Nigerian Journal of Physics (NJP)
ISSN online: 3027-0936
ISSN print: 1595-0611
DOI: https://doi.org/10.62292/njp.v35i3.2026.594
Volume 35(3), September 2026

AI-Driven prediction system for Thermoelectric Properties: Unveiling Novel Materials for Sustainable Energy Solution in Nigeria

Akinola Samson Olayinka

Computational Research Group, Department of Physics, Edo State University Iyamho, Edo State.
Centre for Computational Science (c2snet.org), Nigeria.

*Corresponding Author's Email: akinola.olayinka@edouniversity.edu.ng Phone: +2348062447411

ABSTRACT

Nigeria's persistent energy deficit continues to underscore the need for sustainable alternative energy technologies. Thermoelectric materials convert heat directly into electrical energy and offer a promising pathway for addressing this challenge; however, the discovery of high-performance compounds remains experimentally demanding. This study presents an AI-driven machine learning framework and web-based prediction platform for estimating five thermoelectric properties: Seebeck coefficient (S), electrical conductivity (σ), thermal conductivity (κ), power factor (PF), and figure of merit (zT) directly from chemical formula and measurement temperature. Seven regression models were trained on 5,205 experimental records spanning 880 distinct chemical compositions. An 88-feature engineering framework incorporating compositional descriptors, temperature transformations, complexity metrics, and interaction terms was developed, with variance thresholding reducing the feature set to 61 variables. XGBoost achieved the best predictive performance for zT (R^2 = 0.796, RMSE = 0.154), Seebeck coefficient (R^2 = 0.787, RMSE = 96.8 μV/K), electrical conductivity (R^2 = 0.823), and thermal conductivity (R^2 = 0.797), while Gradient Boosting yielded the highest performance for power factor (R^2 = 0.863).Tree-based ensemble methods consistently outperformed linear models, confirming the non-linear nature of structure–property relationships in thermoelectric systems. Measurement temperature, heavy-element ratio, Shannon compositional entropy, and temperature–composition interaction terms emerged as the most influential predictors. The results demonstrate the effectiveness of compositional machine learning as a rapid and cost-effective screening tool for thermoelectric materials discovery. The validated framework has been deployed as an open-access web application supporting both single-compound prediction and batch screening, with automated identification of promising candidates for experimental validation.

Keywords:
Thermoelectric materials,
Machine learning,
XGBoost,
Feature engineering,
Figure of merit,
Sustainable energy,
Nigeria,
Web application.

INTRODUCTION

The problem associated with energy in Nigeria is a significant one that has affected several sectors of the economy. While the energy solution in Nigeria is still largely focused on the conventional sources like hydro and fossil sources with little effort in the direction of solar sources due to associated high cost. Universities in Nigeria have had a fair share of problems connected with utility supply where some universities were completely cut off from supply due to high cost (Asinya & Ishua, 2025; Orisa et al., 2024; Owebor et al., 2025; Unegbu et al., 2025). This energy cut from universities comes with its consequential effect on the academic and research outputs of these universities with a secondary effect on the national growth. The need to explore alternative sources like thermoelectric energy sources cannot be over-emphasized because the amount heat (thermal energy) that is wasted on daily basis due to various industrial activities and automobile is huge. This idea of thermoelectric energy alternative qualifies as a waste-to-energy (WTE) concept, where waste heat energy from various human and industrial activities can be harnessed for usage in various applications. This directly supports SDG 7 (Affordable and Clean Energy) by improving energy efficiency and recovering otherwise lost energy, SDG 9 (Industry, Innovation, and Infrastructure) by

promoting industrial energy recovery and technological innovation, SDG 11 (Sustainable Cities and Communities) by reducing urban waste heat and energy demand, and SDG 13 (Climate Action) by lowering greenhouse gas emissions associated with additional energy generation (Friedl & Dromann, 2023; Jouhara et al., 2021; UN, 2022). Thermoelectric materials are materials that have the capacity to convert heat energy to electrical energy thereby providing a viable alternative energy source. However, the process of unveiling novel materials for thermoelectric applications is a herculean task due to the complexity of their properties. AI-driven predictive systems can play a significant role the discovery process by predicting thermoelectric properties with high accuracy thereby reducing the challenges in discovering novel materials for thermoelectric applications. This study leverages the potential and innovative AI-driven solution to sustainably discover and optimize novel materials for thermoelectric applications, thereby contributing a sustainable approach to energy challenges in Nigeria.

To mitigate the environmental impact of conventional energy sources, sustainable energy solutions must be adopted urgently given the rising global demand for energy. Energy sustainability may be improved through thermoelectric materials that can change heat energy to electrical energy and vice versa (Al-Fartoos et al., 2023; Muchuweni & Mombeshora, 2023). However, there are challenges such as the costs involved, toxicity and reliance on rare earth elements that hinder widespread adoption of the thermoelectric alternative energy sources (Al-Fartoos et al., 2023). To overcome these hurdles, research now focuses on innovative approaches like one-dimensional nano structure and the development of organic thermoelectric materials including single-walled carbon nanotubes which provide low-cost, non-toxic alternatives to traditional inorganic materials (Muchuweni & Mombeshora, 2023). Additionally, looking into boron-based compounds' thermoelectric properties especially that of boron carbides and intermetallic borides demonstrates their promise for high efficiency in thermoelectric generators operating under harsh environments (Saglik et al., 2023).

Advancement of thermoelectric technology is important in reaching sustainable energy solutions as well as moving away from conventional uses of energy. The material takes waste heat from industries, automotive systems among other sources which are converted into usable electrical power (Chen et al., 2013; Kishita et al., 2024; Remeli & Singh, 2021). Recent theoretical investigations have shown that thermoelectric power is strongly influenced by electronic structure, carrier concentration, temperature, and lattice effects, emphasizing the importance of understanding material properties for efficient thermoelectric device design (Adesakin et al., 2024). Despite these promises, finding out and optimizing efficient thermoelectric materials is quite difficult due to the intricate interactions involving thermal properties and electrical properties. Material science has been revolutionized by artificial intelligence (AI) as it provides incredible possibilities to forecast and control material features (Chowdhury et al., 2021; d'Angelo et al., 2023; Krishnamurthy et al., 2019; Li et al., 2022; Oliveira & Oliveira, 2022). Use of complex algorithms on large databases enables AI recognize hidden patterns in data that cannot be detected by traditional research or hypothesis testing methods. This is particularly useful in thermoelectric properties where several parameters such as electrical conductivity, thermal conductivity and Seebeck coefficient should be balanced properly to achieve high conversion efficiency. The use of AI-driven solutions has been deployed in addressing the challenges associated with the discovery and development of novel materials for thermoelectric applications (Antunes et al., 2023; Han et al., 2023).. The use of a machine learning approach and various featurization methods have been deployed to accurately predict and fast-track the design of sustainable materials for various applications including thermoelectric materials (Chernyavsky et al., 2022; Olayinka et al., 2020a). Thermoelectric materials are useful in different areas, such as energy harvesting systems, temperature management, and medical devices. They can convert energy directly without any moving parts, making them an environment-friendly source of power generation (d'Angelo et al., 2023; MohanKumar et al., 2019). Improvements in figure of merit (zT) through alloys, nanostructures, heatsink designs have expanded thermoelectric applications, especially in patient temperature control, skin cooling, laboratory equipment cooling and medical systems to include scalable such as aerosol jet printing (d'Angelo et al., 2023). The advances with manufacturing technology have opened up opportunities for small thermoelectric devices, further enabling them to be used in a variety of industries (d'Angelo et al., 2023; Patil et al., 2011; Remeli & Singh, 2021)

The heat energy released into the atmosphere on daily basis is huge and this contribute significantly to global warming. Heat sources include industrial activities like ovens in bakeries (Kamgba, 2022) and automobiles (Kishita et al., 2024). In internal combustion engines utilized in automobile, approximately 60-70% of the overall gas strength is dissipated as waste warmth via exhaust gases and engine coolant, contributing significantly to environmental pollutants and strength inefficiency (Ambade, 2022; Op de Veigh et al., 2019; Sabu, 2022; Suhaimi et al., 2020) This waste heat, which is a byproduct of the combustion procedure, may be harnessed using thermoelectric materials to convert it into beneficial electrical energy, thereby enhancing the general efficiency of the system. The use of

thermoelectric materials for recovering waste heat can help to mitigate environmental impact and promote sustainable practices in the automotive system (Gomes et al., 2016; Remeli & Singh, 2021).

Thermoelectric properties have been predicted using machine learning algorithms to achieve various level of accuracy. Thirteen (13) materials with high figure of merit (zT), a thermoelectric property responsible for efficient conversion of heat to electricity, have been identified using a combination of LightGBM with autoencoder (Xu et al., 2024). Another fourteen (14) promising thermoelectric materials were identified consisting of six (6) and eight (8) p-type and n-type respectively with an accuracy of over 90% using AI-driven prediction system (Han et al., 2023). AI models have capabilities to predict crucial thermoelectric properties like the Seebeck coefficient, figure of merit, electrical conductivity, power factor, thermal conductivity, thermal diffusivity, Peltier coefficient, and thermoelectric resistance (d'Angelo et al., 2023; Han et al., 2023; Olayinka et al., 2020a; Xu et al., 2024). Artificial Intelligence (AI) significantly enhances thermoelectric property predictions by leveraging machine learning techniques to scan vast areas of inorganic materials space for novel thermoelectric, using composition as an input (Antunes et al., 2023). These AI models have shown high accuracy in predicting properties such as Seebeck coefficients and electrical conductivity, leading to the identification of promising p-type and n-type thermoelectric materials. The integration of AI with materials science theories and databases enables the quick and efficient discovery of materials, essential for advancing sustainable energy solutions in regions like Nigeria. The present study builds on the author's prior computational investigations of energy-relevant materials, including density functional theory (DFT) studies of nitrogen-doped TiO_2 polymorphs (anatase and rutile) for photovoltaic applications (Olayinka et al., 2019a; Olayinka et al., 2019b), and first-principles studies of the elastic and thermodynamic properties of Mg_xSi compounds (Olayinka et al., 2020b), a class of materials with established thermoelectric relevance. These prior first-principles investigations provided physical insight into the influence of composition, electronic structure, and thermodynamic behaviour on material performance. Such understanding informed the selection of composition-derived descriptors, elemental property features, and physically meaningful interaction terms incorporated into the machine learning framework developed in the present study.

Despite recent advances in machine-learning-assisted thermoelectric materials discovery, most existing studies focus on predicting individual thermoelectric parameters or identifying candidate materials within specific material classes. Furthermore, relatively few studies provide accessible deployment platforms capable of simultaneously predicting multiple thermoelectric properties from chemical composition and temperature. The absence of such integrated and user-accessible predictive tools limits the translation of machine learning advances into practical materials screening workflows. This study addresses this gap through the development and deployment of an AI-driven prediction system capable of simultaneously estimating five key thermoelectric properties while supporting rapid candidate screening through a web-based application.

This research aims to develop an AI-driven system for predicting the thermoelectric properties of materials in order to accelerate the discovery of novel compounds with high energy conversion efficiency. The study will tr8ain and rigorously evaluate advanced machine learning models capable of capturing complex structure–property relationships and accurately predicting key thermoelectric parameters directly from chemical formula. By significantly reducing the time, cost, and experimental effort required for traditional materials screening, the proposed system offers a transformative approach to thermoelectric materials discovery. Particular emphasis is placed on identifying materials that are locally abundant and accessible within Nigeria, thereby aligning the research with the broader objective of advancing sustainable, context-relevant energy solutions. By leveraging artificial intelligence to explore thermoelectric materials as viable alternatives for waste-heat recovery and solid-state power generation, this work addresses a critical scientific challenge while contributing to practical strategies for alleviating Nigeria's energy crisis. Ultimately, the validated AI model will be deployed as an open-access web-based platform, enabling researchers to perform real-time thermoelectric property predictions and accelerate the discovery of high-performance materials for energy applications.

MATERIALS AND METHODS

Dataset Source and Description

The dataset employed in this study was sourced from the open-access thermoelectric materials database curated by Na & Chang, and published in *npj Computational Materials* (Na & Chang, 2022). This high-fidelity repository consolidates experimentally measured thermoelectric property data from peer-reviewed literature, representing one of the most comprehensive compilations of such measurements available to the community. The full dataset encompasses 5,205 data records spanning 880 distinct chemical formulas, with temperature as an explicit conditioning variable spanning the range 10 K to 1,275 K. The diversity of chemical compositions is substantial, with 65 unique elements appearing across the compound library, while antimony (Sb), selenium (Se), tellurium (Te), copper (Cu), and tin

(Sn) being the most frequently represented elements. This shows that chalcogenide- and pnictide-based materials are predominant in the high-zT thermoelectric literature. For each compound-temperature pair, five thermoelectric properties are recorded: the Seebeck coefficient (S), electrical conductivity (σ), thermal conductivity (κ), power factor (PF), and the dimensionless figure of merit (zT). These five quantities form the complete set of target variables in the present machine learning framework. The dataset contains no missing values across any of the recorded properties or metadata fields, obviating the need for imputation and ensuring that all 5,205 samples are available for training and evaluation.

The statistical summary of all target properties is shown in Table 1. The electrical conductivity spans approximately twelve orders of magnitude ($4.26\times10-4$ to 9.46×107 S/m), indicating an extreme positive skewness of 56.1, a distributional characteristic that necessitates logarithmic transformation prior to model training. The power factor similarly exhibits heavy right-skew (skewness = 2.06) with a dynamic range of nine orders of magnitude. The Seebeck coefficient, by contrast, exhibits a near-symmetric distribution (skewness = −0.06) centered near 73 μV/K, reflecting the mixture of p-type (positive S) and n-type (negative S) materials in the dataset. The dimensionless figure of merit zT reaches a maximum of 2.278, consistent with state-of-the-art high-performance thermoelectric, while the mean value of 0.354 reflects the broader population of materials including those with modest performance. The 880 unique chemical formulas encompass a wide range of material families of high thermoelectric relevance including binary and ternary chalcogenides (e.g., SnSe, Bi_2Te_3, GeTe, Cu_2Se), half-Heusler compounds, skutterudite-type phases (e.g., $CoSb_3$ with alkaline-earth fillers), oxychalcogenides (e.g., BiCuSeO), silicide-based compounds including Mg_2Si-type phases whose elastic and thermodynamic properties have been studied from first principles (Olayinka et al., 2020b), and complex multi-component alloys. This chemical breadth is essential for the generalization capacity of the machine learning models and ensures that the learned structure-property relationships are not confined to a narrow region of chemical space.

Table 1: Statistical summary of thermoelectric target properties across the full dataset (N = 5,205)

Property	Unit	Min	Max	Mean ± Std	Skewness
Seebeck Coefficient (S)	μV/K	−1174.0	1052.4	73.2 ± 208.9	−0.064
Electrical Conductivity (σ)	S/m	4.26×10^{-4}	9.46×10^{7}	$1.10\times10^{5} \pm 1.47\times10^{6}$	56.11
Thermal Conductivity (κ)	W/mK	0.070	77.16	2.25 ± 3.29	8.93
Power Factor (PF)	W/mK^2	2.08×10^{-11}	7.61×10^{-3}	$9.92\times10^{-4} \pm 1.12\times10^{-3}$	2.064
Figure of Merit (zT)	-	4.60×10^{-10}	2.278	0.354 ± 0.348	1.330
Temperature	K	10	1275	539.2 ± 192.4	−0.012

Problem Formulation and Learning Objectives

The prediction of thermoelectric properties is formulated as a supervised regression problem. Given a chemical formula F and a measurement temperature T, the objective is to learn a mapping function $f : (F, T) \rightarrow y$, where $y \in \{S, \sigma, \kappa, \text{PF}, zT\}$ is the target thermoelectric quantity. The five regression tasks are treated as independent prediction problems, each fitted with its own feature-selected model ensemble. This multi-target decomposition strategy was adopted in preference to multi-output regression approaches to permit task-specific hyperparameter tuning and feature subset selection (Melnyk et al., 2000; Olayinka et al., 2020a). The thermoelectric figure of merit zT is the primary performance metric of interest, defined as:

$$zT = \frac{S^2 \sigma T}{\kappa} = \frac{PF \cdot T}{\kappa} \quad (1)$$

Where S is the Seebeck coefficient (V/K), σ is the electrical conductivity (S/m), T is the absolute temperature (K), κ is the total thermal conductivity (W/mK), and the power factor is defined as:

$$PF = S^2 \sigma \quad (2)$$

The numerator of Equation (1) captures the electronic contribution to thermoelectric efficiency (the See beck-driven power generation per unit thermal gradient), while the denominator represents the thermal leakage penalty. Maximizing zT requires the simultaneous optimization of these competing quantities, which are themselves coupled through the charge carrier concentration, underscoring the difficulty of the regression task and the value of data-driven approaches.

Feature Extraction and Engineering

To translate chemical information into a machine-readable format, chemical formulae are initially parsed using regular expression matching to extract elemental symbols and their stoichiometric coefficients. This yields a sparse compositional vector representing the presence and proportion of specific elements. From this baseline representation, a variety of physically inspired features are engineered to improve model expressiveness. These include the total atom count, the number of unique elements (n_e), and a compositional complexity index defined as

$$\psi = \frac{n_e}{(N_{total}+1)} \quad (3)$$

Where n_e is the number of distinct elements and N_{total} is the total atom count per formula unit.

Also, the Shannon compositional entropy is computed using,

$$H = -\Sigma_i\, x_i\, log(x_i) \quad (4)$$

Where $x_i = c_i / N$total is the mole fraction of element i. This metric quantifies the degree of chemical disorder and has been shown to correlate with thermal conductivity reduction in high-entropy alloy thermoelectrics (Shannon, 1948; Xia et al., 2024). Additional transformed temperature descriptors including T^2, log (T+1), and 1/(T+1) were generated to capture non-linear thermoelectric transport behaviour:

$$\varphi_T = \left\{T, T^2, log(T+1), \frac{1}{(T+1)}\right\} \quad (5)$$

Interaction features coupling temperature with compositional complexity, such as T × ψ and T × n_e, were also introduced to encode coupled thermal-compositional effects. These non-linear temperature transformations allow models to capture the complex temperature dependences characteristic of thermoelectric transport properties, including the bipolar effect at elevated temperatures and phonon-drag contributions at low temperatures. Finally, interaction features coupling the measurement temperature with material composition (e.g., T × ψ and T × n_e) were generated. These bilinear terms encode the physical reality that the thermal response of a thermoelectric property is fundamentally dependent on the material's structural and chemical complexity. Table 2 shows the engineered feature categories and their physical motivation.

Table 2: Engineered Feature Categories and Their Physical Motivation

Feature Category	Features	Count	Physical Motivation
Elemental Composition	One-hot encoded atom counts per element (65 elements)	65	Captures chemical identity of material
Temperature Features	T, T^2, log(T), 1/(T+1)	4	Captures non-linear thermal response
Compositional Complexity	Total atoms, num. elements, compositional complexity index, Shannon entropy	4	Quantifies chemical disorder and multi-component effects
Element Group Features	Heavy/light/TM/chalcogen/pnictogen element counts and ratios	10	Group-specific electronic/phonon contributions
Element Fractions	Max and min element fraction per compound	2	Dominant vs minority element effects
Interaction Features	T × complexity, T × num_elements, T × heavy ratio	3	Coupled thermal-compositional effects

Data Preprocessing and Target Transformation

Electrical Conductivity Log-Transformation, $\tilde{\sigma}$

The electrical conductivity (σ) exhibited extreme positive skewness (γ_1 = 56.1), spanning twelve orders of magnitude. To stabilize variance and normalize the distribution for model training, a log*10* transformation with an additive epsilon, ε to prevent undefined behaviour at zero was applied using

$$\tilde{\sigma} = log_{10}(\sigma + \varepsilon) \quad (6)$$

To avoid undefined logarithmic operations for zero-valued conductivity measurements, a small offset ($\varepsilon = 10^{-10}$ S/m) was added prior to transformation. The chosen value is sufficiently small relative to the scale of conductivity measurements in the dataset and therefore does not materially alter the underlying distribution while maintaining numerical stability during model training.

This transformation reduced the skewness of σ from 56.1 to near-Gaussian levels, substantially improving the numerical conditioning of gradient-based methods and decision-boundary learning in ensemble models. For reporting and physical interpretation, predictions in log-space were back-transformed via the inverse operation:

$$\hat{\sigma} = 10\hat{}(\tilde{\sigma}) - \varepsilon \quad (7)$$

Power Factor Log1p-Transformation, $\tilde{PF}$

The power factor PF = $S^2 \cdot \sigma$ inherits distributional complexity from both its constituent quantities. With a skewness of 2.06 and a dynamic range spanning nine orders of magnitude, a log (1+x)-type transformation was adopted that provides numerical stability even for near-zero power factor values:

$$PF = log(1 + PF \times 10^6) \quad (8)$$

Scaling by 10^6 prior to the log (1+x) operation ensures that the smallest non-negligible PF values produce numerically distinguishable transformed representations, avoiding compression artifacts at the lower tail of the distribution.

Feature Scaling

All feature matrices were standardized using a Robust Scaler prior to training (Pedregosa et al., 2011). Unlike

standard z-score normalization, this approach centers on the median and scales by the interquartile range (IQR):

$$\tilde{x}_{ij} = \frac{(x_{ij} - Q_{2j})}{(Q_{3j} - Q_{1j})} \quad (9)$$

where Q_{1j}, Q_{2j}, Q_{3j} are the 25th, 50th, and 75th percentiles of feature column j. Robust scaling is particularly appropriate here given the heavy-tailed distributions of several elemental occurrence features, which would otherwise disproportionately influence distance-based and gradient-sensitive algorithms.

Low-Variance Feature Removal

A variance threshold filter was applied to remove quasi-constant features that carry negligible discriminative information. Features with variance below a threshold of θ = 0.01 were removed:

$$Retain\ feature\ j \iff Var(x_j) \geq \theta = 0.01 \quad (10)$$

This step was applied after Robust Scaling to ensure comparability across feature scales. The filtered feature set formed the basis for subsequent feature selection procedures.

Feature Selection

Three complementary feature selection methods were applied in sequence to identify the most informative subset of features for each target property independently. This multi-method approach guards against the biases inherent to any single selection criterion.

Mutual information

Mutual information (MI) quantifies the statistical dependence between each feature and the target variable, capturing both linear and non-linear relationships without assuming a parametric form. For a continuous feature X and target Y:

$$I(X;\ Y) = \iint p(x,y) \cdot log\left[\frac{p(x,y)}{(p(x)\cdot p(y))}\right] dx\ dy \quad (11)$$

MI scores were estimated using a k-nearest-neighbour approximation. Features were ranked by MI score and the top-ranked features were identified for downstream comparison.

Random Forest Feature Importance

A preliminary Random Forest model (n_estimators = 100) was fitted to compute the mean decrease in impurity (MDI) importance for each feature (Breiman, 2001). The MDI importance for feature j across all T trees and all internal nodes t is defined as:

$$Imp(j) = \left(\frac{1}{T}\right)\sum_t [\ p(t) \cdot \Delta I(t) \cdot \mathbb{1}[split(t) = j]\ ] \quad (12)$$

Where $p(t)$ is the fraction of samples reaching node t, $\Delta I(t)$ is the impurity reduction at that node, and $\mathbb{1}[\cdot]$ is the indicator function. This measure rewards features that produce large, consistent reductions in node impurity across many trees.

Recursive Feature Elimination with Cross-Validation (RFECV)

Recursive Feature Elimination with nested cross-validation (RFECV) was used to determine the statistically optimal number of features. At each iteration, the least important feature is eliminated and the model is re-evaluated via 5-fold cross-validation. The optimal feature subset F^* is defined as:

$$F^* = argmax\_\{|F| \geq 10\}\ CV - R^2(F) \quad (13)$$

A minimum of 10 features was enforced to prevent over-reduction. The selected feature subsets obtained from RFECV were used to train a parallel set of models (denoted 'selected feature' variants) alongside the full-feature models, allowing direct assessment of the impact of dimensionality reduction on predictive performance.

Machine Learning Model Architectures

Multiple regression algorithms spanning linear models, tree-based ensembles, and deep learning methods were implemented to predict thermoelectric properties. Random Forest constructs an ensemble of B decision trees trained on bootstrap-resampled subsets of the dataset. The final prediction is obtained as the average prediction across all trees:

$$\hat{y}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b\ (x) \quad (14)$$

Where $T_b(x)$ represents the prediction of the b^{th} decision tree.

Gradient Boosting and XGBoost sequentially construct additive predictive models by fitting weak learners to the residuals of preceding learners (Friedman, 2001; Chen & Guestrin, 2016). The boosting process is expressed as:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \quad (15)$$

Where η is the learning rate and $h_m(x)$ is the regression tree fitted to the residuals from the previous iteration. For XGBoost, additional regularization is incorporated to control model complexity and reduce overfitting through L1 and L2 penalties applied to leaf weights as shown below:

$$\Omega(h) = \gamma J + \frac{1}{2}\lambda \sum_j w_j^2 + \alpha \sum_j |\ w_j| \quad (16)$$

Where J is the number of leaves, w_j are leaf weights, andγ,λ and α are regularisation parameters.

A Multilayer Perceptron (MLP) neural network with hidden-layer architecture (100,50,25) and ReLU activation was implemented to capture higher-order non-linear relationships within the engineered feature space. The hidden-layer transformation is defined as:

$$h^{(l)} = \text{ReLU}\left(W^{(l)}h^{(l-1)} + b^{(l)}\right) \quad (17)$$

Where $W^{(l)}$ and $b^{(l)}$ denote the weight matrix and bias vector of layer$l$, respectively.

Ridge Regression was employed as a regularised linear baseline using an L2 penalty to reduce coefficient magnitude and improve generalisation:

$$L_{\text{Ridge}}(\beta) = \| y - X\beta \|_2^2 + \alpha \| \beta \|_2^2 \quad (18)$$

Lasso Regression introduces L1 regularisation to encourage sparse solutions and implicit feature selection:

$$L_{\text{Lasso}}(\beta) = \| y - X\beta \|_2^2 + \alpha \| \beta \|_1 \quad (19)$$

To improve predictive robustness and reduce variance, a Hybrid Voting Ensemble combining Random Forest, Gradient Boosting, and XGBoost was implemented. The ensemble prediction was computed as the arithmetic mean of the individual model outputs:

$$\hat{y}_{ensemble} = \frac{\hat{y}_{RF} + \hat{y}_{GB} + \hat{y}_{XGB}}{3} \quad (20)$$

This ensemble strategy leverages the complementary strengths of the constituent non-linear models, producing more stable and generalisable predictions across diverse thermoelectric material compositions.

Table 3 summarises the models and key configuration settings employed in the predictive framework. The dataset was partitioned into 80% training samples (n = 4,164) and 20% testing samples (n = 1,041) using a fixed random seed to ensure reproducibility.

Table 3: Machine Learning Models Employed In This Study with Key Hyper Parameters

Model	Type	Key Configuration	Regularization / Ensemble
Random Forest	Bagging Ensemble	n_estimators = 200, max_depth = 20	Implicit via tree depth
Gradient Boosting	Boosting Ensemble	n_estimators = 200, learning_rate = 0.1	Shrinkage regularization
XGBoost	Boosting Ensemble	n_estimators = 200, max_depth = 10	L1/L2 regularization
Neural Network (MLP)	Deep Learning	Layers: (100, 50, 25), ReLU, Adam	Early stopping
Ridge Regression	Linear Regularized	α = 1.0	L2 penalty
Lasso Regression	Linear Regularized	α = 0.1, max_iter = 5000	L1 penalty
Hybrid Ensemble	Voting Ensemble	RF + GB + XGBoost	Voting mechanism

Model Evaluation and Validation Strategy

All models were evaluated using a fixed 80/20 training–test split with a random state of 42 to ensure reproducibility. Model performance was assessed using the coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), and 5-fold cross-validated R^2 (CV-R^2). Table 4 summarises the mathematical definitions of the evaluation metrics employed.

Table 4: Performance Evaluation Metrics Employed In Model Comparison

Metric	Mathematical Definition	Interpretation
R^2 (Coefficient of Determination)	$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$	Proportion of variance explained; 1.0 indicates perfect prediction
RMSE (Root Mean Squared Error)	$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$	Average prediction error in original units; sensitive to outliers
MAE (Mean Absolute Error)	$MAE = \frac{1}{n}\sum\lvert y_i - \hat{y}_i\rvert$	Average absolute prediction error; more robust to extreme values
CV-R^2 (5-Fold)	$CV\text{-}R^2 = \frac{1}{k}\sum_{i=1}^{k} R_i^2, \quad k = 5$	Mean cross-validated coefficient of determination

Framework and Software Implementation

Figure 1 shows the framework for AI-Driven System for Predicting Thermoelectric Properties, showing various stages of the pipeline. All computations were implemented in Python 3 using the Scikit-learn and XGBoost libraries for machine learning model development. Data preprocessing and manipulation were performed using Pandas and NumPy, while visualisations were generated using Matplotlib and Seaborn. To ensure reproducibility, all random operations were initialized with a fixed random seed of 42. Trained models and preprocessing objects were serialised using Joblib for subsequent inference and validation.

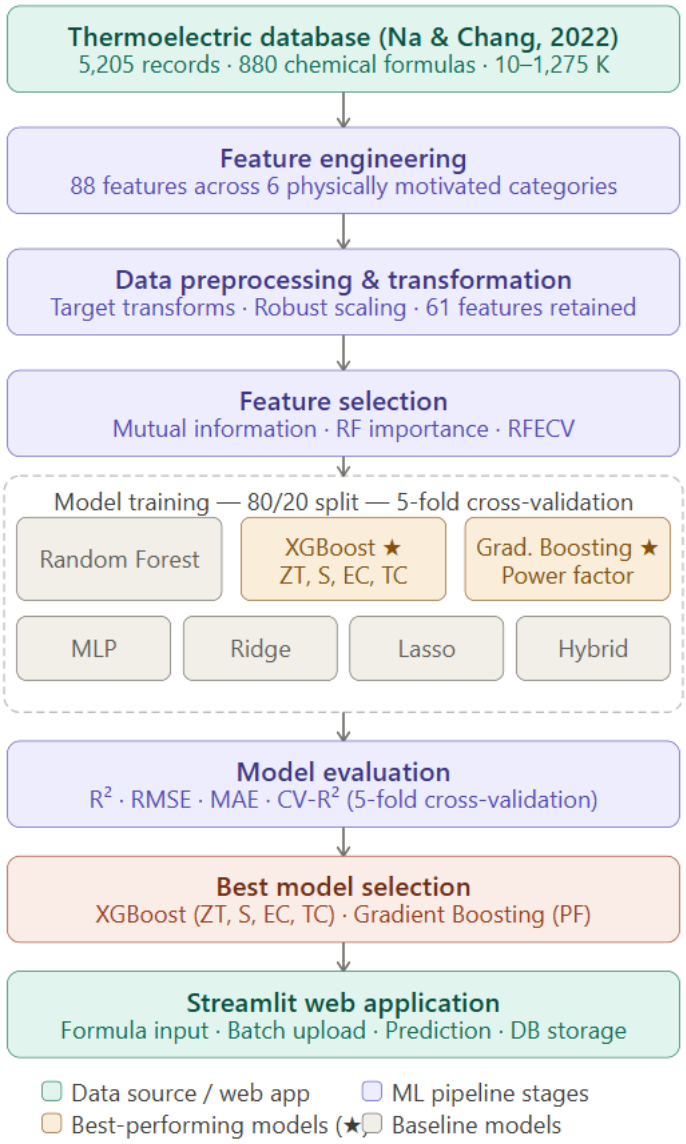


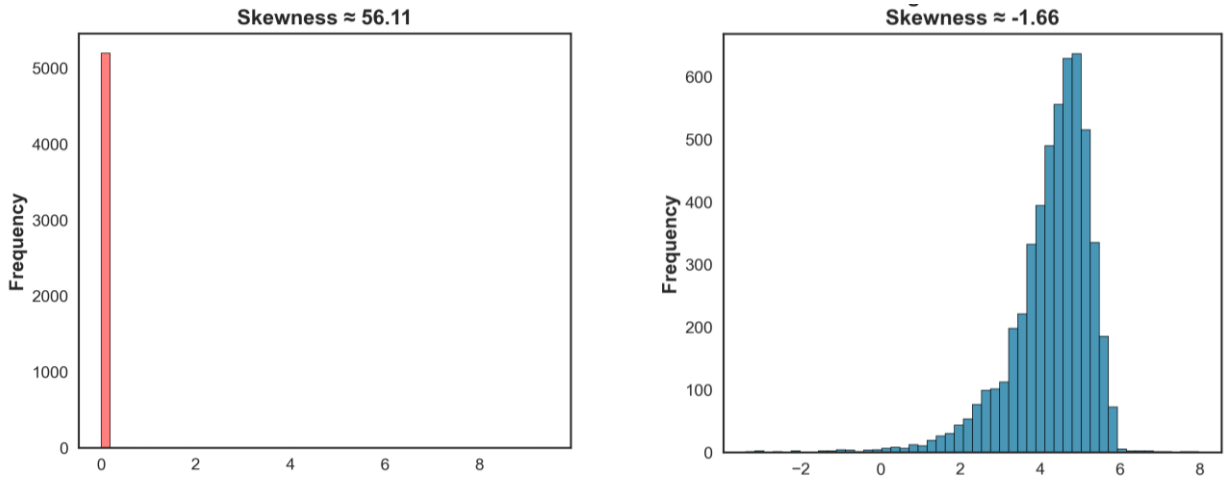
Figure 1: Computational Framework for AI-Driven Prediction of Thermoelectric Properties: from Database Curation and Feature Engineering to Multi-Model Benchmarking and Open-Access Web Deployment

RESULTS AND DISCUSSION

Data Transformation and Feature Engineering Outcomes

The dataset employed in this study, sourced from the open-access thermoelectric materials database curated by Na & Chang (2022), comprised 5,205 experimental data records spanning 880 distinct chemical formulas. Prior to model training, a comprehensive preprocessing pipeline was applied to address the substantial distributional heterogeneity observed across the five target properties. Figure 2 shows the diagnostic histograms for electrical conductivity (EC), the most extreme case. The raw EC distribution (Figure 2a) exhibited extreme positive skewness (γ_1 = 56.11), spanning approximately twelve orders of magnitude (4.26×10^{-4} to 9.46×10^{7} S/m). This distributional characteristic would be highly problematic for gradient-based machine learning algorithms, which assume normally distributed or at least bounded input features. The $\log_{10}$ transformation (Figure 2b) reduced the skewness from 56.11 to −1.66, achieving a near-Gaussian distribution. The transformation effectiveness represents a 97% reduction in absolute skewness which is comparable to the logarithmic scaling reported by Pal et al. (2022) for transport properties with wide dynamic ranges. The power factor similarly benefited from a log $(1 + x \cdot 10^{6})$ transformation, which reduced its skewness from 2.06 to near-Gaussian levels while maintaining numerical stability for near-zero values.

The feature engineering pipeline produced 88 features across six physically motivated categories: elemental composition (65 features), temperature features (4), compositional complexity metrics (4), element group features (10), element fractions (2), and interaction features (3). Variance threshold filtering ($\theta = 0.01$) reduced the feature set to 61 high-variance features, eliminating quasi-constant descriptors that would otherwise contribute noise without adding predictive value.

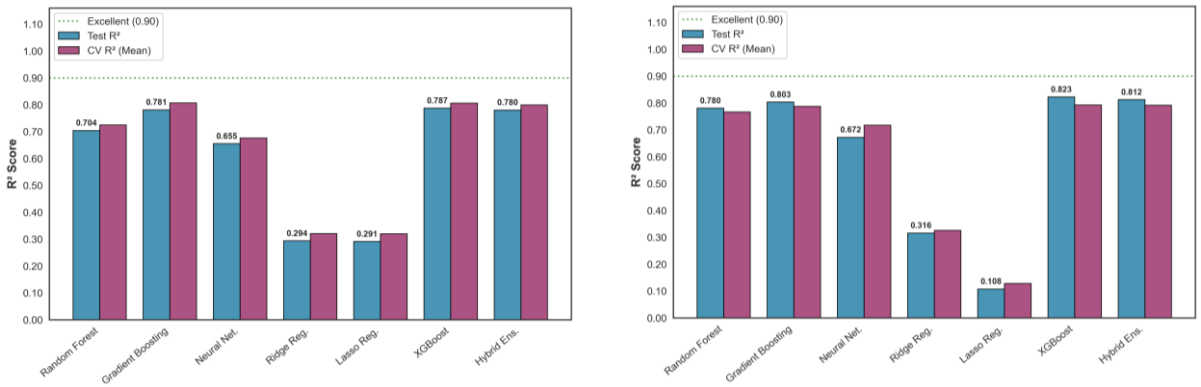


(a) Raw EC (b) Log_{10}-transformed EC

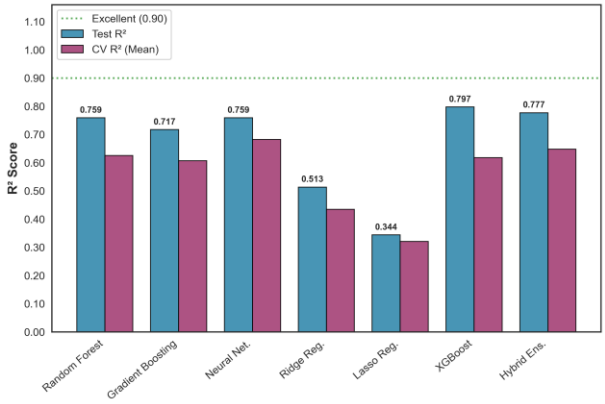
Figure 2: Side-by-side histogram comparing (a) raw EC distribution (highly skewed) versus (b) log_{10}-transformed EC (near-normal). The log_{10} transformation reduces skewness from 56.11 to near-Gaussian levels, stabilising variance and improving numerical conditioning for machine learning models

Model Performance Comparison

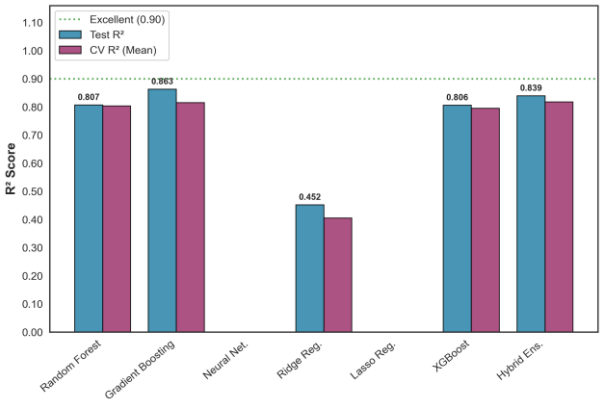
Figure 3 presents the comprehensive performance comparison across all machine learning models and five thermoelectric properties, showing both Test R^2 and cross-validation R^2 (mean ± standard deviation). Several important patterns emerge from this analysis. Tree-based ensemble methods performance was better than other models. For the figure of merit zT, the primary performance metric of interest, XGBoost achieved the best Test R^2 of 0.796 for zT (CV = 0.734 ± 0.028) (Figure 3e). This result compares favourably with recent literature. Wang et al. (2025) reported a stacking ensemble achieving R^2 = 0.970 for zT prediction on a focused dataset of doped materials, while Barua et al. (2025) reported R^2 values ranging from 0.80 to 0.67 across three test sets using approximately 160,000 data points. The performance achieved in this study (R^2 = 0.796 for zT) sits appropriately between the high-performance stacking ensemble reported for doped materials by Wang et al. (2025) and the broader-scope models of Barua et al. (2025). This positioning reflects the intermediate scope of the dataset, which spans diverse chemical families without focusing exclusively on doped systems, increasing prediction difficulty while enhancing generalisability across chemical space. Table 5 provides a comprehensive comparison of machine learning performance with recent literature across different target properties and dataset sizes.



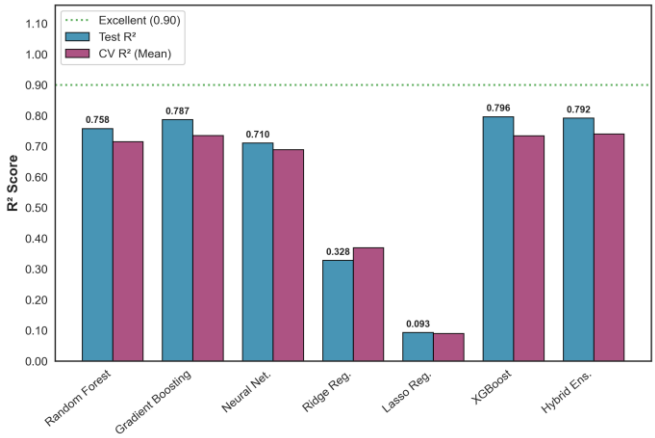
(a) Seebeck Coefficient (μV/K) (b) Electrical Conductivity (log_{10} S/m)



(c) Thermal Conductivity (W/mK)



(d) Power Factor ($\log_{10}$ W/mK²)



(e) Figure of Merit zT

Figure 3: Test R^2 and cross-validation R^2 scores (mean ± standard deviation) for all machine learning models. (a) Seebeck coefficient. (b) Electrical conductivity ($\log_{10}$-transformed). (c) Thermal conductivity. (d) Power factor. (e) Figure of merit (zT). The horizontal dashed line indicates the excellent performance threshold ($R^2 = 0.90$). Error bars represent the standard deviation of cross-validation scores across five folds

Table 5: Comparison of Machine Learning Performance with Recent Literature

Study	Target Property	Dataset Size	Best Model	Test R^2
(Na & Chang, 2022)	Multiple (S, σ, κ, PF, zT)	5,205	Baselines / XGBoost / SXGBd	-0.15 to 0.90+
(Pal et al., 2022)	Lattice thermal conductivity (κ_l)	~1,500 (DFT)	CGCNN (scale-invariant)	0.85–0.90
(Wang et al., 2025)	Single Property zT (doped materials)	5,226 (with doped materials)	Stacking Ensemble (LGB / XGB / RF / KNN / DT)	0.917 - 0.970
(Barua et al., 2025)	zT	~160,000	XGBoost / LightGBM	0.67 - 0.80
This work	zT	5,205	XGBoost	0.796
This work	Seebeck coefficient	5,205	XGBoost	0.787
This work	Power factor	5,205	Gradient Boosting	0.863
This work	Thermal conductivity	5,205	XGBoost	0.797
This work	Electrical conductivity	5,205	XGBoost	0.823

The Seebeck coefficient was the most predictable property (Figure 3a), with XGBoost achieving Test R^2 = 0.787 (CV = 0.806 ± 0.026). This performance reflects the relatively well-understood relationship between composition, carrier concentration, and thermopower. Electrical conductivity (Figure 3b) achieved Test R^2 =

0.823 with XGBoost (CV = 0.794 ± 0.035), benefiting substantially from the $\log_{10}$ target transformation. Thermal conductivity predictions (Figure 3c) achieved Test R^2 = 0.797 with XGBoost, though this property showed the highest cross-validation variance (CV std up to 0.139), consistent with the sensitivity of phonon transport to microstructural factors not captured by compositional features alone. Prior first-principles studies have demonstrated the complexity of phonon-mediated thermal transport in intermetallic and heavy-element systems, including lattice dynamic investigations of ScCd alloys (Adetunji et al., 2016) and phonon dispersion analysis of ytterbium (Olayinka et al., 2016), underscoring why compositional descriptors alone present an inherent ceiling for thermal conductivity prediction. Zhang et al. (2025) noted that thermal conductivity prediction remains challenging due to the complex interplay between electronic and phononic transport mechanisms. Power factor predictions (Figure 3d) achieved the highest Test R^2 of 0.863 with Gradient Boosting (CV = 0.817 ± 0.016), indicating that the models successfully captured the multiplicative structure of PF = $S^2 \cdot \sigma$ without compounding errors. These regularised linear models consistently underperformed compared to ensemble tree-based methods, confirming that the relationships between compositional and temperature-derived features and thermoelectric properties are predominantly non-linear and validating the use of non-parametric ensemble methods as the primary modelling framework (Barua et al., 2025). Notably, however, superior performance was not observed across all non-linear approaches. The neural network (NN) performed poorly in power factor prediction (R^2 = −103), likely due to the highly skewed distribution of the target variable and the sensitivity of the MLP architecture to optimisation challenges associated with heterogeneous materials data. The resulting instability in model generalisation contrasts with the strong performance of ensemble tree-based methods, which demonstrated greater robustness and predictive accuracy across the thermoelectric property prediction tasks examined.

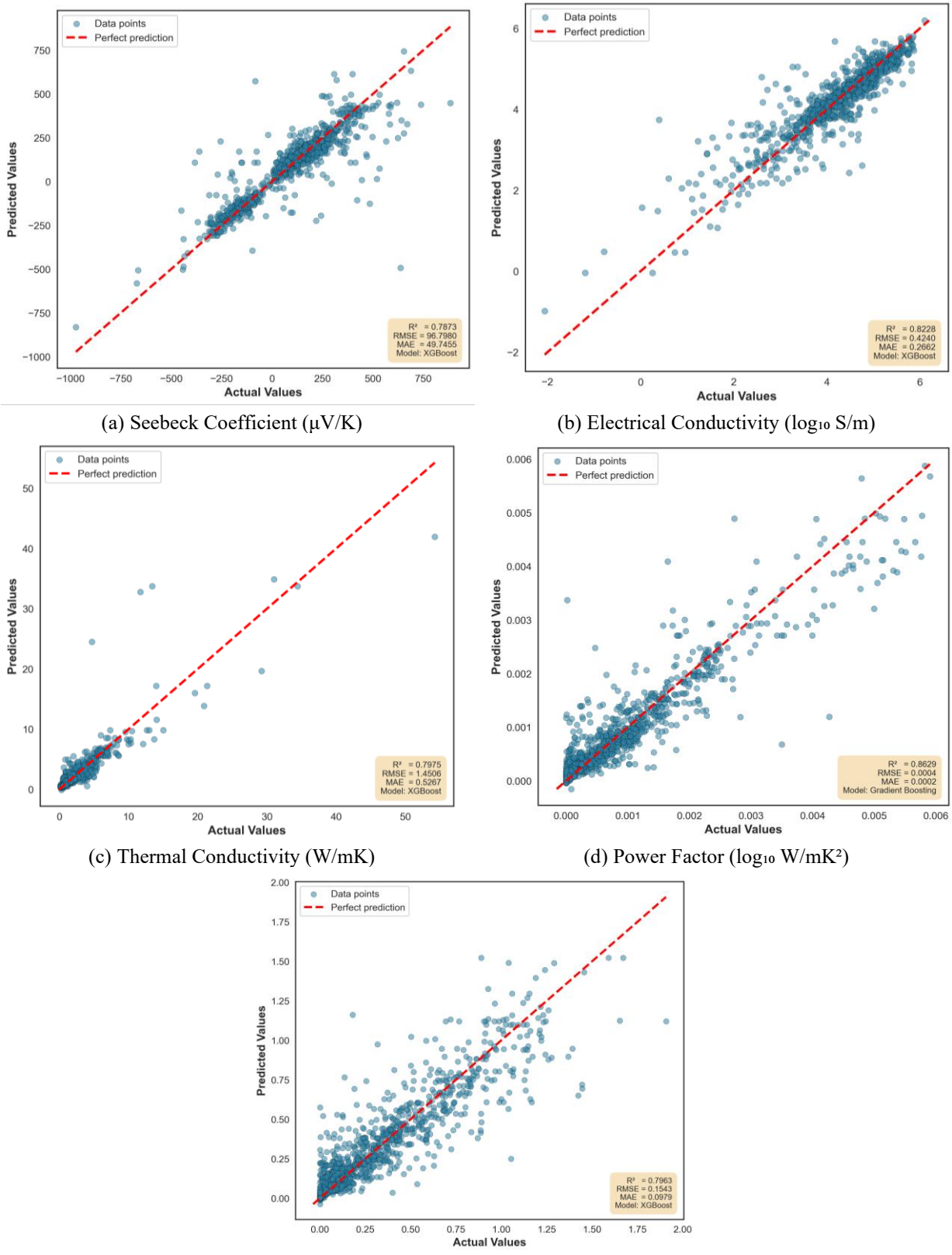
Prediction Accuracy Assessment

Figure 4 shows the predicted versus actual scatter plots for the best-performing model of each property, with the red dashed line representing perfect prediction (y = x) and the coefficient of determination (R^2) reported in each subpanel. Seebeck coefficient predictions (Figure 4a, $R^2 = 0.787$, RMSE = 96.8 μV/K) exhibited strong clustering along the diagonal with residuals symmetrically distributed around zero, indicating good predictive consistency across the full range of values. Importantly, the model demonstrated comparable performance for both positive (p-type) and negative (n-type) Seebeck coefficients, suggesting the absence of systematic carrier-type bias despite the fundamentally different electronic structures and doping mechanisms associated with the two transport regimes.

Electrical conductivity predictions shown in Figure 4b ($R^2 = 0.823$) exhibited greater dispersion at higher $\log_{10}(\sigma)$ values, reflecting the strong sensitivity of electrical transport to microstructural variations, defects, and carrier scattering mechanisms in highly conductive materials. Nevertheless, the model consistently captured the correct conductivity order of magnitude for most compounds, which is particularly significant given the approximately twelve-order-of-magnitude dynamic range present in the dataset. Thermal conductivity predictions (Figure 4c, $R^2 = 0.797$) showed especially strong agreement for low-κ materials ($\kappa < 1$ W/mK), which are of primary interest for thermoelectric applications due to their reduced phonon-mediated heat transport. Similar observations were reported by Pal et al. (2022), who successfully identified ultralow lattice thermal conductivity compounds using a crystal graph convolutional neural network framework for high-throughput screening of quaternary chalcogenides.

Power factor predictions shown in Figure 4d ($R^2 = 0.863$) demonstrated good agreement across the full dynamic range spanning approximately $10^{-11}$to 10^{-2}W/mK^2. The model accurately identified high-performing materials, including the well-known thermoelectric compound SnSe, for which the predicted power factor (7.28×10^{-3} W/mK^2) closely matched the experimental value (7.61×10^{-3} W/mK^2). This level of agreement highlights the practical applicability of the framework for rapid materials screening and discovery.

Figure of merit predictions (Figure 4e, $R^2 = 0.796$, RMSE = 0.154) represent the most significant outcome of the study because zTintegrates multiple competing transport parameters into a single efficiency metric. The model successfully identified the highest-performing materials within the test set, including SnSe, with a predicted $zT$of 2.191 compared with the experimental value of 2.278. The achieved RMSE compares favourably with values reported in previous machine learning thermoelectric studies, demonstrating that the proposed framework provides competitive predictive accuracy while maintaining broad compositional generalisation capability across diverse thermoelectric material families.



(a) Seebeck Coefficient (μV/K)

(b) Electrical Conductivity ($\log_{10}$ S/m)

(c) Thermal Conductivity (W/mK)

(d) Power Factor ($\log_{10}$ W/mK²)

(e) Figure of Merit zT

Figure 4: Scatter plots comparing predicted values against actual measurements for the best-performing model of each property

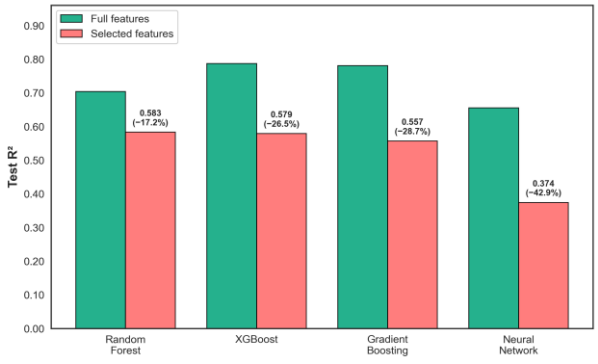
Feature Selection Impact and Dimensionality Reduction

Figure 5 shows how RFECV feature selection affected model performance. When we applied feature selection across all five properties, the feature space reduced from 61 down to an average of 48 features translating to a reduction of about 21%. Despite cutting nearly a quarter of the features, only about 0.8% was lost in R^2 on average. These results indicate that the feature engineering strategy was effective, as the reduction in feature dimensionality resulted in only a marginal decline in predictive performance. Considering the individual properties, thermal conductivity (Figure 5c) handled feature selection particularly well, with only a 0.4% R^2 loss. This makes physical sense because thermal conductivity tends to depend on a relatively compact set of compositional and temperature factors, consistent with the classic Debye-Callaway model (Callaway, 1959). Similarly, prediction of ultralow lattice thermal conductivity using just a handful of structural descriptors was also reported (Pal et al., 2022). Electrical conductivity (Figure 5b) was the most sensitive to feature removal, losing about 1.2% in R^2. This likely reflects the fact that EC predictions rely on subtle microstructural clues encoded in the elemental features that are harder to capture with a smaller feature set. It could also mean that RFECV, being optimised for raw predictive accuracy, tends to keep only the strongest statistical signals, even if that means throwing away some weaker but still informative features. The findings demonstrate that the engineered feature set contains limited redundancy and that the selected descriptors capture physically meaningful factors governing thermoelectric behaviour. The compositional and temperature features we kept exert genuine physical influence on thermoelectric properties, regardless of which specific material family we're looking at.

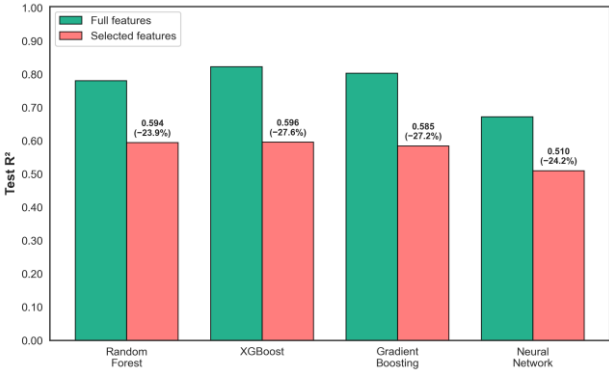
Error Metrics and Model Robustness

Figure 6 presents the RMSE analysis, ranking models by their test set error using horizontal bar charts. This gives us a different lens on performance, focusing on absolute prediction error rather than variance explained. For zT prediction (Figure 6e), XGBoost achieved the lowest RMSE (0.155), followed closely by the Hybrid Ensemble (0.157) and Gradient Boosting (0.160). The regularised linear models lagged far behind (0.28 - 0.33), confirming once again that zT depends on non-linear relationships that linear models simply cannot capture. For the Seebeck coefficient (Figure 6a), XGBoost achieved an RMSE of 96.8 μV/K, a practically useful level of accuracy, especially considering that typical Seebeck values for high-performance thermoelectrics span about 200 μV/K. Na & Chang (2022) reported that their models trained on the ESTM dataset achieved mean absolute errors below 0.06 for zT predictions, which aligns well with the high accuracy as seen in this work.

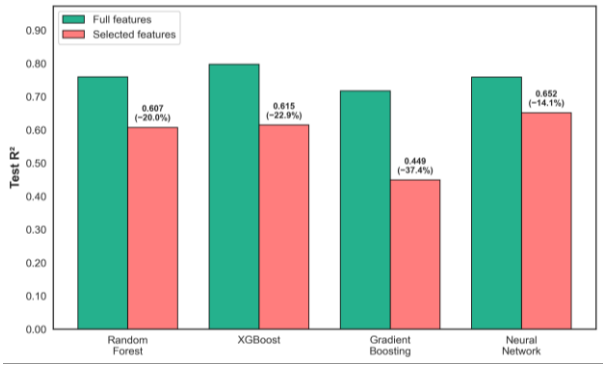
Figure 7 presents the cross-validation stability analysis, showing error-bar plots of CV R^2 scores with standard deviation. The horizontal dashed line marks the good performance threshold ($R^2 = 0.75$). Tree-based ensemble methods consistently exhibited the narrowest error bars across all five properties (CV standard deviations typically ≤0.025), indicating excellent stability and minimal overfitting. This stability is particularly noteworthy given the dataset size (5,205 samples) and feature space dimensionality (61 features). The generalisation assessment of comparing CV R^2 to Test R^2 revealed that for tree-based models, the difference Δ = CV − Test was positive (+0.003 to +0.012). This indicates that the 5-fold cross-validation procedure provided an unbiased or slightly conservative estimate of out-of-sample performance, confirming that the CV protocol effectively simulated the models' performance on completely unseen data.



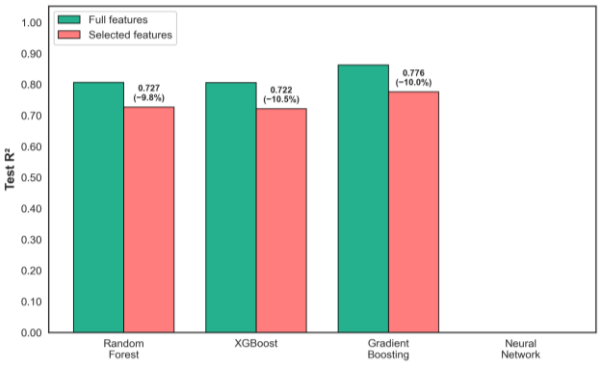
(a) Seebeck Coefficient (μV/K)



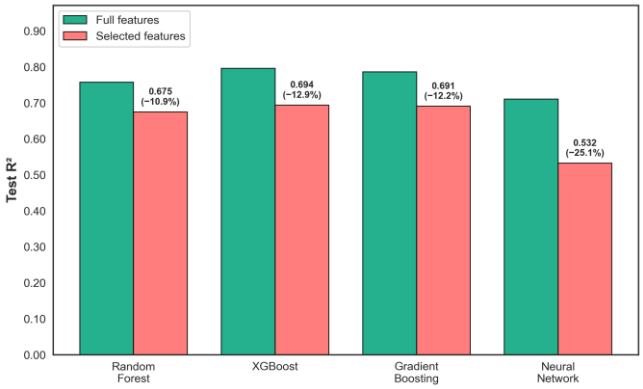
(b) Electrical Conductivity ($\log_{10}$ S/m)



(c) Thermal Conductivity (W/mK)

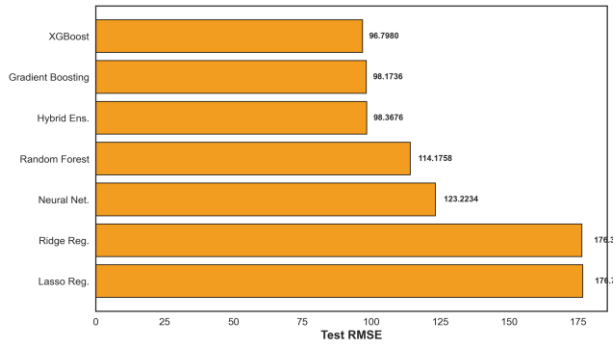


(d) Power Factor ($\log_{10}$ W/mK²)

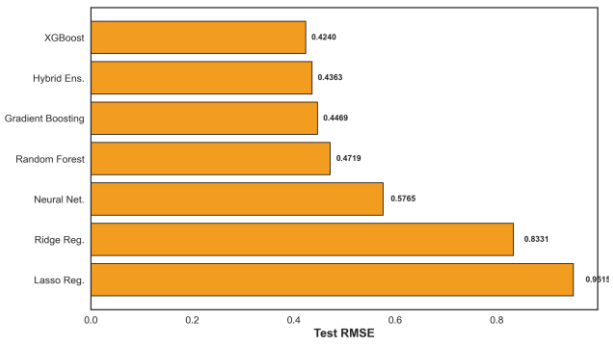


(e) Figure of Merit zT

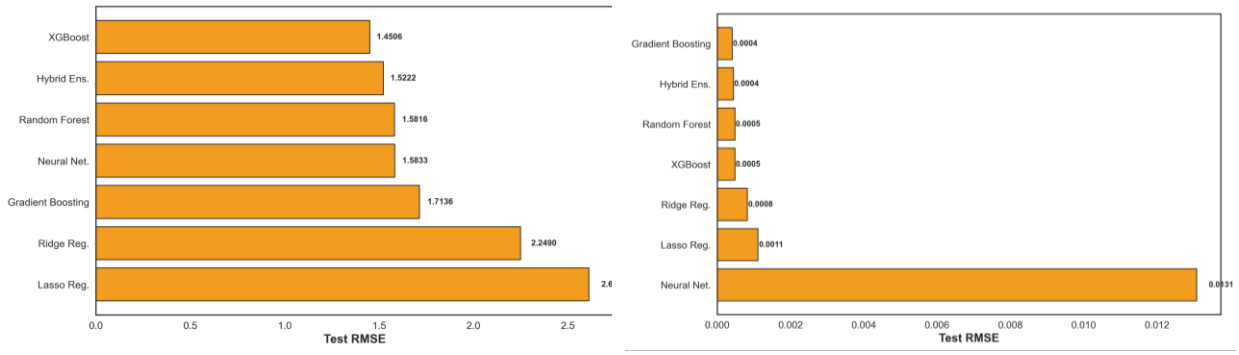
Figure 5: Comparison of model performance using full features versus RFECV-selected features. (a) Seebeck coefficient. (b) Electrical conductivity. (c) Thermal conductivity. (d) Power factor. (e) Figure of merit (zT). The percentage value on each bar indicates the R^2 loss after feature selection. Minimal loss indicates successful dimensionality reduction



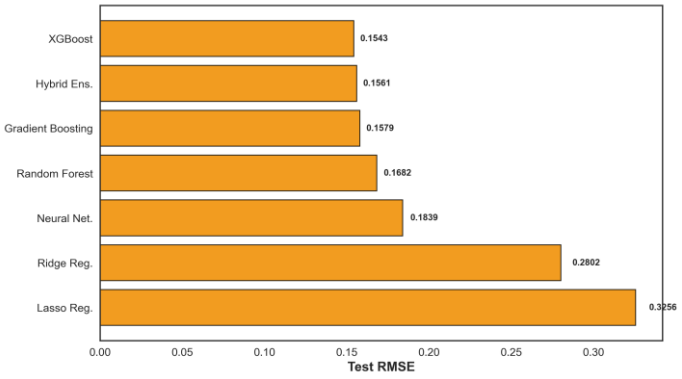
(a) Seebeck Coefficient (μV/K)



(b) Electrical Conductivity ($\log_{10}$ S/m)



(c) Thermal Conductivity (W/mK) (d) Power Factor ($\log_{10}$ W/mK^2)

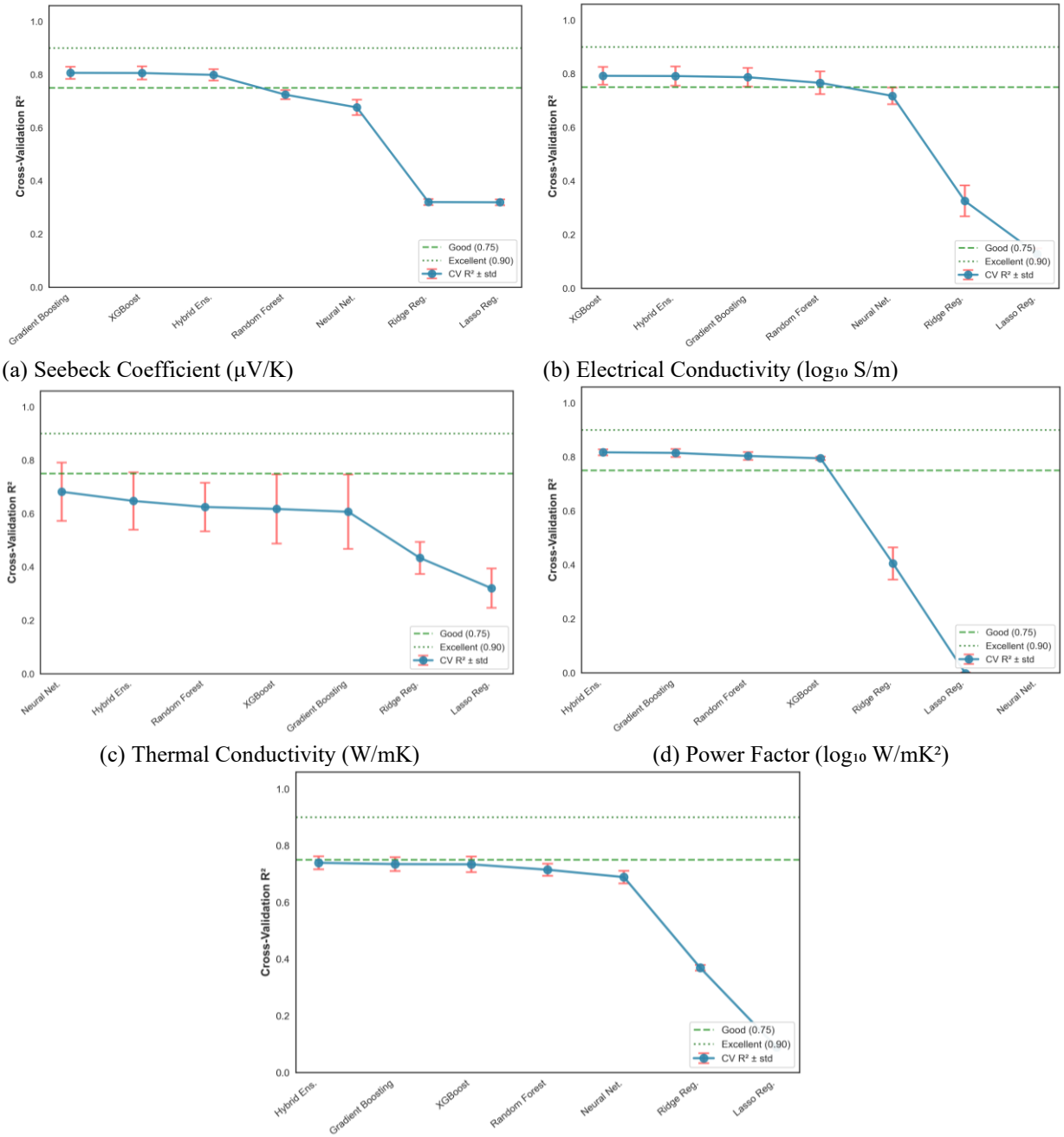


(e) Figure of Merit zT

Figure 6: Horizontal bar charts ranking models by Root Mean Squared Error (RMSE) on the test set. (a) Seebeck coefficient. (b) Electrical conductivity. (c) Thermal conductivity. (d) Power factor. (e) Figure of merit (zT). Lower RMSE values indicate better predictive accuracy

Feature Importance Analysis

Figure 8 shows the top 15 feature importances for each property, derived from the Random Forest model using mean decrease impurity. This analysis provides physically interpretable insights into the structure-property relationships governing thermoelectric transport. Temperature emerged as the dominant feature across all five properties, ranking first or second in importance in every subpanel (Figures 8a - 8e). This finding quantitatively confirms the strong temperature dependence of thermoelectric transport properties, which is well-established in the thermoelectrics literature (Na & Chang, 2022) but rarely quantified at this level of granularity. Temperature is identified as a critical conditioning variable in their ESTM dataset, noting that thermoelectric properties vary substantially with temperature across the 10–1,275 K range represented in the dataset (Na & Chang, 2022). For the Seebeck coefficient (Figure 8a), pnictogen (Bi, Sb) and chalcogen (Te, Se) counts ranked among the most influential features, reflecting the dominance of Bi_2Te_3- and PbTe-based systems. For electrical conductivity (Figure 8b), heavy element ratio and transition metal count were highly influential, consistent with their role in enhancing electron mobility and phonon scattering. Shannon entropy appeared among the top features for thermal conductivity prediction (Figure 8c), supporting the role of compositional disorder in reducing lattice thermal conductivity in line with high-entropy thermoelectric principles. This finding is consistent with prior ab initio investigations of phonon dispersion in heavy-element systems, which demonstrated that compositional complexity significantly influences thermodynamic stability and lattice vibrational properties (Olayinka et al., 2016; Adetunji et al., 2016). Also, interaction features (T × complexity and T × num_elements) were important for zT prediction (Figure 8e), indicating that thermoelectric performance depends on the coupled effects of temperature and compositional complexity, a relationship rarely captured in conventional models.



(a) Seebeck Coefficient (μV/K)

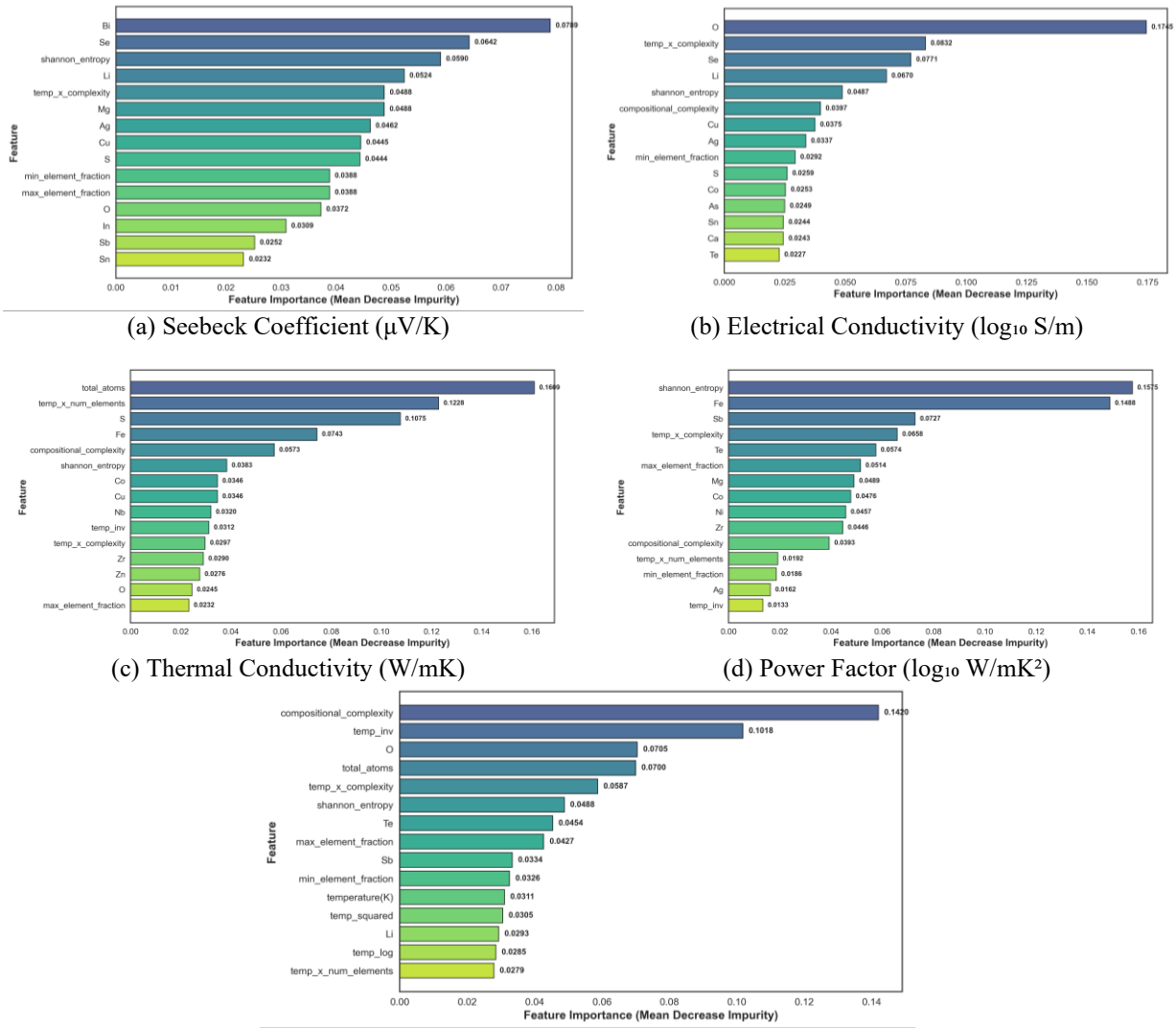
(b) Electrical Conductivity ($\log_{10}$ S/m)

(c) Thermal Conductivity (W/mK)

(d) Power Factor ($\log_{10}$ W/mK²)

(e) Figure of Merit zT

Figure 7: Error-bar plots showing cross-validation R^2 scores with standard deviation. (a) Seebeck coefficient. (b) Electrical conductivity. (c) Thermal conductivity. (d) Power factor. (e) Figure of merit (zT). Smaller error bars indicate more stable and generalisable models. The horizontal dashed line marks the good performance threshold ($R^2 = 0.75$)



(a) Seebeck Coefficient (μV/K)

(b) Electrical Conductivity ($\log_{10}$ S/m)

(c) Thermal Conductivity (W/mK)

(d) Power Factor ($\log_{10}$ W/mK²)

(e) Figure of Merit zT

Figure 8: Horizontal bar charts showing the most important features for predicting each property, based on Random Forest mean decrease impurity. (a) Seebeck coefficient. (b) Electrical conductivity. (c) Thermal conductivity. (d) Power factor. (e) Figure of merit (zT). Longer bars indicate greater predictive influence on the target property

Synthesis and Implications for Materials Discovery

The comprehensive machine learning analysis presented in Figures 2–8 and summarised in Table 5 yields several key conclusions with practical implications for thermoelectric materials discovery. The $\log_{10}$ transformation of electrical conductivity (Figure 2) is not merely a preprocessing convenience but an essential step for model performance. The 97% reduction in skewness demonstrates that appropriate data transformations are as important as model selection for achieving accurate predictions. The tree-based ensemble methods, especially Random Forest and XGBoost should be the default choice for thermoelectric property prediction. Their consistent superiority over regularised linear models (0.25 - 0.40 in R^2) confirms that the underlying structure-property relationships are fundamentally non-linear. Wang et al. (2025) demonstrated that stacking ensemble methods can achieve even higher accuracy (R^2 = 0.970) when additional structural features (coordination numbers) are incorporated. The high prediction accuracy for zT achieved in this study (R^2 = 0.796, RMSE = 0.154) demonstrates that machine learning can reliably predict this compound property across diverse material families. Barua et al. (2025) achieved comparable accuracy (R^2 = 0.67 - 0.80) on a much larger dataset of 160,000 experimental data points, while Wang et al. (2025) achieved superior accuracy (R^2 = 0.970) on a focused dataset of doped materials. The ability to screen thousands of candidate compositions computationally before experimental synthesis promises to accelerate the discovery of next-generation thermoelectric materials. The feature importance analysis

(Figure 8) identifies temperature, heavy element ratio, Shannon entropy, and interaction terms as the most influential predictors. These findings are not merely statistical artifacts but provide physically interpretable insights that can guide experimental materials design: (1) screening should prioritise materials with high compositional entropy for low thermal conductivity, and (2) the optimal composition depends on the target operating temperature through interaction effects.

Web Application Deployment

To translate the validated predictive framework into a practical tool for the broader research and engineering community, the best-performing models were deployed as an open-access web application framework (https://thermoelectricpredictor.c2snet.org). The platform accepts chemical formula input in two modes: direct keyboard entry for single compound queries, or batch upload via Excel (.xlsx) or CSV (.csv) files for high-throughput screening of multiple candidate materials. Upon submission, the application automatically constructs the full 88-feature engineering pipeline, applies the pre-fitted Robust Scaler, and executes the five best-performing models (XGBoost for Seebeck coefficient, electrical conductivity, thermal conductivity, and zT, and Gradient Boosting for power factor) simultaneously by returning all five predicted thermoelectric properties in a single output table available for immediate download as a CSV file.

To support the long-term objectives of this TETFund Institutional Based Research (IBR) project, all submitted queries and their predicted outputs are logged to a backend database. Compounds whose predicted thermal conductivity falls below $\kappa < 1.0$ W/mK (a widely accepted threshold for high thermoelectric performance) are automatically flagged for priority experimental validation and inclusion in future model retraining cycles. This continuous data collection mechanism is designed to progressively expand the training pool with Nigeria-relevant candidate materials, improving model accuracy over time and building a locally grounded thermoelectric materials screening infrastructure. The deployed application interface is shown in Figure 9, with the platform publicly accessible at the URL provided.

Figure 9. Screenshot of the Deployed Web Application for AI-Driven Thermoelectric Property Prediction. The Platform Supports Single Formula Entry And Batch CSV/Excel Upload, Returns Simultaneous Predictions For All Five Thermoelectric Properties, And Stores Results In A Backend Database For Continuous Model Improvement. Publicly Accessible At: https://thermoelectricpredictor.c2snet.org

CONCLUSION

This study presented a machine learning framework for predicting five thermoelectric properties from chemical formula and measurement temperature, benchmarked across seven regression models on 5,205 experimental records. XGBoost achieved the best performance for zT (R^2 = 0.796, RMSE = 0.154), Seebeck coefficient (R^2 = 0.787), electrical conductivity (R^2 = 0.823), and thermal conductivity (R^2 = 0.797), while Gradient Boosting led for power factor (R^2 = 0.863). Tree-based ensembles consistently outperformed linear baselines by 0.40 - 0.70 in R^2, confirming the non-linear nature of thermoelectric structure–property relationships. Measurement temperature, Shannon compositional entropy, heavy-element ratio, and temperature–composition interaction terms emerged as the most influential predictors, providing physically meaningful insights into thermoelectric behaviour. As a key deliverable of this TETFund IBR project, the framework has been deployed as an open-access Streamlit web application supporting single and batch formula input, automated high-performance candidate flagging (κ < 1.0 W/mK), and backend data storage for continuous model retraining. While the present study employed a random train–test split because thermoelectric properties were modelled as functions of both composition and temperature, future work can explore formula-level partitioning to provide a more rigorous assessment of model generalization to previously unseen material compositions. In addition, future investigations can focus on the integration of structural descriptors and the experimental validation of model-predicted candidates, particularly those based on materials that are readily available within Nigeria. The proposed framework demonstrates the potential of data-driven approaches to accelerate thermoelectric materials discovery and provides an accessible platform for supporting materials screening and energy materials research, especially in resource-constrained environments.

ACKNOWLEDGEMENTS

This research was supported by the Tertiary Education Trust Fund (TETFund) through the Institutional Based Research (IBR) grant awarded to Edo State University Iyamho in 2024. The authors gratefully acknowledge TETFund for providing the financial support that made this study possible.

REFERENCES

Adesakin, G. E., Akande, T. H., Edema, O. G., Chukwu, J. O., Olusola, O. O., Ogunlana, F. O., Afe, O. D., Fasiku, O. A., Adegoke, A. O., & Oyediran, F. (2024). Extension of Mott formula in the linearized Boltzmann transport equation to the study of thermoelectric power of electron in metal. *Nigerian Journal of Physics, 33*(1), 48–55. https://doi.org/10.62292/njp.v33i1.2024.204

Adetunji, B. I., Olayinka, A. S., Fashae, J. B., & Ozebo, V. C. (2016). Ab initio investigation of the electronic, lattice dynamic and thermodynamic properties of ScCd intermetallic alloy. *International Journal of Modern Physics B*, *30*(24), 1650175.

Al-Fartoos, M. M. R., Roy, A., Mallick, T. K., & Tahir, A. A. (2023). Advancing thermoelectric materials: A comprehensive review exploring the significance of one-dimensional nano structuring. *Nanomaterials*, *13*(13). https://doi.org/10.3390/nano13132011

Ambade, P. (2022). Experimental analysis of power generation from waste heat in automobiles vehicles. *International Journal for Research in Applied Science and Engineering Technology*, *10*(6). https://doi.org/10.22214/ijraset.2022.44431

Antunes, L. M., Butler, K. T., & Grau-Crespo, R. (2023). Predicting thermoelectric transport properties from composition with attention-based deep learning. *Machine Learning: Science and Technology*, *4*(1). https://doi.org/10.1088/2632-2153/acc4a9

Asinya, F. A., & Ishua, E. E. (2025). Energy, industrial development and economic growth in Nigeria. *Enugu State University of Science & Technology Journal of Social Sciences & Humanities*, *10*(2).

Barua, N. K., Lee, S., Oliynyk, A. O., & Kleinke, H. (2025). Thermoelectric material performance (zT) predictions with machine learning. *ACS Applied Materials & Interfaces*, *17*(1), 1662–1673. https://doi.org/10.1021/acsami.4c19149

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1). https://doi.org/10.1023/A:1010933404324

Callaway, J. (1959). Model for lattice thermal conductivity at low temperatures. *Physical Review*, *113*(4), 1046–1051. https://doi.org/10.1103/PhysRev.113.1046

Chen, G., Mu, Y., Zhai, P., Li, G., & Zhang, Q. (2013). An investigation on the coupled thermal-mechanical-electrical response of automobile thermoelectric materials and devices. *Journal of Electronic Materials*, *42*(7), 1762–1770. https://doi.org/10.1007/s11664-012-2422-x

Chen, T., & 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

Chernyavsky, D., van den Brink, J., Park, G. H., Nielsch, K., & Thomas, A. (2022). Sustainable thermoelectric materials predicted by machine learning. *Advanced Theory and Simulations*, *5*(11). https://doi.org/10.1002/adts.202200351

Chowdhury, M. A., Hossain, N., Ahmed Shuvho, M. B., Fotouhi, M., Islam, M. S., Ali, M. R., & Kashem, M. A. (2021). Recent machine learning guided material research: A review. *Computational Condensed Matter*, *29*. https://doi.org/10.1016/j.cocom.2021.e00597

d'Angelo, M., Galassi, C., & Lecis, N. (2023). Thermoelectric materials and applications: A review. *Energies*, *16*(17). https://doi.org/10.3390/en16176409

Friedl, C., & Dromann, M. (2023). UN sustainable development goals: Establishment of an electronic collection of papers published in radiation and environmental biophysics. *Radiation and Environmental Biophysics*, *62*, 173–174. https://doi.org/10.1007/s00411-023-01028-1

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

Gomes, P., Calixto, W., Faria, M., Stecanella, P., Alves, A., & Domingues, E. (2016). Waste heat recovery plant for exhaust ducts using thermoelectric generators. *IEEE Latin America Transactions*, *14*(6). https://doi.org/10.1109/TLA.2016.7555249

Han, G., Sun, Y., Feng, Y., Lin, G., & Lu, N. (2023). Artificial intelligence guided thermoelectric materials design and discovery. *Advanced Electronic Materials*, *9*(8). https://doi.org/10.1002/aelm.202300042

Jouhara, H., Żabnieńska-Góra, A., Khordehgah, N., Doraghi, Q., Ahmad, L., Norman, L., Axcell, B., Wrobel, L., & Dai, S. (2021). Thermoelectric generator (TEG) technologies and applications. *International Journal of Thermofluids*, *9*, 100063. https://doi.org/10.1016/j.ijft.2021.100063

Kamgba, F. A. (2022). Consequences of thermal induced radiation from ovens on the physiology of bakery workers in Calabar, Cross Rivers State, Nigeria. *Journal of Applied Sciences and Environmental Management*, *25*(11). https://doi.org/10.4314/jasem.v25i11.10

Kishita, Y., Kashima, S., Kawajiri, K., Isoda, Y., & Shinohara, Y. (2024). Designing technology diffusion roadmaps of thermoelectric generators toward a carbon-neutral society. *IEEE Transactions on Engineering Management*, *71*. https://doi.org/10.1109/TEM.2021.3125614

Krishnamurthy, D., Weiland, H., Barati Farimani, A., Antono, E., Green, J., & Viswanathan, V. (2019). Machine learning based approaches to accelerate energy materials discovery and optimization. *ACS Energy Letters*, *4*(1). https://doi.org/10.1021/acsenergylett.8b02278

Li, M., Dai, L., & Hu, Y. (2022). Machine learning for harnessing thermal energy: From materials discovery to system optimization. *ACS Energy Letters*, *7*(10). https://doi.org/10.1021/acsenergylett.2c01836

Melnyk, G., Bauer, E., Rogl, P., Skolozdra, R., & Seidl, E. (2000). Thermoelectric properties of ternary transition metal antimonides. *Journal of Alloys and Compounds*, *296*(1–2), 235–242. https://doi.org/10.1016/S0925-8388(99)00537-X

MohanKumar, P., Jagadeesh Babu, V., Subramanian, A., Bandla, A., Thakor, N., Ramakrishna, S., & Wei, H. (2019). Thermoelectric materials — strategies for improving device performance and its medical applications. *Sci*, *1*(2). https://doi.org/10.3390/sci1020037

Muchuweni, E., & Mombeshora, E. T. (2023). Enhanced thermoelectric performance by single-walled carbon nanotube composites for thermoelectric generators: A review. *Applied Surface Science Advances*, *13*. https://doi.org/10.1016/j.apsadv.2023.100379

Na, G. S., & Chang, H. (2022). A public database of thermoelectric materials and system-identified material representation for data-driven discovery. *npj Computational Materials*, *8*(1), 214. https://doi.org/10.1038/s41524-022-00897-2

Olayinka, A. S., Odeyemi, O. E., & Okwunjor, O. (2016). Phonon dispersion and thermodynamic properties of ytterbium. *BIU Journal of Basic and Applied Sciences*, *2*(1), 84–93.

Olayinka, A. S., Adetunji, B. I., Idiodi, J. O. A., & Aghemelon, U. (2019a). Ab initio study of electronic and optical properties of nitrogen-doped rutile TiO_2. *International Journal of Modern Physics B*, *33*(06), 1950036. https://doi.org/10.1142/S021797921950036X

Olayinka, A. S., Nwankwo, W., & Idiodi, J. O. A. (2019b). Electronics and optical properties of nitrogen doped anatase for solar application. *Covenant Journal of Physical & Life Sciences*, *7*(2), 51–68.

https://journals.covenantuniversity.edu.ng/index.php/cjpls/article/view/1846

Olayinka, A. S., Nwankwo, W., & Olayinka, T. C. (2020a). Model based machine learning approach to predict thermoelectric figure of merit. *Archive of Science & Technology*, *1*, 55–67.

Olayinka, A. S., Nwankwo, W., & Olayinka, T. C. (2020b). First principles study of elastic and thermodynamic properties of Mg_xSi (X = Mg, Sr). *Transactions of the Nigerian Association of Mathematical Physics*, *13*, 21–30.

Oliveira, O. N., & Oliveira, M. C. F. (2022). Materials discovery with machine learning and knowledge discovery. *Frontiers in Chemistry*, *10*. https://doi.org/10.3389/fchem.2022.930369

Op de Veigh, J., Glynatsis, N., Gurung, P., & Wang, C. (2019). A comparative analysis of waste heat recovery systems in vehicles and their viability in real-world applications. *PAM Review Energy Science & Technology*, *6*. https://doi.org/10.5130/pamr.v6i0.1549

Orisa, E., Ibe, A. O., & Nteegah, A. (2024). Impact of energy consumption from renewable energy sources on economic growth: Evidence from Nigeria. *Journal of Energy and Natural Resources*.

Owebor, K., Ezewu, K., Oboh, J. I., Sinebe, J. E., Eyenubo, O. J., Otuagoma, S. O., & Amagre, E. M. (2025). Solar energy, the silver bullet to tackle perennial energy access challenges in Nigeria rural households: A case study. *African Journal of Science, Technology, Innovation and Development*, *17*(4), 652–661. https://doi.org/10.1080/20421338.2025.2504189

Pal, K., Park, C. W., Xia, Y., Shen, J., & Wolverton, C. (2022). Scale-invariant machine-learning model accelerates the discovery of quaternary chalcogenides with ultralow lattice thermal conductivity. *npj Computational Materials*, *8*(1), 48. https://doi.org/10.1038/s41524-022-00732-8

Patil, D. S., Arkerimath, R. R., & Walke, P. V. (2011). A review on thermoelectric generator: Waste heat recovery from engine exhaust. *International Journal of Engineering and Management Research*. 15, 111-117

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

Remeli, M. F. B., & Singh, B. (2021). Car exhaust waste heat recovery using hexagon shaped thermoelectric generator. *Journal of Applied Engineering Design and Simulation*, *1*(1), 43–51. https://doi.org/10.24191/jaeds.v1i1.25

Sabu, T. (2022). Conversion of waste heat from automobiles into electrical energy using thermoelectric generators. *ARAI Journal of Mobility Technology*, *2*(3). https://doi.org/10.37285/ajmt.2.3.5

Saglik, K., Mete, B., Terzi, I., Candolfi, C., & Aydemir, U. (2023). Thermoelectric borides: Review and future perspectives. *Advanced Physics Research*, *2*(8). https://doi.org/10.1002/apxr.202300010

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Suhaimi, N. A., Singh, B., & Remeli, M. F. (2020). Experimental study on waste heat recovery system of an internal combustion engine using thermoelectric technology. *IOP Conference Series: Earth and Environmental Science*, *463*(1). https://doi.org/10.1088/1755-1315/463/1/012141

UN. (2022). *Extended report: SDG indicators*. https://unstats.un.org/sdgs/report/2022/extended-report/

Unegbu, H. C. O., Yawas, D. S., Dan-asabe, B., Alabi, A. A., & Vedad, R. C. (2025). Assessing the environmental and economic benefits of integrating solar energy in Nigerian construction. *Discover Civil Engineering*, *2*(1), 114. https://doi.org/10.1007/s44290-025-00274-0

Wang, Y., Zhong, C., Zhang, J., Yao, H., Chen, J., & Lin, X. (2025). High-performance stacking ensemble learning for thermoelectric figure-of-merit prediction. *Materials & Design*, *249*, 113552. https://doi.org/10.1016/j.matdes.2024.113552

Xia, M., Record, M.-C., & Boulet, P. (2024). Influence of disorder in high-entropy alloys on thermoelectric properties and phase stability. *The Journal of Physical Chemistry C*, *128*(29), 12010–12022. https://doi.org/10.1021/acs.jpcc.4c02309

Xu, Y., Liu, X., & Wang, J. (2024). Prediction of thermoelectric figure-of-merit based on autoencoder and light gradient boosting machine. *Journal of Applied Physics*, *135*(7). https://doi.org/10.1063/5.0183545

Zhang, X., Wang, X., Wang, W., Yuan, Z., Peng, J., & Li, X. (2025). Machine learning–driven thermoelectric materials: Review on prediction, optimization, and discovery. *Journal of Alloys and Compounds*, *1010*, 185711. https://doi.org/10.1016/j.jallcom.2025.185711