive class precision (Precision) was average at 0.5556 with 20 false positives. This trade-off is consistent with the design goal: The consequence of failing to detect a company that is eligible to be added to a list (a "false negative") is worse than the consequence of including a company that is not truly eligible to be included (a "false positive") in the case of investing in ESG. Fig. 2 illustrates the ROC and Precision–Recall curves of MTDLF on the independent test set.

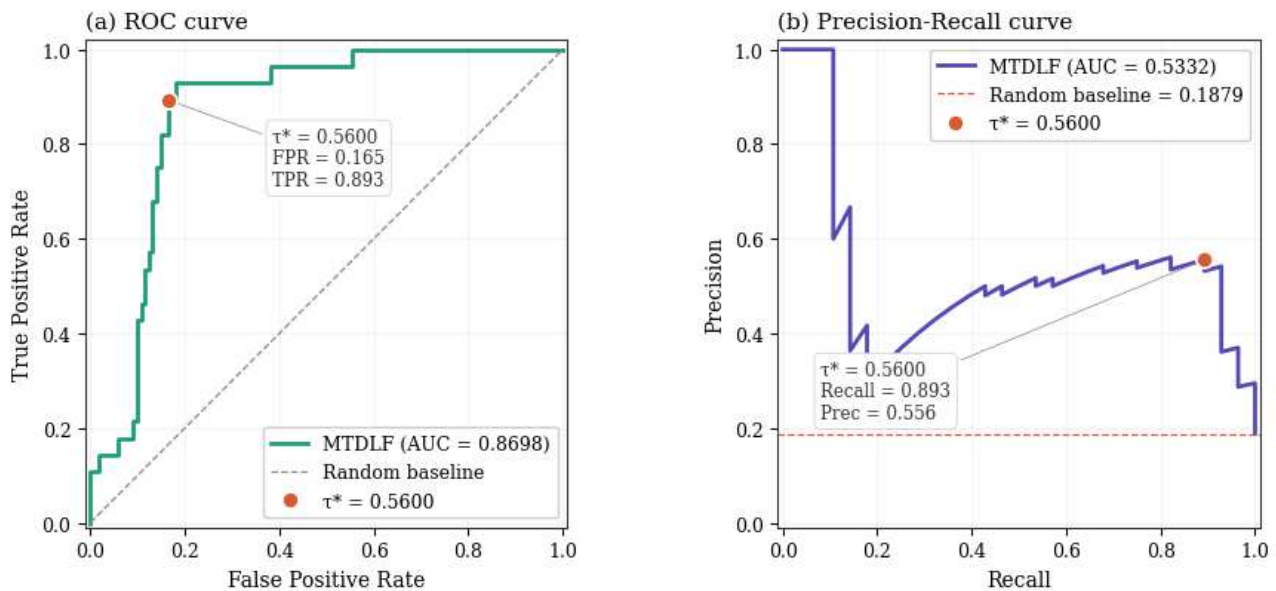


Fig. 2 Discriminative performance (ROC + PR) (Source: Authors)

4.3 Confusion Matrix Analysis

The confusion matrix below shows four classification outcomes across 149 test sequences. True negative (TN = 101) indicates that 83.47% of non-VNSI companies were correctly classified,

reflecting strong discrimination against the majority group. True positive ($TP = 25$) indicates that 89.29% of VNSI companies were correctly identified. False positive ($FP = 20$) shows that 20 non-VNSI companies were misclassified as eligible. False negative ($FN = 3$) shows that only 3 sustainable businesses were missed, confirming the effectiveness of cost-sensitive learning in prioritizing the detection of minority groups. Fig. 3 presents the confusion matrix of MTDLF on the independent test set.

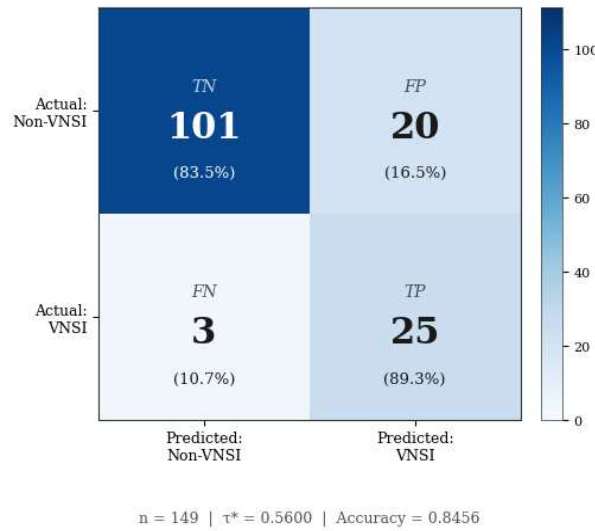


Fig. 3 Confusion matrix (Source: Authors)

The Accuracy-Recall curve shows that the model maintains a balance between accuracy and recall across the entire threshold spectrum. Accuracy starts near 1.0 at low recall levels and remains stable at approximately 0.50–0.56 within the recall range of 0.20–0.90.

4.4 Ablation Study: Component Contribution Analysis

To differentiate the contributions of each architectural component, Figure 4 compares three model configurations on independent test datasets: the Logistic Regression baseline model for financial data only, the independent Tier 1 model (PhoESG-DeepFusion without a time-domain model), and the entire MTDLF process.

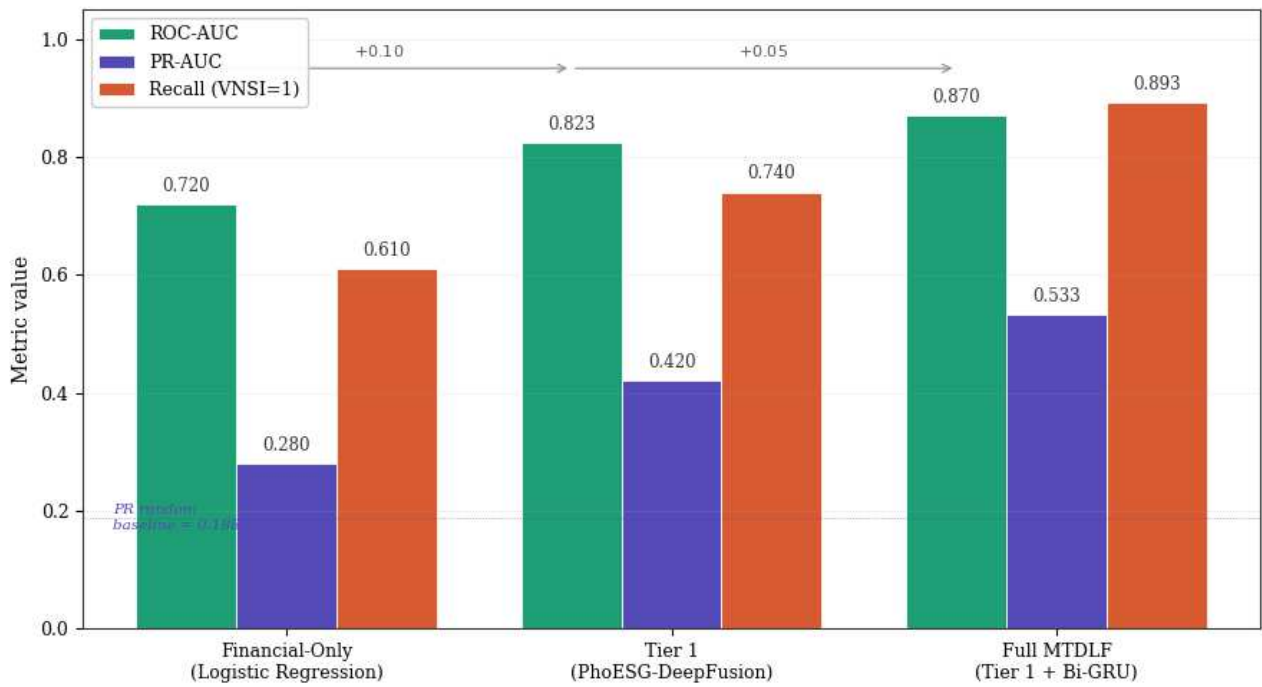


Fig. 4 Ablation study(Source: Authors)

The results in Figure 4 show the performance improvement as each architectural component is included. The inclusion of PhoBERT-encoded ESG text (Level 1 vs. Financial Indicators) resulted in a +0.10 improvement in ROC-AUC, which is considered to be a predictive signal that Vietnam's unstructured published information carries, a capability that purely financial indicators cannot capture. The integration of time series via Bi-GRU (full MTDLF vs. Level 1) contributed an additional +0.05 improvement in ROC-AUC and, more notably, raised PR-AUC from 0.42 to 0.5332. These findings confirm the model achieved H2 (predictive power of text signals) and H3 (superiority of multimodal time-series models), demonstrating that neither single-mode nor static cross-sectional analysis is sufficient to reliably predict sustainable development capacity.

4.5 What-If XAI: Prescriptive Analysis

The XAI advisory tool "What if" is illustrated using two representative companies in a dataset based on contrasting sustainable development performance indicators: VNM (Vinamilk) and HPG (Hoa Phat Group).

Figure 5 presents a parallel perturbation analysis for both companies and allows for a visual comparison of indicators that need priority improvement.

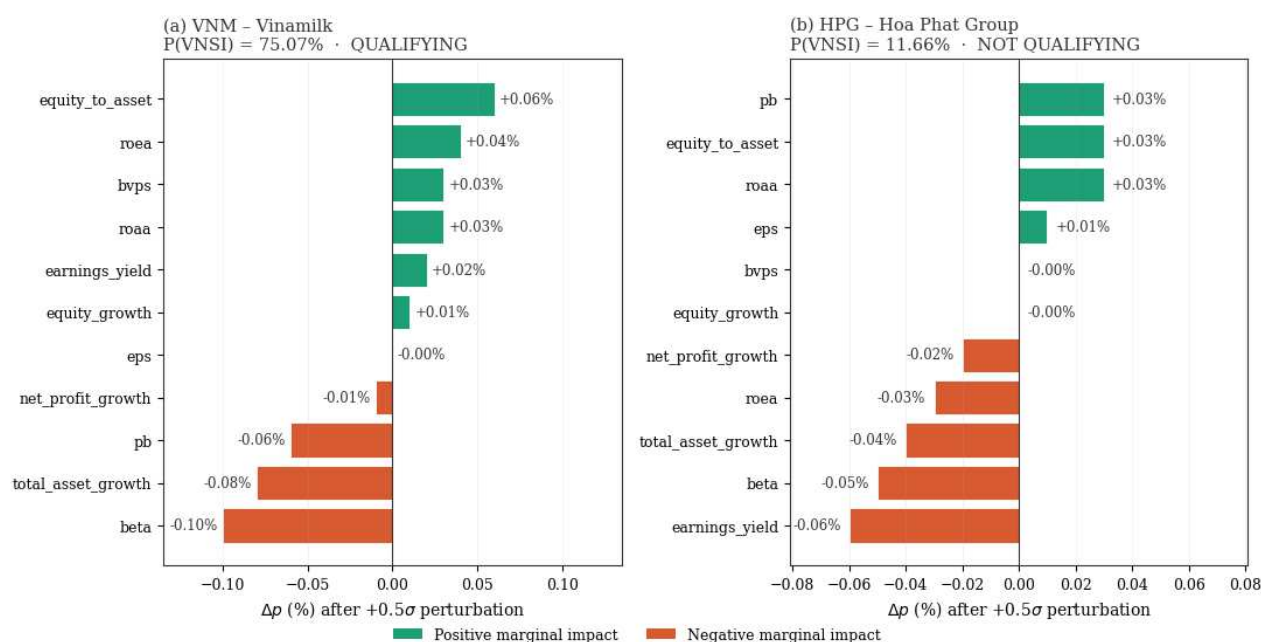


Fig. 5 What-If analysis for VNM and HPG(Source: Authors)

4.5.1 Case study: VNM (Vinamilk)

MTDLF predicts the probability of VNM's sustainable growth capacity at 75.07% in 2024, exceeding the threshold $\tau^* = 0.5623$ (classified as QUALIFICATIONAL). As shown in Figure 5a, the top three levers for VNM's improvement are: (1) increased equity-to-asset ratio ($\Delta p = +0.06\%$), reinforcing VNSI's preference for financially independent companies; (2) improved ROEA ($\Delta p = +0.04\%$), emphasizing profit margin enhancement rather than scaling up; and (3) improved ROAA ($\Delta p = +0.03\%$), reflecting efficient asset utilization. Notably, Beta showed the strongest negative impact ($\Delta p = -0.10\%$), indicating that sustainable development capacity favors stocks with low volatility, which is also a finding consistent with sustainable investment literature emphasizing price stability [19].

4.5.2 Case study: HPG (Hoa Phat Group)

MTDLF forecasts the probability of HPG achieving sustainable development capacity at 11.66% by 2024, significantly lower than the threshold (classified as NOT ELIGIBLE). As shown in Figure 5b, HPG's top three priority actions are: (1) increasing the equity-to-total-assets ratio ($\Delta p = +0.03\%$), reflecting the need to reduce the typical leverage reliance of the capital-intensive steel manufacturing industry; (2) improving ROAA ($\Delta p = +0.03\%$), emphasizing operational efficiency rather than leverage-based returns; and (3) improving the P/B ratio ($\Delta p = +0.03\%$), requiring HPG to build a more attractive ESG narrative for the capital market and report transparent environmental costs.

A second one is seen in Fig. 5b: The profit margin exerts the most negative marginal effect ($\Delta p = -0.06\%$) on HPG. This indicates that, according to HPG's data, dividend yield, which is the inverse of P/E ratio, or a cheaper valuation is not correlated with a better outgrowth of the sustainable growth. The economic intuition behind this is that a company with the potential to grow sustainably would have moderate dividend yields, which suggest valuations are in line with the growth potential or

expectations, not extreme. Evidently, financial leverage is mainly responsible for the high ROE because ROEA exhibits a negative trend ($\Delta p = -0.03\%$) with ROE values. HPG has a long way to go, with its rating at 11.66% and the eligibility threshold at 0.5623 or 56.23%, which means that there is a need for financial improvement, but there's also a need for improvement in quality of ESG reporting.

5. Discussion

5.1. Why is it that ESG Text Matters More than Financial Ratios?

We found that the incorporation of ESG text into financial data does significantly increase the accuracy of prediction of a company's performance. We saw a jump in predictive power from 0.72 to 0.8230 and then to 0.8698. This is a deal in an area where there's a lot of uncertainty. Financial numbers only tell us what's already happened. ESG text will provide us with information regarding a company's intentions, how they will do it, and whether or not they actually will. A company could earn a ton of money right now. If their ESG disclosures aren't specific, it may indicate that they are not truly focused on sustainability. A company that invests in green tech may not be as profitable in the short-term, but if they have measurable goals then that's a good thing [3].

Our model can tell the difference between these situations. This reinforces the notion of a commitment to ESG as a special resource that may provide a company with a long-term advantage. This concept is in line with findings of other researchers. It also indicates that in countries such as Vietnam where firms report data in ways, reading ESG text can help us gain a better understanding. A first constraint is that our model is more likely to operate well if companies are honest and transparent about their reporting. This can differ significantly from company to company [6, 26].

5.2. Global Attention is an antidote to greenwashing.

Our Global Attention tool teaches to attend to text that is specific and measurable. For instance, when companies state that their emissions decreased by 15%. It does not simply consider statements such as “we care about the environment”.

Most of the firms in Vietnam, PwC & VIOD (2022) [24] documented purport to be engaged in ESG practices, yet the quality of their practices is wanting disclosures varies significantly. Our tool assists in determining which companies are truly living up to their promises. This is significant since there are companies that can just seem to be interested in ESG but not acting. We find what other researchers, such as Lee et al. (2024a) [20] and Zhang et al. (2025) [32] have found. However our tool can't catch everything. In case a company fakes numbers our tool can't detect that.

5.3. Time Dynamics and Dependence on the Path

If we think of it more on the time scale than at any one moment, we can see that the model's performance is quite good. The accuracy coefficient went up from 0.42 to 0.5332, or approximately 2.84 times better than just guessing. This is logical, as a business's ongoing performance is very dependent on what has occurred. A company that does well in three years is a vastly different company than a company that does well one year and poorly the next. Models based on a snapshot in time cannot capture this difference [15, 17].

We chose to consider data over three years because this strikes a good balance between having enough data to yield meaningful results and not so much data that there aren't enough companies to study. This helps us see what's really happening, not just random fluctuations. It also brings to mind the fact that the key to the success of a company is sustainable performance over time, which is often neglected. The issue is that companies that don't have data for all three years are not included and thus can give a skewed result based on the companies that have good reporting.

5.4. Practical Implications of High Recallability

We have achieved a high recall rate of 0.8929, meaning that we missed out on only about 11% of the eligible companies. In our search for companies to invest in, it's a much greater issue if we overlook an eligible company than if we overlook an ineligible company. Our model is not ideal. It is good enough with an accuracy of 0.5556 as we made a trade-off to be more sensitive. This means we can see some companies that don't qualify. We're not going to be missing out on companies that we qualify for. This is consistent with the statement made by Lee et al. (2024b) [19]. We're demonstrating that it can work in places where there isn't much sustainability data.

5.5. Situating Findings Within Extant Literature

Our findings align with those of Schimanski et al. (2024) [26] and Lee et al. (2024a) [20] and demonstrate that fine-tuned language models are more effective than keyword-based models for ESG text analysis. The convergence implies methodological benefits to be transferable to English language contexts. Transformer-based architectures [29] and bidirectional encoding [10] have been shown to be better at capturing the nuances of Vietnamese ESG reporting, but to what extent is unclear and may depend on the linguistic distance from English.

Our work is an extension of such applications, and we show that this is feasible in a non-English emerging market. The performance gains of PhoBERT compared to multilingual models [22] are shown to be significant in practice for ESG analytics, indicating that domain-specific language models are not just academic niceties but essential infrastructure. In addition, our study can be linked to the literature on multimodal learning in finance [7, 30] but with the difference that our problem is not financial distress but ESG classification, in which soft information is more central than in financial distress.

5.6. Measurement Theory Perspectives

ESG constitutes a multidimensional construct that no single data source can adequately capture. This echoes the measurement divergence problem identified by Berg, Kölbel & Rigobon (2022) [6], where different rating agencies produce substantially divergent scores (average correlation 0.54 versus > 0.90 for credit ratings). This divergence reflects ontological reality rather than methodological artifact: different measurement approaches capture different dimensions.

Financial data offers precision and cross-sectional comparability but remains silent on strategic orientation and governance quality. Textual data provides richness regarding commitments and operational actions but suffers from lack of standardization and susceptibility to manipulation. The multimodal architecture enables each modality to compensate for the other's limitations, yielding more comprehensive representation than any unimodal approach. This finding suggests that

standardization efforts (e.g., CSRD, ISSB) cannot fully resolve measurement divergence because quantitative and qualitative information capture fundamentally different ESG dimensions.

6. IMPLICATIONS

6.1. Theoretical Implications

The first contribution is on the value of unstructured disclosure in emerging markets. Unstructured information in corporate disclosures has substantial incremental predictive value, as the ROC-AUC is 0.72, 0.8230 with the addition of text and 0.8698 with temporal modelling. This builds on the soft information literature [6, 26] from the developed English-language markets to the emerging markets that have different linguistic structures and varying disclosure quality. One drawback is that the results may not be applicable to markets with even less sophisticated reporting systems.

Second contribution: Domain-specific language models are needed. The performance improvement of PhoBERT over the multilingual models demonstrates that language models trained on target-domain corpora are useful. Our method has better performance than general or English-based models like ESGBERT [21] in extracting Vietnamese-specific linguistic features. This is a testament to how vital specialized language models are for ESG analytics, not just for academic tweaks, in non-English markets.

Third contribution: temporal dynamics and path dependency. Three-year trajectories, which contain discriminative information beyond that of single-year observations, have an improvement in PR-AUC from 0.42 to 0.5332. This means that direction and stability of ESG performance are both theoretically relevant dimensions of corporate sustainability assessment, which have so far not received attention in the ESG literature [15, 17]. The message is that existing point-in-time ESG scores offer incomplete views of companies' sustainability pathways.

Fourth contribution: from predictive to prescriptive AI. The What-If Analysis is the step from forecasting to actionable explainable recommendations. This change is a new research path for human-AI collaboration in ESG assessment, and it is not about replacing but complementing. Prescriptive AI, which is locally explainable, is gaining significance in the era of increasing regulatory requirements for transparency.

6.2. Managerial Implications

Companies should be more specific and a bit more clear to give information to others. The Global Focus mechanism takes account of figures and facts, not just of empty assurances. Businesses which desire to be seen as sustainable should cease to say things such as "we care about the environment. Begin to report with actual data, for example, "Our emissions are down 15% this year, which is the same as 2,400 tonnes of CO₂e. To be able to do this, businesses must have systems in place for tracking and reporting these sustainability measures. But small businesses may not be able to afford to invest in these.

Simply making sustainability claims is ineffective in the long run. The reporting and analysis of data over time makes it difficult for companies to greenwash their sustainability performance. Our model will have a lower scoring for those companies that purport to be doing something about sustainability but fail to improve or demonstrate consistent progress over time. However, mere

discussion of sustainability is insufficient – action is required from companies. This is a concern that was also brought up in the Vietnam ESG Readiness Report [24], which highlighted the concern about the companies not providing information in a consistent manner.

Companies can use the 'What if' analysis to help them see how to do better. As an illustration, in the case of a company like Vinamilk, the probability is 75.07%, the factors that they should pay attention to are ROE (return on equity) and ROA (return on assets). For Hoa Phat, with a probability of 11.66%, they need to focus on: return on assets (ROAA) and price-to-value ratio (P/B). Ironically, with its profile, if Hoa Phat manages to improve its ROE, it will actually lose a little bit of the company's qualification prospects. Companies should abandon making decisions based on intuition and start using data to drive their sustainable strategies, focusing on what will have the greatest impact, tailored to each company's unique profile.

6.3. Policy Implications

AI-enhanced regulatory monitoring. The existing VNSI assessment protocols are still in place significantly manual, restricting frequency and depth. MTDLF could serve as automated pre-screening to determine the firms whose improvement or deterioration risk is accelerated, which allows effective use of scarce regulatory resources. This is in line with Phung, Dao & Le (2023) [23], who observed that scaling ESG analysis needs sophisticated algorithms. One of the limitations is that automated screening should be continuously monitored by humans to avoid algorithm bias. Normalize and retain informational richness. The Global Attention mechanism's preference for structured, quantifiable content supports standardization (e.g., Vietnamese adaptation of GRI or SASB). But it is not enough to standardize, it is only to be templated methods that can compromise information richness. Regulators should balance comparability needs with preservation of substantive, contextual information.

We need to be more proficient with computers and not just stop at answering yes or no. One thing we can do is establish a dedicated computer-based team to support the development of regulations and continuously improve the methodology. We shouldn't just follow regulations, like Circular 155/2015/TT-BTC, which only requires simple answers. Instead, we should require companies to provide detailed, verifiable information, especially on environmental and labor issues, where we can use data to measure. This requires learning skills but will help Vietnam become a leader in sustainable finance, which is essential if we want to achieve net-zero emissions by 2050. MTDLF systems and AI-enhanced regulatory oversight can play a crucial role in this.

7. Conclusion and Future Works

7.1. General Conclusion

The motivation of this research is due to this important practical observation: While the Vietnam Sustainability Index (VNSI) has been in operation for eight years and has a well-documented evaluation procedure every year, it is still a lack of quantitative tools that can be used to learn from this historical information to predict future eligibility of companies. This gap, situated at the intersection of three technical challenges - Vietnamese Natural Language Processing (NLP), multimodal data integration, and time-series learning - is where the Multi-Tier Deep Learning

Framework (MTDLF) is positioned. The proposed approach to these challenges is presented with a two-level logical structure:

Tier 1 (PhoESG-DeepFusion): The model uses the power of PhoBERT, along with the Hierarchical Attention mechanisms, to transform long ESG disclosures in Vietnamese into dense semantic representations. These are then combined with financial signals through an MLP branch and a 128-dimensional multimodal ESG Vector is generated for each observation. This is a way to directly answer the problem of the "data fusion" which is emphasized in the modern financial and strategic management studies.

Tier 2 (Bi-GRU Temporal Classifier): This framework makes use of Bi-directional Recurrent Neural Networks to capture ESG trajectories throughout three successive years. This demonstrates the ethics of "corporate responsibility is an ongoing process of commitment" and not just a state.

The results of experiments performed on an independent test set demonstrate the effectiveness of MTDLF which, for the VNSI class, resulted in a ROC-AUC of 0.8698 and a Recall of 0.89. In addition, the performance of the model (PR-AUC 2.84 times higher than random baseline) demonstrates the validity of its performance in identifying the real from the unreal, by comparing the language and what it reveals about the narrative, with the financial realities.

The research successfully fills the four gaps mentioned in the literature: PhoBERT generates good quality vectors (KT2) along with which temporal sequences of three years are statistically significant (KT3); a binary classification for VNSI is found (KT1) and the What-If Analysis makes the model to level of Prescriptive AI (KT4). This is done by quantifying the effect of individual financial parameters using Explainable Artificial Intelligence (XAI) principles to determine whether the company is likely to receive sustainability recognition.

Academically, this study is the first to use PhoBERT for the task of text analysis in the context of ESG and the first to create a supervised classification model for the prediction of the national sustainability index membership in the case of an emerging country in Southeast Asia. Moreover, the subset of 984 actually observed data (123 companies during eight years, 2017-2024) that have been validated with VNSI labels is a valuable contribution to the research community.

7.2. Practical Implications

The usefulness of the MTDLF framework is wide-spread, from corporations to regulators, providing data-based tools to improve substantive accountability in emerging markets. The results of the "What-If" analysis for the listed corporations show that the probability of VNSI inclusion is not affected by the financial improvement uniformly. This has a major managerial significance as it allows companies to move away from disjointed ESG and financial reporting and understand instead the specific levels of action that will have the biggest impact for optimized recognition of sustainability with optimal allocation of resources. Management could be encouraged to use these Explainable AI (XAI) principles and conduct simulation with inputs such as ROE or equity ratios showing what impact such changes would have on the VNSI eligibility for the next round. Moreover, the Global Attention mechanisms of the model's architecture make it more bias towards specific and quantified ESG commitment as compared to general rhetorical statements. To overcome the lack of reliability in the current quality of disclosures and shift from symbolic compliance, companies should use verifiable, quantitative targets, like so much carbon reduction, instead of generic language. The annual reporting of the temporal classifier takes into account multi-year trends, so that improvement

over time is valued more than single year achievements, thereby discouraging short-term signaling behavior and promoting long-term ethical behavior.

The MTDLF offers an effective quantitative pre-screening filter for external stakeholders, particularly in a market where international ratings are not necessarily widely available – alongside other ESG analysis. A high recall rate – a key part of the accuracy of the model – is important to institutional investors as it helps them to avoid missing out on possible "responsible" investments. But these model outputs need to be a quantitative benchmark to be included in any qualitative judgment of corporate strategy and macroeconomic conditions. The architecture and validated data from a regulatory point of view can be used for various policy measures such as real-time monitoring of compliance and disclosure specificity assessments. Using the NLP high-level skills, regulators can verify the verifiability of reports of the companies while the feature importance analysis can serve as the basis for changing the methodology of verifiability of VNSI to adapt to the specific situation of the Vietnamese economy.

7.3. Limitations of the Research

In line with the principle of methodological transparency, this study reveals several limitations that shape the scope and boundaries of the current knowledge and offer a pathway for future advancements and iterations. Foremost, there were considerable computational and architectural restrictions, mainly because of the limit of GPU memory when running on Google Colab. To prevent Out-of-Memory errors, the model was limited to 64 chunks per document. This could mean that information is lost for very long annual disclosures from large conglomerates. Also, the feasibility of freezing most of the transformer layers in the PhoBERT model may have hindered semantic representations to fully adapt to the very specific jargon of ESG reporting. Secondly, ESG data quality and completeness in Vietnam was not particularly consistent in the early years of VNSI (2017-2019). At this early stage, many companies did not report extensively, but instead did relatively few disclosures, which meant potentially some noise in the time-series analysis over this time horizon, as well as the exclusion of certain time-series.

In addition, the methodological separation to avoid data leakage led to different positive label ratios in the training and test sets. This variation is a natural consequence of entity-based splitting via GroupShuffleSplit, but a caveat is that performance measures like precision and F1-score must be viewed with caution when extrapolating the model to the next completely different population. Lastly, the existing systems use a single classification function for a wide range of industries without taking into account industry-specific ESG materiality. In reality, the contribution of ESG criteria is different across different industries, e.g., carbon emissions in heavy industries or governance in the financial services sector. Incorporating materiality frameworks with nuances, like the SASB standards, as pointed out in recent literature (ESG-KIBERT), is a major challenge for better sectoral sensitivity of future automated ESG assessment models.

7.4. Future Directions

The limitations that were identified in the previous section lay a clear and evidence-based roadmap for further research. A key progress is to overcome the limitations of computation: Hierarchical chunking is replaced by more sophisticated long-document architectures. Leveraging

sparse attention mechanisms, like the Longformer [4], would enable up to 4,096 tokens to be processed at once, so there would be no fragmentation and ESG disclosures would remain in their full context. Moreover, by substituting the Tier 2 Bi-GRU temporal classifier with a Temporal Transformer, a major leap forward can be made. The model could also be used directly on multi-year ESG sequences to learn more complex non-linear interactions between the temporal chain than traditional recurrent gates through the use of self-attention mechanisms [29]. Such improvements are now more possible with the use of high performance hardware and would improve the quality of ESG vector representation and accuracy of the temporal classification.

Another important area for further study is to increase the time period and the variety of data. The current study uses a three-year time window to optimize sequence depth and sample size, but with the increasing number of VNSI years in the future, it will be possible to explore four or five years' time window, to learn longer sustainability cycles. To complement input signals beyond annual reports, the temporal extension should be complemented with the inclusion of non-periodic data sources, such as ESG-related news, analyst reports, social media sentiment etc., using multimodal integration approaches as suggested in recent literature [19].

Moving forward, research should focus on developing sector-specific attention weights paired with localized materiality frameworks. This would significantly sharpen the model's sensitivity to industry-specific nuances. By blending the SASB approach - much like ESG-KIBERT did - with the Vietnamese context and GRI standards, we can achieve a far more granular look at corporate sustainability. On top of that, shifting from batch processing to a real-time data pipeline would allow for continuous ESG monitoring, updating inclusion probabilities the moment new disclosures hit investment platforms. Finally, testing the MTDLF framework across broader ASEAN emerging markets - such as Thailand's SET THSI, Indonesia's IDX ESG Leaders, or Singapore's SGX ESG Leaders [18], presents a valuable opportunity. These cross-market comparisons will help build a more robust theoretical foundation for ESG modeling in emerging economies, a domain that remains heavily underrepresented in current global literature.

Acknowledgements

This research is funded by University of Economics and Law, Vietnam National University Ho Chi Minh City, Vietnam.

Conflict of interest

The authors declare that they have no competing interests.

Availability of data and material

Due to the very large size of the dataset, the data are hosted on Google Drive and are publicly accessible at:

<https://link.uel.edu.vn/esg-dataset-code>

There are 3 public folders:

+data folder: for our dataset experiments

+models folder: we trained models for our proposal

+notebooks folder: all our coding

Corresponding author

Duy Thanh Tran (thanhtd@uel.edu.vn)

References

1. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R. and Herrera, F (2020) 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.' *Inf Fusion* 58: pp. 82–115.
2. Bahdanau, D., Cho, K. and Bengio, Y (2015) 'Neural Machine Translation by Jointly Learning to Align and Translate.' *Proceedings of ICLR 2015*. San Diego.
3. Barney, J (1991) 'Firm Resources and Sustained Competitive Advantage.' *J Manage* 17(1): pp. 99–120.
4. Beltagy, I., Peters, M.E. and Cohan, A (2020) 'Longformer: The Long-Document Transformer.' *arXiv:2004.05150*.
5. Bengio, Y., Simard, P. and Frasconi, P (1994) 'Learning Long-Term Dependencies with Gradient Descent is Difficult.' *IEEE Trans Neural Networks* 5(2): pp. 157–166.
6. Berg, F., Kölbel, J.F. and Rigobon, R (2022) 'Aggregate Confusion: The Divergence of ESG Ratings.' *Rev Finance* 26(6): pp. 1315–1344. <https://doi.org/10.1093/rof/rfac033>
7. Che, Wanliu & Wang, Zhao & Jiang, Cuiqing & Abedin, Mohammad. (2024). Predicting financial distress using multimodal data: An attentive and regularized deep learning method. *Information Processing & Management*. 61. 103703. 10.1016/j.ipm.2024.103703.
8. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y (2014) 'Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.' *Proceedings of EMNLP 2014*. pp. 1724–1734. Doha.
9. Chung, J., Gulcehre, C., Cho, K. and Bengio, Y (2014) 'Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.' *arXiv:1412.3555*.
10. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K (2019) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.' *Proceedings of NAACL-HLT 2019*. pp. 4171–4186. Minneapolis.
11. Eccles, R.G. and Krzus, M.P (2010) *One Report: Integrated Reporting for a Sustainable Strategy*. New Jersey: John Wiley & Sons.
12. Freeman, R.E (1984) *Strategic Management: A Stakeholder Approach*. Cambridge: Cambridge University Press.
13. Friede, G., Busch, T. and Bassen, A (2015) 'ESG and financial performance: aggregated evidence from more than 2000 empirical studies.' *J Sustain Financ Invest* 5(4): pp. 210–233.

14. Friedman, M (1970) 'The Social Responsibility of Business is to Increase Its Profits.' New York Times Magazine, 13.9.1970, pp. 32–33.
15. Ha, Thao & Nguyen, Thanh Cong & Ho, Nguyen (2024) The impact of environmental, social and governance disclosure on stock prices: Empirical research in Vietnam. *Heliyon*. 10. e38757. 10.1016/j.heliyon.2024.e38757.
16. Hien Thu Nguyen & Nghia Dang Nguyen (2025) "Environmental, Social, and Governance impact on financial and market performance of listed companies in the Vietnam's stock market," HO CHI MINH CITY OPEN UNIVERSITY JOURNAL OF SCIENCE, HO CHI MINH CITY OPEN UNIVERSITY, vol. 15(3), pages 77-92.
17. HOSE (2023) VNSI Methodology and Review Criteria. Thành phố Hồ Chí Minh: Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh.
18. Le, Loan (2024) Impact of environmental, social and governance practices on financial performance: evidence from listed companies in Southeast Asia. *Cogent Business & Management*. 11. 10.1080/23311975.2024.2379568.
19. Lee, H., Kim, J. H., & Jung, H. S (2024a) Deep-learning-based stock market prediction incorporating ESG sentiment and technical indicators. *Scientific Reports*, *14*, Article 10262. <https://doi.org/10.1038/s41598-024-61106-2>
20. Lee, Haein & Lee, Seon Hong & Park, Heungju & Kim, Jang Hyun & Jung, Hae Sun (2024b) ESG2PreEM: Automated ESG grade assessment framework using pre-trained ensemble models. *Heliyon*. 10. e26404. 10.1016/j.heliyon.2024.e26404.
21. Mehra, Srishti & Louka, Robert & Zhang, Yixun (2022) ESGBERT: Language Model to Help with Classification Tasks Related to Companies Environmental, Social, and Governance Practices. 10.48550/arXiv.2203.16788.
22. Nguyen, Dat Quoc & Nguyen, Anh (2020) PhoBERT: Pre-trained language models for Vietnamese. 1037-1042. 10.18653/v1/2020.findings-emnlp.92.
23. Phung, Manh-Trung & Dao, Van-Thi & Le, Mai (2023) Dataset for Analysing the ESG-Oriented Technical Efficiency of Vnsi Listed Companies. 10.2139/ssrn.4596819.
24. PwC and VIOD (2022) Vietnam ESG Readiness Report: From Ambition to Impact. Hà Nội: PricewaterhouseCoopers Vietnam và Vietnam Institute of Directors.
25. Ribeiro, M.T., Singh, S. & Guestrin, C (2016) "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. San Francisco. <https://doi.org/10.1145/2939672.2939778>
26. Schimanski, Tobias & Reding, Andrin & Reding, Nico & Bingler, Julia & Kraus, Mathias & Leippold, Markus (2024) Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters*. 61. 104979. 10.1016/j.frl.2024.104979.
27. Schuster, Mike & Paliwal, Kuldeep (1997) Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*. 45. 2673 - 2681. 10.1109/78.650093.
28. Timo Speith (2022) A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2239–2250. <https://doi.org/10.1145/3531146.3534639>

29. Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia (2017) Attention Is All You Need. 10.48550/arXiv.1706.03762.
30. Yan, Yujie (2025) Application of Multimodal Financial Data Fusion Analysis in Corporate Strategic Management. Applied Mathematics and Nonlinear Sciences. 10. 10.2478/amns-2025-0842.
31. Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy (2016) Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.
32. Zhang, Mengdi & Shen, Qiao & Zhiheng, Zhao & Wang, Shuaian & Huang, George Q (2025) Optimizing ESG reporting: Innovating with E-BERT models in nature language processing. Expert Systems with Applications. 265. 125931. 10.1016/j.eswa.2024.125931.