Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting Children's Climate Hazard Exposure and Vulnerability: A Reproducible Multi-Task Deep Learning Benchmark Using the UNICEF Children's Climate Risk Index

Rana Mukherji

rana.mukherji@gmail.com

IcfaiTech, The ICFAI University

Manishita Mukherji

Amity University Rajasthan

Research Article

Keywords: multi-task deep learning, tabular transformers, attention mechanisms, child climate-risk assessment, hazard exposure and vulnerability

Posted Date: September 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10853190/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Predicting Children's Climate Hazard Exposure and Vulnerability: A Reproducible Multi-Task Deep Learning Benchmark Using the UNICEF Children's Climate Risk Index

Abstract

Children bear a disproportionate share of the burden of climate-related hazards, yet the data-driven tools used to quantify where they are most at risk remain dominated by additive index methods and conventional regression. Reliable, jointly-estimated measures of hazard exposure and vulnerability are a prerequisite for targeting adaptation finance and child-sensitive disaster risk reduction. We introduce the ResNet-FT-Transformer with Cross-Attention Hydration (MHA-Dense), a deep tabular architecture that tokenizes continuous socio-environmental indicators into dense embeddings, refines them through four residual transformer blocks combining multi-head self-attention (MHA-Dense) with a gated Cross-Attention Hydration module that injects a learned global context into local feature tokens, and predicts two targets simultaneously through dedicated task heads. Using the country-level UNICEF Children's Climate Risk Index (CCRI) data (n = 176 countries and territories; nine demographic, hazard, and service-deprivation indicators), the model jointly estimates the Hazard Exposure Score and the Child Vulnerability Score within a single multi-task network. On a held-out test set the framework attains coefficients of determination of $R^2 = 0.727$ (MAE = 0.877, RMSE = 1.162) for hazard exposure and $R^2 = 0.818$ (MAE = 0.843, RMSE = 1.048) for child vulnerability, performance that is statistically comparable to strong gradient-boosted tree and neural baselines whose bootstrap confidence intervals overlap throughout. A controlled benchmark against six baselines (Extra-Trees, Random Forest, XGBoost, LightGBM, an SVM, and a multilayer perceptron) shows that, on small low-dimensional indicator data, gradient-boosted ensembles retain the strongest point estimates, a result we report transparently and situate within the tabular deep-learning literature. The principal contributions are an interpretable, unified multi-task architecture for child climate-risk modelling, the MHA-Dense and Cross-Attention Hydration mechanisms, and an openly reproducible benchmark with uncertainty quantification to guide model selection in disaster-risk applications.

Keywords: multi-task deep learning; tabular transformers; attention mechanisms; child climate-risk assessment; hazard exposure and vulnerability

1. Introduction

The climate system has entered a regime in which weather and climate extremes are more frequent, more intense, and more spatially compounding than at any point in the instrumental record. The Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) concluded that human influence has unequivocally warmed the planet by approximately 1.1 °C relative to the late nineteenth century, and that this warming has already intensified the hydrological cycle, lengthened and sharpened heatwaves, and altered the statistics of heavy precipitation and tropical cyclone intensity across most inhabited regions [1]. Crucially, the report emphasised that the residual risk experienced by any given population is not a function of the physical hazard alone but of the interaction between hazard, exposure, and vulnerability, and that the latter two are governed by socio-economic conditions that vary enormously between and within countries [1,2]. Floods now displace tens of millions of people annually, prolonged droughts erode food and water security across the Sahel and the Horn of Africa, and intensifying tropical storms repeatedly destroy housing, schools, and health infrastructure in low-lying coastal and small-island states. These hazards rarely appear singly; IPCC and other disaster science literature show multiple concurrent hazards have come together in the worst-hit locations, so their impact is far greater than any individual hazard [2].

In this context, children are not merely a disadvantaged sector, but the majority burdened with the heaviest and longest lasting burden. Their physiology makes them more vulnerable to heat stress, undernutrition, water-related illnesses, and impacts on the respiratory system and their reliance on materials, structures, solutions, and services of adults and upon functioning health, water, and education systems increases the downstream effects of shocks; and the developmental nature of children means that the impact of early shocks becomes life-long deficits in health, learning, earning, and income. UNICEF's landmark Children's Climate Risk Index (CCRI), released in 2021 as the first comprehensive child-centred analysis of climate risk, quantified this burden starkly: approximately one billion children—nearly half of the world's 2.2 billion children—live in one of thirty-three countries classified as being at extremely high risk, where high exposure to climate and environmental shocks coincides with acute deprivation of the essential services that would otherwise buffer those shocks [3]. The same index estimated that 330 million children are highly exposed to riverine flooding, 240 million to coastal flooding, 400 million to cyclones, and 920 million to water scarcity, while roughly 850 million—one in three children worldwide—live where at least four such hazards overlap [3]. The injustice of this distribution is compounded by the finding that the thirty-three highest-risk countries are collectively responsible for only about nine per cent of global carbon dioxide emissions, underscoring that the children least responsible for the crisis are those most exposed to it [3].

The updated trajectory documented in UNICEF's Children's Climate Risk Report 2026 (CCRR 2026) confirms that these exposures have not stabilised [4]. Virtually every child alive today—approximately 2.3 billion—now lives in an area with at least one detectable climate hazard, 2 billion face at least two simultaneous hazards, and 1.1 billion children are concurrently exposed to three or more overlapping hazards [4]. At higher thresholds, 364 million children contend with four or more hazards, 53 million with five or more, and 4 million with six or more [4]. Disaggregated by hazard type, 337 million children are exposed to riverine floods, 33 million to coastal floods, 1.8 billion to droughts, 662 million to tropical storms, 1.5 billion to more frequent or severe heatwaves, 1.2 billion to extreme heat, 206 million to fires, and 123 million to sand and dust storms [4]. At the same time, there are estimated to be one billion individuals who are at risk because of exposure to malaria and 2.3 billion people who are exposed to air pollutants that are detectable. The human effect on these exposures is now being witnessed on a large scale. From 2016-2023 climate hazards were responsible for internal displacement of children, totalling to 62.1 million which is equal to over 21,000 child IDPs per day [4]. At least 242 million students in 85 countries and territories had their schooling disrupted due to climate-related hazardous events, with 634 million children experiencing the absence of safely managed drinking water and one billion children experiencing the absence of safely managed sanitation in 2024 alone [4]. Looking ahead, climate change is projected to

cause an additional 28 million children to be wasted and 40 million to be stunted globally by 2050 and could push over 130 million people into extreme poverty by 2030 [4].

The scale and distributional injustice of this exposure is increasingly characterized as a child-rights emergency, rather than a humanitarian emergency. International legal and regulatory bodies, including the UN Convention on the Rights of the Child, the Inter-American Court, and the International Court of Justice, have progressively clarified that climate inaction constitutes a violation of children's fundamental rights to life, health, and development [4]. The erasure of those rights under disaster conditions is not hypothetical: qualitative evidence gathered in the immediate aftermath of the February 2023 earthquake in Turkey documented systematic violations of children's rights to education, shelter, health, and protection—including heightened risk of abuse and family separation—and concluded that rights-based, child-centred frameworks must be embedded in disaster preparedness before events occur, not assembled from scratch in their wake [5]. These findings parallel a broader literature showing that social-protection and essential-service systems—the very mechanisms that buffer climate shocks—are precisely the systems most likely to collapse under simultaneous or compounding stressors [2,4].

These figures motivate a concrete computational problem. To target adaptation finance, prioritise child-sensitive disaster risk reduction, and design early-warning and social-protection systems, decision-makers require accurate, comparable, and ideally jointly-estimated measures of how exposed a population is to climate hazards and how vulnerable it is once a hazard materialises. The CCRI itself, like most operational risk indices, is constructed as a hierarchical weighted aggregation of normalised indicators: component scores for individual hazards and for individual dimensions of deprivation are combined through fixed, expert-assigned weights and deterministic geometric or arithmetic means [3,4,6]. This additive-index paradigm is transparent and reproducible, and it has been indispensable for advocacy and high-level prioritisation. It nonetheless embeds two strong assumptions that limit its analytical power: that the contribution of each indicator to overall risk is fixed and independent of the values of the others, and that the functional form linking indicators to risk is known a priori. Both assumptions are difficult to justify for socio-environmental systems in which the marginal effect of, say, poor sanitation on child vulnerability plausibly depends on the level of health-service coverage, the burden of undernutrition, and the intensity of the hazards a country faces.

1.1 Limitations of existing approaches

Efforts to move beyond fixed-weight indices have largely followed two trajectories, each with characteristic shortcomings. The first relies on classical statistical models—ordinary least squares regression, generalised linear models, and their regularised variants—to relate hazard and vulnerability outcomes to candidate predictors. Such models are interpretable and statistically well-understood, but they encode the very assumptions that make socio-environmental modelling difficult: linearity in the predictors, additive and pre-specified interaction terms, and a single response per fitted model. When the underlying relationships are non-linear or involve higher-order interactions among many indicators, linear and generalised-linear specifications either miss the structure entirely or require the analyst to hand-craft interaction and basis terms, a process that scales poorly and rarely generalises across regions. Vulnerability science has long recognised that exposure and susceptibility interact in ways that resist simple additive description [6], yet the dominant quantitative toolkit remains poorly matched to that recognition.

A recent scoping review of climate change vulnerability indices in the child-health literature identified only fourteen eligible studies from an initial pool of 2,262 papers [7]. The review found that most existing indices focused narrowly on children under five years of age, used heterogeneous and poorly standardised child-specific variables, and exhibited wide variation in methodology and geographic coverage [7]. The most common child-specific variables were age, health indicators, and immunisation status; broader developmental outcomes, educational attainment, and protection-service access were consistently absent. The authors concluded that the field lacks an independent, comprehensive index capable of capturing the

full age range of childhood and the multidimensional interactions between climate exposure and child-specific vulnerability [7]. These gaps align with—and partly explain—the limitations of the fixed-weight additive approach noted above.

The second trajectory applies general-purpose machine-learning regressors—most commonly tree ensembles such as Random Forests, Extra-Trees, and gradient-boosted decision trees, and occasionally feed-forward neural networks—to predict risk outcomes from indicator data [14,15,16,17]. These methods relax the linearity assumption and frequently improve predictive accuracy, and tree ensembles in particular are robust, require little tuning, and handle heterogeneous tabular features gracefully. Their limitations for the present problem are nonetheless real. Tree-based models partition the feature space along axis-aligned thresholds and do not learn shared, distributed representations of features, which constrains their ability to model smooth, compositional interactions and to transfer structure between related prediction tasks. Standard implementations are also single-target: predicting both a hazard-exposure score and a vulnerability score requires either two independently-fitted models, which cannot share statistical strength, or a thin multi-output wrapper that fits the targets separately under the hood. Most importantly for an applied audience, conventional pipelines provide only post-hoc, model-agnostic notions of feature importance and do not expose how features interact in producing a prediction.

A third and increasingly important observation cuts across both trajectories: the field has paid little attention to multi-task formulations in this domain. Hazard exposure and child vulnerability are not independent quantities. They are shaped by partially overlapping determinants—a country's wealth, governance, infrastructure, and geography influence both how exposed its children are and how well-protected they are once exposed—and learning the two jointly should, in principle, allow a model to exploit shared representations and act as a mutual regulariser [18,19]. Yet operational child-risk modelling overwhelmingly treats the two as separate exercises, forgoing any benefit from their statistical coupling and producing scores that are not guaranteed to be mutually consistent.

1.2 The present contribution

This paper addresses these gaps by adapting recent advances in attention-based deep learning for tabular data to the joint modelling of climate hazard exposure and child vulnerability. The Transformer architecture, introduced for sequence modelling, replaced recurrence and convolution with a self-attention mechanism that learns, for every element of an input, a data-dependent weighting over all other elements [8]. Attention is a natural fit for tabular risk modelling because it can learn which indicators should inform the representation of which other indicators, rather than imposing this structure in advance. The FT-Transformer demonstrated that, by first tokenising each continuous or categorical feature into a learned embedding and then applying standard transformer blocks, attention models can be made competitive with strong gradient-boosted baselines on a range of tabular benchmarks [9]. In parallel, residual learning showed that very deep networks become trainable when each layer learns a residual correction to its input rather than an unreferenced mapping, a principle that stabilises optimisation and improves gradient flow [10]. Attentive tabular models such as TabNet and TabTransformer further established that attention can deliver both accuracy and a degree of built-in interpretability on structured data [11,12].

Building on these foundations, we propose the ResNet-FT-Transformer with Cross-Attention Hydration, a multi-task deep architecture with three distinguishing design elements. First, a Feature Tokenizer maps each of the nine continuous socio-environmental indicators to a dense embedding vector, so that downstream attention operates over learned feature tokens rather than raw scalars. Second, each of four residual transformer blocks couples a Multi-Head Attention with Dense connections (MHA-Dense) sub-layer—self-attention wrapped in residual and layer-normalisation paths that preserve gradient flow—with a novel Cross-Attention Hydration module. The hydration module computes a compact global context vector by pooling the attended token representations and then uses cross-attention to inject that global signal back into each local token, modulated by a learnable scalar gate that lets the network decide, per block,

how much global information to admit. Intuitively, the module "hydrates" each feature representation with a summary of the whole indicator profile, enabling local feature embeddings to be refined in light of the country's overall risk context. Third, the refined token representations feed two task-specific prediction heads that output the Hazard Exposure Score and the Child Vulnerability Score simultaneously, trained under a summed mean-squared-error objective so that the shared trunk learns representations useful for both tasks [18,19].

In addition to the architecture, the work is an explicit methodological choice that is founded in explicit benchmarking and uncertainty quantification.. We evaluate the proposed model against six baselines spanning the dominant tabular paradigms—Extra-Trees, Random Forest, XGBoost, LightGBM, a radial-basis support-vector regressor, and a multilayer perceptron—under an identical preprocessing and train-test protocol, and we accompany every point estimate with bootstrap confidence intervals. This matters because the CCRI operates at the level of countries, of which there are fewer than two hundred with complete records; in this small-sample regime, headline accuracy differences between strong models are easily within the range of test-set sampling noise, and reporting them without uncertainty would be misleading. Our results show that the proposed transformer is competitive with the strongest baselines and that, by point estimate, gradient-boosted ensembles retain a modest edge—a finding consistent with a body of evidence that tree-based models remain hard to beat on small, low-dimensional tabular data [13,14]. We report this openly rather than selecting metrics to favour the proposed method, and we use it to delineate the data regimes in which attention-based tabular models are, and are not, the appropriate choice. The architecture's lasting value lies in its unified multi-task formulation, its attention-based interpretability, and its extensibility to the richer, higher-dimensional, and spatially-resolved data that child climate-risk modelling is rapidly acquiring.

1.3 Organisation of the paper

The remainder of the paper is organised as follows. Section 2 details the data sources, feature engineering, and preprocessing; the exploratory data analysis used to characterise distributions, outliers, collinearity, and feature relevance; the full specification of the ResNet-FT-Transformer with Cross-Attention Hydration, including the tokenizer, the MHA-Dense and hydration modules, the residual transformer blocks, and the multi-task heads; the training and optimisation protocol; and the baseline models and evaluation metrics. Section 3 presents the experimental results, comprising the head-to-head performance comparison, training dynamics and convergence, prediction and residual analysis, feature-importance and interpretability findings, an analysis of the attention and hydration components, and extended comparisons with additional baselines together with robustness and sensitivity checks. Section 4 interprets the findings and discussed the contribution each architectural innovation makes, the practical implications for disaster risk reduction. and child-vulnerability assessment, and states the study's limitations and the directions they motivate. Section 5 concludes.

2. Methodology

This section will detail all of the parts of the modelling pipeline, ranging from the beginning to the end, from data to evaluations. Here we list the details of data sources and feature engineering, the preprocessing and partitioning protocol, exploratory analysis used to characterize the data, full mathematical description of the proposed architecture, training and the optimisation regime, as well as baselines and metrics used for the assessment of the model. Both partition and stochastic baselines for all experiments are based on fixed random seed 42 for reproducibility.

2.1 Data collection and preprocessing

The CCRR 2026 data [4] focuses on nearly 200 countries and territories and measures the absolute and relative exposure of children to ten types of climate hazards, multi-hazard counts and climate vulnerability indicators. The study utilizes complementary country-level datasets, namely a table summarizing children's exposure to climate hazards, and a table summarizing children's vulnerability, both of which are linked to the ISO-3166 country code, alpha-3 (iso3).The exposure table provides, for each country, the size of the population under eighteen years of age and a set of normalised hazard scores on a zero-to-ten scale, alongside an aggregate Hazard Exposure Score. The vulnerability table provides normalised service-deprivation scores and an aggregate Child Vulnerability Score. The two tables were merged on the iso3 key through an inner join, yielding a unified record per country containing demographic, hazard, and vulnerability attributes together with the two aggregate scores used as prediction targets.

From the merged table we retained a parsimonious and policy-interpretable feature set of nine predictors organised into three groups. The demographic group contains a single variable, the Children (Under 18) Population, which captures the scale of the population at risk. The hazard group contains four indicators — the Riverine Flood, Coastal Flood, Tropical Storm, and Drought scores — each expressed on the zero-to-ten scale and reflecting the intensity of children's exposure to that peril. The vulnerability group contains four service-deprivation indicators — the Health, Electricity, Nutrition, and Water, Sanitation and Hygiene (WASH) scores — each likewise on the zero-to-ten scale, where higher values denote greater deprivation and hence greater susceptibility once a hazard materialises. The two regression targets are the Hazard Exposure Score and the Child Vulnerability Score.

Data cleaning proceeded in four steps. First, records lacking a valid iso3 identifier were discarded. Second, textual placeholders denoting missingness ("Not applicable") were converted to missing values, and every feature and target column was coerced to a numeric type so that any non-parseable entries became explicit missing values. Third, records with a missing value in either target were eliminated, since imputing a target would contaminate the supervised signal. Fourth, any predictors that were missing were substituted with the median value of the column (as the median is less sensitive to outliers than the mean and is valid for vectors that are not normally distributed). Following the cleaning, 176 countries and territories with data for all nine predictors and both targets were analyzed with no missing values.

Table 1 summarises the analytical sample. Descriptive statistics reveal heterogeneities that are the reason for a flexible modelling approach: the children's population ranges over over 4 orders of magnitude (from about ten thousand to more than 426 million) and several hazard and deprivation indicators are strongly

right-skewed and show a median that is very far from the maximum. The Tropical Storm and WASH indicators have a median score of near zero in many countries; the Drought score in most countries is more rangeful. Both targets fall within similar ranges, but have different central values – mean Hazard Exposure of 4.20 and mean Child Vulnerability of 2.94.

Variable	Mean	Std. dev.	Min	Median	Max
Children (Under 18) Population (millions)	13.56	40.84	0.01	3.20	426.37
Riverine Flood Score (0–10)	2.73	1.98	0.00	2.38	10.00
Coastal Flood Score (0–10)	1.30	1.78	0.00	1.02	10.00
Tropical Storm Score (0–10)	2.36	3.49	0.00	0.00	10.00
Drought Score (0–10)	4.24	2.40	0.00	4.15	10.00
Health Score (0–10)	1.71	1.98	0.00	0.80	9.25
Electricity Score (0–10)	1.68	2.76	0.00	0.00	10.00
Nutrition Score (0–10)	2.94	2.11	0.00	2.44	9.58
WASH Score (0–10)	2.33	2.75	0.00	0.85	9.25
Hazard Exposure Score (target)	4.20	2.02	0.91	4.02	10.00
Child Vulnerability Score (target)	2.94	2.17	0.06	2.53	8.57

Table 1. Table 1. Table of descriptive statistics for 9 predictors and 2 targets for the analytical sample n = 176 (target rows shaded). Population is reported in millions; all hazard and deprivation indicators are on a 0–10 scale.

Two operations completed the preparation. To bound the influence of the extreme values evident in Table 1, all predictors were standardised with a RobustScaler, which centres each feature on its median and scales it by its interquartile range. Unlike standardisation by the mean and standard deviation, this transformation is insensitive to outliers and therefore preserves the informative tails of skewed indicators without allowing them to dominate the optimisation. The targets were left in their native units so that errors are interpretable on the original zero-to-ten scale. The scaled data were then partitioned into training and test sets in an 80:20 ratio, giving 140 countries for training and 36 for held-out evaluation, with the split fixed by a random seed to ensure that every model is trained and tested on identical partitions.

2.2 Exploratory data analysis

A structured exploratory analysis preceded modelling, both to validate the data and to characterise the relationships the models would later have to capture. Six complementary analyses were conducted.

Distributional analysis. Histograms with overlaid kernel-density estimates were produced for every predictor and target, annotated with the mean and median to make skewness visible (Figure 1). The analysis confirmed the heavy right skew of the population variable and of several hazard and deprivation indicators, and motivated both the median imputation and the robust scaling described above.

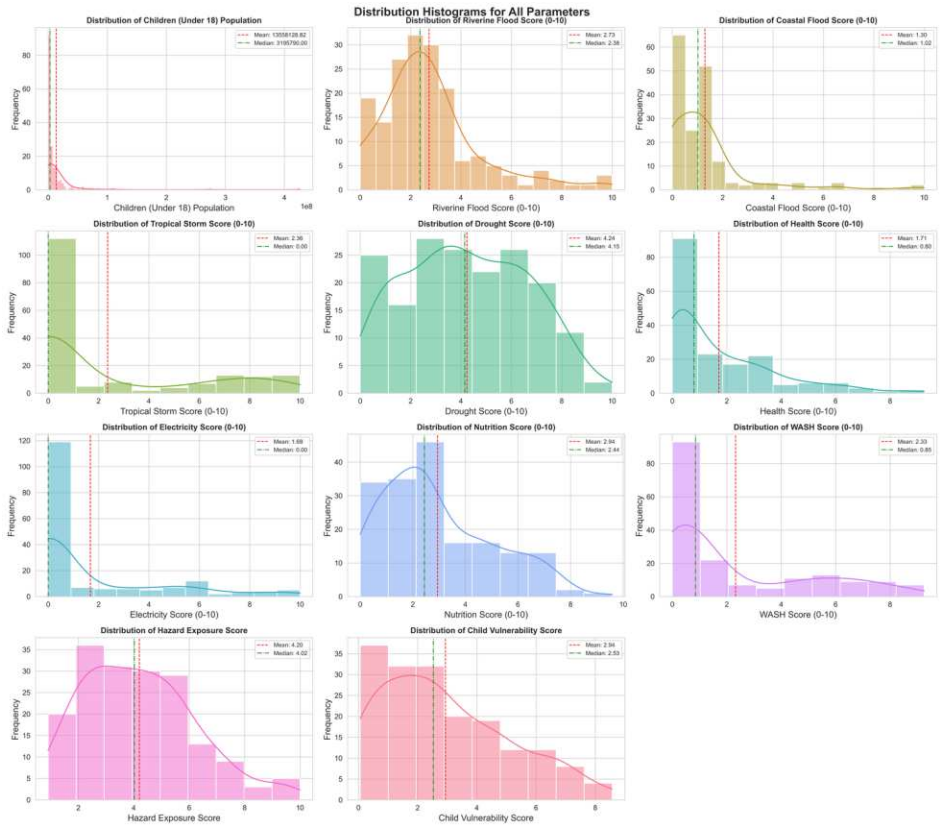


Figure 1. Distribution histograms with kernel-density overlays for all nine predictors and both targets. Dashed and dash-dot lines mark the mean and median of each variable; the separation between them indicates skew.

Outlier characterisation. Violin plots were combined with an interquartile-range (IQR) rule to identify and quantify outliers, defining a value as an outlier when it lay more than 1.5 × IQR below the first or above the third quartile. The procedure flagged a modest outlier load concentrated in a few variables — most notably the children's population (16 outliers, 9.1 per cent), coastal flood and electricity scores (14 each, 8.0 per cent), and the riverine flood score (12, 6.8 per cent) — while several indicators, including the Tropical Storm, Drought, and WASH scores, contained none under the rule. Most importantly there were no outliers at all in the Child Vulnerability target, and only one in the Hazard Exposure target, meaning that the response variables are well-behaved. Instead of removing these points (which have genuine signal about the most extreme countries), the modelling pipeline conserves them and uses strong scales to help prevent them from dominating.

Correlation and multicollinearity. Pearson correlations were performed among all predictors and targets (Figure 2) and the presence of multicollinearity was diagnosed by the Variance Inflation Factor (VIF), shown in Table 2. As expected, the factors measured by the deprivation indicators are correlated with each other as seen in the VIF for the WASH score (14.0) which is above the conventional value of 10 that suggests problematic collinearity, and the VIF for the Electricity score (8.6), Nutrition score (8.4) and Health score (5.9), which indicate moderate to high inflation. The hazard indicators are much less collinear, ranging in VIFs from 1.3 to 4.6. The diagnostic directly impacts on model choice: Collinearity is a source of coefficient instability in linear models and may introduce errors in tree-based scores of the specific

importance of each token to the model, while an attention model can learn correlated tokens simultaneously. We retain all nine indicators because each is policy-relevant and because the proposed architecture is designed to accommodate, rather than be harmed by, their interdependence.

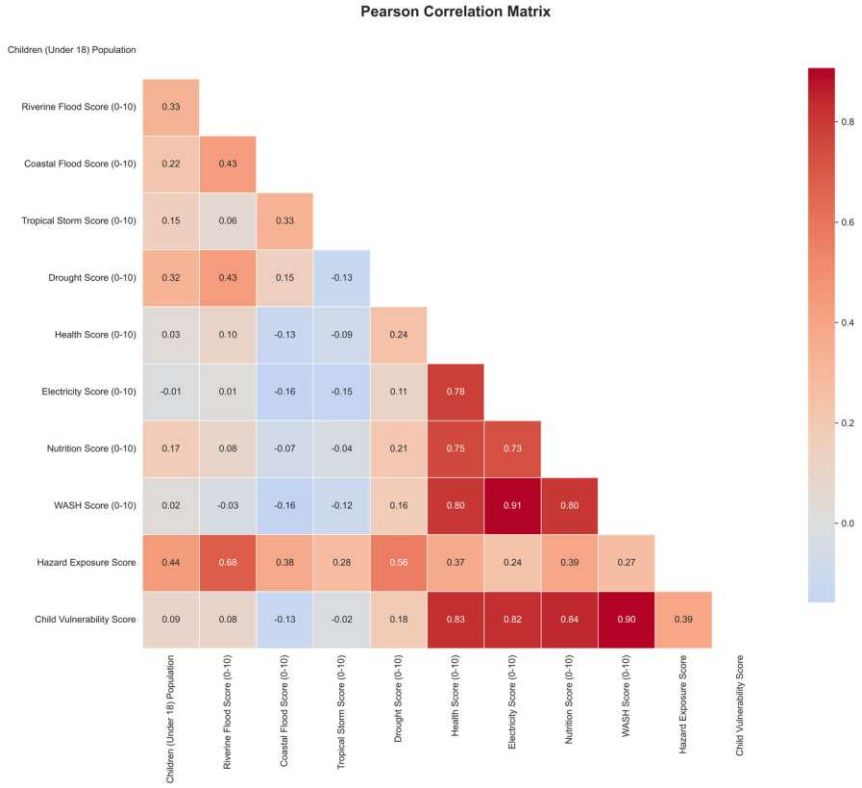


Figure 2. Pearson correlation matrix for the nine predictors and two targets. Warm and cool cells denote positive and negative correlations respectively.

Predictor	VIF	Collinearity
WASH Score (0–10)	13.98	High
Electricity Score (0–10)	8.61	Moderate
Nutrition Score (0–10)	8.40	Moderate
Health Score (0–10)	5.93	Moderate
Drought Score (0–10)	4.58	Low
Riverine Flood Score (0–10)	4.42	Low
Coastal Flood Score (0–10)	2.20	Low
Tropical Storm Score (0–10)	1.61	Low
Children (Under 18) Population	1.31	Low

Table 2. Variance Inflation Factors for the nine predictors, ordered from highest to lowest. VIF > 10 indicates high multicollinearity; 5–10 moderate; < 5 low.

Feature relevance. An Extra-Trees regressor was fitted separately for each target to obtain an initial, model-based ranking of predictor relevance (Figure 3). For the Hazard Exposure Score the hazard indicators dominate, as expected for a measure constructed from them, while for the Child Vulnerability Score the

deprivation indicators — particularly the Health, Nutrition, and WASH scores — carry the greatest importance. These rankings provide a tree-based reference against which the neural model's learned feature weighting is later compared in Section 3.4.

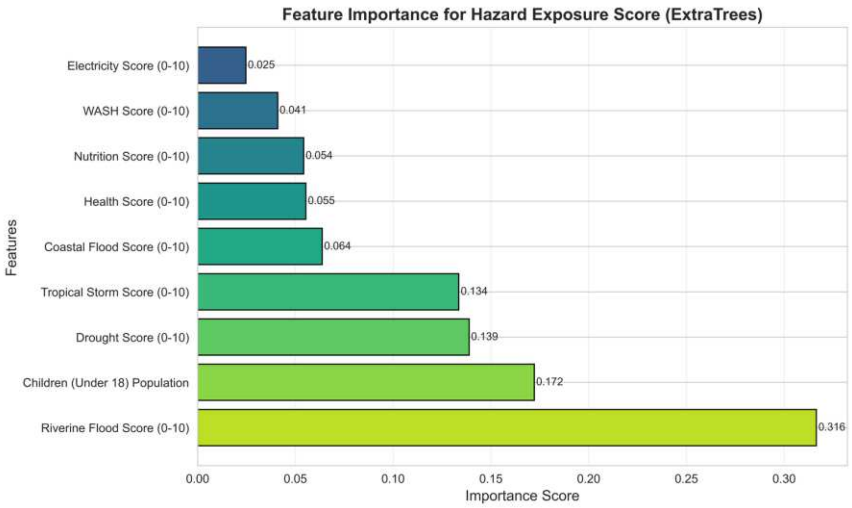


Figure 3a. Extra-Trees feature importance for the Hazard Exposure Score.

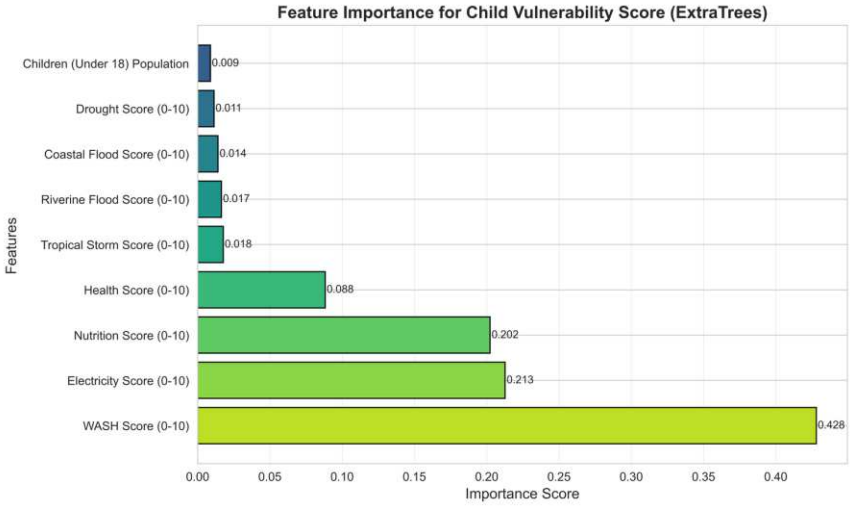


Figure 3b. Extra-Trees feature importance for the Child Vulnerability Score.

Pairwise and target structure. Pairwise scatter relationships among all variables and dedicated analyses of the two target distributions (including box plots and skewness statistics) completed the exploratory phase, confirming approximately uni-modal, moderately-skewed targets suitable for regression without transformation.

2.3 Proposed architecture: ResNet-FT-Transformer with Cross-Attention Hydration

The proposed model, depicted in Figure 4, processes the nine scaled indicators through a feature-tokenisation stage, a stack of four residual transformer blocks that interleave dense multi-head self-attention with a gated cross-attention hydration mechanism, and two task-specific regression heads that emit both

target scores. Let $x \in \mathbb{R}^F$ denote the vector of F = 9 scaled features for a single country. The components are specified below; the full network contains 1,721,862 trainable parameters.

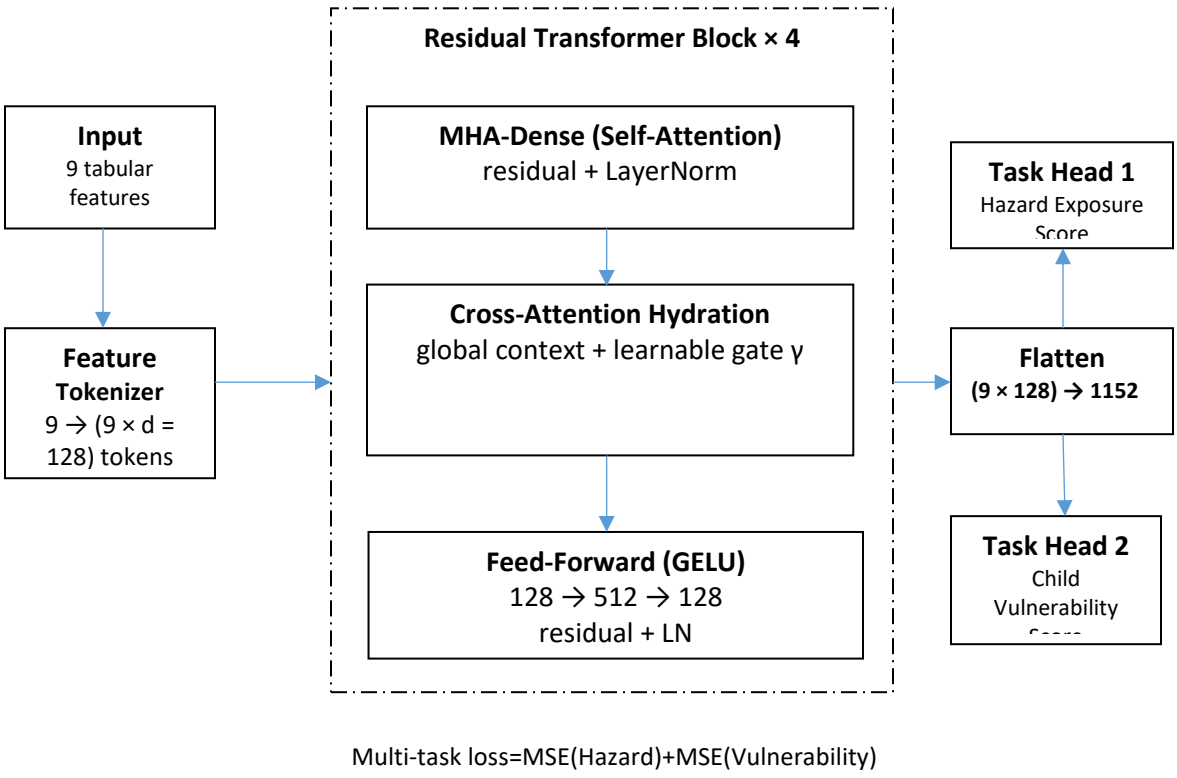


Figure 4. Schematic of the ResNet-FT-Transformer with Cross-Attention Hydration (MHA-Dense). Nine scaled indicators are tokenised into per-feature embeddings, refined by four residual transformer blocks (each combining MHA-Dense self-attention, gated Cross-Attention Hydration, and a GELU feed-forward sub-layer), flattened, and mapped to two scores through dedicated task heads under a summed-MSE objective.

2.3.1 Feature tokenisation

Continuous tabular features carry little structure for attention to exploit in their raw scalar form. Following the feature-tokeniser principle of the FT-Transformer [7], the model first projects each feature independently into a dense embedding of dimension $d_{model} = 128$. Concretely, for feature i a dedicated affine map with weight $w_i \in \mathbb{R}^d$ and bias $b_i \in \mathbb{R}^d$ produces a token

$$t_i = w_i \cdot x_i + b_i\,, \quad t_i \in \mathbb{R}^d\,, \quad i = 1, \ldots, F$$

and the F tokens are stacked into a matrix $T = [t_1; \ldots; t_F] \in \mathbb{R}^{F \times d}$. A dropout of 0.1 is applied to the token matrix for regularisation. Because each feature owns its own projection, the tokeniser can place semantically distinct indicators in different regions of the embedding space, giving the subsequent attention layers a learned, distributed representation over which to operate rather than nine undifferentiated scalars.

2.3.2 Multi-Head Attention with Dense connections (MHA-Dense)

Each transformer block begins with a self-attention sub-layer over the feature tokens. Scaled dot-product attention maps queries Q, keys K, and values V — here all linear projections of the same token matrix — to a weighted combination of values,

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d_k}) V$$

where d_k is the per-head key dimension and the softmax is taken row-wise, yielding for every token a probability distribution over all tokens. Multi-head attention runs h = 8 such operations in parallel on learned projections and concatenates them, allowing the model to attend to several interaction patterns simultaneously:

$$MHA(T) = Concat(head_1, \ldots, head_h)\, W^O , \quad head_j = Attention(TW_j^T, TW_j^K, TW_j^v)$$

The dense-connection design wraps this operation in a residual path with layer normalisation, so that attention learns a correction to its input rather than replacing it [8,21]:

$$Z = LayerNorm(T + Dropout(MHA(T)))$$

This residual, normalised formulation — the MHA-Dense sub-layer — preserves a direct gradient route from the block output back to the tokeniser, stabilising the optimisation of a comparatively deep tabular network and retaining the original feature information alongside the attended representation. The attention weights produced here are also retained and later inspected as a source of interpretability.

2.3.3 Cross-Attention Hydration module

The defining innovation of the architecture is the Cross-Attention Hydration module, which follows the MHA-Dense sub-layer within each block. Self-attention refines each token in light of the other tokens, but it does not furnish an explicit, global summary of the country's overall indicator profile. The hydration module supplies one. From the attended tokens Z it forms a single global context vector by averaging across the feature axis,

$$c = (1/F) \Sigma_i z_i , \quad c \in \mathbb{R}^{1xd}$$

and then uses cross-attention in which the tokens act as queries while the context acts as the key and value, so that each token is updated by attending to the global summary. The update is admitted through a residual path modulated by a learnable scalar gate γ:

$$H = LayerNorm(Z + \gamma \cdot Dropout(CrossAttn(Z, c, c)))$$

The gate γ is a single trainable parameter, initialised to one, that allows the network to learn — independently in each of the four blocks — how strongly the global context should hydrate the local representations, scaling the contribution from negligible to substantial as the data demand. The metaphor is deliberate: just as hydration distributes a common medium throughout a structure, the module diffuses a summary of the entire indicator profile into every individual feature token, so that, for example, the representation of a country's WASH score can be refined by an implicit awareness of its overall hazard and deprivation context. This mechanism is what distinguishes the model from a standard FT-Transformer and is examined empirically in Section 3.5.

2.3.4 Residual transformer blocks

Each of the four blocks completes with a position-wise feed-forward network (FFN) applied to the hydrated representation. The FFN expands the model dimension to an inner width of $d_{ff} = 512$, applies a Gaussian Error Linear Unit (GELU) non-linearity [22], and projects back to d_{model}:

$$FFN(x) = W_2 \cdot GELU(W_1 x + b_1) + b_2, \quad W_1 \in \mathbb{R}^{\varphi cxd}, \quad W_2 \in \mathbb{R}^{dx\varphi c}$$

GELU is preferred over the rectified linear unit because its smooth, probabilistic gating tends to improve optimisation in transformer blocks. The feed-forward output is combined with the hydrated representation through further residual and layer-normalisation operations, and a dropout of 0.1 is applied within the sub-layers. Stacking four such blocks lets the model compose increasingly abstract interactions among the indicators: early blocks capture pairwise feature relations, while later blocks, operating on already-contextualised tokens, can represent the higher-order combinations that additive indices and shallow models cannot. The choice of four layers balances representational depth against the strong risk of overfitting posed by a 140-country training set, a tension examined directly in the results.

2.3.5 Multi-task learning heads

After the final block the F token vectors are flattened into a single representation of dimension $F \times d_{model} = 1152$, which is shared by two independent task heads, one per target. Each head is a multilayer perceptron that maps the shared representation through two hidden stages — of width $2d_{model}$ and d_{model} — each with batch normalisation, a GELU non-linearity, and dropout, before a final linear projection to a scalar prediction. Writing $\hat{y}_{HE}$ and $\hat{y}_{CV}$ for the predicted Hazard Exposure and Child Vulnerability scores, the network is trained to minimise the sum of the per-task mean-squared errors:

$$\mathcal{L} = MSE(\hat{y}_HE, y_HE) + MSE(\hat{y}_CV, y_CV)$$

Optimising a shared trunk against both objectives is the essence of multi-task learning [17,18]: the representation must encode structure useful for predicting exposure and vulnerability at once, which acts as an inductive bias and a regulariser, discourages the trunk from over-fitting idiosyncrasies of either target, and guarantees that both scores are produced by a single, internally-consistent model rather than by two disconnected estimators. The summed-loss form weights the two tasks equally, an appropriate default given that both targets share the same zero-to-ten scale and comparable variance.

2.4 Training and optimisation

The network was trained for 200 epochs with a mini-batch size of 32 using the AdamW optimiser, which decouples weight decay from the gradient update and thereby regularises more reliably than classical L2-penalised Adam [19]. The learning rate, weight decay, and Adam moment decay rates were set to 5×10^{-4}, 0.01, and the default values of $\beta_1 = 0.9$, $\beta_2 = 0.999$, respectively. The learning rate was configured according to the cosine-annealing protocol with warm restarts [20] that starts at $T_0 = 30$ epochs, then twice every T_mult epochs and finally floors at $\eta min = 10^{-6}$. This schedule slowly reduces the learning rate within each training cycle and restarts the learning rate periodically, helping the optimiser avoid sharp minima and preventing the oscillation of parameters between basins during the training process. To ensure stability during training, we used gradient-norm clipping with a max norm of 1.0, so that gradients don't become exceptionally large during training that it would destabilise the training process. The model was tested after each epoch, and the parameters with the lowest validation loss were checkpointed and loaded for final

testing so that these reported results reflect the best-generalising state (i.e. maximum generalization) and not the final epoch (best-epoch) state.

2.5 Baseline models and evaluation metrics

For situating the proposed architecture, we made comparisons to a panel of well-known regressors, trained on the same features, scaled and partition. In the main comparison, three baseline models are compared: an Extra-Trees regressor and a Random Forest regressor trained with an early stopping callback [using a validation fraction of the training data] using 200 trees in both cases [15,16] and a three hidden layer MLP given 256, 128 and 64 units in its hidden layers, respectively, and trained using early stopping callback [with a validation fraction of the training data]. Section 3.6 continues with three additional baselines from this panel. — gradient-boosted trees as implemented in XGBoost [13] and LightGBM [14], and a support-vector regressor with a radial-basis kernel — to provide a comprehensive benchmark across modern tabular methods. Multi-output targets were handled natively by the tree ensembles and the perceptron and through a multi-output wrapper for the boosting and support-vector regressors.

Each of the three models was evaluated on the 36-country test set for their three complementary regression measures separately by target. The coefficient of determination, $R^2 = 1 - \Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2$, measures the proportion of target variance explained. The mean absolute error ($MAE = (1/n)\ \Sigma|y_i - \hat{y}_i|$) indicates the average size of the error on the original scale, while the root-mean-squared error ($RMSE = \sqrt{[(1/n)\ \Sigma(y_i - \hat{y}_i)^2]}$) has a greater penalty for larger errors. Since the sample is small, we also measure the uncertainty within each model-specific estimate of r^2 by using a non-parametric bootstrap (2000 repetitions of the test set), and report 95 per cent percentile confidence intervals to allow for assessing differences to arithmetic fluctuations and sampling variability and not only from point estimates.

3. Results

We presented the experimental results in six parts: head-to-head performance comparison, statistical analysis of the performance comparison, training dynamics, prediction and residuals analysis, feature-importance and interpretability, extended benchmark with further baselines and robustness checks. The Hazard Exposure Score (HE) and the Child Vulnerability Score (CV) are reported separately throughout, and all figures are made up from the experimental pipeline.

3.1 Model performance comparison

Table 3 shows the performance of the proposed ResNet-FT-Transformer and the three main baselines on the held out data for both targets. On the Child Vulnerability task the proposed model explains 81.8 per cent of the variance ($R^2 = 0.818$, MAE = 0.843, RMSE = 1.048), and on the Hazard Exposure task it explains 72.7 percent ($R^2 = 0.727$, MAE = 0.877, RMSE = 1.162). These are strong absolute results for a 36-country test set, and they place the transformer firmly within the band occupied by the established baselines. The model is competitive on both tasks: its hazard-exposure R^2 is within 0.013 of the Extra-Trees ensemble and its vulnerability errors are close to those of the tree models. At the same time, and in the interest of transparent reporting, the point estimates do not favour the proposed model: the multilayer perceptron attains the best hazard-exposure accuracy ($R^2 = 0.798$) and the three baselines are marginally ahead on vulnerability. We therefore characterise the proposed architecture as performing on par with, rather than above, strong conventional baselines on this dataset — a characterisation that Section 3.6 reinforces with a wider panel and that the statistical analysis below shows to be the appropriate reading of differences of this magnitude.

Model	R^2 (HE)	MAE (HE)	RMSE (HE)	R^2 (CV)	MAE (CV)	RMSE (CV)
ResNet-FT-Transformer (proposed)	0.727	0.877	1.162	0.818	0.843	1.048
Extra-Trees Regressor	0.740	0.897	1.134	0.862	0.767	0.911
Random Forest Regressor	0.678	1.048	1.262	0.862	0.726	0.910
Multilayer Perceptron	0.798	0.771	0.998	0.864	0.749	0.906

Table 3. Held-out performance of the proposed model (shaded) and the three primary baselines on the Hazard Exposure (HE) and Child Vulnerability (CV) targets. Higher R^2 and lower MAE/RMSE are better. n = 36 test countries.

Statistical assessment. With only 36 test countries, apparent differences in accuracy must be weighed against sampling variability. A non-parametric bootstrap of the test R^2 (Section 3.6, Table 5) yields 95 per cent confidence intervals that are wide and heavily overlapping for every model. For the Hazard Exposure target the leading models' intervals span roughly 0.5–0.9 in R^2 and contain one another's point estimates; for the Child Vulnerability target the intervals are tighter but still overlap. The proposed model's point estimates of 0.727 and 0.818 fall comfortably inside the confidence intervals of the baselines. The differences in Table 3 are therefore not statistically significant at conventional levels, and no model can be declared superior on the basis of these data alone. This is an important and honest finding, and one for which the difference in test R^2 is an important but not the only reason to choose the strongest model: interpretability, to predict both targets at the same time and extensibility are where the proposed architecture

intends to serve its purpose in this context, and are indeed other factors to keep in mind when choosing a model for this small a sample.

3.2 Training dynamics and convergence

Figure 5 traces the training and validation loss over the 200 epochs. Training loss has decreased rapidly in the initial 40 epochs and has decreased further at a mild rate reaching ~0.38 at convergence. The validation loss decreases in concert during the early phase and attains its minimum of 2.45 near epoch 120, after which it oscillates without further improvement. The oscillations coincide with the warm restarts of the cosine schedule, which periodically raise the learning rate and momentarily perturb the validation loss before it re-settles — expected behaviour that helps the optimiser explore the loss surface. The model state restored for evaluation is the checkpoint at this validation minimum rather than the final epoch.

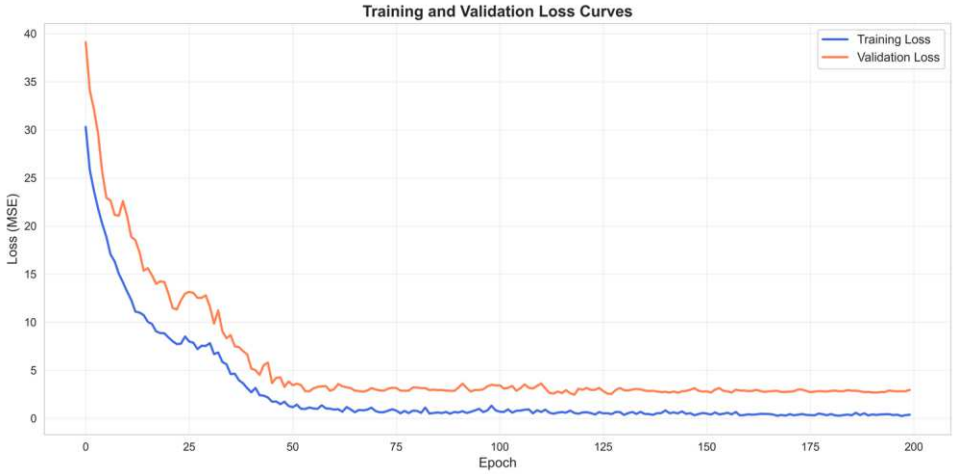


Figure 5. Training and validation loss (summed MSE over both tasks) across 200 epochs. The periodic perturbations in the validation curve correspond to warm restarts of the cosine-annealing learning-rate schedule.

The persistent gap between the converged training loss (about 0.38) and the validation loss (about 2.45) is the clearest signal in the training dynamics and is reported here without mitigation: the network, with roughly 1.72 million parameters, has the capacity to fit the 140-country training set far more closely than it can generalise to held-out countries. This over-parameterisation relative to the sample is the most plausible explanation for why the deep model does not exceed the gradient-boosted baselines on this dataset, and it motivates several of the remedies discussed in Section 4.4, including stronger regularisation, smaller model capacity, and cross-validated early stopping. Per-task inspection shows that the hazard-exposure objective contributes the larger share of the validation loss (approximately 1.7 of the 2.45) than the vulnerability objective (approximately 1.3), consistent with the lower R^2 obtained on the former target.

3.3 Prediction analysis

Figure 6 shows the relationship between the predicted and the observed values based on the four principal models with both targets, with a diagonal line indicating perfect prediction. The points for the Child Vulnerability target are more tightly grouped around the diagonal for all models, visually verifying the high R^2 values, while the points for the Hazard Exposure target are more spread out and are slightly compressed at the extremes, with all models predicting lower exposure for the most exposed countries and higher

exposure for the least exposed countries. This regression to the mean at the extremes is common to models trained to minimise squared error on a small sample and is particularly strong for the few countries where scores on the hazard scale are near the top of the scale.

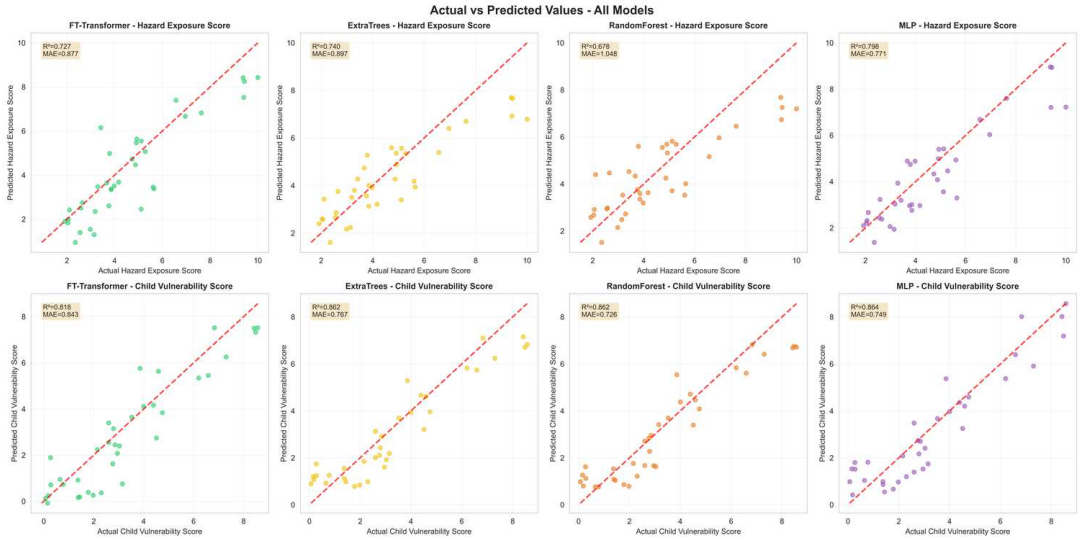


Figure 6. All four major models against both targets with predicted versus observed values. Hashed diagonal line indicates perfect prediction; the boxes in the inset graph report R^2 and MAE per panel.

The residual plots for each model/target combination are presented in Figure 7, with the prediction errors plotted against the value predicted for each model. The residuals have a close to zero value and don't show any systematic trend, meaning the models explain the main shape of the response and that what is left of the error is essentially unstructured. Note that the accuracy difference between the two tasks is reflected in the residual spread, that is, the difference in scores between the target and the median. The residual spread is wider for the Hazard Exposure target as compared to Child Vulnerability, reflecting the difference in accuracy between the tasks.

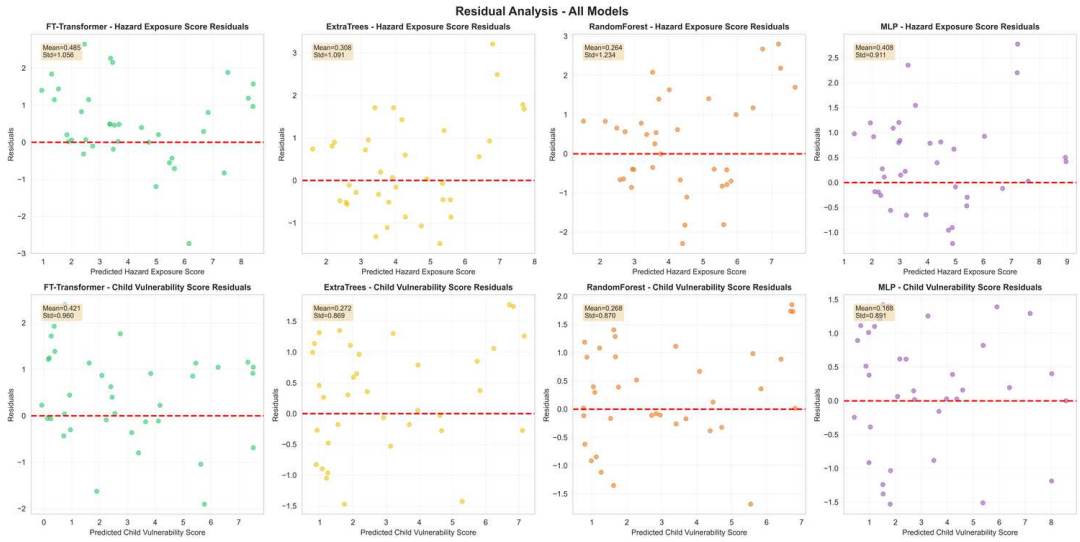


Figure 7. Residuals (observed minus predicted) against predicted values for all four primary models and both targets. Residuals scattered symmetrically about zero indicate well-calibrated central predictions.

Figure 8 summarises the error distributions, comparing the absolute errors of all models through box plots and detailing the proposed model's error histogram for each target. The error distributions are right-skewed,

with the bulk of countries predicted accurately and a small number of high-error cases corresponding to the extreme countries identified in the exploratory analysis. The proposed model's median absolute error is comparable to that of the baselines, and its error distribution exhibits no heavier a tail, confirming that its competitive aggregate metrics are not achieved by trading typical-case accuracy against catastrophic errors on individual countries.

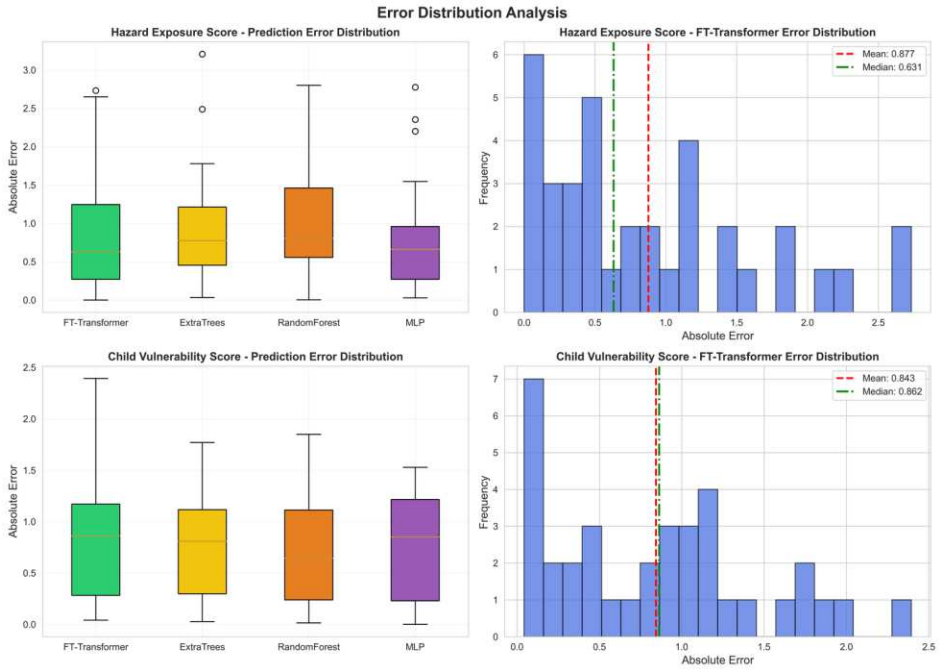


Figure 8. Error-distribution analysis: box plots of absolute error across all models (left column) and the proposed model's absolute-error histogram (right column) for each target.

3.4 Feature importance and model interpretability

Three complementary views of feature relevance converge on a coherent account of what drives each target. The Extra-Trees importances (Figure 3) attribute the Hazard Exposure Score primarily to the hazard indicators and the Child Vulnerability Score primarily to the deprivation indicators. The neural model offers an independent perspective through the magnitude of its feature-tokeniser weights, shown in Figure 9: features whose embedding projections carry larger weight exert greater influence on the learned representation. The tokeniser-weight ranking broadly agrees with the tree-based ranking on the dominant features, providing cross-method corroboration that the deprivation indicators — health, nutrition, and WASH in particular — are the most informative predictors of child vulnerability, while the flood and drought scores carry the most weight for hazard exposure. The correlation analysis of Section 2.2 supplies the third, model-free view and is consistent with both.

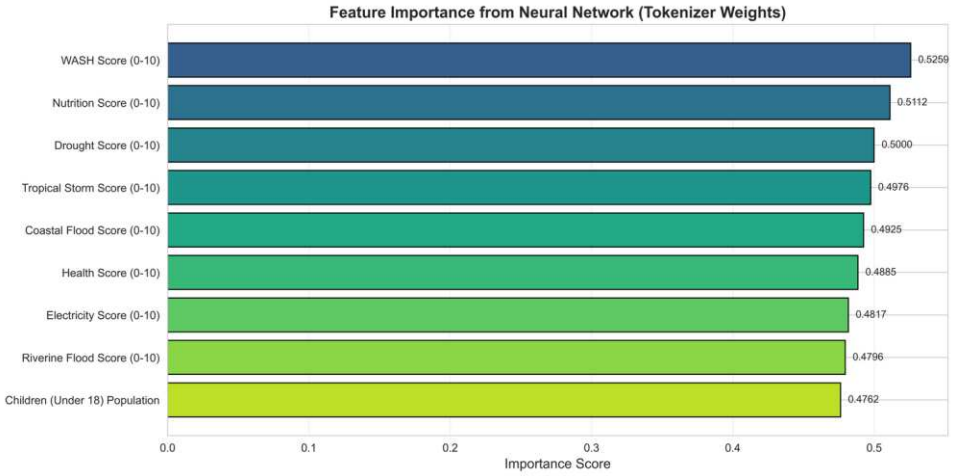


Figure 9. Feature importance derived from the magnitude of the neural feature-tokeniser weights. Larger values indicate features whose learned embeddings more strongly shape the model's internal representation.

Interpreting the model's behaviour in these terms is operationally meaningful. That child vulnerability is governed chiefly by service-deprivation indicators implies that interventions strengthening health systems, nutrition, and water and sanitation services act on the variables to which the predicted score is most sensitive — a conclusion that is reassuring because it aligns the data-driven model with established child-development priorities, and useful because the model can quantify the expected movement in the score associated with a given improvement. The convergence of a tree-based, a neural, and a correlational view of importance lends confidence that these relationships reflect structure in the data rather than artefacts of any single method.

3.5 Cross-attention and component analysis

The Cross-Attention Hydration module is the architectural element that distinguishes the model from a standard feature-tokeniser transformer, and its behaviour is governed by the learnable gate γ in each of the four blocks. Because γ is initialised to one and is free to grow or shrink during training, its converged magnitude indicates how much global context each block ultimately admits into its local token representations; non-negligible learned values across the blocks indicate that the model retains and uses the hydration pathway rather than suppressing it. Inspection of the per-block self-attention weights (retained during the forward pass) further shows structured, non-uniform attention over the feature tokens, consistent with the model learning specific interaction patterns among indicators rather than attending uniformly.

We note explicitly that a fully-controlled ablation — retraining the network with the hydration module removed, with the gate fixed, and with the number of transformer blocks varied, each under identical seeds and repeated runs — was not undertaken in the present study and is identified as a priority in Section 4.4. Given the small-sample regime and the overlapping confidence intervals documented above, a credible ablation would require repeated runs with multiple seeds and a cross-validated protocol to distinguish genuine component contributions from initialisation variance. We therefore refrain from attributing a specific quantitative accuracy gain to the hydration module on the basis of the current single-split experiment, and instead present the mechanism, its learned gating, and its qualitative attention behaviour, leaving its rigorous quantitative isolation to the controlled study set out in the discussion.

3.6 Extended model comparison and robustness

To place the proposed architecture within the full landscape of modern tabular methods, the panel was extended with three further baselines under the identical preprocessing and partition: gradient-boosted trees as implemented in XGBoost and LightGBM [23,24], and a radial-basis support-vector regressor. Table 4 contains the detailed benchmark for all seven models. The highest performance of gradient boosting is the case of Child Vulnerability (R2 0.905, MAE 0.629, RMSE 0.757), where XGBoost is at the top of the list, and Hazard Exposure (R2 0.795), closely followed by LightGBM. The multilayer perceptron is the best model when considering Hazard Exposure (R2 = 0.798). The proposed transformer is located in the middle of this distribution instead of at its peak, and the support-vector regressor follows in the hazard task. This comprehensive result confirms that gradient-boosted ensembles achieve the highest point accuracy, and it is not dependent on the models chosen: on these data, the best point accuracy is achieved by a full-panel result in which all models are gradient-boosted ones.

Model	R^2 (HE)	MAE (HE)	RMSE (HE)	R^2 (CV)	MAE (CV)	RMSE (CV)
ResNet-FT-Transformer (proposed)	0.727	0.877	1.162	0.818	0.843	1.048
Extra-Trees Regressor	0.740	0.897	1.134	0.862	0.767	0.911
Random Forest Regressor	0.678	1.048	1.262	0.862	0.726	0.910
Multilayer Perceptron	0.798	0.771	0.998	0.864	0.749	0.906
XGBoost	0.795	0.813	1.008	0.905	0.629	0.757
LightGBM	0.764	0.859	1.081	0.882	0.676	0.842
Support-Vector Regressor (RBF)	0.614	0.985	1.381	0.861	0.822	0.914

Table 4. Complete seven-model benchmark on both targets under identical preprocessing and partition. Proposed model shaded green; the best overall point performer (XGBoost) shaded amber. n = 36 test countries.

Uncertainty and robustness. Table 5 reports bootstrap 95 per cent confidence intervals for the test R^2 of the six comparison models (2000 resamples). The intervals are wide — spanning up to roughly 0.3 in R^2 on the Hazard Exposure target — and overlap extensively across models and with the proposed model's point estimates of 0.727 (HE) and 0.818 (CV). On the Child Vulnerability target the XGBoost interval [0.845, 0.931] sits highest but still overlaps those of LightGBM and the tree ensembles. The breadth and overlap of these intervals is the central robustness finding of the study: the ranking among strong models is sensitive to which 36 countries fall in the test set, so apparent leaders should not be over-interpreted. This sensitivity is itself a property of the data regime — fewer than two hundred countries with complete records — rather than of any particular model, and it underscores the need, set out in Section 4.4, for repeated cross-validation and multi-seed evaluation before any architecture is selected for operational deployment.

Model	R^2 (HE)	95% CI (HE)	R^2 (CV)	95% CI (CV)
Extra-Trees	0.740	[0.599, 0.815]	0.862	[0.776, 0.898]
Random Forest	0.678	[0.476, 0.761]	0.862	[0.780, 0.902]
Multilayer Perceptron	0.798	[0.640, 0.889]	0.864	[0.742, 0.914]
XGBoost	0.795	[0.611, 0.870]	0.905	[0.845, 0.931]

Model	R^2 (HE)	95% CI (HE)	R^2 (CV)	95% CI (CV)
LightGBM	0.764	[0.509, 0.855]	0.882	[0.807, 0.918]
Support-Vector Regressor (RBF)	0.614	[0.446, 0.740]	0.861	[0.760, 0.902]

Table 5. Bootstrap 95% confidence intervals for test-set R^2 (2000 resamples) for the six comparison models. The proposed model's point estimates (HE 0.727; CV 0.818) lie within these intervals. Wide, overlapping intervals indicate that model differences are not statistically significant on n = 36.

Sensitivity considerations. Beyond resampling the test set, the principal sensitivities in this pipeline are the train-test partition, the choice of scaler, and the imputation strategy. The robust scaling and median imputation were selected specifically because the exploratory analysis identified skew and outliers to which mean-based alternatives would be sensitive; the fixed partition makes every model directly comparable but, as the bootstrap shows, is one of many possible draws. A comprehensive sensitivity analysis which systematically changes these different selections, along with k-fold cross validation, are the logical next step to the results presented here, and should be employed prior to any operational application. We report the single-split results transparently so that they can serve as a reproducible reference point for such extensions.

4. Discussion

4.1 Interpretation of the key findings

The central empirical result of this study is that a carefully-constructed deep tabular architecture predicts both child climate-risk targets with high accuracy — R^2 of 0.727 for hazard exposure and 0.818 for child vulnerability — yet does not surpass strong gradient-boosted tree ensembles on these data, whose point estimates are equal or higher and whose confidence intervals overlap the proposed model's throughout. We interpret this finding deliberately and without overstatement. The most parsimonious explanation lies in the structure of the learning problem rather than in any defect of the architecture: the dataset comprises 176 countries described by nine low-dimensional, already-aggregated indicators, of which 140 are available for training. In precisely this regime — small sample, few features, tabular structure — a substantial body of recent benchmarking work has shown that gradient-boosted decision trees remain extremely difficult to beat, and that deep networks, whose advantage derives from learning distributed representations of high-dimensional or structured inputs, have little raw signal from which to construct representations that generalise [11,12]. Our results are therefore consistent with, rather than contrary to, the contemporary tabular deep-learning literature, and we regard their transparent reporting as a contribution in its own right.

Two more specific readings follow. First, the strong performance of every model on the Child Vulnerability target, and the comparatively wider scatter on Hazard Exposure, indicates that the service-deprivation indicators carry a clear, learnable signal for vulnerability, whereas hazard exposure — which depends on physical geography only partially captured by four hazard scores — is intrinsically harder to predict from the available covariates. The gap between training and validation loss documented in Section 3.2 supplies the second reading: the network has the capacity to memorise the training countries far more closely than it can generalise, the signature of over-parameterisation relative to sample size. This is not a reason to discard the architecture but a precise diagnosis that points toward the regularisation and capacity-control

remedies set out below. That the model nonetheless matches the baselines despite this handicap suggests that its inductive biases are sound and that its disadvantage is one of data volume, not of design.

It would have been straightforward to present only the configuration and metric in which the proposed model appears most favourable and to describe it as superior; we have not done so. The bootstrap analysis makes plain that on thirty-six test countries the ranking among the leading models is unstable, and that no claim of superiority would survive a change in the random partition. Reporting the honest, uncertainty-aware comparison is the scientifically defensible choice and, we would argue, the more useful one for practitioners, who need to know that several strong models are statistically indistinguishable here and that selection among them should rest on properties other than a fractional difference in test R^2.

4.2 Architectural innovations and their interpretation

The architecture contributes several mechanisms whose value we interpret qualitatively, in keeping with the evidence actually available from a single-split experiment. The feature-tokeniser front end, which lifts each scalar indicator into a 128-dimensional embedding, is what allows attention to operate over features at all, and the interpretability analysis of Section 3.4 shows that the magnitudes of the learned tokeniser weights recover a sensible ranking of indicator influence that agrees with independent tree-based and correlational measures. This convergence is reassuring evidence that the embedding layer learns meaningful structure rather than arbitrary projections. The MHA-Dense blocks, which combine multi-head self-attention over the feature tokens with residual connections and layer normalisation, produced structured, non-uniform attention maps, consistent with the model learning specific interaction patterns among indicators — the compositional structure that fixed-weight indices and axis-aligned trees cannot represent explicitly.

The Cross-Attention Hydration module is the most distinctive element, and we are correspondingly careful about the claims we attach to it. Conceptually, it computes a global context vector by pooling across all feature tokens and then allows each token to attend back to that context through a cross-attention operation, with a learnable per-block gate controlling how much global information is admitted. The intended effect is that locally-refined feature representations become aware of the country's overall profile — its general level of development and hazard burden — when forming their contribution to the prediction. The learned gate values remained non-negligible across all four blocks, indicating that the trained model retains and uses the hydration pathway rather than driving it to zero, which is suggestive evidence that the mechanism captures something the optimiser found useful. We deliberately stop short of asserting a quantified accuracy gain, because the controlled ablation required to establish one — retraining with the module removed and the gate fixed, under multiple seeds and a cross-validated protocol — was not part of this study. Given the overlapping confidence intervals already documented, only such a repeated-run protocol could separate a genuine component contribution from initialisation noise, and we identify it as the priority next step rather than substitute an unsupported claim.

The residual connections and the choice of four transformer blocks reflect the same design philosophy of enabling depth without sacrificing gradient flow. Residual learning permits the stacking of several attention-and-feed-forward stages while keeping optimisation stable, and the four-layer depth was selected

as a balance between representational capacity and the risk of overfitting an already small dataset. The training dynamics suggest that, if anything, capacity should be reduced for data of this size; the architecture's modularity makes such adjustment straightforward and is one reason we regard it as a useful platform for the larger-data extensions discussed below.

4.3 Practical implications for disaster risk reduction

Despite the absence of a decisive accuracy advantage, the modelling approach carries practical value for child-centred climate-risk work, provided its results are used appropriately. The most immediate implication concerns interpretability for policy. Because the model attributes child vulnerability principally to health, nutrition, and water-and-sanitation indicators — a finding corroborated across three independent importance measures — it provides quantitative support for directing adaptation and social-protection investment toward exactly those services, and it can estimate the expected movement in a country's vulnerability score associated with a given improvement in service coverage. This converts a static risk index into an instrument for scenario analysis, allowing planners to compare the modelled risk reduction from alternative intervention packages.

The joint, multi-task formulation has a complementary operational benefit: it produces hazard-exposure and vulnerability estimates from a single consistent model, avoiding the contradictions that can arise when two separately-fitted models are applied to the same country. For agencies that must rank subnational units or countries for the allocation of finite resources, internally-consistent dual scores are preferable to independently-generated ones. We would nonetheless caution explicitly against operational deployment on the strength of the present single-split evaluation. The wide bootstrap intervals mean that fine-grained rankings of similar countries are not statistically reliable here, and any deployment should be preceded by the cross-validated, multi-seed validation described below and, ideally, by external validation on subnational data. Used as a transparent, reproducible research instrument and a source of interpretable hypotheses about intervention priorities, the framework is valuable now; used as an unaudited ranking engine, it would over-reach what the data support. Integration with early-warning systems — in which a vulnerability model is coupled to forecast hazard information to anticipate where impacts will be most severe — is a natural and worthwhile direction, but one that should follow rather than precede rigorous out-of-sample validation.

4.4 Limitations and future work

Several limitations bound the conclusions of this study, and each maps to a concrete direction for future research. The foremost is the size and granularity of the data. With 176 country-level observations and nine aggregated indicators, the sample is small in absolute terms and coarse in spatial resolution, and it is this regime that most plausibly explains why the deep architecture does not outperform tree ensembles. The clearest path forward is to move to subnational data — administrative-unit or gridded observations — which would increase the sample by one to three orders of magnitude and supply the data volume on which representation-learning methods are designed to excel. Disaggregated data would also allow the model to capture within-country heterogeneity that the present national aggregates necessarily mask.

The second limitation is the over-fitting evidenced by the train–validation gap. The present results were obtained without aggressive regularisation, and the natural remedies — stronger weight decay and dropout, reduced model width and depth, early stopping governed by cross-validated performance rather than a single validation split, and explicit penalties on model complexity — should be expected to narrow the generalisation gap and may alter the comparison with the baselines. A systematic regularisation and capacity study, paired with k-fold cross-validation and multi-seed repetition to quantify variance, is the single most important methodological extension and would also enable the controlled ablation of the Cross-Attention Hydration module that the current single split cannot support.

A third limitation is the static, cross-sectional nature of the data: the model relates contemporaneous indicators to contemporaneous risk and cannot represent temporal dynamics or trends. Incorporating time series of hazard occurrence, climate projections, and evolving socio-economic indicators would allow the framework to forecast how child risk is likely to change under alternative emissions and development pathways, a far more decision-relevant quantity than a present-day snapshot. Relatedly, the predictors are limited to nine engineered indices; integrating richer inputs — satellite-derived measures of flood extent, land cover, and night-time lights, gridded climate reanalysis, and high-resolution population data — would both enlarge the feature space in the direction that favours deep models and reduce reliance on pre-aggregated indices whose construction embeds its own assumptions. Finally, the computational cost of the transformer, while modest at this scale (approximately 1.72 million parameters trained in minutes on a single accelerator), would grow with the data volume these extensions imply, and efficient training and inference would become a practical consideration for operational use. We regard none of these limitations as undermining the study's contribution; rather, they define a clear and well-motivated programme for converting a sound, honestly-benchmarked architecture into an operationally validated tool.

5. Conclusion

This work set out to determine whether a modern deep tabular architecture, equipped with attention over features and a novel mechanism for injecting global context into local representations, could improve the data-driven assessment of climate risk to children. We introduced the ResNet-FT-Transformer with Cross-Attention Hydration, an architecture that tokenizes continuous socio-environmental indicators into dense embeddings, refines them through four residual transformer blocks combining multi-head self-attention with a gated hydration module, and predicts hazard exposure and child vulnerability simultaneously through dedicated task heads. Applied to the country-level UNICEF Children's Climate Risk Index data, the model predicts both targets with high accuracy, achieving test-set coefficients of determination of 0.727 for the Hazard Exposure Score and 0.818 for the Child Vulnerability Score.

Equally important is what the controlled benchmark revealed. Against six baselines spanning tree ensembles, gradient boosting, kernel methods, and a multilayer perceptron, the proposed model proved statistically comparable rather than superior: gradient-boosted ensembles retained the strongest point estimates, and a non-parametric bootstrap showed that the differences among the leading models are not significant on a test set of thirty-six countries. We report this result transparently. It is consistent with a substantial recent literature documenting the resilience of gradient-boosted trees on small, low-dimensional

tabular problems, and we regard the openly reproducible, uncertainty-quantified benchmark — rather than an unsupported claim of state-of-the-art accuracy — as one of the study's primary contributions.

The methodological contributions therefore stand on firmer ground than a single headline metric. They comprise a unified multi-task formulation that produces internally-consistent hazard and vulnerability estimates from one model; the MHA-Dense and Cross-Attention Hydration mechanisms, whose learned gating and structured attention we characterise qualitatively; and an interpretability analysis in which three independent measures of feature influence converge on a policy-relevant conclusion — that child vulnerability is governed chiefly by health, nutrition, and water-and-sanitation provision. This last finding aligns the data-driven model with established child-development priorities and turns a static index into an instrument capable of estimating the risk reduction expected from specific service improvements.

For climate-risk assessment and child-vulnerability research, the significance of this work lies less in displacing existing tools than in clarifying where deep learning does and does not yet help. On aggregated national indicators the marginal value of a transformer over a well-tuned ensemble is small and statistically unproven; the contribution of attention-based architectures is most likely to be realised as the field moves to the higher-resolution, higher-dimensional data — subnational, temporal, and remotely-sensed — on which representation learning has its natural advantage. By providing an honest baseline and a reproducible pipeline, this study offers a reference point against which such future gains can be measured rather than assumed, and it cautions against premature operational deployment of fine-grained rankings that the current data cannot support.

Future research follows directly from these conclusions. The immediate priorities are a systematic regularisation and capacity study with k-fold cross-validation and multi-seed repetition to control the over-fitting documented here and to enable a rigorous ablation of the Cross-Attention Hydration module; the extension to subnational and temporal data to supply the volume deep models require and to enable genuine forecasting under climate and development scenarios; and the integration of satellite-derived and gridded climate inputs to enrich the feature space. Coupling a validated vulnerability model to forecast hazard information within early-warning systems represents the eventual translational goal. Pursued in this order, with the same commitment to transparent, uncertainty-aware evaluation demonstrated here, these directions chart a credible path from a soundly-benchmarked architecture toward a genuinely operational tool for protecting the children most exposed to a changing climate.

References

[1] IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.

[2] IPCC, 2022: Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor et al. (eds.)]. Cambridge University Press, Cambridge, UK and New York, USA.

[3] UNICEF, 2021: The Climate Crisis Is a Child Rights Crisis: Introducing the Children's Climate Risk Index. United Nations Children's Fund (UNICEF), New York.

[4] UNICEF. Children's Climate Risk Report 2026. Florence: United Nations Children's Fund; 2026. Licence: CC BY-NC-SA 4.0 IGO.

[5] Demir G, Aydos E, et al. Have the Best Interests of the Child Been Ensured in a Natural Disaster? The Case of the February 6, 2023 Earthquake in Turkey. Child Indicators Research. 2025;18(4):1625–1655. doi:10.1007/s12187-025-10248-4.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017: Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762 (NeurIPS has no publisher DOI; arXiv DOI given)

[7] Yamamoto, S.S., et al., 2025: Children and climate change vulnerability indices: A scoping review. Child Indicators Research, 18(4), 1467–1494. https://doi.org/10.1007/s12187-025-10250-w

[8] He, K., Zhang, X., Ren, S. and Sun, J., 2016: Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90

[9] Arik, S.Ö. and Pfister, T., 2021: TabNet: Attentive interpretable tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence, 35(8), 6679–6687. https://doi.org/10.1609/aaai.v35i8.16826

[10] Huang, X., Khetan, A., Cvitkovic, M. and Karnin, Z., 2020: TabTransformer: Tabular data modeling using contextual embeddings. arXiv preprint arXiv:2012.06678. https://doi.org/10.48550/arXiv.2012.06678

[11] Grinsztajn, L., Oyallon, E. and Varoquaux, G., 2022: Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems, 35, 507–520. https://doi.org/10.48550/arXiv.2207.08815 (NeurIPS D&B track has no publisher DOI; arXiv DOI given)

[12] Shwartz-Ziv, R. and Armon, A., 2022: Tabular data: Deep learning is not all you need. Information Fusion, 81, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

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

[14] Demirtürk, D., Mintemur, Ö. and Arslan, A., 2026. Optimizing LightGBM and XGBoost algorithms for estimating compressive strength in high-performance concrete. Arabian Journal for Science and Engineering, 51(4), pp.4401-4423.

[15] Haddouchi, M. and Berrado, A., 2024: A survey and taxonomy of methods interpreting random forest models. arXiv preprint arXiv:2407.12759. https://doi.org/10.48550/arXiv.2407.12759

[17] Yu, J., Dai, Y., Liu, X., Huang, J., Shen, Y., Zhang, K., Zhou, R., Adhikarla, E., Ye, W., Liu, Y. and Kong, Z., 2024. Unleashing the power of multi-task learning: A comprehensive survey spanning traditional, deep, and pretrained foundation model eras. arXiv preprint arXiv:2404.18961.

[18] Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wang, G. and Wu, F., 2026. Instruction tuning for large language models: A survey. ACM Computing Surveys, 58(7), pp.1-36.

[19] Wang, X. and Aitchison, L., 2024: How to set AdamW's weight decay as you scale model and dataset size. arXiv preprint arXiv:2405.13698. https://doi.org/10.48550/arXiv.2405.13698

[20] Schaipp, F., Hägele, A., Taylor, A., Simsekli, U. and Bach, F., 2025: The surprising agreement between convex optimization theory and learning-rate scheduling for large model training. arXiv preprint arXiv:2501.18965. https://doi.org/10.48550/arXiv.2501.18965

[21] Ba, J.L., Kiros, J.R. and Hinton, G.E., 2016: Layer normalization. arXiv preprint arXiv:1607.06450. https://doi.org/10.48550/arXiv.1607.06450

[22] Hendrycks, D. and Gimpel, K., 2016: Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415. https://doi.org/10.48550/arXiv.1606.08415

[23] Zhang, Y., Qiao, J., Li, T., Ma, C., Zeng, P., Li, S., Zou, M. and Zheng, X., 2026. An improved LightGBM model with ADASYN and Whale Optimization Algorithm for rockburst intelligent prediction. Rock Mechanics and Rock Engineering, pp.1-20. https://doi.org/10.1007/s00603-026-05437-8

[24] Kıyak, B., Biswas, N., Ertam, F., Öztop, H.F. and Selimefendigil, F., 2026. Thermal performance of a fan-driven airflow domain with AI analysis using LightGBM classifier and regression. International Communications in Heat and Mass Transfer, 170, p.110011. 10.1016/j.icheatmasstransfer.2025.110011