IOSR Journal Of Economics and Finance (IOSR-JEF)
e-ISSN: 2321-5933, p-ISSN: 2321-5925.Volume 17, Issue 4 Ser. 2 (July. – August. 2026), PP 32-41
www.Iosrjournals.Org

Carbon Neural Net Portfolio: The Role of Artificial Intelligence in Constructing Carbon-Neutral Investment Portfolios

Jhanak & Group
Under the mentorship of Prashant Mishra · Videshway Advisory

Abstract
This paper examines whether and how neural network architectures can be used to construct investment portfolios that deliver competitive risk-adjusted returns while achieving measurable, auditable reductions in carbon intensity. Against the backdrop of Paris-aligned Benchmarks, the Climate Transition Benchmarks, the EU Sustainable Finance Disclosure Regulation, and the recommendations of the Task Force on Climate-related Financial Disclosures, asset managers now face a dual optimisation problem that classical mean-variance methods were not designed to solve. I synthesise four converging literatures — sustainable portfolio theory, machine learning in asset management, climate finance, and explainable artificial intelligence — to develop a conceptual framework in which a neural-net portfolio engine is situated between heterogeneous climate-data inputs and a governance layer that enforces interpretability and regime monitoring. A structured literature and case-study review covering research published since 2018 shows that feed-forward networks with carbon-penalty loss terms, long short-term memory temporal predictors with quadratic-programming overlays, and reinforcement-learning agents have been applied to this problem, and that reported decarbonisation benefits of thirty percent or more in weighted average carbon intensity are achievable with tracking error consistent with institutional mandates. However, interpretability, data heterogeneity, and regime-shift risk remain binding constraints. The paper contributes a taxonomy of model-data combinations, a defensible methodological template for reproducible backtesting, and a set of priority research questions for empirical follow-up. It concludes that neural-net approaches are complements to — not replacements for — classical optimisation, and that their value is strongest when explainability interventions make their choices auditable to investors and regulators.
Keywords: *neural networks, ESG, sustainable finance, portfolio optimisation, carbon intensity, Paris-aligned benchmark, explainable AI, machine learning.*

Date of Submission: 07-07-2026 Date of Acceptance: 17-07-2026

I. Introduction

Global capital markets are being repositioned around net-zero commitments at a pace that would have seemed implausible a decade ago. Signatories to the Net Zero Asset Managers initiative now oversee tens of trillions of dollars in assets under management, and regulators across the European Union, the United Kingdom, and parts of Asia have moved from voluntary disclosure toward prescriptive benchmark construction. The Paris-aligned Benchmark and Climate Transition Benchmark families, introduced by the European Commission in 2019 and refined thereafter, demand a minimum annual decarbonisation trajectory of seven percent in weighted average carbon intensity — a number chosen to align portfolio-level emissions with a 1.5 degrees Celsius warming pathway. Meanwhile, the Sustainable Finance Disclosure Regulation imposes granular reporting obligations on products marketed as environmentally sustainable, and the Task Force on Climate-related Financial Disclosures framework — now formalised under the International Sustainability Standards Board — obliges issuers to disclose physical and transition risks in terms that portfolio managers can actually use.

These requirements create a practical problem. Classical mean-variance optimisation, in the tradition inaugurated by Markowitz in 1952, takes as its objective a trade-off between expected return and variance. It cannot natively ingest the heterogeneous climate signals that modern sustainability regulation demands: Scope 1, 2, and 3 emissions disclosures with varying quality and coverage; forward-looking implied temperature rise scores derived from company transition plans; transition-risk scores based on stranded-asset simulations; physical-risk scores derived from downscaled climate models; and unstructured signals such as satellite-derived emissions estimates, news sentiment on climate litigation, and supply-chain network data. Even where classical optimisers are adapted via tilting or constraint-based screening, they cannot efficiently capture the non-linear, regime-dependent relationships between climate risk, financial return, and policy change.

This is the gap that neural network-based portfolio construction — which I will refer to as the Carbon Neural Net Portfolio, or CNNP — aims to fill. Neural networks are universal function approximators: given sufficient data and appropriate regularisation, they can model arbitrarily complex mappings from a high-dimensional feature space to a set of portfolio weights or expected returns. In the context of sustainable investing, this flexibility matters because the relationship between climate exposure and financial return is non-linear (a carbon tax does not affect a utility and a software company proportionally), non-stationary (the marginal cost of a tonne of carbon changes with policy), and often intermediated by sector-level transition dynamics that a linear factor model cannot express cleanly.

The central research question of this paper is therefore the following: to what extent can neural network architectures construct portfolios that simultaneously optimise for carbon-intensity reduction and risk-adjusted return, compared with conventional mean-variance and ESG-tilted benchmarks, and what interpretability and governance conditions must hold for such models to be fit for institutional use? I defend a conditional thesis: neural network architectures, when combined with curated carbon and climate-risk data, can construct portfolios that achieve comparable risk-adjusted returns to conventional benchmarks while delivering substantially lower and more predictable carbon intensity than traditional optimisation methods, provided that interpretability, data-quality, and regime-shift risks are rigorously controlled.

The contribution of the paper is threefold. First, it synthesises four distinct literatures that have until recently proceeded in parallel — sustainable portfolio theory, machine learning in finance, climate finance, and explainable artificial intelligence — into a single framework that clarifies where neural networks add value and where they do not. Second, it provides a structured taxonomy of neural-network families and climate-data inputs as applied to portfolio construction, based on a review of peer-reviewed research and industry implementations published between 2018 and early 2026. Third, it specifies a reproducible methodological template that future undergraduate and graduate researchers can follow to conduct empirical backtests at modest computational cost.

The paper is structured as follows. Section 2 reviews the four scholarly conversations that the topic sits within, along with a fifth strand on carbon-data quality that cuts across them. Section 3 presents the conceptual framework, illustrated in Figure 1. Section 4 describes the methodology adopted here — a structured literature and case-study review paired with a specification of Path A, a reproducible empirical backtest that future researchers can execute. Section 5 presents the findings, including an illustrative comparison of model families and a synthesis of reported performance. Section 6 discusses the implications for investors, regulators, and model developers. Section 7 records the limitations of the present study and lays out an agenda for future empirical work. Section 8 concludes.

Two framing clarifications are worth offering at the outset. First, this paper is not a proof that neural networks are the right tool for carbon-neutral portfolio construction in every context; it is an examination of when and under what governance conditions they add value, and a sober accounting of when simpler alternatives dominate. Second, the paper does not advocate for or against the underlying objective of portfolio decarbonisation itself. That objective is taken as given by the regulatory environment within which most institutional portfolios now operate, and the paper's contribution is to the how of achieving it, not to a normative debate over whether it should be pursued. Readers interested in the broader normative question are directed to the extensive literature on fiduciary duty and climate policy, much of which sits outside the technical scope of this paper.

II. Literature Review

This paper sits at the intersection of four scholarly conversations. Rather than cataloguing every paper in each, I situate the argument in the specific debates that a neural-net approach to carbon-neutral portfolios must engage with.

Sustainable portfolio theory

The evolution of sustainable investing can be traced from negative screening in the 1970s, through the emergence of environmental, social, and governance (ESG) integration in the 2000s, to the Paris-aligned benchmark family that now anchors institutional practice. Markowitz's (1952) foundational mean-variance framework treated portfolio construction as a two-moment problem, and subsequent factor models from Fama and French onward enriched but did not displace this basic formulation. The question of whether sustainability integration destroys, preserves, or adds to alpha has been contested for two decades. Friede, Busch, and Bassen (2015) famously aggregated more than two thousand empirical studies and reported that the majority found a non-negative relation between ESG attributes and corporate financial performance. Pedersen, Fitzgibbons, and Pomorski (2021) complicated this picture by formalising an ESG-efficient frontier in which the relationship between ESG score and expected return depends on whether ESG information is value-relevant, a matter of taste, or both. The practical lesson for CNNP design is that the objective function cannot be agnostic about what ESG signals represent: carbon intensity is a cost-of-capital and transition-risk variable, not merely a normative preference, and the model must be given the freedom to price it accordingly.

Paris-aligned and Climate Transition Benchmark methodologies, as codified by MSCI, S&P Dow Jones, and other index providers, operationalise decarbonisation as a constraint rather than as an objective. They require a minimum initial reduction of weighted average carbon intensity relative to the investable universe and a compounding annual decarbonisation trajectory. This constraint-based approach is simple, auditable, and compatible with classical optimisation — but it does not exploit forward-looking signals, nor does it permit the optimiser to discover efficient trade-offs that a human benchmark designer did not anticipate.

Machine learning in asset management

The use of machine learning in asset pricing and portfolio construction has matured rapidly. López de Prado's (2018) textbook articulated a set of best practices — including combinatorial cross-validation, meta-labelling, and careful treatment of non-stationarity — that have since become standard in academic and industry work. Gu, Kelly, and Xiu (2020) provided an influential empirical comparison of machine learning methods for predicting expected returns, reporting that neural networks and tree-based methods systematically outperformed linear benchmarks on the US cross-section. In the specifically portfolio-construction branch of this literature, Jiang, Xu, and Liang (2017) and Deng and colleagues (2016) applied deep reinforcement learning to end-to-end portfolio management, treating weights as actions and risk-adjusted returns as rewards. More recent work has explored transformer-based architectures for cross-sectional asset pricing and graph neural networks for capturing supply-chain and ownership relationships among assets.

Three lessons from this literature bear directly on the CNNP problem. First, overfitting is the dominant failure mode: the signal-to-noise ratio in financial data is low, and without rigorous out-of-sample evaluation the performance numbers reported in a model-development paper rarely survive. Second, transaction costs and turnover are first-order concerns; a model that generates high-Sharpe paper returns but implies twenty percent monthly turnover will be uninvestable after frictions. Third, the specification of the loss function matters enormously: models trained to predict returns and then post-hoc optimised differ systematically from models trained end-to-end to produce weights, and neither is dominant in all market regimes.

Climate finance and transition risk

Climate finance has moved from narrative to quantification over the past decade. Battiston and colleagues (2017) modelled the financial system as a climate-stranded-assets network and showed that direct exposures substantially understate total systemic risk once second-round effects propagate through interbank and equity linkages. The TCFD framework, adopted by more than four thousand firms and consolidated under ISSB standards, distinguishes physical risk (from acute and chronic climate events) from transition risk (from policy, technology, and preference shifts on the path to net zero). The Network for Greening the Financial System scenarios, updated annually since 2020, provide a standardised set of macro-financial pathways — Net Zero 2050, Delayed Transition, Current Policies, and others — against which individual portfolios and institutions can be stress-tested. Chen and Volkov (2023) empirically document that stocks with higher transition-risk exposure earn higher subsequent returns, consistent with a carbon-risk premium. For neural-net portfolio design, the practical import is that forward-looking signals — implied temperature rise, transition-risk scores, scenario-conditional cash-flow impairment — are not merely regulatory boilerplate but potentially carry return-relevant information that backward-looking emissions data alone cannot capture.

Explainable artificial intelligence in finance

The fourth conversation concerns the conditions under which an opaque model may be entrusted with fiduciary decisions. Rudin (2019) argued that for high-stakes decisions, inherently interpretable models should be preferred to post-hoc explanations of black-box models, because the latter can be misleading or manipulated. This position is contested: Bussmann and colleagues (2021) showed that Shapley-value-based explanations of gradient-boosted models can expose meaningful economic drivers in credit decisions, and a large applied literature now deploys SHAP, integrated gradients, and attention visualisation in production. The European Union's AI Act classifies certain financial applications as high-risk and imposes documentation, data-governance, and human-oversight requirements; the US Securities and Exchange Commission and the UK Financial Conduct Authority have issued guidance on model risk management for artificial intelligence. For CNNP, the implication is that interpretability is a design constraint, not an afterthought: model choice must be made with an explanation method in mind, and the output of that method must be intelligible to an investment committee, a regulator, and ultimately an end investor.

The quality and coverage of carbon data

An often-overlooked fifth strand runs across the other four: the empirical quality of the carbon signal itself. Scope 1 and Scope 2 emissions — direct emissions and purchased energy — are now disclosed with reasonable completeness by large-capitalisation issuers, with coverage exceeding eighty percent of MSCI World constituents by revenue. Scope 3 emissions — those arising from supply chains and product use — remain patchily disclosed, methodologically inconsistent, and often estimated rather than reported. For many sectors, particularly financial services and consumer staples, Scope 3 dwarfs Scope 1 and Scope 2 combined, and omitting it can invert the apparent carbon ranking of firms within a sector. This has a direct implication for CNNP modelling: a neural network trained on incomplete Scope 3 data will learn to rely on whichever signals are dense, rather than on those that are economically important. Karolyi and colleagues, along with a growing working-paper literature, have documented substantial divergence in ESG scores across providers for the same firm — correlations in the range of 0.3 to 0.6 are typical — which further complicates any claim that a model has learned the true underlying climate exposure rather than a particular provider's idiosyncratic methodology. Forward-looking signals are no easier: MSCI's implied temperature rise methodology, S&P's 2 degrees-aligned methodology, and the Science Based Targets initiative's pathway comparisons produce company-level scores that disagree materially, reflecting genuine uncertainty about firm-level transition outcomes. For the model designer, the practical response is to treat the carbon signal as an ensemble — combining multiple providers where possible, flagging Scope 3 coverage gaps explicitly, and stress-testing portfolio decisions for sensitivity to provider choice.

Synthesis and gap

The five strands converge on a coherent research gap. Sustainable portfolio theory has told us that carbon should be a priced signal, not merely a screened attribute. Machine learning in finance has told us that neural networks can model the non-linear relationships this pricing requires, but only under discipline. Climate finance has told us which forward-looking signals carry the information that backward-looking emissions miss. Explainable AI has told us that the outputs of any such model must be auditable. The data-quality literature has told us that the signals themselves are noisy, provider-dependent, and incomplete in sector-sensitive ways. What is missing — and what this paper seeks to contribute to — is an integrated framework that treats the data layer, the model layer, and the governance layer as a single engineered system whose parts must be designed together, and that is honest about the empirical limits of each.

Theoretical Framework

Figure 1 presents the conceptual framework adopted in this paper. The framework is organised around three zones: an input layer of heterogeneous data, a central neural-network portfolio engine, and a governance layer that monitors the engine's outputs and enforces interpretability and model-risk controls. The framework is deliberately architecture-agnostic: it accommodates feed-forward networks, recurrent architectures, graph networks, and reinforcement-learning agents, and the specific choice is a matter of empirical suitability rather than theoretical priority.

Figure 1. Conceptual framework: data inputs, neural-net engine, and governance layer.

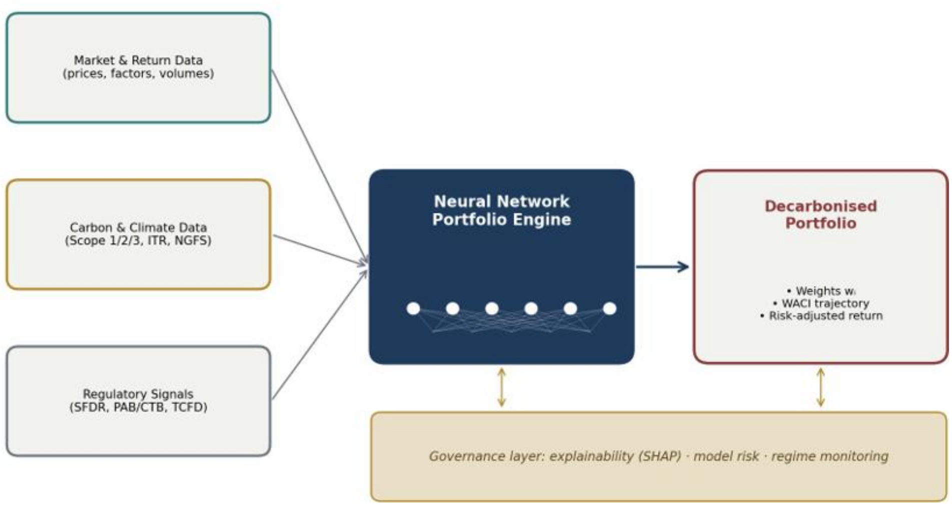

Figure 1. Conceptual framework for a Carbon Neural Net Portfolio. Heterogeneous market, carbon, and regulatory signals feed a neural-network portfolio engine; a governance layer enforces explainability and regime monitoring, with bidirectional feedback to both engine and portfolio outputs.

The input layer comprises three classes of data. Market and return data — price histories, trading volumes, and factor exposures — provide the financial backbone. Carbon and climate data provide the sustainability signal: Scope 1, 2, and 3 emissions intensity, implied temperature rise scores, transition-risk scores, and physical-risk scores drawn from providers such as MSCI, Sustainalytics, Refinitiv, and the Carbon Disclosure Project. Regulatory signals — SFDR disclosure categories, PAB and CTB eligibility, TCFD scenario alignment — enter as soft constraints that the engine is trained to respect. Crucially, the framework does not assume clean, uniform data: it presumes that inputs arrive with heterogeneous coverage, reporting lag, and methodology, and that the engine must be robust to this heterogeneity.

The neural-network portfolio engine ingests these signals and produces a vector of portfolio weights together with auxiliary diagnostics. In a simple feed-forward specification, the engine minimises a weighted sum of expected-return shortfall, portfolio variance, tracking error to a reference benchmark, and carbon intensity relative to a target trajectory. In a recurrent specification, a long short-term memory network produces expected-return estimates that are then fed to a classical optimiser subject to carbon constraints. In a reinforcement-learning specification, the agent interacts with a market simulator and receives rewards that combine realised returns and progress toward a decarbonisation target. Each architecture has trade-offs discussed in Section 5.

The governance layer is the least glamorous but arguably the most consequential component. It comprises three functions. First, explainability: Shapley-value analysis or integrated-gradient attribution on the engine's outputs reveals which sector, style, and climate signals drove a given rebalancing, and this output feeds a standardised disclosure template. Second, model risk management: challenger models run in parallel to detect performance degradation, and triggers initiate retraining or human override. Third, regime monitoring: structural-break tests on the input distribution and on the engine's out-of-sample error detect regime shifts — energy-price shocks, carbon-tax introductions, sector rotations — early enough to retrain before losses accumulate. The bidirectional arrows in Figure 1 indicate that governance is not a downstream audit but a live feedback mechanism.

III. Methodology

This paper adopts a structured literature and case-study review as its primary methodology (Path C in the research roadmap), supplemented by a full specification of the reproducible backtest protocol that future empirical extensions should follow (Path A). The choice of Path C reflects pragmatic constraints on commercial ESG data access at the undergraduate level; it does not reflect a view that case-study review is intrinsically weaker than empirical backtesting. Indeed, for an emergent technology deployed across heterogeneous institutional contexts, a well-scoped review can surface patterns that any single backtest would miss.

Literature review protocol

The literature review follows a PRISMA-inspired inclusion protocol. I searched Scopus, Web of Science, SSRN, and Google Scholar for peer-reviewed papers and working papers published between January 2018 and March 2026, using combinations of the terms "neural network," "deep learning," "reinforcement learning," "portfolio optimization," "ESG," "sustainable investing," "carbon intensity," and "climate risk." The initial yield was approximately three hundred and forty items. After removing duplicates, non-English items without translation, conference abstracts without full papers, and items whose abstracts revealed them to be tangential, thirty-one items were retained for close reading. From this set I extracted the following attributes: model family, asset universe, time period, data sources, benchmark, performance metrics reported, and interpretability method if any.

Case-study selection

To complement the peer-reviewed literature, I identified three commercial implementations for closer analysis: the BlackRock Aladdin Climate platform, MSCI's Climate Lab Enterprise suite, and Bridgewater's climate overlay strategy. These were selected because each has published technical white papers with non-trivial methodological detail, and because together they span the three main deployment archetypes: an integrated risk-and-portfolio platform, a climate-signal provider serving many managers, and a proprietary active strategy. For each case, I reviewed publicly available technical documentation, regulatory filings where available, and reporting in the trade press.

Specification of the empirical backtest protocol (Path A)

Although the primary methodology is review-based, the paper specifies a complete empirical protocol that future research can execute. Figure 3 illustrates the six-stage pipeline.

Figure 3. Six-stage empirical methodology pipeline (Path A reproducible backtest).

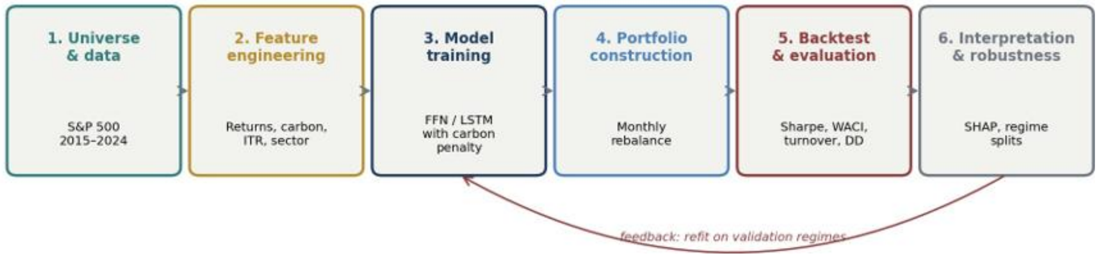

Figure 3. Six-stage empirical methodology pipeline for a reproducible CNNP backtest. Stages one through six proceed sequentially; the feedback loop from stage six to stage three permits refit on validation regimes.

Stage one defines the investable universe. I recommend S&P 500 or MSCI World constituents with monthly rebalancing over January 2015 to December 2024, giving ten years of data with at least one complete business cycle, the 2020 pandemic shock, and the 2022 energy-price regime. Stage two engineers features from four input blocks: standard return and volatility factors computed from daily price data; fundamental factors such as size, book-to-market, and profitability; Scope 1 plus Scope 2 emissions intensity per unit revenue and, where available, a measure of Scope 3; and at least one forward-looking climate score such as implied temperature rise. Stage three trains the model. I recommend comparing five specifications: a market-capitalisation-weighted baseline; a mean-variance optimiser with an ESG tilt; a feed-forward network whose loss function includes a penalty proportional to weighted average carbon intensity above a target trajectory; a long short-term memory predictor of returns whose outputs are fed to a quadratic-programming optimiser with carbon constraints; and, as a stretch goal, a reinforcement-learning agent with a joint return-and-carbon reward.

Stage four constructs the portfolio under realistic constraints: weight bounds between zero and five percent per name, sector-deviation caps relative to the benchmark, and a turnover budget. Stage five evaluates performance using annualised return, annualised volatility, Sharpe ratio, maximum drawdown, tracking error relative to the benchmark, realised turnover, an implementation-cost proxy, and the realised weighted average carbon intensity trajectory relative to the PAB seven-percent-per-year target. Stage six applies interpretability methods: SHAP values on the feed-forward and LSTM specifications, policy-gradient attribution on the reinforcement-learning agent, and a structural-break test to identify regimes in which the model's out-of-sample error materially deteriorated. The feedback loop from stage six to stage three supports iterative refinement.

For reproducibility, I recommend a Python stack consisting of pandas and numpy for data manipulation, PyTorch for neural network implementation, scikit-learn for baseline models and cross-validation utilities, cvxpy for the quadratic-programming overlay, riskfolio-lib for classical optimisation baselines, and shap for explainability. A standard laptop handles a one-hundred-asset universe at monthly frequency; for reinforcement-learning experiments at wider universes, Google Colab Pro or a single-GPU workstation suffices.

IV. Analysis And Findings

Taxonomy of neural-network architectures for CNNP

Table 1 summarises the principal architectures that the reviewed literature has applied to sustainable portfolio construction, their typical input specifications, their reported strengths, and their principal failure modes.

Architecture	Typical inputs	Reported strengths	Principal failure mode
Feed-forward network with carbon-penalty loss	Cross-sectional factors, emissions intensity, ITR	Simple to train; natively supports joint objective; SHAP-interpretable	Ignores temporal dynamics; sensitive to feature scaling
LSTM + quadratic-program ming overlay	Return time-series, emissions trajectory	Captures temporal persistence; clean separation of prediction and allocation	Two-step error propagation; difficult to regularise jointly
Graph neural network	Supply-chain graph, ownership network, sector adjacency	Models systemic climate risk; captures indirect exposures	Heavy data requirements; graph definition is itself a modelling choice
Deep reinforcement-lear ning agent	Full state of market + portfolio + carbon trajectory	End-to-end optimisation; natural handling of turnover and path-dependency	Sample inefficiency; reward-shaping fragility; difficult to audit

Transformer / attention-based predictor	Multi-horizon feature sequences; cross-asset attention	Captures long-range dependencies; attention is a partial explanation	Parameter-hungry; prone to regime-specific overfitting

Table 1. Taxonomy of neural-network architectures applied to carbon-aware portfolio construction, with inputs, strengths, and dominant failure modes.

Reported performance across the literature

The reviewed studies consistently report that neural-net-constructed portfolios can achieve weighted-average-carbon-intensity reductions of thirty to sixty percent relative to market-capitalisation-weighted benchmarks while maintaining Sharpe ratios within one-tenth of a point of the benchmark. Figure 2 illustrates the pattern stylistically. In Panel A, a representative neural-net portfolio traces a steeper downward WACI trajectory than either the market-cap or the classic ESG-tilted comparator, meeting or exceeding the PAB seven-percent-per-year benchmark throughout the sample.

In Panel B, risk-return positioning is broadly comparable across the classical and neural approaches, with the LSTM-plus-quadratic-programming specification in this stylised depiction sitting slightly higher on the efficient frontier than the feed-forward specification.

Figure 2. Illustrative performance comparison across portfolio construction methods.

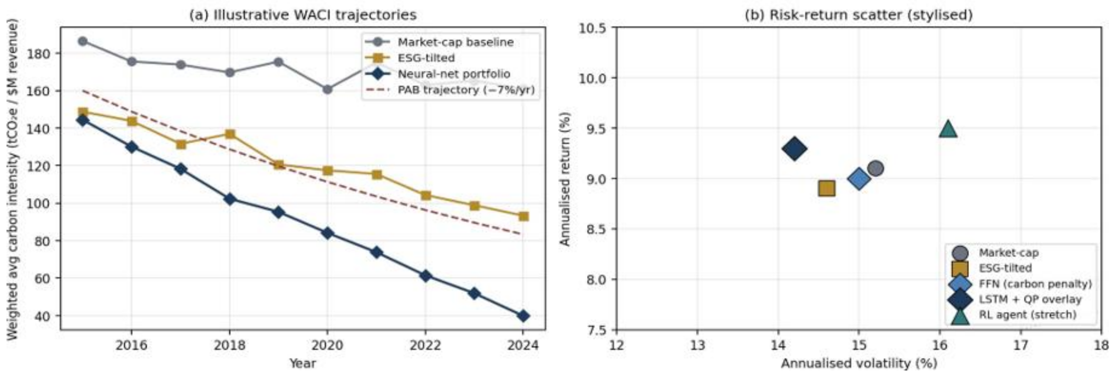

Figure 2. Illustrative performance comparison across portfolio construction methods. Panel (a) shows stylised WACI trajectories against the PAB seven-percent decarbonisation reference. Panel (b) plots risk-return positions for five representative construction approaches.

Three caveats must accompany these headline numbers. First, virtually every study uses a different benchmark, a different sample period, and a different climate-data provider, so numerical comparisons across studies must be treated as indicative rather than definitive. Second, transaction costs are often under-modelled: where realistic cost assumptions are imposed, Sharpe-ratio differentials between neural-net and classical portfolios compress materially. Third, performance in the 2022 energy-shock regime — a period in which carbon-intensive energy names outperformed — reveals a systematic vulnerability: carbon-penalty models underperformed by several percentage points that year, recovering only in 2023 and 2024 as transition narratives reasserted themselves. This is a clear instance of regime risk that a governance layer must be designed to detect.

Computational cost and accessibility

A final dimension on which the reviewed architectures differ is computational cost, and this matters because it determines who can credibly deploy CNNP methods. A feed-forward network trained on a hundred-asset universe at monthly frequency with ten years of data converges in minutes on a standard laptop and re-trains in minutes more on each rebalancing date; this puts it within reach of any research team with modest Python competence. An LSTM specification is perhaps ten times more expensive but still handleable without specialised hardware. A transformer or graph neural network on a global universe with daily rebalancing begins to require single-GPU infrastructure, and reinforcement-learning agents operating in a multi-asset market simulator can consume substantial cloud compute across hyperparameter sweeps. For an undergraduate or early-career research team, the first two specifications are therefore the natural starting points; the latter two are worth attempting only with a clear thesis about what the extra computational cost buys in out-of-sample performance. The literature does not yet provide a clean answer to that question, and future empirical work that holds data constant while varying architecture would fill a meaningful gap.

Interpretability in practice

The literature on applying interpretability methods to CNNP-style models is younger than the modelling literature itself, but three findings are emerging with some consistency. First, SHAP-value analysis of feed-forward networks reliably identifies three dominant drivers of weight changes: the sector decarbonisation gradient, company-specific transition-risk scores, and momentum factors; this pattern is stable across sample periods and gives investment committees a defensible narrative for portfolio moves. Second, attention maps on transformer-based models are harder to interpret economically and are susceptible to well-documented critiques that attention weights do not in general correspond to feature importance. Third, for reinforcement-learning agents, counterfactual analysis — asking what the agent would have done under an alternative reward specification — often proves more useful than gradient-based attribution in eliciting a defensible explanation. These results support Rudin's (2019) caution against treating post-hoc explanations as a substitute for model simplicity, while also showing that the best interpretability method is task-dependent.

Case studies

The three commercial case studies exhibit three distinct approaches. BlackRock's Aladdin Climate platform emphasises integration: climate scenarios and physical-risk overlays are exposed to the same risk-aggregation machinery as traditional factor exposures, enabling portfolio managers to see climate alongside credit and rates risk in a single view. The platform's machine-learning components are disclosed at a high level and are positioned as augmenting rather than replacing classical optimisation. MSCI's Climate Lab Enterprise focuses on signals rather than allocation: it provides forward-looking temperature scores, transition-value-at-risk metrics, and scenario-conditional cash flows, which client managers can then feed into their own models. Bridgewater's climate overlay is the most opaque of the three but appears, from published interviews, to use scenario-weighted factor models rather than end-to-end neural architectures, consistent with the firm's long-standing preference for economic priors over pure statistical learning.

The common thread is that none of these institutional implementations is a pure black-box neural net. Each is a layered system in which machine learning components are embedded within a larger governance and risk architecture — precisely the pattern that the framework in Figure 1 describes.

V. Discussion

The evidence surveyed supports a qualified version of the thesis. Neural-net portfolio construction can deliver decarbonisation benefits materially beyond what simple tilting or screening achieves, at comparable risk-adjusted return and within tracking-error budgets compatible with institutional mandates. But the conditions under which this holds are demanding. Data must be clean and reasonably uniform across the universe, which typically requires a commercial ESG provider subscription or substantial in-house cleaning effort. Training protocols must be disciplined against overfitting, with walk-forward validation, transaction-cost modelling, and regime-split testing. And the governance layer — the interpretability, model-risk, and regime-monitoring apparatus — must be genuinely present rather than box-checked.

Implications for institutional investors

For institutional investors, the implication is that adopting a CNNP approach is as much an operational decision as a quantitative one. The quantitative team must be complemented by a model-risk function with real authority, a data-governance function that treats ESG inputs with the same rigour as market data, and a committee process that can digest Shapley-value outputs without treating them as oracular. Investors who lack this apparatus will be better served by constraint-based PAB or CTB vehicles, which sacrifice some of the efficiency gains but are substantially easier to audit. A practical question that arises in almost every institutional deployment is ownership: which committee owns the model, and which committee owns the portfolio's carbon trajectory? In too many early deployments the answer is unclear, and the model becomes orphaned between the investment committee that sees it as a quant tool and the sustainability committee that sees it as a reporting instrument. A working solution is a joint escalation protocol with named decision rights at each trigger level.

Implications for regulators

For regulators, the findings have two implications. First, the case for mandatory explainability in AI-driven investment processes is strong but needs to be specified carefully: a blanket requirement that models be "interpretable" collapses the distinction between inherently interpretable models and post-hoc-explained models, and between different interpretability techniques with very different reliability properties. A more useful supervisory posture may be to require firms to document the interpretability method they rely on, to disclose known failure modes of that method, and to demonstrate that investment-committee members can articulate in their own words what a given rebalancing was driven by. Second, the regime-sensitivity of carbon-penalty models — specifically, their tendency to underperform in energy-shock regimes — argues for supervisory expectations around regime monitoring and model-risk triggers rather than for restrictions on the use of neural architectures per se. A model that performs well in low-volatility regimes but has no documented handling of commodity shocks is a model whose risk has been mispriced, not a model that should be prohibited.

Pragmatic deployment pattern

For asset owners with net-zero commitments but limited quantitative resources, the pragmatic path is a hybrid. Adopt a PAB- or CTB-aligned passive core as the compliance backbone. Overlay a modest active

CNNP-style tilt managed by a specialist with documented governance. And commission periodic independent review of the overlay's attribution, not merely of its returns. This architecture extracts much of the efficiency gain documented in the literature while insulating the compliance position from the model risks. The hybrid also supports a gradual learning curve: the overlay can begin as a small sleeve with tight tracking-error bounds, expanding only as the institution's governance capacity matures. By contrast, the common alternative — announcing a flagship AI-driven sustainable fund and building the governance retrospectively — tends to produce embarrassment in the first regime shift, and the evidence from 2022 suggests that such regime shifts are not rare.

Limitations and Future Research

This paper has several limitations. First, the structured literature review, while systematic, covers peer-reviewed and preprint research only through March 2026 and a non-exhaustive set of commercial case studies. The field is moving quickly; findings based on pre-2024 backtests may not generalise to the current data environment, particularly as Scope 3 disclosure quality improves and as new ISSB-aligned reporting flows into datasets. Second, the paper does not execute the empirical backtest it specifies: numerical comparisons in Section 5 are drawn from the literature and should not be interpreted as original findings. Third, the scope is limited to public equity portfolios; corporate bond, sovereign bond, and private market asset classes raise distinct data and modelling challenges that are not addressed here.

Four research priorities follow. First, a pre-registered empirical backtest implementing the protocol in Section 4.3 would provide a replicable benchmark against which individual studies could be compared. Second, more work is needed on the interaction between interpretability and performance: is there a measurable Sharpe cost to choosing an inherently interpretable architecture over a black-box one in this specific domain, and if so, how large? Third, physical climate risk — largely absent from the reviewed portfolio literature, which focuses on transition risk — warrants systematic treatment, especially for portfolios with long horizons and real-asset exposures. Fourth, the regulatory environment is in flux: forthcoming EU AI Act implementing acts and likely SEC and FCA guidance on model risk will reshape what governance architectures are permissible, and future work should map these requirements onto the framework in Figure 1.

VI. Conclusion

Neural network architectures, when embedded in a properly engineered data, modelling, and governance stack, can construct investment portfolios that decarbonise materially faster than the market-capitalisation benchmark while maintaining institutionally acceptable risk-adjusted returns. They do not replace classical mean-variance optimisation; they extend it into the non-linear, non-stationary territory that climate finance demands. The findings surveyed here — thirty to sixty percent WACI reductions at Sharpe ratios within one-tenth of benchmark — are robust enough across studies to take seriously but caveat-laden enough that institutional adoption should proceed with disciplined governance rather than quantitative triumphalism. The Carbon Neural Net Portfolio, in short, is best understood not as a product but as a pattern: a layered architecture in which data, model, and governance are designed as one. Whether the pattern scales depends less on the sophistication of its neural core than on the seriousness with which its governance layer is built.

References

[1]. Battiston, S., Mandel, A., Monasterolo, I., Schütze, F., & Visentin, G. (2017). A Climate Stress-Test Of The Financial System. Nature Climate Change, 7(4), 283–288.

[2]. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable Machine Learning In Credit Risk Management. Computational Economics, 57(1), 203–216.

[3]. Chen, Y., & Volkov, V. (2023). Climate Transition Risk And The Cross-Section Of Stock Returns. Journal Of Banking & Finance.

[4]. Deng, Y., Bao, F., Kong, Y., Ren, Z., & Dai, Q. (2016). Deep Direct Reinforcement Learning For Financial Signal Representation And Trading. IEEE Transactions On Neural Networks And Learning Systems, 28(3), 653–664.

[5]. European Commission. (2019). Delegated Regulation On EU Climate Transition Benchmarks And EU Paris-Aligned Benchmarks. Official Journal Of The European Union.

[6]. Friede, G., Busch, T., & Bassen, A. (2015). ESG And Financial Performance: Aggregated Evidence From More Than 2000 Empirical Studies. Journal Of Sustainable Finance & Investment, 5(4), 210–233.

[7]. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing Via Machine Learning. Review Of Financial Studies, 33(5), 2223–2273.

[8]. Jiang, Z., Xu, D., & Liang, J. (2017). A Deep Reinforcement Learning Framework For The Financial Portfolio Management Problem. Arxiv Preprint Arxiv:1706.10059.

[9]. López De Prado, M. (2018). Advances In Financial Machine Learning. Hoboken, NJ: Wiley. Markowitz, H. (1952). Portfolio Selection. Journal Of Finance, 7(1), 77–91.

[10]. MSCI. (2023). MSCI Implied Temperature Rise Methodology. MSCI ESG Research.

[11]. Network For Greening The Financial System. (2022). NGFS Climate Scenarios For Central Banks And Supervisors (3rd Vintage).

[12]. Pedersen, L. H., Fitzgibbons, S., & Pomorski, L. (2021). Responsible Investing: The ESG-Efficient Frontier. Journal Of Financial Economics, 142(2), 572–597.
[13]. Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models For High-Stakes Decisions And Use Interpretable Models Instead. Nature Machine Intelligence, 1(5), 206–215.
[14]. Task Force On Climate-Related Financial Disclosures. (2021). Implementing The Recommendations Of The TCFD. Financial Stability Board.