Peer review

Hung, Gordon. 2025. “Advanced Machine Learning Approach for CO2 Emission Forecasting: Leveraging Global Patterns to Overcome Data Insufficiency and Improve Model Accuracy.” *Journal of High School Science* 9 (1): 31–51.

1.Multicollinearity is undesirable in any statistical model; including in a machine learning model. Yet you have expressly chosen explanatory variables that are highly correlated with each other (R> 0.93). Why ? Usually, models go to great lengths to eliminate multicollinearity such as using heat-maps or correlograms, using VIF values, using Lasso regression or ridge regression and dropping highly correlated variables by using PCA. This seems to me a fundamental error in your models and analysis. You may have confused feature importance with Correlation. Please discuss and describe in the manuscript. Present correlogram, or heat-maps and VIF values.

2.I do not understand the correlations in Table 1. Why is fossil fuels not correlated to CO2 emissions? Please provide some level of detail in your variable naming such as “fossil fuel consumption”, or “fossil fuel import” etc. Why do cropland and grazing land have high correlation with CO2 emisssion? Does this imply that the greater the area of cropland or grazing land, the greater the CO2 emissions? How does that make sense? Should renewables and nuclear not have negative correlations of greater magnitude with CO2 emissions ? Please describe and discuss in the manuscript.

3.What is the effect of weighing Taiwan 50:1 on the model’s CO2 predicted emissions worldwide? It seems like you have artificially reduced the Taiwan class imbalance of <0.008% to approx, 40% due to this adjustment. How do you justify this? Class imbalances or unbalanced datasets are usually corrected by SMOTE or data autmentation or using inverse weighing class cross-entropy loss functions. Why were these accepted methods of reducing class imbalance not used? Please explain and describe in the manuscript.

4.You state “……we implemented a novel approach that utilizes world data for training and Taiwan data for testing.” This is only possible if the feature importance (not Pearson coefficient, see point 1) are the same for World and Taiwan data. Are they? Actually, this is not the procedure you are following in your explanation in the manuscript after table 3 where you train with 70% of world data and 30% of taiwan data in training and world data only or Taiwan data only for testing. This means that the dataset is trained on 70% of Taiwan data. Pleae remove any inconsistencies in explanation in the manuscript.

5.you state that the optimum training/testing split ratio is given by the square root of the parameters included. In this case your ratio should have been 4.3:1; i.e 77:23 split. Why then, did you choose a 70:30 split? Discuss and explain in the manuscript.

6.What did you use to estimate feature importance (not Pearson coefficient) in your models? Gini? SHAP? reverse-one-at-a-time? Bootstrapping? K-fold-cross validation? Discuss in the manuscript. Present features and feature importance for all the models used.

7.I do not understand Table 4. What do True and False mean?

8.In figures 2 through 4, where does the actual CO2 emission data come from (is this worldwide or Taiwan specific?). For example, where do the specific data points originate from? Is this an autocorrelation graph that is increasing with respect to time ?also need units tons?gigatons?

9.Lastly, why do you need complicated ML models if a simple multiple linear regression of electric power, transporation and industry versus CO2 emissions will explain > 90% variance in the data (i.e. R2 > 0.9) see: https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions#:~:text=In 2021 and 2022%2C the,of the COVID-19 pandemic.

Open response questions

Comments to author

1. *Multicollinearity is undesirable in any statistical model; including in a machine learning model. Yet you have expressly chosen explanatory variables that are highly correlated with each other (R> 0.93). Why ? Usually, models go to great lengths to eliminate multicollinearity such as using heat-maps or correlograms, using VIF values, using Lasso regression or ridge regression and dropping highly correlated variables by using PCA. This seems to me a fundamental error in your models and analysis. You may have confused feature importance with Correlation. Please discuss and describe in the manuscript. Present correlogram, or heat-maps and VIF values.*

To resolve the issue of multicollinearity, we composed a new dataset that consists of features that are not highly correlated to each other. This new dataset contains the following features: Year, Population Density, Primary Energy Consumption (TWh), and Land Use Per Capita. We explain the definition of these features in Section *3.2.1*, titled *Acquistion*. As suggested, we calculated the VIF value of each feature and concluded that there is no multicollinearity since all the VIF values obtained are below 5. The in-depth explanation of the steps taken to address multicollinearity is included in Section *3.2.3*, titled *Multicollinearity*. We also included a heatmap, created by calculating the average correlation values between all variables, including both features and the target, across 56 countries; the values were averaged to represent the overall correlation. A detailed explanation of the heatmap can be found under Section *3.2.4*, titled *Correlation.*

2. *I do not understand the correlations in Table 1. Why is fossil fuels not correlated to CO2 emissions? Please provide some level of detail in your variable naming such as "fossil fuel consumption", or "fossil fuel import" etc. Why do cropland and grazing land have high correlation with CO2 emisssion? Does this imply that the greater the area of cropland or grazing land, the greater the CO2 emissions? How does that make sense? Should renewables and nuclear not have negative correlations of greater magnitude with CO2 emissions ? Please describe and discuss in the manuscript.*

   3. "Fossil fuels" specifically refers to "percentage of electricity generated from fossil fuels" rather than total fossil fuel consumption. This distinction explains why its correlation with CO2 emissions appears counterintuitive - the percentage of electricity source does not directly reflect the absolute consumption of fossil fuels, which would indeed show a stronger correlation with CO2 emissions. Similarly, "renewables" and "nuclear" represent the percentage of electricity generated from these sources, not their absolute contribution to energy production. Therefore, their correlations with CO2 emissions may not show the expected negative relationship that would be seen with absolute measurements.

Regarding cropland and grazing land, their high correlation with CO2 emissions can be attributed to agricultural activities. Livestock grazing, in particular, is a significant contributor to greenhouse gas emissions through various mechanisms including methane production and land use changes. However, we have decided to exclude these features from our analysis due to issues with multicollinearity. A detailed explanation of the new variables employed and their measurements can be found in Section 3.2 (Data), where we provide detailed definitions and context for each feature in our dataset.

3. *What is the effect of weighing Taiwan 50:1 on the model's CO2 predicted emissions worldwide? It seems like you have artificially reduced the Taiwan class imbalance of*
<0.008% to approx, 40% due to this adjustment. How do you justify this? Class imbalances or unbalanced datasets are usually corrected by SMOTE or data autmentation or using inverse weighing class cross-entropy loss functions. Why were these accepted methods of reducing class imbalance not used? Please explain and describe in the manuscript.

To mitigate the effects of artificially adjusting the dataset, we employed a filter on our dataset that filters out extreme values that can negatively influence the models. Since CO2 emission values can vary significantly among countries, we filtered out countries with CO2 emissions more than 50 percent higher or lower than Taiwan's to prevent our models from being overly influenced by extreme values; for example, this prevents our models from predicting extremely high CO2 emissions for Taiwan after processing data samples from countries with high CO2 emissions like India and China. After applying this filter, we no longer need to apply additional techniques to exclude outliers and adjust the dataset. A more detailed explanation can be found under section *3.2.2*, named *Preprocessing*.

You state "……we implemented a novel approach that utilizes world data for training and Taiwan data for testing." This is only possible if the feature importance (not Pearson coefficient, see point 1) are the same for World and Taiwan data. Are they? Actually, this is not the procedure you are following in your explanation in the manuscript after table 3 where you train with 70% of world data and 30% of taiwan data in training and world data only or Taiwan data only for testing. This means that the dataset is trained on 70% of Taiwan data. Pleae remove any inconsistencies in explanation in the manuscript.

In the old manuscript, we trained the models on the world dataset, which also included Taiwan's data samples. The purpose of training the models on the world dataset, which consists of many countries, is to provide the models with a greater variety of samples to facilitate their learning and generalization. However, we adjusted our dataset to resolve some of the concerns made by the reviewer. Our new train dataset now only contains 56 countries, including 38 Taiwan samples, which have similar CO2 trends as Taiwan. More information about the dataset can be found under section *3.2*, named *Data*. In all our standard experiments, we trained our models on this dataset and tested our models on the dataset that contains only 20 randomly selected Taiwan samples. For comparison, we also trained our models on a separate dataset with 38 randomly selected Taiwan samples and tested them on the same 20-sample test set. A more detailed explanation of this comparison can be found under section *3.4*, named *Establishing Baseline Models*. In our updated analysis, we focused on a dataset of 56 countries with CO2 emissions comparable to Taiwan. Our evaluation of feature importances revealed similar patterns between this multi-country dataset and the Taiwan-only dataset. The performance improvements demonstrated in Section 3.4 using the 56-country dataset sufficiently validate the effectiveness of our proposed methodology.

4. *you state that the optimum training/testing split ratio is given by the square root of the parameters included. In this case your ratio should have been 4.3:1; i.e 77:23 split. Why then, did you choose a 70:30 split? Discuss and explain in the manuscript.*

We rounded it to a 70:30 split as it is the conventional practice. However, to ensure consistency with our research, we decided to use the exact values suggested by Joseph's paper (1). For our new dataset, the recommended train test split would be 67: 33, which we used to evaluate our models. More details of the steps taken can be found in Section *3.4*, named *Establishing Baseline Models.*

(1) V. R. Joseph, Optimal ratio for data splitting. Statistical Analysis and Data Mining: The ASA Data Science Journal. 15, 531–538 (2022). https://doi.org/10.1002/sam.11583

5. *What did you use to estimate feature importance (not Pearson coefficient) in your models? Gini? SHAP? reverse-one-at-a-time? Bootstrapping? K-fold-cross validation? Discuss in the manuscript. Present features and feature importance for all the models used.*

We estimated feature importance using SHAP values. SHAP is a comprehensive framework, derived from game theory, for assessing each feature's contribution to the prediction. We calculated SHAP values for our top three models—Gradient Boosting Regressor, FFNN, and Random Forest Regressor—as these models are the focus of this research. A more detailed explanation of our evaluation of feature importance can be found in Section 4.1, titled *Feature Importance*.

6. *I do not understand Table 4. What do True and False mean?*

In Linear Regression, "calculate intercept" and "copy x" are boolean hyperparameters that take in either True or False. If "calculate intercept" is set to True, the model will include an intercept term in the regression equation. This allows the model to account for the possibility that the relationship between the features and the target does not pass through the origin. Otherwise, the linear regression model assumes a relationship where the equation line passes through the center. On the other hand, if "copy x" is set to True, the model will copy the input data to ensure that the original data remains unchanged but also adds a computational burden to the model. However, after removing features that suffer from multicollinearity and training our models on the new dataset, Linear Regression is no longer one of our top models; thus its optimizations are no longer included in the manuscript.

7. *In figures 2 through 4, where does the actual CO2 emission data come from (is this worldwide or Taiwan specific?). For example, where do the specific data points originate from? Is this an autocorrelation graph that is increasing with respect to time*

   ?also need units tons?gigatons?

The original Figures 2 to 4 have been updated to Figures 5 to 7. In Figures 5 through 7, the actual CO2 emission data comes from the test set for Taiwan. The models were trained on a dataset including 56 countries, Taiwan included, and then tested on the test set, which consists of 18 samples of Taiwan's data. The graphs show how well the models' predictions match the actual values, with points closer to the ideal fit line indicating better predictions. It is not an autocorrelation graph. The unit is in million tons. We have modified the figures to include the unit and clarified the source of the CO2 emission data in Section 4 (Results and Discussion).

8. *Lastly, why do you need complicated ML models if a simple multiple linear regression of electric power, transporation and industry versus CO2 emissions will explain > 90% variance in the data (i.e. R2 > 0.9) see: https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions#:~:text=In 2021 and 2022%2C the,of the COVID-19 pandemic.*

The results of our experimentation show that while Linear Regression achieves an R2 score of 0.95, it is still relatively inaccurate compared to more complex models like Gradient Boosting Regression, FFNN, and Random Forest Regressor. Furthermore, out of all the models we tested, Linear Regression ranks 6th out of 7. Although the difference between an R2 score of 0.95 and 0.99 may not seem significant, it represents the gap between a prediction that is off by millions of tons of CO2 and one that is off by only a few thousand or even hundreds of tons. Therefore, more complex models like FFNN, Gradient Boosting, and Random Forest successfully minimize this inaccuracy because they can account for finer details in the dataset that are often overlooked by simpler models like Linear Regression. These improvements in the models can offer policymakers the most comprehensive predictions.

Important Additional Changes

We added a comparison between the models trained on the Taiwan-only dataset and the models trained on the dataset with 56 countries. This comparison shows that our proposed methodology is superior compared to the traditional methodology as the models trained on the dataset with 56 countries achieved higher results. A more detailed explanation of this comparison can be seen under section *3.4*, named *Establishing Baseline Models*.

We also added Figures 1 and 3 to more comprehensively show the procedures of our methodology.

Thank you for addressing my comments. Upon reviewing the revised manuscript, I have the following comments:

1.Your figure 3 a and b do not seem to align with your method. For example, in figure 3a, after the Training box, the box designations should be Taiwan only dataset and 56 country dataset (not Taiwan only testset). For 3b, after the testing box, the 56 country dataset box should be removed and arrows should directly go to the results boxes. The final box should be split into two boxes designated: Top models for Taiwan only and Top models for 56 country dataset.

2.You will need to make substantial changes to content distributed in incorrect categories. For example, you have presented Results under the “Materials and Methods” category. Please move all the results from the study and present them in the Results category. This includes the heatmap and its results, the results from multicollinearity……. etc. Note that you will still need to provide verbiage in the Materials and Methods section as to how these experiments were performed, justification etc.

3.Please include your response from point 9 of my earlier comments into the manuscript in the discussion section. The response "The results of our experimentation show that while Linear Regression achieves an R2 score of 0.95, it is still relatively inaccurate compared to more complex models like Gradient Boosting Regression, FFNN, and Random Forest Regressor. Furthermore, out of all the models we tested, Linear Regression ranks 6th out of 7. Although the difference between an R2 score of 0.95 and 0.99 may not seem significant, it represents the gap between a prediction that is off by millions of tons of CO2 and one that is off by only a few thousand or even hundreds of tons. Therefore, more complex models like FFNN, Gradient Boosting, and Random Forest successfully minimize this inaccuracy because they can account for finer details in the dataset that are often overlooked by simpler models like Linear Regression. These improvements in the models can offer policymakers the most comprehensive predictions." should hence be included in the discussion section of the manuscript.

4.Put down under "Limitations" that the results can be influenced by the selection of the different country dataset and by the arbitrary selection of 50% above or 50% below CO2 emissions when compared to Taiwan. Include in the verbiage that other models that explicitly deal with domain transfer such as Transfer Learning ML models with domain adaptation may be more suitable for your study.

5.Write the manuscript in third person, past perfect tense. Note that incorrect use of grammar and composition will significantly delay your manuscript processing at the copyediting stage, if accepted.

Open response questions Comments to author Thank you for addressing my comments. Upon reviewing the revised manuscript, I have the following comments:

1. Your figure 3 a and b do not seem to align with your method. For example, in figure 3a, after the Training box, the box designations should be Taiwan only dataset and 56 country dataset (not Taiwan only testset). For 3b, after the testing box, the 56 country dataset box should be removed and arrows should directly go to the results boxes. The final box should be split into two boxes designated: Top models for Taiwan only and Top models for 56 country dataset.

Yes, thank you for pointing that out. I have made the according changes to the figure. After reorganizing the manuscript, this figure is changed to figure 2. 2.

You will need to make substantial changes to content distributed in incorrect categories. For example, you have presented Results under the "Materials and Methods" category. Please move all the results from the study and present them in the Results category. This includes the heatmap and its results, the results from multicollinearity……. etc. Note that you will still need to provide verbiage in the Materials and Methods section as to how these experiments were performed, justification etc.

The research paper has been reorganized, with all the results, including the VIF values, heatmap, model performances, hyperparameter tuning results, and more, moved to the results section. In the Materials and Methods section, detailed explanations for each step and their justifications are still provided.

3. Please include your response from point 9 of my earlier comments into the manuscript in the discussion section. The response "The results of our experimentation show that while Linear

Regression achieves an R2 score of 0.95, it is still relatively inaccurate compared to more complex models like Gradient Boosting Regression, FFNN, and Random Forest Regressor. Furthermore, out of all the models we tested, Linear Regression ranks 6th out of 7. Although the difference between an R2 score of 0.95 and 0.99 may not seem significant, it represents the gap between a prediction that is off by millions of tons of CO2 and one that is off by only a few thousand or even hundreds of tons. Therefore, more complex models like FFNN, Gradient Boosting, and Random Forest successfully minimize this inaccuracy because they can account for finer details in the dataset that are often overlooked by simpler models like Linear Regression. These improvements in the models can offer policymakers the most comprehensive predictions." should hence be included in the discussion section of the manuscript.

This explanation has been added under the Performance Analysis section of our paper. We explained that while Linear Regression achieved a high R^2 score, it is still relatively inaccurate compared to our top models. Furthermore, although the gap between the models may not seem significant, it represents the difference between a prediction off by millions of tons of CO2 and one that is off by only a few thousand or even hundreds of tons. Therefore, more complex models successfully minimize this inaccuracy because they can account for finer details in the dataset that are often overlooked by simpler models like Linear Regression.

4. Put down under "Limitations" that the results can be influenced by the selection of the different country dataset and by the arbitrary selection of 50% above or 50% below CO2 emissions when compared to Taiwan. Include in the verbiage that other models that explicitly deal with domain transfer such as Transfer Learning ML models with domain adaptation may be more suitable for your study.

In the conclusion section, we mentioned that the results can be influenced by the arbitrary selection of countries with emissions ranging from 50% more to 50% less than those of Taiwan and stated that future work can address this problem by utilizing models specifically designed for domain transfer. We provide Domain-Adversarial Neural Networks (DANN), Transfer Learning, and Generative Adversarial Networks (GANs) as examples.

5. Write the manuscript in third person, past perfect tense. Note that incorrect use of grammar and composition will significantly delay your manuscript processing at the copyediting stage, if accepted.

We have changed the grammatical structure of the manuscript to the third person and past perfect tense.

Thank you for addressing my comments. Accept.