Exploring the Key Drivers of CO_2 Emissions in Australia

Safa Ghannam

University of Technology Sydney

Safa.ghannam@uts.edu.au

***Abstract*—Accurately identifying the key drivers of carbon dioxide (CO_2) emissions is crucial for developing effective mitigation strategies. This study develops and evaluates a range of statistical methods, machine learning models, and interpretability techniques namely, Spearman correlation, mutual information, random forest, XGBoost, and SHAP, to analyze the impact of 38 economic, energy, and environmental factors on CO_2 emissions in Australia between 1982 and 2022. The study captures both linear and non-linear relationships with SHAP applied to enhance transparency and explain the contribution of individual features. The results indicate that consumption, gas usage, and transport are the most significant contributors to CO_2 emissions, followed by total greenhouse gas emissions (excluding land use change and forestry), electricity supply, energy consumption, and coal. Additionally, oil consumption, GDP, and AUS Energy Growth-QLD play significant roles in shaping emission patterns in Australia. Of these models, random forest outperforms XGBoost, achieving the highest accuracy (R^2 = 0.99) with the lowest error, demonstrating its superior ability to capture complex interactions. These findings demonstrate the strength of ensemble models combined with interpretable analysis and provide actionable insights for policymakers, enabling data-driven strategies for carbon reduction and sustainable economic planning.**

***Keywords* — CO_2 emissions, machine learning, XGBoost, SHAP, random forest, energy, economic factors, renewable energy, policy, sustainability.**

I. INTRODUCTION

Climate change remains one of the most pressing global challenges, primarily driven by rising CO_2 emissions from human activities, particularly fossil fuel combustion [1], [2]. These emissions intensify the greenhouse effect, leading to a rise in global temperatures and severe environmental consequences [3]. According to the Quarterly Update of Australia's National Greenhouse Gas Inventory, CO_2 emissions account for approximately 63% of total greenhouse gas emissions, underlining the need for effective mitigation strategies. Understanding the drivers of CO_2 emissions is complex, as many contributing factors interact in non-linear and interdependent ways. Traditional models often struggle to capture these complexities, making it difficult to generate clear and actionable insights for policymakers. To address this challenge, we develop and evaluate statistical and machine learning (ML) techniques, integrated with interpretability methods, to identify the key drivers of CO_2 emissions in Australia and gain deeper insights into their dynamics. Leveraging long-term national data across economic, energy, and sectoral variables, this study aims to determine which factors most strongly influence emissions. The findings provide data-driven insights to support policy decisions focused on improving energy efficiency, reducing fossil fuel dependence, and promoting renewable energy. Ultimately, this research contributes practical strategies for emissions reduction in Australia and offers a framework that can be adapted to other national contexts.

II. RELATED WORK

Feature selection is a key aspect of emissions modeling, given the wide range of economic, energy, and environmental variables influencing CO2 levels [9]. Previous studies have examined the effects of economic growth, energy consumption, trade, industrial activity, and demographic changes on emissions [3], [5]. The application of machine learning (ML) techniques has become increasingly prominent due to their ability to model complex, nonlinear relationships [3], [10]. For instance, a study in Saudi Arabia demonstrated that the choice of input features significantly affects CO2 prediction accuracy [22]. Advanced ML models such as artificial neural networks (ANN), support vector machines (SVM), and ensemble-based approaches have shown high predictive performance [11]. To enhance model transparency, interpretability methods such as SHAP and permutation-based importance scores have been adopted, enabling more informed policymaking [12]. Additionally, feature selection techniques help reduce dimensionality, improving computational efficiency and minimizing over-fitting risks [21]. Building on these findings, our study applies advanced ML models with SHAP interpretation to assess the main drivers of CO_2 emissions in Australia. We consider a comprehensive set of economic, energy, and environmental variables, including GDP, fossil and renewable energy sources, transport, manufacturing, mining, residential and commercial energy use, as well as emissions from CO_2, methane, and nitrous oxide, and land-use changes. Table I summarizes key contributions from related studies, highlighting their focus, methods, and relevance.

TABLE I

SUMMARY OF RELATED WORK ON CO_2 EMISSION MODELLING

Ref.	Focus Area	Methodology	Key Findings	Relevance
[3], [5]	Economic, energy, trade, demographics	Multi-variable analysis	Multiple factors jointly influence emissions	Supports broad variable selection
[22]	Input feature selection	Feature selection	Input choice significantly affects model accuracy	Highlights importance of input design
[11]	CO_2 prediction accuracy	ANN, SVM, ensemble	ML models provide high predictive accuracy	Justifies use of advanced ML methods
[12]	Model interpretability	SHAP, permutation scores	Transparent models improve policy relevance	Supports use of SHAP in this study
[21]	Dimensionality reduction	Feature selection analysis	Reducing features enhances performance	Enables efficient model development

III. LITERATURE REVIEW

Accurate CO_2 emissions forecasting is critical for effective policy and decision making. Traditional statistical models rely on historical emissions data, whereas modern ML techniques offer enhanced predictive capabilities [11]. Studies have demonstrated that incorporating external influencing factors, such as economic growth and energy consumption, improves model accuracy and reliability [14]. Recent advancements in deep learning, particularly long short-term memory (LSTM) networks, have further refined emissions forecasting by capturing temporal dependencies in emissions data [13]. However, model interpretability remains a significant challenge. To address this, explainable AI techniques such as SHAP and local interpretable model-agnostic explanations (LIME) have been integrated into predictive models to provide clearer insights into how input variables impact emissions [12]. Sectoral variations also play a crucial role in emissions modeling. Research suggests that sector-specific factors, such as industrial activity and land-use changes, should be explicitly incorporated into predictive models for improved accuracy [14]. Many authors have emphasized that incorporating multiple factors as inputs improves the performance of prediction models [6], [8], [7]. The interplay between air pollution, economic expansion, and industrial activities related to iron and steel in China was examined using monthly data from 2000 to 2017 to determine their interconnections [18]. The association among CO_2 emissions, economic development, and energy consumption in Turkey and the South Caucasus region was explored. The study employed a three-variable vector autoregressive model to analyze panel data [19]. In a later study, the authors investigated the causal relationships between renewable energy sources (solar and wind), coal consumption, economic performance, and CO_2 emissions. The research focused on the three largest energy consumers and CO_2 emitters: China, India, and the USA [20]. To improve computational efficiency and minimize the risk of prediction model overfitting, conducting a feature selection analysis is recommended. This approach reduces the number of input variables required for each model [21]. Our study builds upon these insights by employing ensemble learning and explainable AI techniques to analyze the key drivers of CO_2 emissions in Australia, providing a comprehensive framework for exploring the influencing factors for CO_2 emissions in Australia. Australia was selected for this research due to the limited focus on CO_2 emission studies, particularly regarding the influence of various economic, energy, and environmental factors. To the best of our knowledge, there has been no research that examines multiple key determinants from various sectors shaping CO_2 emissions in Australia, highlighting the need for a data-driven approach to inform sustainable policy decisions. To analyze the influencing factors on CO_2 emissions, ML techniques were implemented using Python, following best practices outlined in [4].

IV. METHODOLOGY

This section describes the methodologies employed to analyze CO_2 emissions in Australia. The study utilizes multiple analytical approaches, including Spearman correlation, mutual information analysis (MI), XGBoost, random forest, and SHAP analysis, to capture both the linear and non-linear relationships among influencing factors. Additionally, preprocessing steps and cross-validation techniques were applied to ensure data consistency and model reliability. Unlike prior studies that rely solely on correlation analysis or black-box ML models, we apply an interpretable artificial intelligence approach by leveraging SHAP to quantify each factor's impact on CO_2 emissions. Statistical methods (Spearman correlation, MI) help capture both linear and non-linear relationships, while ML models (XGBoost, random forest) provide robust feature importance rankings. SHAP further enhances interpretability, making our approach suitable for data-driven policymaking.

A. Datasets and Exploratory Data Analysis

This study integrates two datasets to analyze CO_2 emissions in Australia. The first dataset was obtained from Our World in Data's (OWID) Greenhouse Gas Emissions, while the second dataset, Australian Energy Update 2024, provides detailed in- sights into Australia's energy consumption. The OWID dataset was filtered to retain only data for Australia, and records prior to 1982 were removed. The datasets were merged based on the Year column, resulting in a combined dataset spanning from 1982 to 2022 with thirty-eight features related to economic, energy, and environmental factors. To gain insights into the dataset, summary statistics and visualizations were conducted. Trends in CO_2 emissions over time were analyzed, and relationships between CO_2 emissions and other variables were examined to identify potential correlations and dependencies.

Time series plots of key features were generated to observe their distinct patterns over time, providing insights for further analysis as shown in Fig. 1.

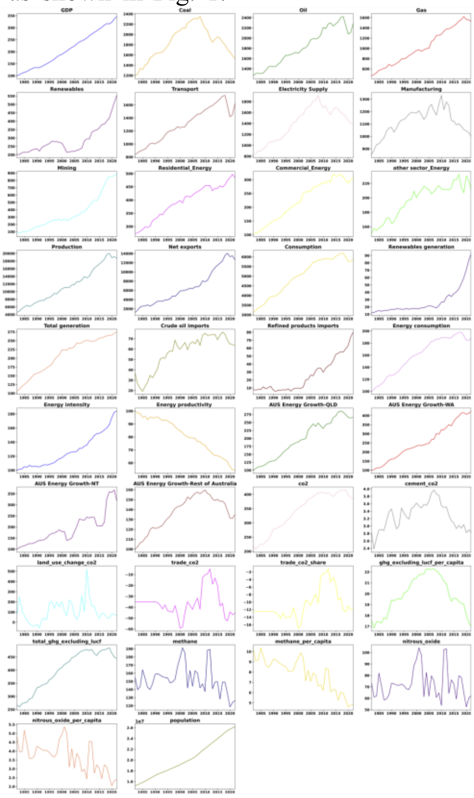

Fig. 1. Feature visualization – Time series plot of CO_2 emissions and influencing factors (1982–2022)

Table II presents a statistical summary of CO_2 emissions in Australia from 1982 to 2022, providing key insights into historical trends and variations. The mean CO_2 emission was 339.36, while the median was higher at 362.54, indicating a left-skewed distribution where earlier years had lower emissions compared to recent distribution where earlier years had lower emissions compared to recent years. The standard deviation of 69.47 suggests significant fluctuations, particularly in recent years, reflecting economic shifts and policy changes. The trend analysis, shown in Fig. 3, highlights a steady increase in emissions until 2015, followed by a slight decline, aligning with the negative skewness (-0.59) that suggests an initial rise followed by stabilization or reduction. The negative kurtosis (-1.12) confirms that emissions have generally followed a smooth trend rather than experiencing sharp spikes. The seasonal decomposition further reveals periodic fluctuations, suggesting recurring cycles that may be linked to seasonal energy demand, industrial activity, or policy interventions. The residual component highlights short-term variations, potentially driven by economic events or external shocks. The minimum emission recorded was 207.65, while the maximum reached 415.77, indicating a substantial rise over four decades despite recent stabilization. The standard error of 10.85 demonstrates the reliability of the mean as a representative measure. These statistical insights, combined with trend and seasonal de-composition, provide a comprehensive understanding of CO_2 emissions in Australia, helping to identify long-term patterns and external influences affecting emission levels. The histograms of the selected influencing factors in the dataset are shown in Fig. 2.

TABLE II

DESCRIPTIVE STATISTICS OF CO_2 EMISSIONS (1982–2022)

Statistic	Value
Mean	339.362
Median	362.537
Standard Deviation	69.474
Standard Error	10.850
Kurtosis	-1.121
Skewness	-0.586
Minimum	207.645
Maximum	415.770

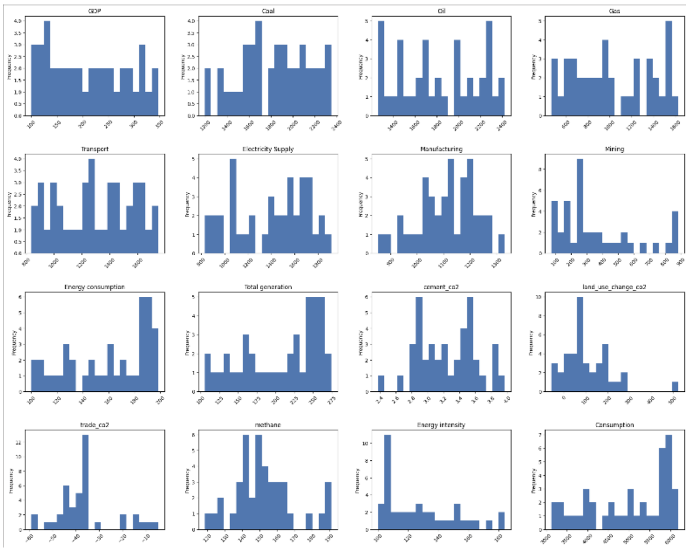

Fig. 2. Histograms of selected factors in the dataset, illustrating their distributions

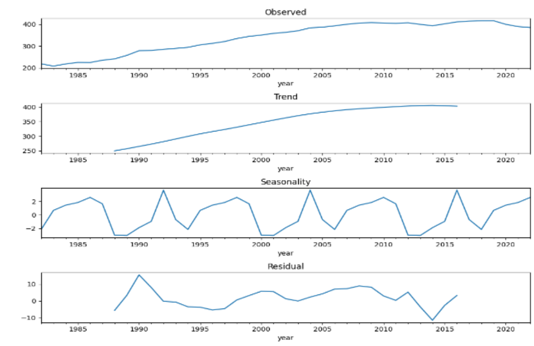

Fig. 3. Seasonal decomposition of CO_2 emissions in Australia (1982–2022)

B. Data Preprocessing

Preprocessing techniques commonly used in the ML literature were applied to enhance data reliability and integrity. This included identifying and handling outliers, removing duplicates, addressing inaccuracies, and managing noisy and missing data to minimize noise impact and ensure data quality. Specifically, missing values, primarily at the beginning of the dataset, were handled using backward filling to propagate the next valid observation backward. Complex column names were also simplified to improve clarity. Furthermore, a log transformation was applied to mitigate skewness, normalize the data, and stabilize variance, thereby enhancing model performance.

C. Feature Importance Analysis

1) Statistical Analysis - Spearman Correlation Analysis: Statistical analysis has been widely used in the literature to evaluate the influence of various factors on predictive models [17]. Correlation matrix analysis plays a vital role in examining relationships between variables in a dataset [15]. In this study, Spearman correlation was employed to assess the monotonic relationships between CO_2 emissions and various influencing factors. This method ranks strength and direction of relationships, helping to identify dominant influencing factors. However, it cannot detect non-linear dependencies. To address this limitation, MI is used, as it quantifies both linear and non-linear relationships, providing a more comprehensive view of feature relevance.

2) Statistical Analysis - MI Analysis: MI measures the shared information between CO2 emissions and predictor variables, capturing both linear and non-linear relationships. This approach is particularly useful in identifying dependencies that are not evident in correlation analysis. Higher MI values indicate stronger associations between the predictors and CO_2 emissions. In this study the mutual_info_regression function from scikit-learn was utilized to calculate MI scores.

The insights gained from these feature importance analyses inform the development of predictive frameworks, aiming to capture both linear and non-linear dependencies between CO_2 emissions and influencing factors. To gain deeper insights into CO_2 emissions and identify key driving factors, we utilize ML algorithms namely random forest and XGBoost to explore feature importance and capture complex relationships. Additionally, SHAP analysis is applied to the models to enhance interpretability, providing a detailed understanding of how individual features influence CO_2 emissions.

3) ML Model - Based Feature Importance: ML algorithms provide valuable insights into the key factors influencing CO_2 emissions. In this study, two ML techniques were employed: XGBoost and random forest. Traditional regression models struggle with high-dimensional interactions between emission drivers, whereas XGBoost is efficient in handling structured data and minimizing overfitting. Random forest, known for its ability to model non-linear relationships and reduce variance, enhances model robustness. Before training, CO_2 emissions were set as the target variable and an 80/20 train-test split was applied for evaluation. Hyperparameter tuning was conducted using 5-fold cross-validation with RandomizedSearchCV over 50 iterations. The models were then retrained using optimized configurations, and their performance was evaluated using multiple metrics. To enhance interpretability, SHAP analysis was applied. Unlike traditional feature importance scores that only rank variables, SHAP provides a detailed decomposition of each feature's impact, ensuring transparency in decision making. This is particularly crucial for policymakers seeking clear insights into the drivers of CO_2 emissions.

- XGBoost Model

XGBoost, a gradient boosting algorithm, was implemented due to its ability to handle high-dimensional data and capture complex feature interactions [16]. The model's performance was assessed using four evaluation metrics. Feature importance scores were derived from the trained XGBoost model to rank the influencing variables impacting CO_2 emissions.

Feature importance scores were derived from the trained XGBoost model to rank the influencing variables impacting CO_2 emissions.

- Random Forest

Random forest, an ensemble learning method based on decision trees, was used for its robustness in handling non-linearity and feature interactions. The model was trained using the same dataset as XGBoost, and its performance was evaluated using different evaluation metrics. Feature importance was determined through SHAP analysis, which provided insights into the most influencing factors.

D. Evaluation Metrics

The models' performance was evaluated using four widely recognized metrics, namely mean squared error (MSE), R-squared (R^2), root mean squared error (RMSE), and mean absolute error (MAE).

V. Results and Discussion

This section presents the results obtained from the four analytical approaches: Spearman correlation, MI, XGBoost, and random forest. Each method is analyzed individually, followed by a comparative discussion to evaluate their effectiveness in identifying the key influencing factors of CO_2 emissions in Australia.

- Spearman correlation analysis

The results from Spearman correlation analysis indicate strong positive correlations between CO_2 emissions and several key features as shown in Fig. 4. The results show that total greenhouse gas emissions excluding land use change and forestry (LUCF) have the highest correlation with CO_2 emissions (0.996). Other factors that exhibit strong correlations include energy consumption (0.975), transport (0.955), and oil consumption (0.949). These findings underscore the dominant role of the energy and transportation sectors in driving CO_2 emissions. In contrast, methane and nitrous oxide emissions exhibit weaker correlations, suggesting distinct influencing patterns. However, it is important to note that correlation analysis does not capture non-linear dependencies, which may limit its predictive capability.

- MI analysis

MI analysis quantifies the amount of shared information between CO_2 emissions and each predictor variable, as shown in Fig. 5. The results reveal that total GHG emissions excluding LUCF (representing annual greenhouse gas emissions from fossil fuels and industry) has the highest MI score of 2.087, indicating the strongest dependency on CO_2 emissions. This suggests that greenhouse gas emissions, excluding land use and land-use change, play a central role in determining overall CO_2 levels. Following closely, GDP (1.592) and total generation (1.574) also exhibit high dependency, indicating that economic activity and energy generation are crucial drivers of CO_2 emissions. Other prominent features include consumption (1.564), energy consumption (1.562), and sector-specific energy consumption metrics such as residential energy (1.533) and commercial energy (1.529). These factors underscore the significance of both residential and commercial energy usage in influencing CO_2 emissions. Additionally, population (1.514) and energy growth indicators such as AUS Energy Growth- QLD (1.460) highlight the relationship between demographic factors, regional energy production, and CO_2 emissions. Transport (1.431) and production (1.397) also emerge as important contributors, highlighting the significant role of the industrial and transportation sectors in CO_2 emissions. Lower-scoring features, such as land use change CO_2 (0.206) and nitrous oxide (0.309), suggest that while they have some influence, their impact on CO_2 emissions is less significant compared to energy related and economic factors. Unlike Spearman correlation, MI captures non-linear relationships, providing a more comprehensive understanding of feature relevance. Interestingly, variables such as population and crude oil imports, which exhibited moderate correlations in Spearman analysis, had higher MI values, indicating potential non-linear dependencies.

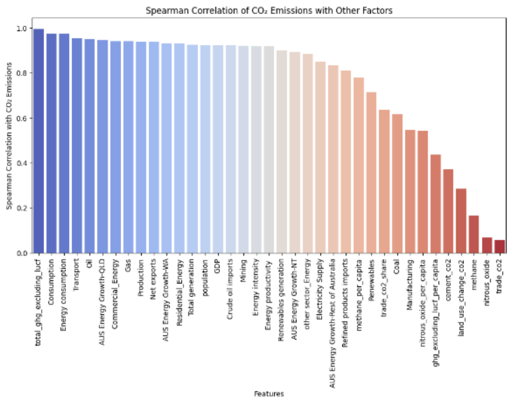

Fig. 4. Spearman correlation of CO_2 emissions with other influencing factors

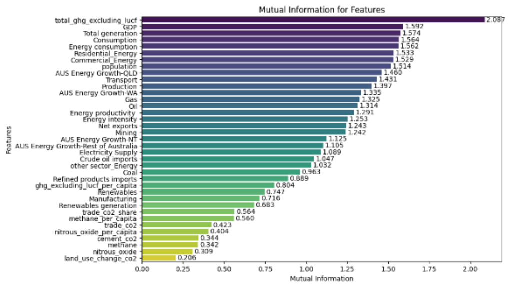

Fig. 5. Mutual information scores highlighting the significance of features and their relationship with CO_2 emissions

- XGBoost Model Performance and Interpretation

The XGBoost model achieved an R^2 score of 0.958, with an RMSE of 0.0396. The final evaluation results for the trained model are presented in Table III. The performance metrics of the XGBoost-developed model indicate strong performance. The feature importance scores were extracted and visualized to identify the most influencing variables in the model, as shown in the left panel of Fig. 6, which include GDP, coal consumption, oil consumption, and total GHG excluding LUCF, aligning with both the correlation analysis and MI results. These findings align with expectations, as economic activities and fossil fuel consumption are major contributors to greenhouse gas emissions. Gas and electricity supply also ranked high, suggesting the crucial role of energy sources in emissions trends. The SHAP analysis further confirms the high impact of GDP and fossil-fuel-related features (coal, oil, and gas), which contributed substantially to CO2 levels, reinforcing their significance. However, some lower-ranked features in correlation analysis, such as land use change CO2, gain significance in the XGBoost model, suggesting complex interactions not captured by linear methods. SHAP analysis results are visualized in Fig. 7.

TABLE III

MODEL EVALUATION RESULTS: XGBOOST AND RANDOM FOREST

Metric	XGBoost	Random Forest
MSE	0.001569	0.000349
RMSE	0.039604	0.018688
R^2	0.958183	0.990689
MAE	0.022896	0.0127608

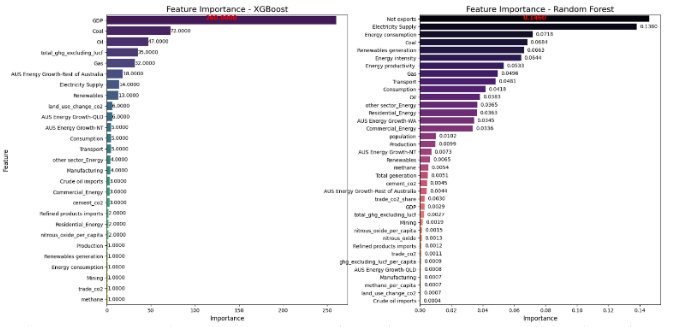

Fig. 6. Feature importance comparison for XGBoost and random forest, highlighting the most influencing variables affecting CO_2 emissions

- Random Forest Model Performance and Interpretation

The random forest model outperformed XGBoost, achieving an R² of 0.990, an RMSE of 0.0187, and a lower MAE, indicating superior generalization capability as shown in Table III. The feature importance analysis results, presented in Fig. 6, reveal that net exports are the most significant driver of CO_2 emissions, followed by electricity supply, energy consumption, and coal usage. These findings emphasize the critical role of trade-related emissions and energy production in shaping Australia's CO_2 emissions. Additionally, energy intensity and energy productivity are identified as key contributors, highlighting the importance of energy efficiency and overall consumption patterns in determining emissions. Transport and total greenhouse gas (GHG) emissions (excluding LUCF) also made notable contributions, reinforcing the interconnected nature of economic and environmental factors. The SHAP analysis shown in Fig. 8 further confirms the dominance of net exports, electricity supply, energy consumption, and transport as primary drivers of CO2 emissions. The SHAP visualizations enhance interpretability by illustrating how individual feature values contribute to the predicted CO_2 emissions across observations. Rather than providing a single importance score, SHAP shows the direction and magnitude of each feature's effect. The color gradient red for high feature values and blue for low helps indicate whether a variable increases or decreases the predicted emissions. For instance, a red point far to the right suggests that a high feature value contributes strongly to a higher prediction, while a blue point to the left indicates a negative contribution. High values of these features were generally associated with higher emission predictions, reinforcing their roles as primary drivers of CO_2 output. This level of detail provides a more detailed understanding of model behavior and supports clearer, evidence-based insights for policymakers. The lower error rates of the random forest model demonstrate its robustness in handling high-dimensional data and complex feature interactions, making it a highly reliable predictive model for CO_2 emissions analysis. The notable performance gap between the two models can be attributed to the dataset's limited size and annual resolution. With only 42 observations and relatively smooth temporal trends as shown in Table II, RF being less sensitive to hyperparameter tuning was better able to generalize. In contrast, XGBoost typically requires larger datasets and finer-grained data to fully leverage its boosting mechanism.

- Comparative Analysis of All Methods

To assess the consistency and reliability of the models, we compare the rankings of the influencing features across all methods. The top 10 influencing factors based on a combination of all approaches are presented in Fig. 9. and the bubble chart in Fig. 10 visually represents the frequency of feature appearances across all the techniques. The size of each bubble reflects the frequency with which a feature appears in the top 10 across these methods, emphasizing the most consistent influencing factors.

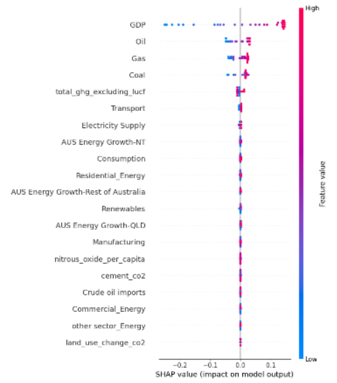

Fig.7. SHAP analysis for XGBoost, demonstrating feature contributions to CO_2 emission relationships

Fig. 8. SHAP analysis for random forest, illustrating the primary drivers of CO_2 emissions and their respective contributions

Consumption, gas, transport and total greenhouse gas emissions (excluding LUCF) consistently rank as the factors with the most influence across all methods. Spearman correlation and MI highlight the strength and type of relationships between CO_2 emissions and various influencing factors. While these methods provide valuable preliminary insights, they do not measure direct impact as effectively as the feature importance scores from the ML models. XGBoost and random forest provide valuable insights into the relationships between CO_2 emissions and influencing factors, with random forest achieving the lowest error rates and superior performance. The higher interpretability and robustness of random forest in capturing complex interactions among CO_2 emission drivers contribute to a more comprehensive understanding of the relationships between influencing factors. SHAP analysis confirms the dominant influence of GDP and fossil fuel consumption while revealing additional interactions not captured by traditional statistical methods. The role of renewable energy is consistently identified by all models, signaling its emerging influence on emissions trends, although fossil fuels remain the dominant contributor. It is important to clarify that in this study, consumption refers to total final energy use across all sectors in Australia, encompassing all energy types (oil, gas, coal, renewables), whereas energy consumption specifically refers to energy used within particular sectors or forms, such as electricity consumption within the National Electricity Market. Overall, while correlation analysis and MI provide valuable insights, random forest has the lowest error rates making it a more reliable model for identifying the key drivers of CO_2 emissions in Australia. SHAP analysis further enhances interpretability, emphasizing the role of trade-related emissions, energy production, and fossil fuel consumption in shaping Australia's CO_2 emissions.

Rank	Mutual Information	Spearman Correlation	XGBoost Importance	Random Forest Importance	SHAP Analysis (XGBoost)	SHAP Analysis (Random Forest)
1	total_ghg_excluding_lucf	total_ghg_excluding_lucf	GDP	Net exports	GDP	Net exports
2	GDP	Consumption	Coal	Electricity Supply	Oil	Electricity Supply
3	Total generation	Energy consumption	Oil	Energy consumption	Gas	Energy productivity
4	Consumption	Transport	total_ghg_excluding_lucf	Coal	Coal	Energy consumption
5	Energy consumption	Oil	Gas	Renewables generation	total_ghg_excluding_lucf	Consumption
6	Transport	AUS Energy Growth-QLD	AUS Energy Growth-Rest of Australia	Energy intensity	Transport	Transport
7	Residential_Energy	Commercial_Energy	Electricity Supply	Energy productivity	Electricity Supply	Renewable generation
8	Commercial_Energy	Gas	Renewables	Gas	AUS Energy Growth-NT	Energy intensity
9	population	Residential_Energy	land_use_change_co2	Transport	Consumption	Gas
10	AUS Energy Growth-QLD	Total generation	AUS Energy Growth-QLD	Consumption	Residential Energy	Coal

Fig. 9. Top 10 factors influencing CO_2 emissions in Australia across various methods, providing a comparative view of key determinants

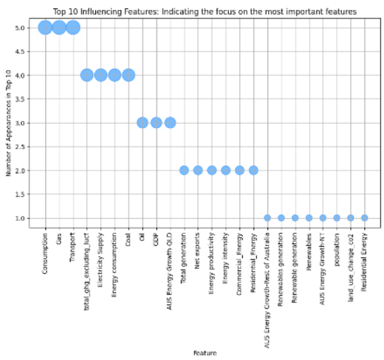

Fig. 10. Bubble chart depicting the frequency of feature appearances in the top 10 across various methods

VI. CONCLUSION AND LIMITATIONS

This study explored the relationship between CO_2 emissions in Australia and thirty-eight influencing factors across various sectors using a combination of statistical and ML techniques. Spearman correlation and MI provided foundational insights into feature relevance, while XGBoost and random forest enhanced our understanding of the key drivers of CO_2 emissions through SHAP analysis. Random forest identified net exports, electricity supply, energy consumption, and coal usage as the most significant factors. The findings underscore the critical role of economic activity and energy consumption in shaping emissions. By employing multiple methodologies, we gained a comprehensive understanding of the key factors influencing emissions, which is crucial for developing effective policies and strategies to mitigate urban CO_2 emissions. This multi-faceted approach not only strengthens confidence in the identified drivers but also provides a solid foundation for policy and decision making aimed at reducing emissions and promoting sustain- able urban development in Australia. Additionally, the methodologies and insights from this study can be applied to other countries, enabling the development of models tailored to different regions. This adaptability enhances the global relevance of the research, offering valuable tools for policymakers and researchers worldwide to address urban CO_2 emissions and contribute to global sustainability efforts. This study provides valuable insights but has some limitations. The use of annual data may overlook short-term or seasonal variations, thereby limiting the model's ability to capture finer temporal dynamics and certain aspects of non-stationarity in emissions trends. While SHAP enhances model interpretability, it cannot establish causal relationships. The dataset, although comprehensive, does not include policy factors, international trade factors, or carbon pricing data, which constrains the model's capacity to reflect broader economic and policy interactions. Moreover, the model's accuracy depends on the quality and granularity of the data, and improvements in data availability could further enhance predictive precision. Future work could explore deep learning techniques and advanced modeling approaches to better capture temporal variations and complex interactions influencing CO_2 emissions.

REFERENCES

[1] H. A. Kazem, A. Quteishat, and M. A. Younis, "Techno-economical study of solar water pumping system: optimum design, evaluation, and comparison," *Renew. Energy Environ. Sustain.*, vol. 6, p. 41, 2021.

[2] C. Mukonza and G. Nhamo, "Trade implications of grid emission factors under climate change and the green economy: comparative study of African and Asian continents," *Journal of Economic and Financial Sciences*, vol. 9, no. 1, pp. 13–27, 2016.

[3] S. Adams, F. Adedoyin, E. Olaniran, and F. V. Bekun, "Energy consumption, economic policy uncertainty and carbon emissions; causality evidence from resource-rich economies," *Economic Analysis and Policy*, vol. 68, pp. 179–190, 2020.

[4] J. Brownlee, "Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End," Machine Learning Mastery, Australia, 2016.

[5] L.-T. Zhao, J. Miao, S. Qu, and X.-H. Chen, "A multi-factor integrated model for carbon price forecasting: Market interaction promoting carbon emission reduction," *Sci. Total Environ.*, vol. 796, p. 149110, 2021.

[6] M. Zounemat-Kermani, E. Matta, A. Cominola, X. Xia, Q. Zhang, Q. Liang, and R. Hinkelmann, "Neurocomputing in surface water hydrology and hydraulics: A review of two decades retrospective, current status and future prospects," *Journal of Hydrology*, vol. 588, p. 125085, 2020.

[7] B. N. Nguyen, C. Prevedello, M. Cools, and J. Teller, "Factors influencing residential water consumption in Wallonia, Belgium," *Utilities Policy*, 2021.

[8] M. T. Timotewos, M. Barjenbruch, and B. M. Behailu, "The assessment of climate variables and geographical distribution on residential drinking water demand in Ethiopia," *Water*, vol. 14, no. 11, p. 1722, 2022.

[9] W. Xu, J. Wang, Y. Zhang, J. Li, and L. Wei, "An optimized decomposition integration framework for carbon price prediction based on multi- factor two-stage feature dimension reduction," *Annals of Operations Research*, 2022.

[10] C. Zhang, Y. Wu, and Y. Yang, "The Influencing Factors of sCER Price Dynamics Under the Clean Development Mechanism: Theory and Econometric Analysis," *Journal of Systems Science and Complexity*, vol. 31, no. 5, pp. 1244–1272, 2018.

[11] X. Li and X. Zhang, "A comparative study of statistical and machine learning models on carbon dioxide emissions prediction of China," *Environmental Science and Pollution Research*, vol. 30, pp. 117485-117502, 2023, doi: 10.1007/s11356-023-30428-5.

[12] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[13] S. Kumari and S. K. Singh, "Machine learning-based time series models for effective CO2 emission prediction in India," *Environmental Science and Pollution Research*, vol. 30, pp. 116601–116616, 2023, doi: 10.1007/s11356-022-21723-8.

[14] Y. Zhang, B. Nie, J. Du, J. Chen, Y. Du, H. Jin, X. Zheng, X. Chen, and Z. Miao, "Feature selection based on neighborhood rough sets and Gini index," *PeerJ Computer Science*, vol. 9, p. 1711, 2023, doi: 10.7717/peerj-cs.1711.

[15] Evans, J.D.: Straightforward Statistics for the Behavioral Sciences, pp. 135– 149. Thomson Brooks/Cole Publishing Co, Pacific Grove (1996)

[16] Zheng, S., Huang, Y., Sun, Y.: Effects of urban form on carbon emissions in China: Implications for low-carbon urban planning. Land **11**(8), 1343 (2022). MDPI

[17] Ghannam, S., Hussain, F.: Investigating the influence of Australia Day and Christmas Day on water demand in the Greater Sydney region. Water Supply **24**(10), 3540–3567 (2024). IWA Publishing

[18] Mele, M., Magazzino, C.: A machine learning analysis of the relationship among iron and steel industries, air pollution, and economic growth in China. J. Clean. Prod. **277**, 123293 (2020). Elsevier

[19] Magazzino, Cosimo. "Economic growth, CO2 emissions and energy use in the South Caucasus and Turkey: a PVAR analyses." International Energy Journal 16.4 (2016).

[20] Magazzino, C., Mele, M., Morelli, G., Schneider, N.: The nexus between information technology and environmental pollution: Application of a new machine learning algorithm to OECD countries. Util. Policy **72**, 101256 (2021). Elsevier

[21] Davoodi, S., Thanh, H.V., Wood, D.A., Mehrad, M., Rukavishnikov, V.S.: Combined machine-learning and optimization models for predicting carbon dioxide trapping indexes in deep geological formations. Appl.Soft Comput. **143**, 110408 (2023). Elsevier

[22] Nassef, Ahmed M., et al. "Application of artificial intelligence to predict CO2 emissions: critical step towards sustainable environment." Sustainability 15.9 (2023): 7648.