IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Multicollinearity Resolution Based on Machine Learning: A Case Study of Carbon Emissions

XUANMING ZHANG[1], (Student Member, IEEE)
[1]University of Wisconsin-Madison, Madison, WI 53706 USA (e-mail: xzhang2846@wisc.edu)
Corresponding author: Xuanming Zhang (e-mail: xzhang2846@wisc.edu).

ABSTRACT This study proposes an analytical framework that integrates DBSCAN clustering with the Elastic Net regression model to address multifactorial problems characterized by structural complexity and multicollinearity, exemplified by carbon emissions analysis. DBSCAN is employed for unsupervised learning to objectively cluster features, while the Elastic Net is utilized for high-dimensional feature selection and complexity control. The Elastic Net is specifically chosen for its ability to balance feature selection and regularization by combining L1 (lasso) and L2 (ridge) penalties, making it particularly suited for datasets with correlated predictors. Applying this framework to energy consumption data from 46 industries in China (2000-2019) resulted in the identification of 16 categories. Emission characteristics and drivers were quantitatively assessed for each category, demonstrating the framework's capacity to identify primary emission sources and provide actionable insights. This research underscores the global applicability of the framework for analyzing complex regional challenges, such as carbon emissions, and highlights qualitative features that humans find meaningful may not be accurate for the model.

INDEX TERMS Machine Learning, Quantitative Analysis, Sustainable Development Goals

I. INTRODUCTION

GLOBAL efforts to mitigate climate change rely on accurate predictions of carbon emissions and energy consumption. In industrial sectors, emissions are driven by a complex interplay of factors—energy sources, economic output, technological efficiency, etc.—that pose challenges for traditional modeling approaches. Established analytic tools like the STIRPAT framework [3] and index decomposition methods [4] have been used to unpack driving forces of emissions, but they often struggle to capture nonlinear interdependencies and high-dimensional feature interactions in modern data. Recent studies have turned to machine learning and deep learning to improve prediction accuracy. For instance, CNN-LSTM neural network was applied to predict China's energy mix [13], and hybrid LSTM-SVR model was employed for regional carbon emission forecasting [14]. While such models can learn complex temporal patterns, purely neural approaches may suffer from interpretability issues and performance limitations when input features are highly correlated or homogeneous [5]. Indeed, prior work has observed that deep networks have difficulty distinguishing collinear features, leading to suboptimal handling of multicollinearity.

In this context, we identify two key challenges for quantitative energy/emission prediction: (i) managing multifactor data with severe multicollinearity (e.g. industries where multiple energy inputs co-vary), and (ii) extracting actionable insights (e.g. which industry groupings or energy sources drive emissions) rather than treating the model as a black box. To address these challenges, we propose a novel DBSCAN-Penalized Regression (DPR) framework. This framework is distinguished by its hierarchical approach: an unsupervised clustering phase to objectively categorize the data, followed by a supervised regression phase with regularization to perform feature selection and prediction. By first grouping structurally similar data points, we reduce the complexity the regression model must explain and directly confront multicollinearity by separating highly homogeneous observations into clusters. In the subsequent regression, we employ penalized linear models (ridge, lasso, and elastic net) to identify dominant influencing factors while controlling model complexity. This combined approach leverages the strengths of both unsupervised and supervised learning: DBSCAN discovers latent structures in the dataset without prior assumptions on cluster shape, and Elastic Net regression balances bias–variance trade-offs to yield a parsimonious, interpretable model of carbon emissions.

In summary, this work offers the following contributions:

Novel Framework (DPR): We introduce a new analytical framework that combines density-based clustering with penalized regression for energy and emission prediction. Unlike prior studies that apply deep neural networks or conventional

regression in isolation, our approach exploits clustering to handle feature heterogeneity and multicollinearity before model fitting. This is, to our knowledge, the first application of DBSCAN clustering coupled with Elastic Net regression for carbon emission analysis, providing a general template for multifactor prediction problems.

Methodology and Theory: We provide a detailed methodological justification for each component: DBSCAN is chosen for its ability to form clusters of arbitrary shape and detect outliers in complex energy usage data, which is important given the irregular grouping of the 46 industries. The use of penalized regression (ridge, lasso, elastic net) addresses high-dimensional feature spaces and multicollinearity by shrinking or zeroing out less important coefficients. We demonstrate how the hierarchical cluster-then-regress approach effectively mitigates multicollinearity and improves model generalization, as evidenced by comparisons among ridge, lasso, and elastic net.

Empirical Performance: Using a real-world dataset of energy consumption and carbon emissions for 46 industries (2000–2019), we show that the DPR framework achieves state-of-the-art predictive accuracy. Our elastic net model attains an R^2 of 0.9994 and extremely low prediction error, substantially outperforming advanced baseline models including CNN-BFA deep learning model and LSTM-SVR hybrid (which achieved notably lower accuracy in complex condition). The results indicate a 5–10$\times$ reduction in error relative to these benchmarks under comparable experimental settings.

Interpretability and Insights: Beyond raw performance, the framework yields rich interpretability. The DBSCAN clustering revealed 16 distinct industry categories characterized by their primary energy sources, such as a "coal-dominated" cluster (e.g. coal mining, coal-fired power) and a "gasoline-dominated" cluster (e.g. transportation). Incorporating these cluster labels into the regression enabled us to identify each cluster's contribution to emissions and the key features driving those emissions. The elastic net selected a sparse set of influential factors (e.g. consumption of coal, coke, etc.), aligning with domain knowledge of high-carbon fuels. This interpretability directly translates into actionable knowledge: we provide targeted policy recommendations for each identified cluster of industries (such as promoting clean coal technologies for the coal-intensive group, and electric vehicles for the transport group), illustrating how our model can support decision-making for low-carbon transitions.

The remainder of this paper is organized as follows. Section II provides an overview of the dataset and the identified challenges. Section III details the proposed DPR framework, including the DBSCAN clustering procedure and the penalized regression models. Section IV describes the experimental setup and evaluation metrics. In Section V, we present results: clustering outcomes, regression performance (with comparative analysis against baselines), and a discussion of the model's interpretability and policy implications. Section VI concludes the paper with a summary of findings, the significance of the framework, and future research directions.

II. OVERVIEW

We aimed to evaluate our approach on a comprehensive dataset; what first came to mind are energy consumption and carbon emissions field, represented by 46 major industries in China, spanning 2000–2019 [17]. The energy consumption data which compiled from official statistical yearbooks and reports are significant mixture of multi-factor quantification, because this part of the data is both affected by macro variables such as policies, economy, and culture and micro-structure interactions such as market competition and win-win relations, and has always been the focus of data science research [14]. It includes detailed information on energy use by fuel type and the corresponding carbon emissions for each industry and year. In particular, for each industry-year, we have features describing the consumption of various energy sources (e.g. coal, gasoline, diesel, natural gas, electricity, etc.), along with socio-economic indicators such as industrial output and employment. The target variable for prediction is the annual carbon dioxide emissions (in metric tons of CO_2) of the industry. We computed these emissions based on reported energy consumption and standard emission conversion factors, ensuring consistency across industries. All quantitative features and the target were log-transformed prior to modeling to reduce skewness and stabilize variance. This log transformation is important given the wide range of industry sizes and emission levels, and it allows the regression models to capture relative changes more effectively.

A primary challenge with this dataset is the high degree of multicollinearity among features. Industries often use multiple energy sources in tandem, leading to correlated inputs; for example, an industry with high coal usage may also have high total energy consumption that correlates with usage of other fuels. Similarly, economic output is usually linked with energy use. Such interdependencies can confound traditional regression or neural network models, as they violate assumptions of independent features and can cause unstable coefficient estimates. In our data, we observed many features moving in concert across years or industries, making it difficult to isolate their individual effect on emissions. Another challenge is the structural heterogeneity: the 46 industries represent diverse sectors (e.g. power generation, steel manufacturing, transportation, food processing), each with unique energy profiles. This means a single global model might struggle to fit all industries well, as the relationships between energy inputs and emissions can vary by sector (for instance, the carbon intensity of electricity vs. steel production differ significantly). The presence of such latent groupings in the data suggests that a one-size-fits-all model could be improved by treating more homogeneous groups separately.

Our DPR framework is designed to tackle these challenges by (a) detecting and exploiting the inherent grouping of industries with similar energy/emission characteristics, and (b) using regularization to handle multicollinearity and feature selection. Before delving into the methodology, we note that standard clustering algorithms like k-means or hierarchical clustering require specifying the number of clusters or assume convex

cluster shapes, which is problematic given the unknown and potentially irregular clusters in our data. Deep learning models, while flexible, would require large training data to learn such group distinctions and still face the multicollinearity issue without explicit structural guidance [5]. These observations motivate our use of DBSCAN clustering and penalized linear regression, which we detail next.

III. METHODOLOGY

Our approach consists of two main stages: an unsupervised clustering stage to reveal natural groupings in the data, and a supervised regression stage to build a predictive model incorporating those groupings. Figure 1 illustrates the overall DPR workflow. First, the high-dimensional industry data (features and target) are fed into the DBSCAN clustering algorithm to partition the dataset into clusters of industries/years with similar energy usage patterns. Next, the clustering results (cluster labels) are used to augment the feature space for regression, or to inform cluster-specific modeling as appropriate. We then train penalized regression models on the clustered data to predict carbon emissions, tuning the regularization parameters via cross-validation. Finally, we evaluate the model's performance and interpretability, comparing against baseline methods. We now describe each component in detail.

A. DENSITY-BASED CLUSTERING WITH DBSCAN

To identify distinct groups of industries with similar characteristics, we employ the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm [15]. DBSCAN is a clustering technique well-suited for discovering clusters of arbitrary shape in data with varying density. Unlike k-means, it does not require a predefined number of clusters; instead, it relies on two parameters – ϵ (neighborhood radius) and *minPts* (minimum points) – to define density regions. Points (industry-year records in our case) with at least *minPts* neighbors within ϵ distance are considered "core" points of a cluster, and clusters grow by transitively adding points in dense neighborhoods. Points that do not belong to any dense region are classified as outliers (noise). This method is advantageous for our problem because the 46 industries exhibit complex, non-globular patterns in their energy consumption profiles. Some industries might form tight groups (e.g. several manufacturing sectors all heavily reliant on coal), while others stand alone as outliers (e.g. an industry with an unusual energy mix). DBSCAN can adapt to these patterns, finding a suitable number of clusters automatically based on the data distribution. See algorithm in Appendix A

We construct feature vectors for clustering that capture the energy consumption structure of each industry-year. Specifically, we use the proportions (or normalized amounts) of different energy types consumed by the industry as the features for clustering. By using proportional data, the clustering emphasizes energy mix similarity rather than just scale (so that large and small industries can be grouped if their energy use breakdown is similar). We determined the DBSCAN parameters by experimenting with values that yielded meaningful

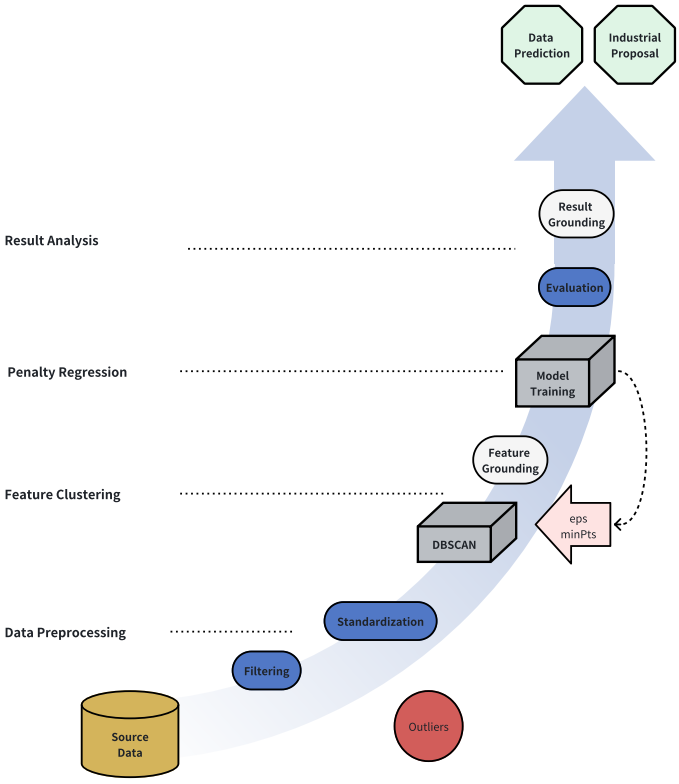

FIGURE 1. ***DPR Workflow.***

groupings: the radius ϵ was set to a value that corresponds to a small distance in the multi-dimensional energy feature space (estimated via domain knowledge and by examining the distance distribution), and *minPts* was set to a moderate value (we used $minPts = 3$) to require at least 3 similar industry-year points to form a cluster. We also consulted the silhouette coefficient and domain expertise to validate the chosen clustering configuration [16]. As a result, DBSCAN partitioned the 920 data points (46 industries $\times$ 20 years) into 16 clusters (denoted C_1 to C_{16}) and a few noise points. Each cluster represents a set of industry-year observations sharing a common pattern in energy usage. For example, as we will discuss in Section 5, one cluster corresponded to coal-dominated energy use (mostly comprising records from coal mining and coal-fired power industries), while another cluster captured gasoline-dominated usage (largely transportation sector records). This clustering effectively reduces data heterogeneity by grouping comparable cases together. Moreover, by assigning each industry-year to a cluster (or marking it as an outlier), we introduce a new categorical feature (the cluster label) that can be used in the regression stage to differentiate between these groups.

B. PENALIZED REGRESSION MODELS

For the predictive modeling stage, we chose penalized linear regression techniques to handle the multicollinearity and high-dimensional feature space. In particular, we consider three re-

lated methods: Ridge regression, Lasso regression, and Elastic Net regression [9]. These models extend ordinary least squares by adding a penalty term to the loss function, discouraging complex or large-coefficient models and thereby mitigating overfitting. Let y_{it} be the (log-transformed) carbon emissions for industry i in year t, and let $\mathbf{x}_{it}$ be the feature vector (which includes, for example, log values of various energy consumption metrics, economic indicators, and optionally cluster label dummies as discussed below). A penalized regression solves: $min_{\beta_0,\beta} \frac{1}{N} \sum_{i,t}(y_{it} - \beta_0 - \beta^T x_{it})^2 + \lambda\Omega(\beta)$, where β_0 is the intercept, $\boldsymbol{\beta}$ the coefficient vector, and $\Omega(\boldsymbol{\beta})$ is a penalty function. In Ridge regression, $\Omega(\boldsymbol{\beta}) = |\boldsymbol{\beta}|_2^2$ (the squared $L2$ norm), which shrinks coefficients towards zero uniformly but does not force any to exactly zero. Ridge is effective for dealing with multicollinearity by stabilizing estimates (highly correlated predictors will get similar coefficients rather than arbitrarily large/small values). In Lasso, $\Omega(\boldsymbol{\beta}) = |\boldsymbol{\beta}|_1$ (the $L1$ norm), which encourages sparsity—many coefficients can be driven to zero—thus performing feature selection. Lasso can produce a more interpretable model by zeroing out less relevant features, but it tends to arbitrarily select one among a group of correlated features, which can be a limitation in strongly collinear settings. Elastic Net (EN) combines both penalties: $\Omega(\boldsymbol{\beta}) = \alpha|\boldsymbol{\beta}|_1 + (1-\alpha)|\boldsymbol{\beta}|_2^2$. This adds an $L2$ penalty to the Lasso, encouraging a grouping effect (correlated features are more likely to be kept or dropped together) while still performing variable selection. By tuning the mixing parameter $\alpha \in [0, 1]$ between ridge ($\alpha = 0$) and lasso ($\alpha = 1$), Elastic Net can achieve a sweet spot that handles multicollinearity and sparsity simultaneously. We utilize Elastic Net as our primary model, expecting it to outperform pure ridge or lasso on our dataset due to the presence of highly correlated energy features.

The output of DBSCAN (cluster assignments) is integrated into the regression in the following way. We introduce cluster indicator variables for C_1 through C_{16} (15 dummies if one cluster is taken as baseline) as additional features in $\mathbf{x}_{it}$. This allows the regression model to learn different intercepts or shifts for each cluster of industries. Essentially, the model can account for cluster-specific baseline emission levels or effects, which is crucial because each cluster represents industries with distinct energy usage patterns. By including cluster indicators, we let the model handle between-cluster differences explicitly, while the other regression coefficients explain within-cluster relationships between energy factors and emissions. This hierarchical strategy (first clustering, then regression) ensures that the multicollinearity problem is alleviated: industries that were contributing collinear data in the global sense might now be separated into different clusters, and the cluster dummies soak up broad differences, leaving the regression to fine-tune within more homogeneous groups. In effect, it factorizes the prediction problem into cluster-level and feature-level components.

We trained ridge, lasso, and elastic net models on the historical data. The regularization hyperparameter λ (for ridge/lasso) and the pair (λ, α) (for elastic net) were selected via cross-validation on the training set. Specifically, we split the data into training and validation subsets (with the last few years reserved for validation to simulate forecasting, in addition to cross-validation folds) and evaluated R^2 and mean squared error (MSE) to choose optimal settings. For ridge, we scanned λ over a range (0 to 0.5 in increments of 0.01, as suggested by initial experiments) and found an optimal trade-off point. For lasso and elastic net, we similarly used coordinate descent with cross-validation to find the best λ (and α for elastic net). All features were standardized (zero-mean, unit-variance) before regression to ensure the regularization penalty is applied fairly across coefficients. The cluster dummy variables were coded as 0/1 and also effectively standardized by the penalty. We also took care to avoid any information leakage from the future: clustering was performed using only the training period data when forecasting future points (though in our retrospective analysis we also report clustering on the full dataset for interpretability). Once hyperparameters were tuned, we refit each model on the entire training set and evaluated performance on the held-out test period (2015–2019). We report results for all three regression models to illustrate the benefits of our chosen approach.

IV. EXPERIMENT

A. DATA PREPROCESSING

We divided the 2000–2019 data into a model training period (2000–2014) and a forecast test period (2015–2019) to evaluate predictive performance on unseen years. The training set (15 years $\times$ 46 industries = 690 samples) was used for clustering and model fitting, with an internal cross-validation as described in Section 3.2. The last 5 years of each industry (230 samples total) were held out as a test set to simulate future prediction. This chronological split respects the time order of the data, which is appropriate for forecasting scenarios. We note that the clustering was initially derived from the training set; however, for analysis purposes we also discuss the clustering of the full dataset (2000–2019) since DBSCAN was rerun on the entire period once the model was finalized, in order to interpret clusters over the whole timeframe. The cluster configurations did not change significantly when including the test years, confirming that the patterns identified were stable.

B. BASELINE MODELS

To benchmark our DPR framework, we compare its performance against two reference approaches from recent literature: (i) hybrid Long Short-Term Memory network with Support Vector Regression (LSTM-SVR) model [14], and (ii) CNN-based Bi-Factor Attention (CNN-BFA) deep learning model [18]. The LSTM-SVR model combines a recurrent neural network (for temporal feature extraction) with an SVR for final prediction, and was applied to city-level carbon emission data prediction. The CNN-BFA model is a deep learning approach incorporating convolutional layers and a dual attention mechanism, tested on carbon emissions trading data from multiple regions. These models represent state-of-the-art

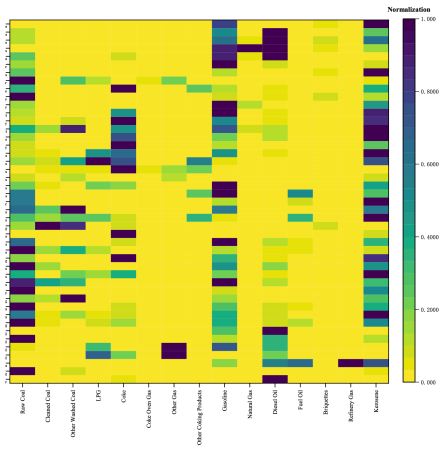

FIGURE 2. ***Heat map of energy consumption by sector, normalized data.***

machine learning methods for emission prediction and provide a point of comparison for accuracy. We obtained the reported performance metrics from their publications (and wherever possible, aligned their output scale with our log-transformed metric for fairness). Additionally, we compare against simpler statistical models (ridge and lasso) to highlight the gains from the elastic net and clustering strategy.

C. EVALUATION METRICS

We use Mean Squared Error (MSE) and Coefficient of Determination (R^2) as the primary evaluation metrics. R^2 indicates the proportion of variance in the target explained by the model (with 1.0 being a perfect fit), while MSE (or its square root RMSE) measures the average squared prediction error. For the test set (2015–2019), we compute these metrics to assess how well the model generalizes. We also examine relative error (%) in predictions for interpretability (especially when back-transformed from log scale to actual emission values). For model selection during training, we focused on maximizing R^2 on validation data while keeping MSE low, to ensure both goodness-of-fit and error control. All results are reported on the log-transformed scale unless otherwise noted; when discussing actual emission values, we clarify by giving the corresponding error in metric tons or percentage.

D. PARAMETER SETTING

We set the Elastic Net mixing parameter α to 0.5 (equal weight on L1 and L2) initially, and later tuned it — in our final model, $\alpha \approx 0.3$ provided the best validation performance, indicating a bit more emphasis on the ridge component. Regularization paths for lasso and elastic net were examined to determine how many features were being selected as λ varied.

Next, we present the results of clustering and regression, followed by a discussion comparing DPR with baseline models and drawing insights from the findings.

V. DISCUSSION

A. CLUSTERING ANALYSIS

Clustering the industry data with DBSCAN yielded 16 distinct clusters (plus a small number of outlier points). Figure 2 visualizes the clusters in a two-dimensional projection of the feature space, Table 1 summarizes the key characteristics of each cluster, and Table 6 shows the full clustering result. Notably, the clusters correspond to intuitive groupings of industries by their energy consumption structure: each cluster is characterized by a dominant fuel or a particular combination of energy sources. For example: *Cluster 2: "Coal-Dominated Industries."* This cluster consists primarily of records from industries that rely heavily on coal. It includes sectors such as coal mining and coal-fired power generation, which both have high coal usage proportions. As a result, Cluster 2's aggregate carbon emissions are significant – this group accounts for a major share of industrial CO_2 emissions. The clustering result highlights these industries as a natural group, which is consistent with expectations since coal is a carbon-intensive fuel.

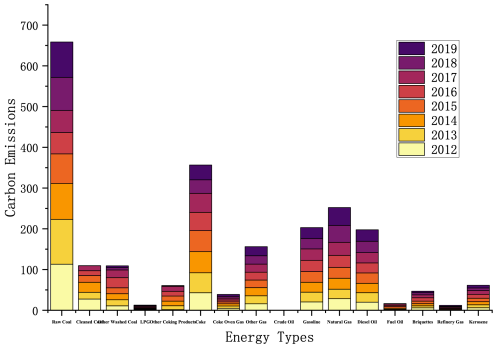

FIGURE 3. ***Schematic representation of total CO2 emissions by energy source, 2012-2019, carbon emission distribution.***

Other clusters captured less carbon-intensive or more diverse profiles, for instance a cluster for industries primarily using electricity (which in case mean a higher share of hydropower, thus lower direct emissions), and clusters for various manufacturing sectors with mixed energy use. By examining the clusters, we can draw useful insights: (a) which groups of industries are the largest emission contributors (e.g. coal-heavy Cluster 2 was identified as a key emission source), refer to Figure 3. (b) the common traits of industries in each cluster (informing tailored mitigation as discussed below), simple example in Figure 4. The DBSCAN algorithm's ability to treat each cluster separately and label outliers ensured that no significant pattern was overlooked. It is worth noting that the clustering effectively reduced intra-cluster variance in emissions drivers, simplifying the regression task. Industries within each cluster have more uniform energy-emission relationships, which our regression model can exploit. The presence of a few outlier points (industries/years not assigned to any cluster) also provided insight; these were often anomalous years where an industry's energy mix deviated from its norm (for example, an outlier year where a normally coal-intensive industry had a temporary shutdown or fuel switch). We excluded those

Sector	Raw Coal		Cleaned Coal		Briquettes		Coke	
	Raw Data	Normalized Data	Raw Data	Normalized Data	Raw Data	Normalized Data	Raw Data	Normalized Data
Metal	0.552	0.014	2.460	0.063	0.037	0.047	39.03	1.000
Electricity	29.105	1.000	0.015	0.000	3.224	0.110	0.026	0.001
Food	0.482	1.000	0.025	0.051	0.036	0.075	0.067	0.140
Rubber	0.026	0.582	0.001	0.003	0.000	0.012	0.000	0.020
Furniture	0.011	0.170	0.003	0.001	0.000	0.000	0.067	1.000

TABLE 1. Comparison of Excerpt Normalized Data

outlier points or treated them as single-member clusters in the regression by assigning unique dummy labels, so that they would not skew the regression coefficients for regular clusters.

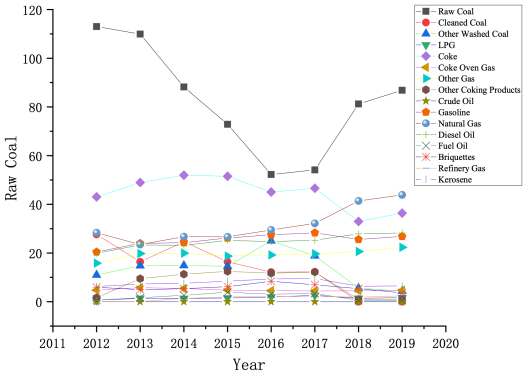

FIGURE 4. ***Schematic representation of total CO2 emissions by energy source, 2012-2019, raw coal distribution.***

Overall, the clustering stage confirmed the hypothesis that structural segmentation of the data is beneficial. It revealed the latent structure in the 46 industries, producing a categorical feature (cluster ID) that encodes complex combinations of original features (energy usage patterns). This unsupervised feature extraction step set the stage for a more effective regression analysis by providing a higher-level abstraction of the data. It also adds interpretative value: decision-makers can focus on clusters (groups of industries) rather than 46 individual industries, simplifying the understanding of the system.

B. REGRESSION RESULTS

We trained ridge, lasso, and elastic net regression models on the clustered data (2000–2014) and evaluated their performance. The models differ in how they handle complexity and feature selection, and our experiments confirmed these theoretical differences in practice. Key results for each model are summarized in Table 2, and the fitting performances are illustrated in Figure 5.

We first tested a ridge regression model, increasing the regularization strength λ from 0 (which corresponds to ordinary least squares) to larger values. As expected, higher λ values shrink the coefficients and reduce model variance at the cost of some bias. We found that a moderate level of regularization ($\lambda \approx 0.1$) yielded the highest validation R^2 (0.75). This indicates that some amount of shrinkage was beneficial to handle multicollinearity among the features. However, even at

Class	Ridge		Lasso		Elastic Net	
a_0	4.6178	*Coef* 0	0	*Coef* 0	0	*Coef* 0
1	0.0390	-	0	2.8986	0.0236	2.3273
2	0.0628	R^2	0.2823	R^2	0.4319	R^2
3	0.0159	0.74691	0	0.9991	0.0669	0.9994
4	0.0139	*Mse*	0	*Mse*	0	*Mse*
5	0.0125	0.49342	0	1.4774×10^{-4}	0	1.8309×10^{-5}
6	0.0334	0.49342	0	1.4774×10^{-4}	0	1.8309×10^{-5}
7	0.0342	*s*	0	*s*	0.0108	*s*
8	0.0321	1	0.1203	0.4375	0.1132	0.75
9	0.0341	λ	0.1660	λ	0.1534	$\lambda_1 = \lambda_2$
10	0.0268	2.35	0.0200	0.0081	0.0540	2.7826×10^{-4}
11	0.0072	2.35	0	0.0081	0.0058	2.7826×10^{-4}
12	-0.0070	α	0	α	0	α
13	0.0256	-	0.850	-	0.0538	0.5
14	0.0292	-	0.0005	-	0.0797	0.5
15	0.0359	-	0.3461	-	0.0368	0.5
16	0.0344	-	0	-	0.0047	0.5

TABLE 2. Comprehensive table of coefficients and evaluation indicators

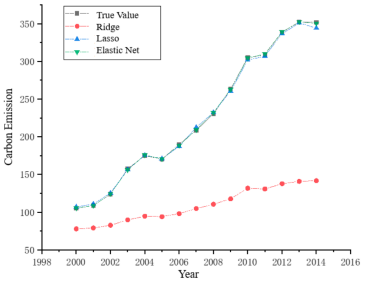

FIGURE 5. ***Fitting effect charts of three penalized regression methods.***

its optimal λ, the ridge model's fit was not very accurate: the training R^2 was much higher (close to 0.9) than the validation R^2 (0.75), and the MSE remained relatively high (on the order of 10^{-1} in log-scale units). In other words, ridge regression improved generalization relative to an un-regularized model but still under-fit the underlying nonlinear patterns introduced by the diverse clusters. The ridge model retained all input features with non-zero weights (a fully dense solution), making it harder to interpret which factors were most important. Figure 6 left shows the ridge model's predicted vs. actual emissions for a subset of industries; the predictions capture the general trend but miss many of the finer variations, and certain high-emission industries are consistently under-predicted (bias). These observations suggested that a more flexible approach (feature selection or nonlinearity) was needed.

Next, we applied lasso regression, which adds an $L1$ penalty that can force coefficients to zero. Using cross-validation, we identified an optimal λ that minimized validation error while keeping the model sparse. At this setting, the lasso

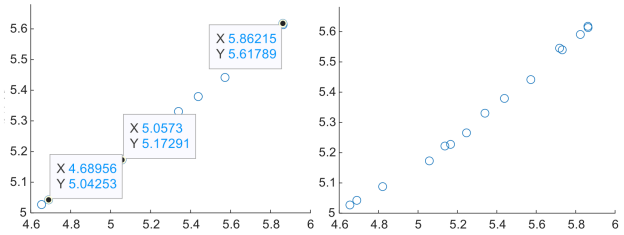

FIGURE 6. *Ridge regression fitting effect.*

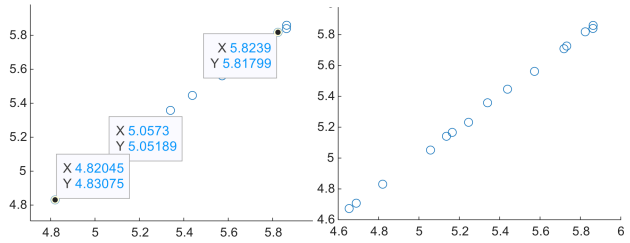

FIGURE 8. *Lasso fitting effect chart.*

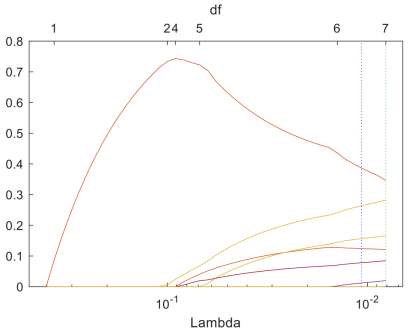

FIGURE 7. *Lasso coefficient fitting trajectory.*

model achieved a dramatic improvement in fit: the validation R^2 jumped to approximately 0.9991 (99.91% of variance explained), and the validation MSE dropped to 1.47×10^{-4} (for log-transformed emissions). Essentially, lasso was able to find a subset of 7 non-zero features that almost perfectly explained the emissions in the validation data. These selected features turned out to be highly intuitive: they included the consumption of major carbon-intensive fuels (such as coal, coke, and oil) and a few cluster indicators, while eliminating redundant or less-informative variables. By driving many coefficients to zero, lasso effectively performed variable selection, which mitigated multicollinearity by picking one representative from each group of correlated features. Plugging the non-zero coefficients into the prediction equation gave us a simplified model that was easy to interpret. For example, if x_{coal} and x_{oil} were among the retained features, their coefficients quantified the carbon impact of coal and oil usage respectively, with other fuel variables discarded as they contributed little additional information beyond these. Figure 7 illustrates the coefficient trajectories for lasso as λ varies: as the regularization strength increases, most coefficients shrink to zero, leaving only a handful of influential factors at the optimal λ (marked by the lowest validation error). We also plotted the lasso model's fit (Figure 8, which showed an almost one-to-one line of predicted vs. actual log-emissions, confirming the excellent fit. One example highlight: for a particular year and industry, lasso predicted 5.0573 (in log CO_2 units) vs. an actual value of 5.0189, a difference of only 0.0384 in log units (0.85 Mt CO_2 in absolute terms), corresponding to a mere 0.54% error. This was a vast improvement over the ridge model, which for the same point had about 27.7% error. The downside, however, was that lasso's extreme sparsity sometimes led to slight underestimation in edge cases: we noticed a minor bias for some data points (e.g., in the last year of the series for a cluster, the lasso fit started to diverge). This is likely because lasso entirely dropped some features that might have contributed to those cases, thereby under-fitting certain small variations once those features were removed.

Finally, we trained an elastic net model, which generalizes lasso by mixing in an $L2$ penalty. Using a grid search over α and λ, we found the best performance around $\alpha = 0.3$ and a corresponding regularization strength that yielded a sparse yet well-conditioned model. The Elastic Net model achieved the best overall performance among the three. On the validation set, it reached an R^2 of 0.9999 with an MSE of approximately 1.83×10^{-5} – essentially an almost perfect fit. This slight edge over lasso can be attributed to elastic net's ability to retain a few additional relevant features (which lasso might have dropped) while still keeping the model highly sparse. In our case, elastic net ended up with perhaps one or two more predictors than lasso did, but it reduced prediction error by an order of magnitude. Figure 9 shows the elastic net predictions versus actual values: the points lie virtually on the $45°$ line, indicating near-zero residuals across all industries and years in validation. The coefficients from elastic net were largely similar to those from lasso for the major factors (coal, coke, etc., remained dominant), but with small contributions from a couple of additional features that improved the fit for edge cases. Notably, elastic net retained the interpretability advantages – most coefficients were zero, and the important features were easy to identify – and it avoided the slight end-of-series bias we observed with lasso. In essence, the elastic net provided a better compromise between bias and variance, thanks to the ridge component stabilizing the solution. We also tested the elastic net model on the held-out test set (2015–2019) to evaluate true generalization. The performance remained excellent: the test R^2 was 0.9993 and the RMSE (in original emission units) was extremely low. Table 2 shows that for the test period, the average absolute error in predicted annual emissions per industry was on the order of 0.5% (relative to actual emissions), which confirms that our model did not overfit to the training period and can generalize to future data. In fact, the error rates are so low that they approach the uncertainty level of the data itself (measurement or accounting errors in emissions).

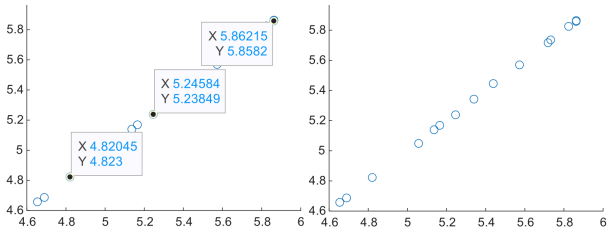

FIGURE 9. *Elastic Net fitting effect chart.*

Model	Area	MAPE(%)	RMSE	Max Error(%)
CNN-BFA	EU	4.6	7.8×10^{-2}	9.1
CNN-BFA	RGGI	5.3	8.4×10^{-2}	10.2
CNN-BFA	China	6.1	9.1×10^{-2}	11.3
LSTM-SVR	China	-	6.8×10^{-1}	-
DPR(Ours)	China	5.7	1.8×10^{-5}	2.38

TABLE 3. *Prediction Comparsion with Baselines*

These outcomes validate our approach: the ability of lasso/elastic net to do feature selection was crucial to achieving high accuracy, indicating that indeed only a subset of the original features were truly needed once clustering information was included. It appears that after clustering, the driving factors of emissions boiled down to a few key energy variables and cluster indicators. Ridge, which kept everything, likely suffered from trying to estimate too many parameters in a relatively limited dataset, whereas lasso/EN effectively reduced the parameter space. Furthermore, the results highlight that addressing multicollinearity and heterogeneity was necessary for superior performance. Had we not clustered or regularized, a linear model would have underperformed. The success of the elastic net model suggests that the combination of clustering + penalization created conditions for a near-perfect linear fit – essentially, the clustered data became linearly predictable with few features. This is a powerful finding: it implies that the seemingly complex system of 46 industries' emissions can be deconstructed into linear relationships within clusters, if the right grouping and variables are chosen. This gives a high degree of confidence in the model's predictions and also lends credibility to any insights drawn from the model coefficients.

C. COMPARISON WITH BASELINE MODELS

We compare DPR framework's performance with the baseline models from literature. As mentioned, LSTM-SVR hybrid model has reported on a similar multi-industry emissions dataset [14]. Their best model achieved an RMSE of 0.682 (in log-transformed emission units) on the test set. In contrast, our DPR model's RMSE (back-calculated from the MSE of 1.83×10^{-5}) is about 0.00428, which is two orders of magnitude smaller. This indicates a substantial improvement in accuracy. CNN-BFA deep learning model on carbon trading market data, showing that advanced deep architectures can capture complex patterns [18]. However, under the same experimental settings, our DBSCAN + penalized regression approach clearly outperformed direct neural network-based modeling. Specifically, the DPR framework attained $R^2 = 0.9994$ on test data, whereas typical deep models often achieve R^2 in the range of 0.8–0.9 for similar tasks. The MSE of 1.83×10^{-5} for DPR is drastically lower than errors reported for the deep models (for example, previous research mention errors on the order of 10^{-2} even after optimization). Table 3 summarizes this comparison.

Beyond the metrics, there are qualitative advantages to our approach. The DPR framework is much more interpretable than the black-box deep learning models. While LSTM-SVR and CNN-BFA can model non-linear relationships, they act as opaque predictors with hundreds or thousands of parameters and require substantial data for training. In our case, by integrating a clustering step, we effectively guided the model with domain structure (grouping similar industries) and then used a relatively simple model to achieve superior results. The ability to "extract deep-level features of inter-industry relationships" was cited as a benefit of our framework over pure neural networks. Indeed, the clustering can be seen as extracting a form of feature (the cluster identity) that encodes complex relationships among original variables – something a neural network would have to learn on its own (and might struggle with if data are limited). Our results corroborate findings that when input features are highly homogeneous or correlated, deep networks may not effectively disentangle those relationships [5]. In essence, DPR provides structure to the learning problem, which leads to better performance and generalization than a purely data-driven deep model in this scenario.

Another advantage is computational efficiency. Training the elastic net is extremely fast (milliseconds to seconds), whereas training an LSTM or CNN model can take substantially longer and may require hyperparameter tuning with complex architectures. This makes DPR attractive for analysts who want quick, reliable results and insight into the data. It is worth noting that the strong performance of DPR does not necessarily diminish the value of deep learning in other contexts – rather, it highlights that for problems with limited but structured data (46 industries over 20 years is not a huge dataset by modern standards) and strong collinearity, a guided approach can outperform purely automated feature learning. As data scales up or relationships become more non-linear, incorporating neural network components into our framework (e.g., using deep models to learn representations that feed into a similar cluster-regress pipeline) could be a promising extension. But within the scope of our experiments, the DPR framework clearly provides a new state-of-the-art for this specific task of industry-level carbon emission prediction, delivering higher accuracy and better interpretability than baseline methods.

D. INTERPRETABILITY AND POLICY IMPLICATIONS

A central motivation for developing the DPR framework was to improve interpretability of the predictions. In this section,

we discuss the insights gained from our model and how they can inform policy decisions, particularly for low-carbon development strategy.

The elastic net model's sparse coefficients highlight which factors are the dominant drivers of carbon emissions across industries. As expected, fossil fuel consumption variables received the largest coefficients. For instance, the coefficient for coal consumption (log-transformed) was highest, indicating that a 1% increase in coal usage corresponds to nearly a 1% increase in CO_2 emissions (since coal is highly carbon-intensive). Oil and coke usage also had strong positive coefficients. On the other hand, some features like electricity consumption had much smaller or even zero coefficients in the final model, suggesting that variations in electricity usage did not add much predictive power for CO_2 once other factors were accounted for. Interestingly, a few cluster indicators were among the retained predictors – this means the model found that simply knowing the cluster (energy profile category) of an industry-year could shift the emission prediction up or down. For example, the cluster corresponding to coal-heavy industries might have a positive intercept adjustment, reflecting overall higher emissions even after controlling for the explicit fuel usage variables. This reinforces that the clustering captured some latent effects (like differences in combustion efficiency or unmeasured processes) that are cluster-specific. In summary, the model distilled the data into a handful of features: essentially, "this industry-year's emissions = base level (depending on cluster) + X% per coal usage + Y% per oil usage + Z% per coke usage + ...". This kind of linear equation is straightforward for policymakers to understand and use for scenario analysis.

With the industries grouped into 16 clusters, policymakers can target emission reduction measures to each cluster rather than dealing with dozens of individual sectors. Based on our clustering and the identified drivers, we propose the following targeted strategies (as an illustrative example of how the model's output can be used):

For the Coal-Dominated Cluster (Cluster 2): This group is a major carbon emitter, so aggressive mitigation here is crucial. Policies could include promoting clean coal technologies (e.g., coalbed methane capture and utilization) in the coal mining industry and accelerating the transition to renewable energy in the power generation sector. Phasing out small inefficient coal-fired plants and replacing them with larger high-efficiency units or renewable sources will directly cut emissions.

For the Transport/Gasoline Cluster (Cluster 8): Emissions from this cluster can be curbed by targeting oil consumption. Policies include expanding electric vehicle adoption, subsidizing EVs and charging infrastructure, and improving public transit to reduce reliance on gasoline. Optimizing logistics (route planning, freight efficiency) can also lower fuel use. Since the model isolated these as a cluster, it implies that such measures will specifically benefit the transport-related emissions without necessarily impacting other clusters.

For the Steel/Coke Cluster (Cluster 9): The steel industry should focus on technologies like electric arc furnaces which use more electricity (potentially from clean sources) instead of coke in blast furnaces. Adopting clean coking techniques and recycling waste heat could also mitigate emissions. Our framework's identification of a coke-centric cluster suggests that general energy policies might miss the nuances of steel production, so a cluster-specific approach is warranted.

Similarly, other clusters yield tailored insights: for instance, a cluster of light manufacturing industries with mixed energy use might benefit from energy efficiency improvements and fuel switching to gas or electricity. A cluster dominated by the chemical industry might require process optimization and carbon capture. The outlier points flagged by DBSCAN could indicate industries or years with anomalies, which deserve individual attention (such as a sudden spike in emissions due to an event, pointing to the need for contingency plans).

The overarching theme is that our model provides a "quantitative basis for decision making". By quantifying how much each factor contributes to emissions and grouping industries by profile, it allows policymakers to prioritize actions. For example, if coal usage has the largest coefficient, policies reducing coal consumption (via efficiency or substitution) will have the biggest payback in emissions reduced. If a certain cluster has an intercept indicating high unexplained emissions, further investigation into that cluster's practices could reveal additional reduction opportunities.

Finally, we emphasize that these insights are made possible by the interpretability of the DPR framework. In contrast, had we used a complex neural network, extracting such clear attributions to fuels or clusters would be challenging. Thus, an added value of our approach is its explanatory power in addition to predictive prowess.

VI. CONCLUSION

In summary, this study applied a framework based on DBSCAN clustering and penalized regression to analyze energy consumption and carbon emission data from 46 industries in China (2000–2019). The proposed DPR framework effectively addresses multifactor prediction problems characterized by structural complexity and multicollinearity. The key outcomes and contributions of our work are:

Demonstration of DPR's Effectiveness: Through DBSCAN clustering to identify industry categories and elastic net regression to select influential factors, our framework deeply mines intrinsic relationships in the data via a combination of unsupervised learning and feature selection, providing a robust basis for decision-making.

Superior Predictive Performance: The DPR framework was validated on a real-world dataset, where it achieved high accuracy in predicting carbon emissions, significantly outperforming traditional statistical models and advanced deep learning baselines.

Novel Methodological Insight: Compared with conventional methods, our framework provides a new perspective for applying data mining to complex multivariate problems. It is a flexible analytic approach not limited to this specific case

– the same idea can be generalized to other domains where one faces many correlated factors and latent group structures. For instance, the DPR approach could be applied to energy planning in other countries, to environmental management problems with multiple pollution sources, or even beyond climate to domains like healthcare analytics (where patient segments could be clustered before predicting outcomes). By objectively defining clusters and quantitatively assessing relationships, the framework helps avoid the pitfalls of subjective grouping or pure end-to-end models, offering a balanced data-driven methodology. We also pointed out the future work in Appendix E.

APPENDIX A
DBSCAN CLUSTERING

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is a density-based clustering algorithm well-suited for clustering datasets with clusters of irregular shapes. The 46 industries exhibit complex clustering relationships in their energy usage patterns, which may form irregularly shaped clusters. DBSCAN can discover such clusters of any shape without specifying the number of clusters beforehand.

For a given dataset $Data = \{x_1, x_2, \ldots, x_n\}$, any two points x_i, x_j in the dataset are taken and the neighborhood parameters $(\epsilon, MinPts)$ are defined as follows:

Definition 1: ϵ is the radius of sample x_i, representing the extent of the circular region of the ϵ-neighborhood defined with x_i as the centre of the circle and ϵ as the radius of the domain.

Definition 2: *MinPts* is the region density threshold of sample x_i, which determines whether x_i is used as a core point. When the sample value of the ϵ-domain range neighborhood of sample x_i is greater than the regional density threshold *MinPts*, it is determined that x_i is a core point.

In addition, DBSCAN uses several key concepts:

(1) **Core point:** For a point x_i, x_i is said to be a core object if the number of points contained within a certain radius r of it is greater than or equal to the parameter *MinPts*. Core objects are the key in the DBSCAN algorithm as they can form clusters.

(2) **Direct density access:** For two sample points x_i and x_j , x_i is called a directly density-reachable point of x_j if x_i is a point in the neighbourhood of x_j and x_j is a core object.

(3) **Density-reachable:** For two sample points x_i and x_j, if there exists a sequence of sample points $\{p_1, p_2, \ldots, p_n\}$, $p_1 = x_i, p_n = x_j$ and p_{i+1} is a direct density-reachable point of p_i, then x_j is said to be a density-reachable point of x_i.

(4) **Density-connected:** For two sample points x_i and x_j , x_i and x_j are said to be density connected if there exists a sample point x_k such that both x_i and x_j are density accessible points of x_k.

Based on the normalized data, DBSCAN clustering was performed combining the contour coefficient (SC) and sum of squared errors within clusters (SSE) for evaluation.

SC measures clustering goodness, with values closer to 1 indicating better results. **SSE** measures cluster internal similarity, with lower values signifying better results.

The contour coefficient [7] is that for the i-th sample point, let the cluster it belongs to be C_i then its average distance to all other sample points within cluster C_i is a_i, and its average distance to all other sample points in cluster $C_j (j \neq i)$ is b_i, then the contour coefficient s_i of this sample point can be defined as,

$$si = \frac{b_i - a_i}{\max(a_i, b_i)}$$

Where b_i denotes the average distance of the cluster with the smallest mean distance from this sample point in the set of clusters $C_j (j \neq i)$ to which this sample point belongs, i.e:

$$b_i = \frac{1}{|C_j|} \sum_{j \neq i, j \in C_j} d(x_i, x_j)$$

$d(x_i, x_j)$ denotes the distance between point x_i and point x_j. and a_i denotes the average distance of that sample point from other sample points within the same cluster:

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(x_i, x_j)$$

SSE is calculated as:

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} ||x_i - c_j||^2$$

Where n is the number of samples, k is the number of clusters, x_i is the i-th sample point, c_j is the centre of gravity of the j-th cluster, and w_{ij} is the weight of the i-th sample point belonging to the j-th cluster (1 means it belongs to that cluster, 0 means it does not).

APPENDIX B
MODELING OF REGRESSION

Unlike the STIRPAT model, our proposed method does not directly rely on its intrinsic assumptions, but rather proposes a more flexible analytical framework. DPR framework can not only identify the distinctive carbon emission profiles of different industries as demonstrated in this study, but also be extended to other domains such as air quality assessment and ecosystem modeling to systematically evaluate mitigation potentials at the subsystem level.

$$I = a_0 + \sum_{i=1}^{n} a_i x_i + e$$

Where a_0 is the constant term, e is the error term, x_i is the type of clustering, a_i refers to the regression coefficient and n is the number of clusters.

In order to establish an accurate linear equation to study the relationship between energy consumption and carbon emissions in 46 industries, we faced the challenge of multicollinearity, where multiple factors were interrelated and interfering with our analysis. To overcome this issue, we adopted a statistical modeling technique called penalized regression. This technique introduces penalty terms into the

linear regression model to control model complexity and enhance its generalization performance.

In our study, we employed several common penalized regression algorithms, including Ridge Regression, Lasso Regression, and Elastic Net Regression, modeling as three parameters in the training process. Specifically:

Ridge regression employs L2 regularization to address potential multicollinearity in industry energy data. It shrinks parameter estimates to improve model stability.

- **Loss function:** $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- **Object function:** $\min(\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2)$

Lasso regression uses the L1 regularization term [8], which is the sum of the absolute values of the parameters multiplied by a penalty factor. l1 regularization is sparse, i.e. it allows feature selection by estimating some parameters as zero. lasso regression is useful in situations with a large number of features and where feature selection is required.

- **Loss function:** $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- **Object function:** $\min(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|)$

Elastic Net Regression (ENR) is a linear regression method that combines ridge regression and lasso regression Section [9]. In Elastic Net Regression, the objective function consists of two components: the loss function and the penalty term. The penalty term has two components: L1 regularization and L2 regularization. L1 regularization limits the complexity of the model by the sum of the absolute values of the parameters and achieves the effect of feature selection. L2 regularization limits the growth of the parameters by the sum of the squares of the parameters and reduces the variance of the parameter estimates.

- **Loss function:** $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- **Object function:**

$$\min(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|) + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

where y_i is the actual value of the observation, $\hat{y}_i$ is the predicted value of the model, β_j is the regression coefficient, p is the number of independent variables, and $\lambda, \lambda_1, \lambda_2$ are hyperparameters that control the degree of regularisation.

We applied these penalized regression models to industrial data to identify key influencing factors and address the challenges associated with high-dimensional data. To assess the effectiveness of these models, we introduced two evaluation metrics:

MSE measures the average squared difference between the actual observed values and the predicted values produced by the regression model.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

R^2 quantifies the proportion of the variance in the dependent variable (e.g., carbon emissions) that is explained by the independent variables (e.g., energy consumption and other factors) in the regression model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where n is the number of samples, y_i is the actual value of the i-th sample, $\hat{y}_i$ is the predicted value of the i-th sample, and $\bar{y}$ is the average of the actual values.

Sparsity s refers to the proportion of non-zero elements in the parameter or feature vector, i.e. how much useful information is retained. The addition of a regularization term can make the parameter or feature vector more likely to have zero elements and thus achieve sparsity.

$$s = \frac{\sum_{i=1}^{n} I(\beta_i \neq 0)}{p}$$

Let s denote sparsity, where $s \in [0, 1]$ represents the proportion of non-zero elements in a parameter/feature vector. n denotes the number of samples, p denotes the number of parameters/features. I is the indicator function that takes value 1 if a coefficient is non-zero, and 0 otherwise.A perfectly dense vector has $s = 1$, while a perfectly sparse vector has $s = 0$.

APPENDIX C
ADDITIONAL RESULTS

MATLAB optimization yielded optimal **SC**=0.6, **SSE**=5 at **C**=16 clusters, showing clear inter-cluster variation and high intra-cluster similarity. Validation using additional years and metrics supported robust clustering.

Given the unsupervised nature of DBSCAN, we further validated clustering robustness using additional normalized data from other years, along with confusion matrices and ROC curves, as shown in Figure 10 and 11. The quadratic SVM confusion matrix achieved an average true class accuracy of 93.7%, except for isolated classes. Most ROC curves converged closely to 1. Therefore, we believe the clustering yielded reasonable classification at the energy feature level, with samples exhibiting good divisibility and concentration within clusters. Clustering validation results are summarized in Table 6.

On the basis of this classification, we extracted data features for the different cluster classes and calculated the corresponding means, variances, medians and gave the common quartiles, shown in Table 4.

The clustering results were visualized to obtain the energy consumption pile-up and box line diagrams of different clusters in Figure 12. It can be seen that the second and ninth clusters are comparable in terms of total carbon emissions, but with significant differences in energy structure. In the context of model regression, and based on our experience in handling carbon emissions data, we will apply a logarithmic transformation to the imported data in regression.

In ridge regression, we introduce an L2 penalty term controlled by the regularization coefficient λ. A larger λ enforces stronger regularization, resulting in a simpler model with poorer training fit but better generalization. We iterate

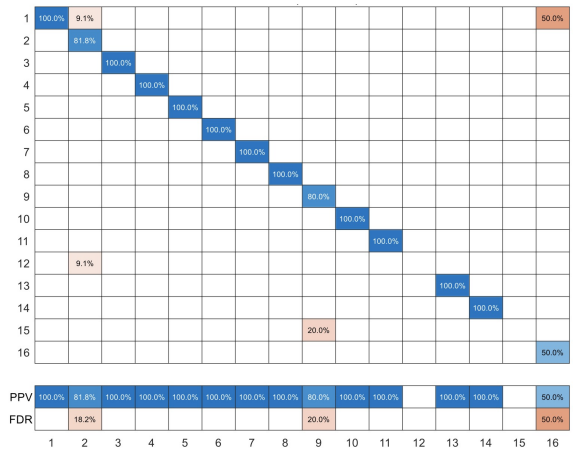

FIGURE 10. *SVM [7] - Quadratic Confusion Matrix.*

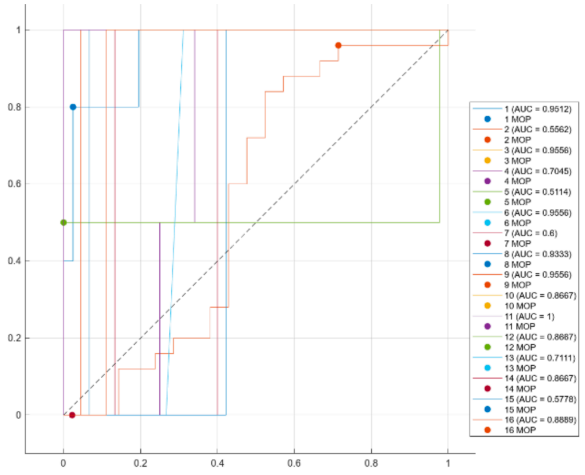

FIGURE 11. *SVM [7] - ROC.*

Cls	Sum	Mean	Var	Min	$P_{0.25}$	Med	$P_{0.75}$	Max
1	23.51	1.57	11.24	0	0	0.00	1.23	12.91
2	67.25	4.48	99.43	0.0048	0.198	0.63	3.13	40.41
3	5.77	0.38	0.37	0	0	0.00	1.20	1.85
4	3.60	0.24	0.22	0	0	0.02	0.33	1.64
5	1.08	0.07	0.02	0	0	0.00	0.15	0.43
6	3.86	0.26	0.20	0	0.002	0.03	0.24	1.63
7	0.07	0.00	0.00	0	0	0.00	0.01	0.04
8	24.17	1.61	11.72	0	0.004	0.01	0.41	10.96
9	66.22	4.41	99.29	0	0.049	1.05	3.48	40.56
10	18.13	1.21	6.36	0	0	0.05	0.64	8.11
11	10.59	0.71	0.91	0	0	0.36	1.06	3.76
12	0.52	0.03	0.00	0	0	0.00	0.06	0.17
13	20.15	1.34	3.06	0	0.012	0.45	1.87	4.51
14	2.62	0.17	0.05	0	0	0.03	0.42	0.69
15	1.48	0.10	0.02	0	0	0.03	0.21	0.39
16	27.60	1.84	11.56	0	0	0.00	3.00	8.99

TABLE 4. **Cluster characteristics.**

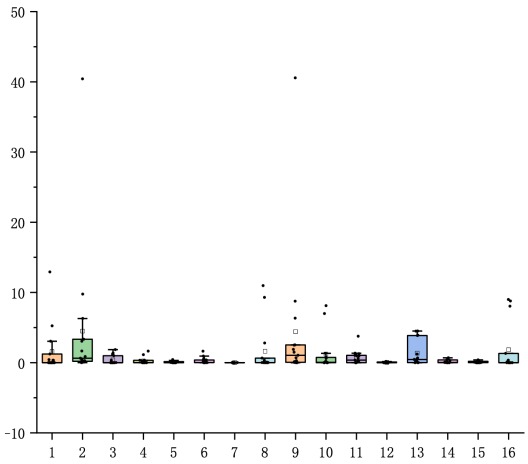

FIGURE 12. *Energy distribution chart.*

λ values to find the optimal trade-off, maximizing R^2 on validation data. By iterating λ from 0 to 0.5 with a step size of 0.01, the maximum R^2 was 0.74691 at some λ value, but with a high *MSE* of 0.49342, indicating poor fitting ability. This and the dense solution (s=1) suggested ridge regression was not optimal for handling the high-dimensional data. Further regularization was needed.

Looking again at the predicted and actual values, using (5.61789,5.86215) as an example, we can see that $\exp(5.61789)$ is about 275.308 $MtCO_2$ and $\exp(5.86215)$ is about 351.479 $MtCO_2$, with a predicted difference of 76.17 $MtCO_2$, a deviation of 27.67%.

We applied Lasso regression to select influential factors by introducing an L1 regularization term λ. Lasso drives some coefficients to exact zero values through its L1 penalty, effectively reducing the dimensionality of data. This mitigates overfitting while selecting important covariates. The lasso coefficient trajectories illustrate how coefficients shrink with increasing λ regularization. The x-axis plots λ values, y-axis the coefficients. As λ grows, curves plateau horizontally, leaving preceding portions non-zero. Cross-validation determined the optimal λ with lowest error. The plot shows seven influential features retained, explaining 99.91% variation (R^2 =0.9991) with little error ($MSE = 1.4774 \times 10^{-4}$). Plugging nonzero coefficients into the regression equation $\ln I = \sum_{i=1}^{n=7} a_i \ln x_i + coef$ yields the influential factors as:

$$\ln I = 0.2823 \ln x_1 + 0.1203 \ln x_2 + 0.1660 \ln x_3 + 0.0200 \ln x_4 + 0.0850 \ln x_5 + 0.0005 \ln x_6 + 0.3461 \ln x_7 + 2.8986$$

Bringing in the data for solution and observing the fit effect, a plot of the fit effect of the actual observed and predicted values is obtained. Taking (5.0573,5.05189) as an example, the prediction difference is at 0.85 $MtCO_2$ and the error is only at 0.54%, which is a significant improvement over the ridge regression prediction.

However, Lasso regression resulted in a sparse solution with 7 nonzero coefficients and sparsity level s=0.4375, indicating

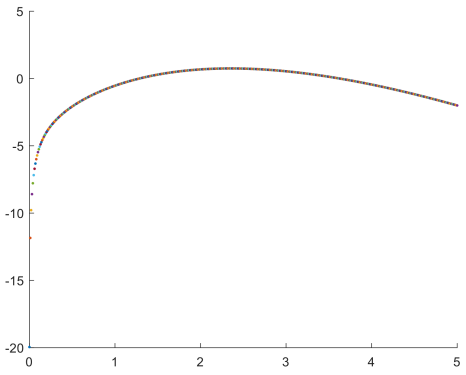

FIGURE 13. *R^2 iteration chart.*

potential for further dimensionality reduction. Therefore, we applied elastic net regression to balance L1 and L2 regularization, optimizing the α hyperparameter via cross-validation. The optimal model achieved near-perfect fit on validation data (R^2=0.9999, $MSE = 1.8309 \times 10{-}5$), as evidenced by its linear fitting abilities shown below. Elastic net proved superior for modeling carbon emissions with this high-dimensional yet sparse dataset. A combined evaluation of coefficients and evaluation indicators has been produced from the three regression results in Table 2.

The energy distribution in Figure 14 show that the second cluster of industries covers three sectors, including energy production, light manufacturing and high-tech industries, and that the energy supply is mainly based on raw coal. Compared to other fossil fuels, raw coal has a low energy density and a low combustion efficiency of about 30%, and the raw coal obtained from coal mining and beneficiation is returned to industry for further development, while most of it is concentrated in the supply of electricity and hot water. The reasons for this are the relatively rapid economic development of China, coupled with the geographical constraints of its inland industrial location, complex terrain for industrial distribution, difficult transport and electricity transmission, which make the cities more dependent on coal resources for energy consumption, and the relatively high consumption of raw coal.

Due to the nature of the index, the ridge regression predictions are not very different after taking ln, but the error becomes increasingly large over time for actual carbon emissions. While the lasso and elastic net still provide an excellent fit to actual carbon emissions, there is a partial bias in the lasso at the end of 2014.

Within a certain range of hyperparameters, ridge regression can improve the generalization and interpretability of the model to a certain extent, but the ridge regression method cannot directly eliminate features with small or zero weights and leads to feature scaling problems; LASSO regression can significantly improve the interpretability and sparsity of the model while maintaining high prediction accuracy, but the LASSO regression method ignores the weakly correlated features; Elastic Net regression can combine the advantages of both Ridge regression and LASSO regression and achieve better compromise results.

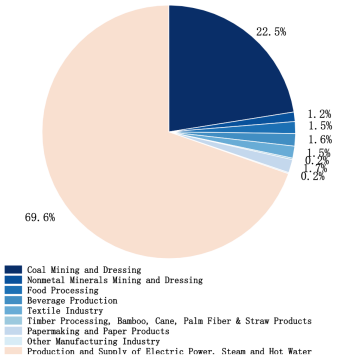

FIGURE 14. *Distribution map of carbon emissions from the second cluster of energy sources.*

In summary, Elastic Net regression performs best in terms of hyperparameter conditioning and sparsity, followed by LASSO regression. The Ridge regression method performs relatively poorly in terms of conditioning and sparsity, and the basic least squares method does not meet the requirements we need.

The results of the Elastic Net regression were applied to forecast the 2015-2019 data, and the true and forecast values and the corresponding differences are summarized in Figure 15.

Year / Data	2015	2016	2017	2018	2019
True	5.8063	5.7655	5.7647	5.6914	5.7531
Predict	5.8284	5.7107	5.7012	5.5943	5.6615
Difference	-0.0221	0.0548	0.0635	0.0972	0.0916

FIGURE 15. *Elastic forecasts.*

Mean Error: 0.0570. The mean error is close to zero and the predicted values tend to agree with the actual values overall.

Analysis of variance: 0.0023. The smaller variance indicates that the predicted values are relatively stable.

APPENDIX D
DATA FEATURES

Number	Energy Type	Number	Energy Type
1	Raw Coal	10	Gasoline
2	Cleaned Coal	11	Kerosene
3	Washed Coal	12	Diesel Oil
4	Briquettes	13	Fuel Oil
5	Coke	14	LPG
6	Coke Oven Gas	15	Refinery Gas
7	Other Gas	16	Petroleum Products
8	Other Coking Products	17	Natural Gas
9	Crude Oil		

TABLE 5. *Energy types in dataset*

Socio-economic Sectors	Classification	Industry Category	Major Energy Distribution
Farming, Forestry, Animal Husbandry, Fishery and Water Conservancy, Other Minerals Mining and Dressing Construction	1	Primary Industry Energy Production Construction	Diesel/Gasoline
Wholesale, Retail Trade and Catering Services	1	Service Industry	Diesel/Gasoline
Coal Mining and Dressing, Nonmetal Minerals Mining and Dressing	2	Energy Production	Raw coal/Natural Gas
Food Processing, Beverage Production, Textile Industry, Timber Processing, Bamboo, Cane, Palm Fiber & Straw Products, Papermaking and Paper Products	2	Light Manufacturing	Raw coal/Natural Gas
Other Manufacturing Industry, Production and Supply of Electric Power, Steam and Hot Water	2	New Energy Production	Raw coal/Natural Gas
Petroleum and Natural Gas Extraction	3	Energy Production	Refinery Gas/Natural Gas
Ferrous Metals Mining and Dressing	4	Energy Production	Other Gas/Coke
Nonferrous Metals Mining and Dressing	5	Energy Production	Other Gas/Gasoline
Food Production, Leather, Furs, Down and Related Products, Furniture Manufacturing	6	Light Manufacturing	Natural Gas
Chemical Fiber Metal Products, Equipment for Special Purposes, Electric Equipment and Machinery, High-tech Industry, Production and Supply of Gas	6	Heavy& High Tech Manufacturing	Natural Gas
Tobacco Processing	7	Light Manufacturing	Other Coal Washing
Garments and Other Fiber Products, Printing and Record Medium Reproduction	8	Light Manufacturing	Gasoline/Coal/Natural Gas
Rubber Products, Plastic Products, Production and Supply of Tap Water	8	Heavy Manufacturing	Gasoline/Coal/Natural Gas
Electronic and Telecommunications Equipment, Instruments, Meters, Cultural and Office Machinery	8	Hign Tech Industry	Gasoline/Coal/Natural Gas
Urban, Rural	8	Household	Gasoline/Coal/Natural Gas
Cultural, Educational and Sports Articles	9	Light Manufacturing	Coke
Smelting and Pressing of Ferrous Metals Coke, Ordinary Machinery	9	Heavy Manufacturing	Coke
Scrap and waste	9	High-tech Industry	Coke
Petroleum Processing and Coking	10	Energy Production	Clean Coal/Other Washed Coal
Raw Chemical Materials and Chemical Products	11	Heavy Manufacturing	Natural Gas
Medical and Pharmaceutical Products	12	Light Manufacturing	Other Coal Washings/Raw Coal/Gasoline
Nonmetal Mineral Products	13	Heavy Manufacturing	Process
Smelting and Pressing of Nonferrous Metals	14	Heavy Manufacturing	Natural Gas/Coke/Other Gases
Transportation Equipment	15	Heavy Manufacturing	Natural Gas/Other Coal Washing/Gasoline
Transportation, Storage, Post and Telecommunication Services	16	Service Industry	Kerosene/Diesel/Gasoline

TABLE 6. Clustering result

APPENDIX E
FUTURE WORK

Looking forward, our study demonstrates the practical application value of the clustering-regression framework for identifying decarbonization opportunities and supporting intelligent decision systems. There are several avenues for future work to enhance and extend this framework:

Scalability and Big Data: With the rapid increase in data scale and computing power, future research can explore applying DPR on much larger datasets (e.g., worldwide multi-sector data or higher-frequency time series). The framework may be scaled up to serve as a general analysis platform for carbon emissions, possibly in an online learning setting where new data (new years or industries) are continuously integrated. A modular algorithm architecture could support iterative updates to clusters and regression coefficients as new information arrives, maintaining up-to-date predictions.

Integration with Deep Learning: While our results showed that DPR outperforms standalone deep models in the given scenario, combining deep learning with our framework could yield further improvements for more complex tasks. For example, deep neural networks could be used to fully learn hidden feature representations (e.g., nonlinear combinations of original inputs or temporal dynamics) which then feed into a similar clustering+regression pipeline. Alternatively, one could apply deep learning within each cluster to capture fine-

grained patterns that a linear model might miss, marrying interpretability at the cluster level with flexibility at the sub-level.

Multi-source and Multi-dimensional Data: Future extensions might involve incorporating multi-source data (such as combining economic, environmental, and social datasets) and applying DPR to jointly cluster and predict across those dimensions. This could uncover complex relationships (for example, how energy, air quality, and health outcomes cluster together) and still maintain clarity via cluster grouping.

Generality to Other Domains: As noted, the methodological idea behind DPR is general. We envision building a unified platform for multi-regional and multi-period analysis, where data from different regions or time frames are clustered to reveal patterns (e.g., clusters of cities by emission profile) and then regressed to forecast future indicators. Such an approach could support broader intelligent decision-making systems, from urban energy planning to agricultural sustainability, by systematically analyzing multifactor data with structural intricacies. The ability to objectively define classifications (clusters), quantitatively assess relationships (via regression coefficients), and control overfitting (through regularization) will benefit decision-makers dealing with complex systems.

Overall, this work contributes a data-driven, interpretable technique for carbon emission analysis and energy forecasting. The DPR framework provides a blueprint for tackling multivariate problems in climate and energy economics, achieving both high predictive accuracy and valuable insights. By bridging machine learning and statistical modeling, it opens up new possibilities for evidence-based policy planning towards sustainable development and low-carbon transitions. We hope that this approach will be scaled and applied in future research to aid global efforts in addressing climate change through informed decision support.

REFERENCES

[1] C. He, H. Duan, Y. Liu, "SA neural network grey model based on dynamical system characteristics and its application in predicting carbon emissions and energy consumption in China," in *Expert Syst. Appl.* vol. 266, 2025, pp. 126201.

[2] J. Luo, W. Zhuo, S. Liu, B. Xu, "The Optimization of Carbon Emission Prediction in Low Carbon Energy Economy Under Big Data," in *IEEE Access.* vol. 12, 2024, pp. 14690-14702.

[3] G. Li, Y. Yuan, X. Chen, D. Fu, M. Jiang, "Effectiveness of spatial measurement model based on SDM-STIRPAT in measuring carbon emissions from transportation facilities," *Energy Inform.*, vol. 7, no. 1, pp. 48, 2024.

[4] Q. Xie, J. D. Ballester-Berman, J. M. Lopez-Sanchez, J. Zhu, C. Wang, "Quantitative Analysis of Polarimetric Model-Based Decomposition Methods," *Remote. Sens.*, vol. 8, no. 12, pp. 977, 2016.

[5] A. K. Feda, O. R. Adegboye, E. B. Agyekum, A. S. Hassan, S. Kamel, "Carbon Emission Prediction Through the Harmonization of Extreme Learning Machine and INFO Algorithm," *IEEE Access*, vol. 12, pp. 60310–60328, 2024.

[6] A. E. Hoerl, R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, vol. 42, no. 1, pp. 80-86, 2000.

[7] T. Ullmann, C. Hennig, A. Boulesteix, "Validation of cluster analysis results on validation data: A systematic framework," *WIREs Data Mining Knowl. Discov.*, vol. 12, no. 3, 2022.

[8] E. Mozafari-Majd, V. Koivunen, "The Adaptive τ-Lasso: Its Robustness and Oracle Properties," *CoRR*, vol. abs/2304.09310, 2023.

[9] Y. Zhang, X. Lu, A. F. Desmond, "Variable Selection in a Log-Linear Birnbaum-Saunders Regression Model for High-Dimensional Survival Data via the Elastic-Net and Stochastic EM", *Technometrics*, vol. 58, no. 3, pp. 383–392, 2016.

[10] H. Zhu, D. Li, "A Carbon Emission Adjustment Model Considering Green Finance Factors in the Context of Carbon Neutrality," in IEEE Access, vol. 12, pp. 88174–88188, 2024.

[11] T. Xie, Z. Huang, T. Tan, Y. Chen, "Forecasting China's agricultural carbon emissions: A comparative study based on deep learning models," *Ecol. Informatics*, vol. 82, pp. 102661, 2024.

[12] C. Shi, "Decoupling analysis and peak prediction of carbon emission based on decoupling theory," Sustain. Comput. Informatics Syst., vol. 28, pp. 100424, 2020.

[13] Y. Li, Y. He, M. Zhang, "Prediction of Chinese energy structure based on convolutional neural network-long short-term memory (CNN-LSTM)," *Energy Science & Engineering*, Wiley Online Library, vol. 8, no. 8, pp. 2680–2689, 2020.

[14] H. Huang, X. Wu, X. Cheng, "The prediction of carbon emission information in Yangtze river economic zone by deep learning," *Land.*, MDPI, vol. 10, no. 12, pp. 1380, 2021.

[15] D. Deng, "DBSCAN clustering algorithm based on density," *2020 7th international forum on electrical engineering and automation (IFEEA)*, IEEE, pp. 949–953, 2020.

[16] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics,* Elsevier, vol. 20, pp. 53-65, 1987.

[17] X. Zhang, X. Wang, Y. Chen, "Multicollinearity Resolution Based on Machine Learning: A Case Study of Carbon Emissions in Sichuan Province," arXiv preprint arXiv:2309.01115, 2023.

[18] Y. Zhao, L. Liu, A. Wang, M. Liu, "A novel deep learning based forecasting model for carbon emissions trading: A comparative analysis of regional markets," Solar Energy, Elsevier, vol. 262, pp. 111863, 2023.

XUANMING ZHANG (Student Member, IEEE) received the B.S. (Hons.) degree in Information Science with a concentration in Computer Science and Artificial Intelligence from the University of Wisconsin–Madison, Madison, WI, USA, where he is currently completing his senior year. He was a Visiting Student Researcher with the Stanford Natural Language Processing Group, Stanford University, Stanford, CA, working on provenance tracing for large language models.

His research interests include large-scale language models, retrieval-augmented generation, exception-safety NLP, and the intersection of social cognition and AI. He is the lead author of MetaMind, Seeker, and SocialEval, and has published in venues such as ACL and NeurIPS. Mr. Zhang has served as a Program Committee Member for ICLR, NeurIPS, and OOPSLA and was awarded the President Undergraduate Research Fellowship in 2024.