Using Basic Machine Learning Methods to Predict TGA Data Points and Plots

Ivan Matyushov

I. Research Objective

In materials science, thermogravimetric analysis (TGA) is a common thermal analysis tool used in many labs for different applications. TGA uses a furnace to burn and degrade any small sample piece or powder that is placed. A balance in the tool then measures the sample's weight throughout the temperature ramping process. TGA is useful for understanding the degradation of a material under temperature, thermal stability, reaction rates, material composition, and comparing materials' performances. The data is often rather simple as it is shaped like a sigmoid with some possible bumps where other materials start being degraded. Differentiating this data and finding critical points can be a tedious process for researchers.

By utilizing a large language model (LLM), one can create a way to quickly process TGA datasets, create a plot, and perform standard analyses such as: finding onset temperatures, temperatures of peak degradation, activation energies, and temperatures of phase transitions. Creating this tool will simplify the data analysis process by enabling scientists to instantaneously extract data points rather than needing to parse through the data and run analyses. With future developments, such a model could expand to predicting TGA curves of unknown materials and predicting the composition of an unknown material based on its TGA profile.

In this project, I primarily focused on providing a starting point of data that can be predicted from TGA using basic machine learning methods. I primarily focused on using variants of SVMs to guide predictions of basic data. The goal is to test some basic predictions, gauge their accuracy, and understand whether more complex predictions could be made. Initially, I hoped to build a model that would be able to draw some useful analyses from TGA data but a lack of large datasets and the complexity needed to have an accurate system made it difficult to build a robust model in the project's timeline.

II. Related Work

LLMs have an important place in materials science and lots of research is being done to implement them. Some of the most powerful applications researchers envision involve predicting properties, automating processes, and extracting text from articles much like ChatGPT[1,2]. Researching these papers gave me some initial ideas for potential projects to start with.

Afterward, I delved into specific research papers on machine learning methods that were already applied to TGA datasets. This helped guide me along the methodologies and specific predictions to make in my machine-learning model. The research published by Dr. Zaiffullizan et al[3], they used artificial neural networks (ANNs), support vector machines (SVMs), and decision trees (DTs) to draw predictive data plots of the data, find degradation amounts at certain temperatures, and predict activation energies. The research separately trained the three types of models and tested them against each other with the experimental results. For each type, they trained six models for different conditions of the experiment. All their data was split with 80% for training

and 20% for testing. Dr. Zaiffullizan's work provided the clearest path to follow for my project. I primarily adopted the use of SVMs.

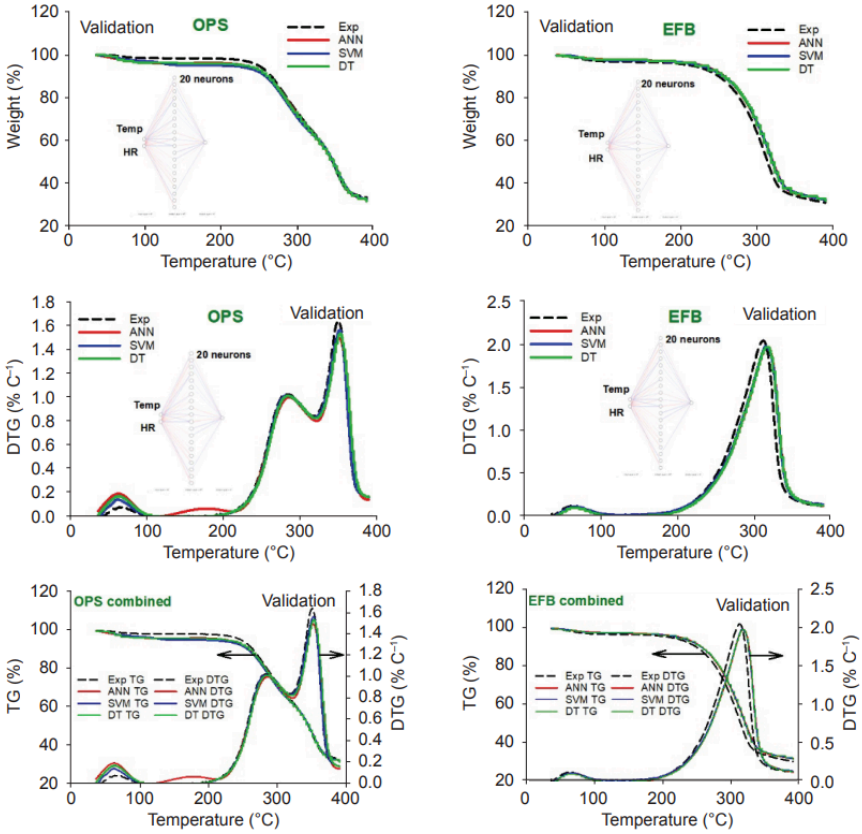

Figure 3. Comparison of ANN, SVM and DT validation models to predict the TG and DTG (derivative thermogravimetric) profiles of OPS and EFB biomass pyrolysis at a maximum temperature of 400°C and heating rate of 20°C min⁻¹.

Figure 1: Plots of data that Zaiffullizan et al were able to develop from their predictive machine learning models.

Dr. Watts' paper[4] presented the results of TGA to then predict how certain particles would act in a torrefication reactor. His paper presented the possibility of using MATLAB machine learning tools for my project but I ultimately went with Python. Dr. Florentino de Carvalho's research[5] implemented, ANNs, random forests (RFs), and DTs to predict the ignition, burnout, and activation energies of poultry manure samples. They proved that AI could be used to accurately optimize conditions for use in combustion without the need for constant experimentation. Dr. Ali's[6] research looked at using ML methods, specifically random forests and gradient boosting regressions, to predict TGA plots for various electronic waste products in an attempt to remove the need for electronics and experimental analysis to derive the TGA plots.

For this project, I attempted to primarily match some of their analytical results for a set of biochar samples data that I found online that came from work by Ariena Villalobos and the group she worked with[7]. They primarily tested whether soil integrated with certain types of biochar would have its thermal and hydrophysical properties improved. My project primarily attempted to see if certain key temperature data points of

the biochar's TGA data could be predicted by AI and whether the TGA plots could be accurately predicted with experimental variations.

III. Hypothesis/Research Questions

Most TGA data needs to be analyzed using specific software designed for the tool that it came with. This software requires manually clicking through different sections of data and clicking on different tools to extract useful information from TGA data. This becomes tedious when having numerous samples to comb through and extract data from. Building a model that automates the data analysis process will cut out tedious time for researchers who can then focus on preparing future experiments and writing out papers. This model could be trained to be more accurate than researchers or their analysis software making this process both more efficient and accurate. Additionally, like some of the TGA papers mentioned, accurately predicting the TGA data of future samples could make future experimentation unnecessary and sample parameters could be precisely optimized based on predictions.

For the biochar data I found, I noticed that they structured their data key data points around certain temperature points where key points in the soil process occur. Reducing the degradation at these points means that the soil is becoming more robust and able to withstand harsher thermal conditions. I wanted to see if these temperature points could be accurately classified for future data based on the trends in previous data. Predicting this data accurately could then, based on inputted parameters of new biochar samples, could then allow for faster optimization of the biochar samples. I also looked at predicting the plots of future TGA data to attempt to see if, from parameter changes, future TGA data could be feasibly predicted. Much of this work still falls along the ideas of what other papers have done but this work is adding to the broader literature by tackling it with soil samples that have biochar added to them.

In order to conduct the research I had to guide myself with the research questions below:

Research Questions:

1. How accurate will this model be?
2. How should I structure and parse the data for analysis?
3. Will this model be able to accurately predict any temperature points?
4. Will this model be able to draw up its own TGA data and plots based on prior data?
5. Will this model need substantial data or will it be able to work with small datasets?
6. What will be the most effective libraries and tools for automating this process?
7. What sort of methods or actions will I need to take to improve the analysis of the data?

IV. Methodology

1. Data Collection and Processing

The data I used for this project came from Dr. Nathan Howell[8] as part of a research project done by Ariena Villalobos[7] testing soil quality after adding biochar under certain conditions to the soil. The online dataset consists of 16 samples of TGA data with each sample varying by pyrolysis temperature (450m or 600 C) and the type of wash performed on them (DI or Acid). For the classification of key temperature points, I used all 16 data sets and parsed out the mass percent and percent loss measured at these key temperature points (0, 220, 315, 400, 900, 1000 C). I created a separate Excel sheet that recorded the mass and temperatures given along with the experimental parameters that changed from sample to sample.

For the prediction of TGA data, I mostly just utilized the first set of samples (1-4) to see whether predicting TGA data would yield any value. Not much postprocessing of the data was needed. I mostly just extracted the mass percent and temperature data of these samples into a separate excel sheet that made for easier analysis and plotting. The data was all placed in just two columns with the sample sets 1-4 data coming one after another.

2. Model Design

Similar to what Dr. Zaiffullizan[3] did, I primarily focused on using SVM tools for my predictive models. The SVM tools I used were support vector classifiers (SVCs) and Support Vector Regressions (SVRs). With the classifier I incorporated data of mass percent, percent loss, temperature, wash type, and the pyrolysis time. I implemented a basic SVC found in the sklearn.svm package of Python. The data was organized like the testing set shown below:

[75 rows x 5 columns]

	Temp	Mass perc	perc_loss	p_time	wash
74	315	52.51066	1.5120	60	1
75	1000	10.96435	3.6910	60	1
76	900	76.36124	15.4340	60	1
77	315	88.87181	2.6100	20	0
78	220	92.67370	7.3260	60	0
79	400	85.49078	3.8790	40	1
80	400	91.79500	0.9470	60	1
81	1000	60.30823	1.8310	10	1
82	315	92.95380	0.6990	20	1
83	220	91.60250	8.3975	10	1
84	900	44.08716	39.9220	20	0
85	900	76.36124	15.4340	60	1
86	0	100.00000	0.0000	60	0
87	900	69.48259	20.9320	20	0
88	900	13.39272	21.8230	60	1
89	315	89.09322	2.5090	10	1
90	315	87.50231	4.1310	10	0
91	400	88.59649	2.3960	60	0
92	0	100.00000	0.0000	10	0
93	400	84.00985	4.8620	20	0
94	315	92.02618	0.8570	20	0
95	0	100.00000	0.0000	40	0
96	220	91.39228	8.6080	40	1
97	1000	73.49411	2.8670	60	1
98	220	92.88312	7.1170	20	0

Figure 2: Testing dataset used to show how appearance and structure of data used.

For the prediction of TGA data, I used SVRs from the sklearn.svr package that performed regressions on the TGA data in the radial basis function (RBF) mode. In SVRs, the RBF mode refers to the use of the Radial Basis Function (RBF) kernel to map the input data into a higher-dimensional feature space. This is particularly useful for capturing complex, non-linear relationships between the input data and the target variable. TGA data is often non-linear so I looked to capture this non-linearity with the RBF mode. I first tested the SVR with random data points of the TGA data plot and see if a representative plot could be drawn out. I then looked at using SVRs to test whether segments of the plot or full TGA plots could be predicted from the prior data that was fed. One potential lagging point of my SVRs was that they only incorporated temperature and mass percentage columns for TGA data rather than multiple columns of conditions like the SVC model did.

3. Training and Analysis

For these models, I will break the data into a ratio of 75:25 of training to testing data. Since the validation needed is not very in-depth and complicated, there will not need to be as many samples dedicated to testing. Once the proper procedures are performed on the data, we will compare the model results to the real values that we expect to see from the data. I took into account the R^2 values, and mean bias errors (MBEs) to evaluate the accuracy and level of bias in the trained model.

To evaluate the accuracy of the predicted temperature classifications, I took the predicted data and matched it to the actual data on a bar graph along with calculating the general accuracy. The bar graph provides a visual representation of the accuracy.

To evaluate the accuracy of the predicted TGA data, I used the previously mentioned R^2 values and mean bias errors to calculate the data's errors. I also visually plotted the predicted data against the training and testing sets to give clear visuals comparing the training, testing, and predicted datasets.

V. Experimental Setup

To set up the data for temperature classification prediction, I parsed all the data from the downloaded set of data into a separate sheet. In the original data, the targeted data for this section appeared like so:

		Sample 1		Sample 2		Sample 3		Sample 4		Reference	
	Temp C	Mass %	% Loss	Mass %	% Loss	Mass %	% Loss	Mass %	% Loss	Mass %	% Loss
mois+vol-hemi	0	100	8.36634	100	8.51781	100	6.90576	100	7.3263	100	22.1354
	220	91.63366		91.48219		93.09424		92.6737		77.8646	
hemi-cell divide	220	91.63366	4.13135	91.48219	2.61038	93.09424	1.71638	92.6737	1.68072	77.8646	25.35394
	315	87.50231		88.87181		91.37786		90.99298		52.51066	
cell-lignin	315	87.50231	7.60772	88.87181	4.86196	91.37786	2.83263	90.99298	2.39649	52.51066	21.36083
	400	79.89459		84.00985		88.54523		88.59649		31.14983	
ash only	400	79.89459	52.91727	84.00985	39.92269	88.54523	25.41555	88.59649	19.87688	31.14983	17.75711
	900	26.97732		44.08716		63.12968		68.71961		13.39272	
Complete	900	26.97732	10.60915	44.08716	7.61891	63.12968	3.86243	68.71961	2.67462	13.39272	2.42837
	1000	16.36817		36.46825		59.26725		66.04499		10.96435	

Figure 3: Original dataset showing mass % at key temperature points.

I took this data from all 16 of the samples and placed it into an Excel sheet of the structure shown in Fig. 2. At first, I started my analysis on the data with it being ordered like it is by temperature and time organized increasingly. I then realized that this would not create a random analysis and prediction of the data. To fix this, I implemented a shuffling function to the data prior to analysis by the SVC. I made the wash type into binary data since its either DI or Acid wash with DI being denoted as 0 and acid being denoted as 1. I also took out the pyrolysis temperature as I found that it often reduced the accuracy of the predicted classification. Throughout the analysis, I started with only two columns of data then added in pyrolysis time and wash type to improve the classification accuracy. The SVC had some relative accuracy but not enough to conclude the SVC was optimal for making predictions. Likely, more data and more iterations on all the datasets were needed for better accuracy.

For the SVR analysis performed, as mentioned I split the data into simply mass and temperature data points where the data stems from the gradual TGA measurements generally taken. This data is structured as so:

Sample 1			Sample 2			Sample 3			Sample 4		
Temp, C	Time, min	Mass, %	Temp, C	Time, min	Mass, %	Temp, C	Time, min	Mass, %	Temp, C	Time, min	Mass, %
22.539	0	100.0038	22.54	0	99.99729	29.935	0	99.99892	30.378	0	99.99831
27.539	1.51742	98.398	27.54	1.31119	98.38669	34.935	2.40219	97.82341	35.378	2.37618	97.78803
32.539	2.25264	98.07737	32.54	2.10783	98.05232	39.935	3.18584	97.46774	40.378	3.16906	97.41587
37.539	2.83514	97.89439	37.54	2.73228	97.86749	44.935	3.77459	97.3038	45.378	3.76551	97.2601
42.539	3.34513	97.75075	42.54	3.26873	97.68539	49.935	4.27962	97.17159	50.378	4.27407	97.14646
47.539	3.81389	97.60597	47.54	3.75236	97.5078	54.935	4.73652	97.04776	55.378	4.73539	97.00569
52.539	4.25462	97.44007	52.54	4.20207	97.3272	59.935	5.16375	96.91651	60.378	5.16681	96.87381
57.539	4.67647	97.21968	57.54	4.62781	97.10909	64.935	5.57049	96.73194	65.378	5.57706	96.71224
62.539	5.08247	96.97322	62.54	5.03665	96.86133	69.935	5.96441	96.55231	70.378	5.97427	96.51904
67.539	5.47875	96.71329	67.54	5.43417	96.58874	74.935	6.35049	96.37239	75.378	6.36332	96.32317
72.539	5.86786	96.43835	72.54	5.8237	96.29296	79.935	6.73023	96.15711	80.378	6.74582	96.08314
77.539	6.25258	96.129	77.54	6.20888	95.98352	84.935	7.10493	95.93975	85.378	7.12321	95.83328
82.539	6.63395	95.76479	82.54	6.59042	95.64066	89.935	7.47596	95.71607	90.378	7.49602	95.58309
87.539	7.01238	95.41215	87.54	6.97011	95.32904	94.935	7.844	95.47842	95.378	7.86573	95.34924
92.539	7.38818	95.08503	92.54	7.34839	95.01281	99.935	8.21138	95.28865	100.378	8.23458	95.11564
97.539	7.76433	94.74721	97.54	7.727	94.70624	104.935	8.58023	95.09427	105.378	8.60495	94.9024

Figure 4: Structure of dataset for TGA prediction. Data goes further until 1000 C.

To outline the clear split use of SVRs, the first was for random data pulled across the data, the second SVR was for predicting a segment of TGA data, and the third SVR

was for predicting a full TGA plot with an SVR. For the SVRs, I had to reshape the data using x_t.values.reshape(-1,1) function to put the data into 2 column array as SVRs do not take single-column arrays as inputs. For the first SVR where I used it to predict the shape of the graph from random data points, I precisely did a 75:25 split of the data while, with the other SVRs, I used index ranges of data that were ~75:25 splits. The prediction with the random scattered points of data proved to be accurate but this was the least necessary prediction to make. The SVRs that predicted a segment of data and a full set of TGA data had no notable accuracy to report.

VI. Results and Comparison

1. Temperature Classification:

   The accuracy for the temperature classification varied with each shuffle iteration of the data so the accuracy was also inconsistent. The accuracy ranged from 0.28 to 0.52 with usual accuracies seen being 0.36, 0.40, and 0.44. The model did best at predicting the high and low-temperature points given a mass value. These data points were more extreme and had more defined value ranges that the model could more accurately classify. Once the classifications became more blurred and experimental conditions shifted the mass percent values differently, there was more difficult in classifying the temperatures accurately. The SVC also tended to over predict one or two temperature values very often compared to how present it was in the dataset. Below is the bar graph depicting where the classifications fall into:

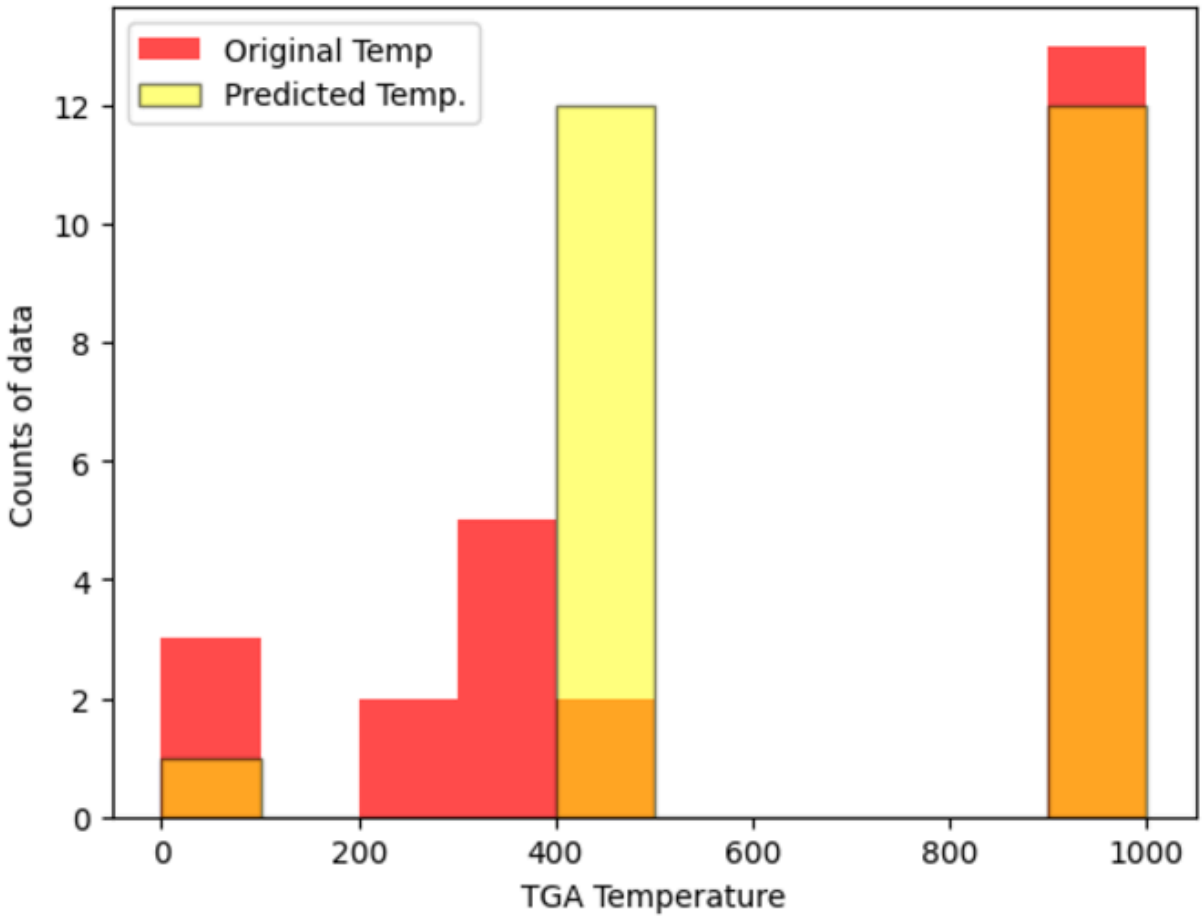

Figure 5: Bar graph of predicted data points compared to actual data points in the dataset. Orange shows areas of overlap.

The high yellow bar often appears for one of the temperature sets where the SVC is wrongly predicting many of those points to exist. That could be due to the binary nature of SVCs where they need to be adjusted differently for multi-class datasets like this one. Also, there is likely a lack of data for this model to be trained accurately. I believe the accuracy figure give some promise but I am only dealing with 100 data points total across the 16 TGA samples. With double or triple the data points of the same nature, this SVC could show more consistent accuracy. Perhaps varying the C and epsilon parameters could have yielded better results too.

2. TGA Plot prediction with SVRs:

As mentioned, I used three SVRs for slightly different applications to test the accuracy for different purposes. When asking the SVR to predict the upcoming set of data from all the prior data, the SVR proved to be the least accurate. Only when taking random data points across the dataset did the SVR prove to have some accuracy.

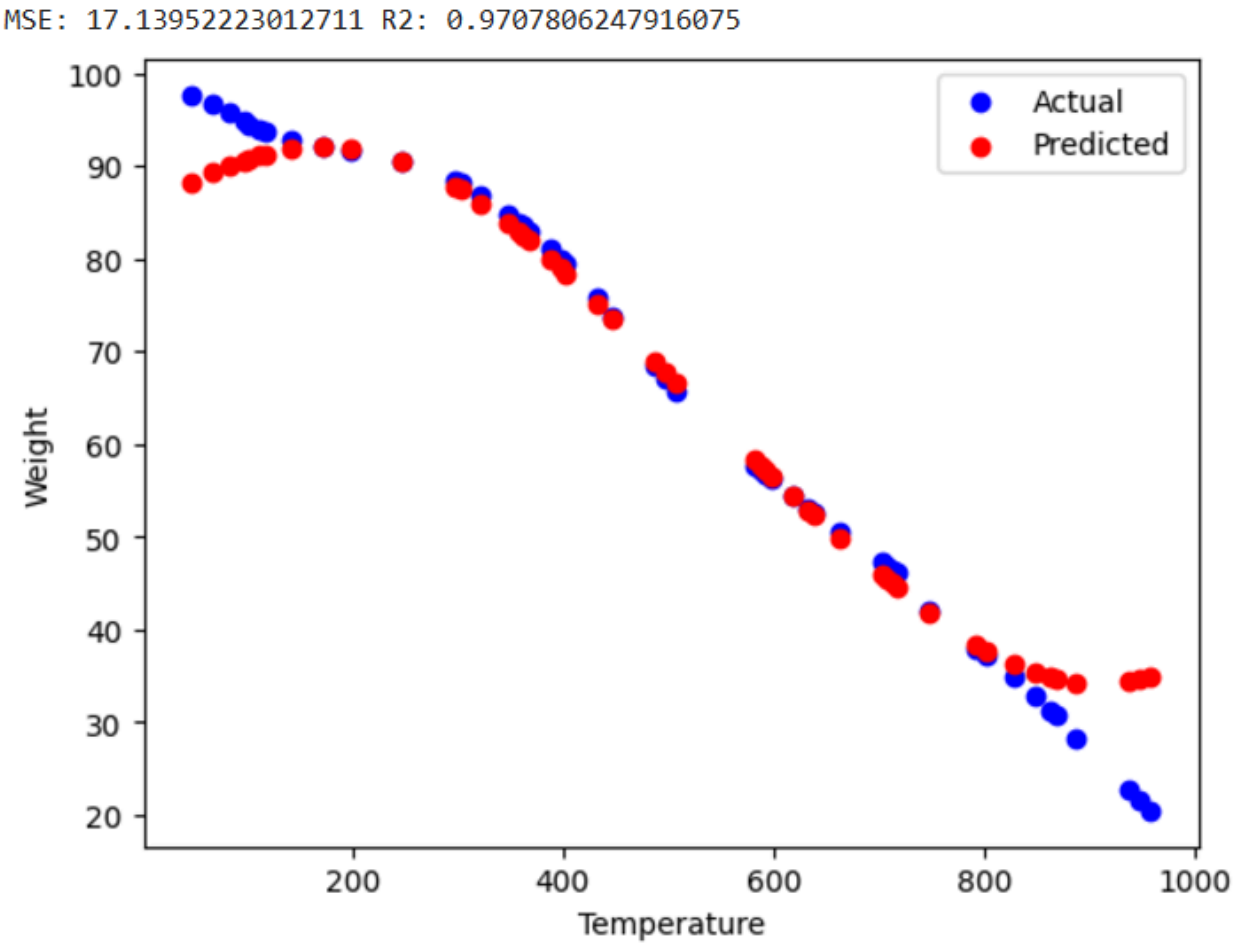

Figure 6: Plot of predicted data given 75% of training data. R^2 and MSE error values are depicted above the plot.

The R^2 value is close to 1 so this reinforces that the SVR here is rather accurate. Below are the plots from predicting segments of data.

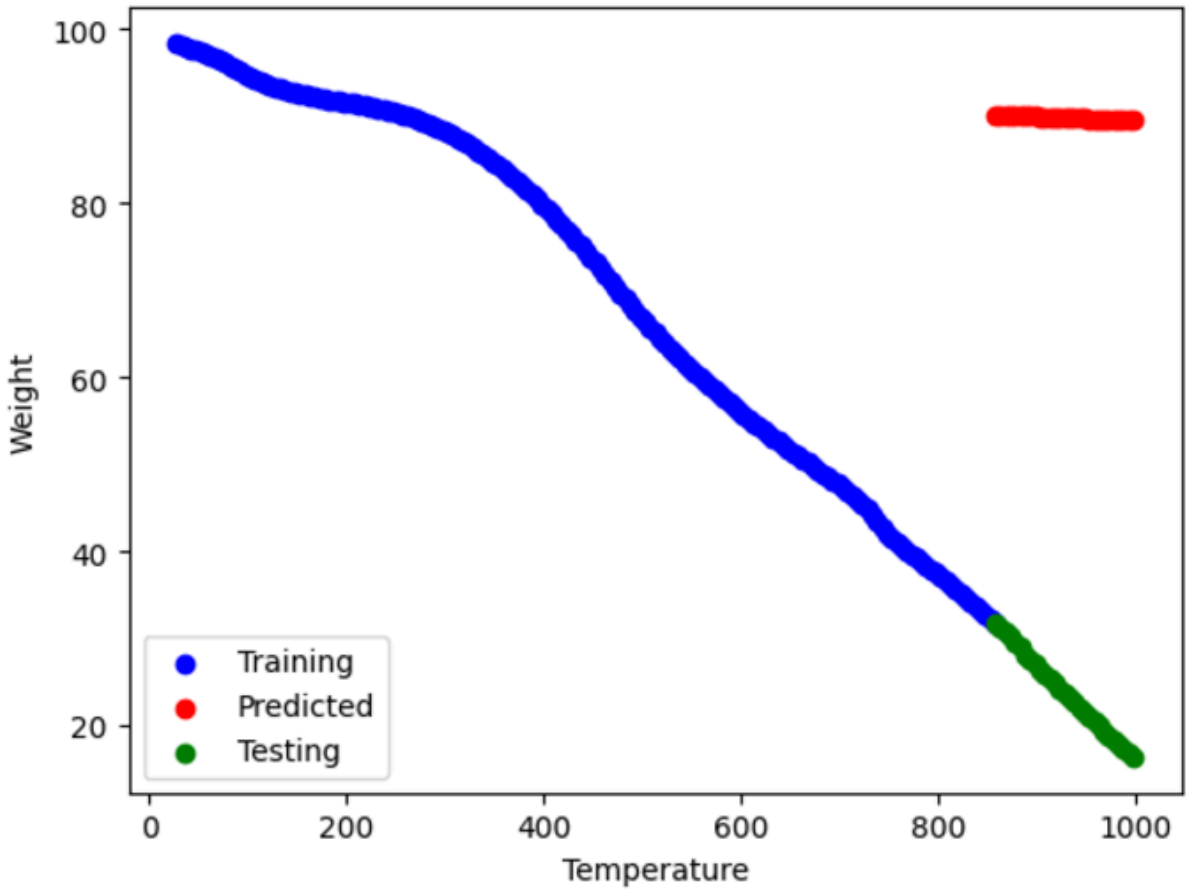

Figure 7: Plot of predicted data vs. tested data for a specific segment that was trying to be predicted.

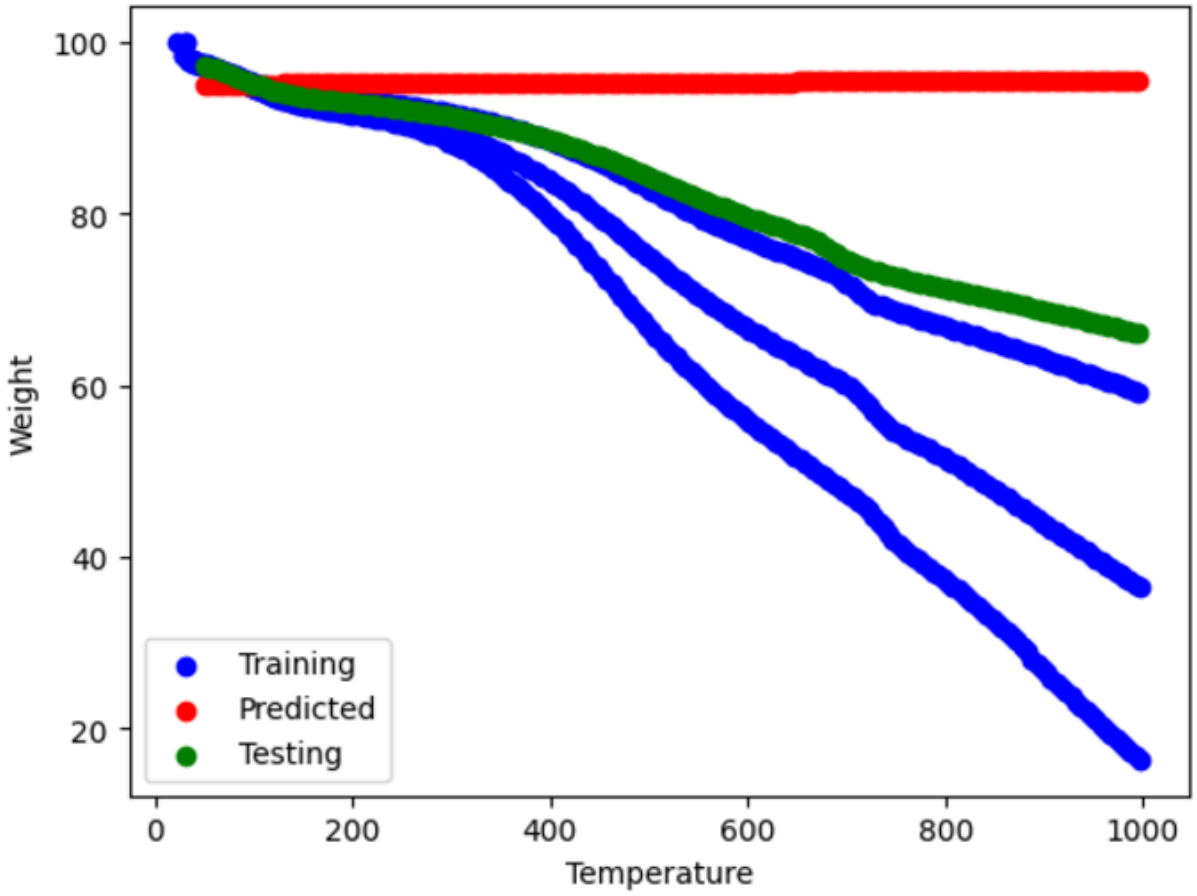

Figure 8: Plot of predicted data vs. tested data for an entire segment of TGA data

Clearly, the SVRs trying to predict segments of TGA data from prior data were quite inaccurate. The SVR seemed to mostly the initial linear patterns through and not adopt any of the datasets' patterns going downwards after certain temperatures. Despite using the RBF kernel for the fits, the SVRs remained linear and flat. I am unsure of what exactly resulted in the SVR's failings.

There could be issues with the scaling of the data as I only reshaped it to fit the prediction. There could also be problems with the lack of other data involved and the way the data is structured with all 4 of the samples going into one dataset. The pattern and shape of the data may be difficult for the SVR to accurately predict. Initially, I expected the prediction of a full dataset to be inaccurate due to the prediction demand being heavier, but I thought the segmented prediction would work with some accuracy. Both seem to have the same issues that need to be resolved for proper prediction making.

Since the papers I read did not include much information about the SVM used or precisely how they developed it to work within reason so I had little clear guidance here. I believe that making my SVM incorporate more columns of data rather than only the mass percent and temperature columns like derivative loss percent data or pyrolysis time could make the predictions more accurate. Also, many of the researchers used and applied

SVMs to different conditions of the reaction so having more specificity could improve the SVRs' predictions.

VII. Conclusions

Overall, I believe that the SVC model gave the most promise for yielding successful. The SVC showed some level of accuracy that I believe could be improved with more data being collected. Increasing the number of samples so the number of data points increases to at least 300 data points could improve the accuracy. The added data points could also shrink the variability in error when rerunning the SVC. With the accuracy of the classifiers resolved, the SVC could be useful in performing quick analyses of future TGA datasets with reasonable accuracy.

The SVRs clearly proved to be the least accurate. More substantial work needs to be done to develop these further to reasonable accuracy and use for the future. As mentioned previously, more columns of data, data reorganization, and a better understanding of SVMs would help to modify the model accordingly. Tweaking the C and epsilon parameters in Python could also be a next stage development for the model. Tweaking these parameters can reduce the overfitting or underfitting of the model. Currently, the SVRs seems to be overfitting for the initial part of the graph and extrapolating that out across the whole temperature range rather than noticing the actual shape of the plot.

Going forward, more features need to built onto the model for it to be any level of useful. In this project, I did not actually test out ANNs or DTs like many researchers have done with TGA. Dr. Zaiffullizan's work found that DTs made the most accurate predictions of data points and TGA plots for their dataset. I did not fully understand how to implement a decision tree for the purpose of TGA analysis. With more time spent, I could likely see a way to code in a DT and also an ANN then compare their results. Additionally, more features than just predictive plotting or temperature classification could be added like predicting activation energies, onset degradation temperatures, and peak degradation points simplifying the analysis process.

The research that the other papers referenced have done provides a lot of promise in terms of using AI to lessen the need for frequent, precise experimentation in order to optimize the parameters for an experiment. With this model's features and accuracy expanded according to the lab group's work, the lab group could optimize their experiments and more easily find optimal conditions for any benefactors of their research. Not only are more samples needed but different conditions and setups are needed so that there is a broad base of expected conditions that are then lightly tweaked by variations in parameters and more easily predicted.

VIII. Acknowledgments

I would like to thank Dr. Nathan Howell for publishing his dataset publicly for use in my project to better learn machine learning applications. I would like to thank Dr. Haiyan Wang at Arizona State University for teaching this class and providing the necessary resources to create this project and model. Doing this project provided me with wonderful exposure to machine learning methods and how they can be used in my field.

IX. Team Roles and Responsibilities

I will be the sole person responsible for all the work on this project and will do it entirely on my own. I found the data myself and wrote the entire code for the project.

X. Data Availability

The dataset used for this project comes from research done by Ariena Villalobos et al[7] and it was published by Dr. Nathan Howell[8]. The dataset was found at: https://figshare.com/articles/dataset/Thermogravimetric_analysis_TGA_original_data_files_biochar_samples_1-16/25513396'

The code for the project can be seen at: https://github.com/bigboivan/MAT422/blob/main/MAT422_project.ipynb

XI. References

1. Jablonka, K. M. et al (2023). 14 examples of how LLMs can transform materials science and chemistry: A reflection on a large language model hackathon. *Digital Discovery*, *2*(5), 1233–1250. https://doi.org/10.1039/D3DD00113J
2. Lei, G., Docherty, R., & Cooper, S. J. (2024). Materials science in the era of large language models: A perspective. *Digital Discovery*, *3*(7), 1257–1272. https://doi.org/10.1039/D4DD00074A
3. Zaifullizan, Y. M., Kuan, L. M., Salema, A. A., & Ishaque, K. (2023). Comparison of artificial intelligence models to predict oil palm biomass pyrolysis and kinetics using thermogravimetric analysis. *Journal of Oil Palm Research*, *35*(1), 86-99. https://doi.org/10.21894/jopr.2022.0048
4. Watts, J., Potter, A., Mohan, V., Kumari, P., Thengane, S.K., Sokhansanj, S., Cao, Y. and Kung, K.S. (2023), Proxy quality control of biomass particles using thermogravimetric analysis and Gaussian process regression models. Biofuels, Bioprod. Bioref., 17: 1274-1289. https://doi.org/10.1002/bbb.2504
5. Carvalho, R. F. de, Pasolini, V. de H., Breciani, J. G. F., Costa, A. B. S., & Sousa, R. C. de. (2024). Poultry manure combustion parameters to produce bioenergy: A thermogravimetric analysis by isoconventional models and machine learning. *Case Studies in Thermal Engineering*, *53*, 103757. https://doi.org/10.1016/j.csite.2023.103757
6. Ali, L., Sivaramakrishnan, K., Kuttiyathil, M. S., Chandrasekaran, V., Ahmed, O. H., Al-Harahsheh, M., & Altarawneh, M. (2023). Prediction of Thermogravimetric Data in the Thermal Recycling of e-waste Using Machine Learning Techniques: A Data-driven Approach. *ACS Omega*, *8*(45), 43254–43270. https://doi.org/10.1021/acsomega.3c07228
7. Villalobos, A., Kafle, A., Todd, J., Singh, M., Howell, N., & Deb, S. K. (2024) Effect of Cotton Gin Waste Biochar Amendment on Soil Hydrophysical Properties and Cotton Responses Under Deficit Subsurface Drip Irrigation [Abstract]. ASA, CSSA, SSSA International Annual Meeting, San Antonio, TX. https://scisoc.confex.com/scisoc/2024am/meetingapp.cgi/Paper/159204

8. Howell, Nathan (2024). Thermogravimetric analysis (TGA) original data files, biochar samples 1-16. figshare. Dataset. https://doi.org/10.6084/m9.figshare.25513396.v1