Machine Learning Algorithms for Consumer Plastics Identification and Sorting

By Alex Ko

October 13, 2022

Abstract:

Seeking a solution to maintain plastic recycling costs while increasing the output of reusable material, this paper poses the question "what type of machine learning algorithm is most suitable for a plastic identification system in consumers' homes?" It investigates five total machine learning algorithms to determine which one best balances accuracy, management of resources, and time efficiency, ultimately arriving at the conclusion that a Support Vector Machine that uses a Polynomial Kernel is the best algorithm and serves to demonstrate that algorithms such as the ones analyzed have become advanced enough for more efficient, AI driven systems to replace those of the status quo.

Introduction:

Plastics are extremely robust and durable substances sourced from fossil fuels, a quality that makes them extremely popular for building, packaging, etc… But after these plastics have served their purpose or need to be replaced, this same longevity that made them good for so many years also makes them extremely bad for the environment, becoming litter for up to 500 years (Chariot Energy).

In order to combat these negative effects, many will point to recycling as the solution, recognizing that the reuse of plastics (in different forms) will keep them both out of landfills and

oceans. And while the thought behind recycling is a very promising one, current recycling systems within the US often fail to recycle even the majority of recyclable plastic waste. According to the United States Environmental Protection Agency, only 8.7% (EPA) of the total plastics generated in 2018 were recycled (similar to the international recycling rate of 9% (Parker)) and only 18.5% (EPA) of PET plastics (a type of plastic that is 100% recyclable (PET Resin Association)). While these low numbers stem from the meager 35% of people in the US who actually recycle, there is no reason that these two statistics should be separated by almost 50% relative to each other (Morgan). The reasoning behind this massive difference has to do with how plastic is recycled.

In order to recycle plastic, recycling plants need to melt plastics at very specific temperatures to break apart distinct chemical bonds. The issue with this system, however, is that different plastics have different melting points, meaning that plastics must be sorted by type before they can be recycled. Most often, recycling plants have human employees conducting this job; picking out recyclable pieces of plastic and sorting them together as they pass by on a conveyor belt. Unfortunately, these workers often leave much recyclable material on these conveyor belts, which send any remaining items (intended only to be trash) to incinerators or landfills. And even though recently developed sorting machines significantly reduce the amount of recyclable material that is missed within recycling facilities, these machines are often expensive and out of reach for many smaller recycling plants.

With all this in mind, it becomes clear that sorting plastics is one of the biggest areas of difficulty in the recycling process, justifying a look into potential methods using AI to simplify this step.

Background:

To address this issue, we propose that plastics are sorted within consumers' homes using machine learning computer vision models. By doing so, much of the plastics arriving at recycling plants will already be sorted, meaning that human sorters will have an easier time in sorting out recyclable vs. unrecyclable materials.

With this solution however, the question arises of how these artificial intelligence algorithms geared towards recycling will be accessed by everyday people and for a reasonable price. As an example, recycling companies could deploy special recycling cans where plastics are scanned by a camera powered by artificial intelligence before being sorted into smaller bins within the recycling device for future pickup by recycling trucks, as shown in the figure below.

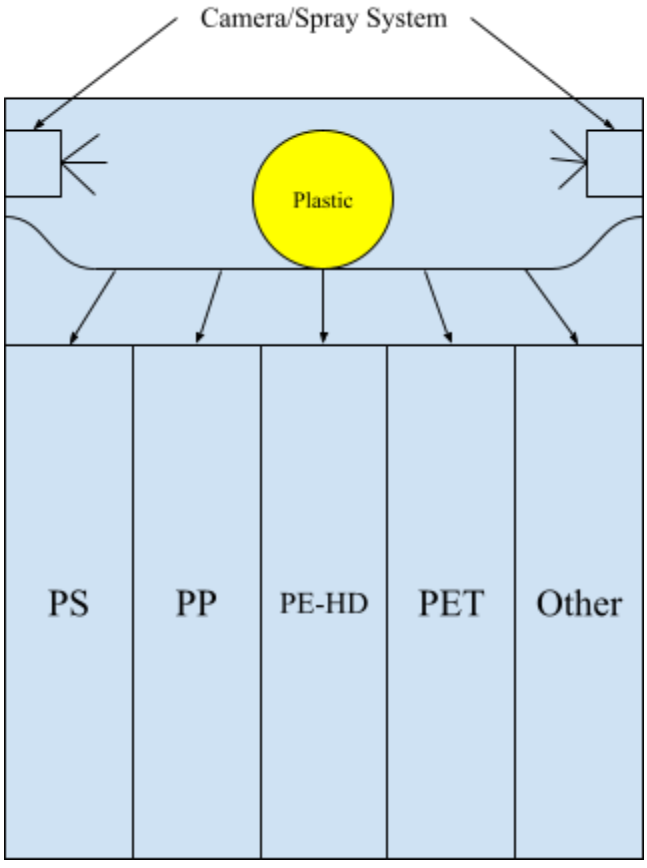

Figure 1: Sketch of Potential AI Powered Recycling Can

But with a system such as the one described, it quickly becomes evident that it will not be cheap to implement due to the complicated recycling cans and additional costs of setting up new recycling infrastructure for the novel system, such as modified recycling trucks. While this may be true in the short run, we believe that the ultimate costs of integrating a system that makes use of the proposed solution will result in an equal if not great profit margin for recycling companies stemming from a reduced need for human sorters and a greater volume of recycled plastics being produced that can then be sold to plastics manufacturers. And not to mention the potential benefits that keeping plastics out of landfills and incinerators has for the environment.

With all of that said however, the goal of reducing the cost of the system is still a high priority. Particularly, the recycling cans that consumers will be using have a high potential for cost reduction. On the one hand, the hope is that as close to 100% of the plastics deposited in the cans will be sorted correctly. However, given the exponential increase in required memory, compute power, and time for machine learning algorithms as the accuracy approaches 100%, the goal for the remainder of this paper is to investigate a few different machine learning algorithms and weigh the overall effectiveness of them depending on how well they balance these factors.

Dataset and Data Preprocessing:

This project utilized the WaDaBa Images Database that features 4000 images of 100 different objects, each photographed 40 times in different conditions. Each image was classified based on the type of plastic, its color, the type of light that it was taken in, how deformed it was, its dirtiness, whether it had a cap or lid, whether it had a plastic ring, and the random position that it was in. All of these features were described in the names of the images that they were tied to.

Upon incorporating the images into Google Colaboratory, a few key preprocessing steps needed to be taken before the data could be reasonably used. We first compressed each of the images to a resolution of 224 by 337 pixels before cropping them to a resolution of 224 by 224 pixels (square) in order to decrease the amount of RAM necessary throughout the training and testing of the machine learning algorithms. Next, the images were converted into ".npy" files while simultaneously being added to a numpy array of size [4000, 224, 224, 3], allowing us to easily access the data later. Finally, given that the algorithms we worked with required two dimensional array inputs, we reshaped the current array to an array of size [4000, 150528].

Additionally, since this project revolves strictly around identifying the type of plastic that something is regardless of other factors, we extracted only the plastic type from the names of the files before adding them to a different numpy array, thus giving us the labels for our data, which are represented in figure 2 below.

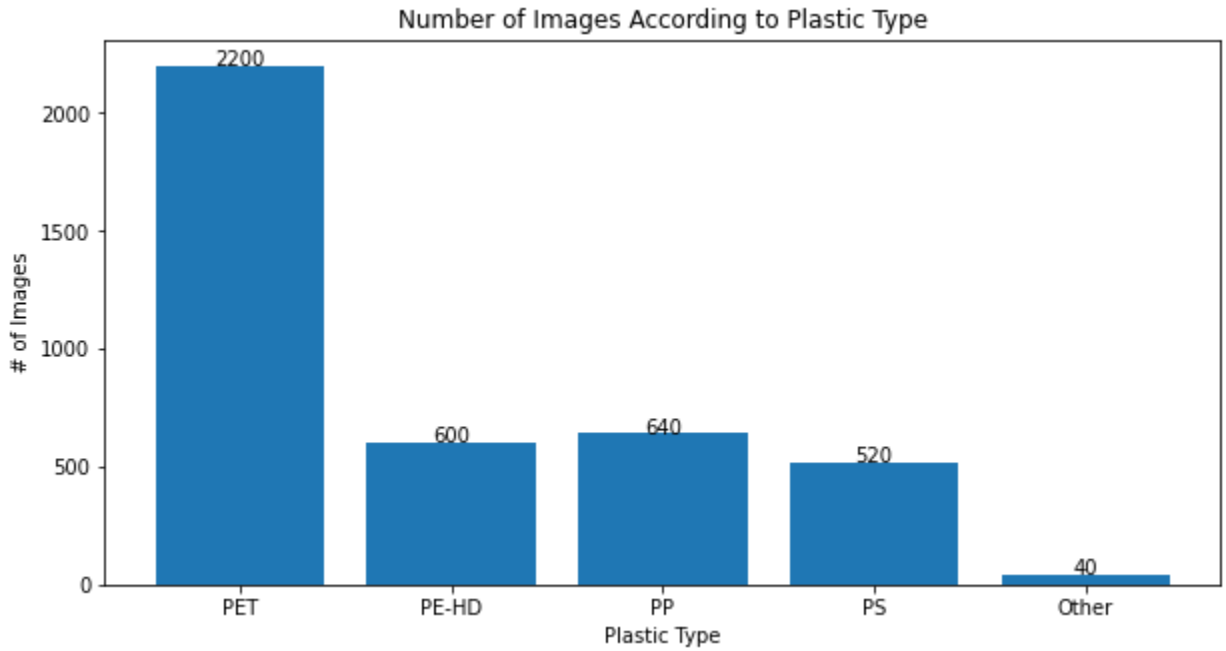

Figure 2: Number of Images According to Plastic Type

Methodology/Models:

Using the train_test_split() method, the data sourced from the preprocessing stage was split 80% training to 20% test and set to a common random state for each trial so that the comparison of the models would be guaranteed to be on the same data. Using this split data, three different machine learning models (Logistic Regression, Support Vector Machine, and K-Nearest Neighbors Classifier) were trained and tested before the classification_report() method was used to report the F1 accuracy score for each model on the data.

Additionally, the time that each model required to train and test on the dataset was recorded.

Logistic Regression

Our implementation of the logistic regression model was very straightforward in that the model was trained on the training portion of the data before being tested on the other portion of the data. Logistic Regression is often regarded as an easy model to work with given that it is simple to set up and has a high training efficiency. However, it should also be kept in mind that one of the drawbacks of Logistic Regression is that models have a tendency to overfit themselves, resulting in data that may not be remotely replicable in the real world.

Support Vector Machine

Support Vector Machines (SVM) make use of tools known as kernels in order to help reduce the time taken to compute complex calculations. For the purposes of this project's data, there are three different kernels that can be executed on SVMs to obtain valid results: Gaussian, Polynomial, and Sigmoid. In this project, we investigated all three models.

Furthermore, it should be noted that the Polynomial Kernel requires the specification of a degree, changing how flexible the decision-boundary is. For our experiment, the degree was set to a value of five.

K-Nearest Neighbors Classifier

For the K-Nearest Neighbors Classifier, the parameter “n_neighbors” determines that the algorithm will look at “n_neighbors” number of values/points closest to the one it is currently analyzing, determine the relative distances, and then classify the point based on which point it is the closest to. Because of the way that this algorithm functions, it must be kept in mind that the value of “n_neighbors” can drastically affect how the model reacts, especially at smaller values as it tends to overfit when this is the case (the model is more generalized when the value of “n_neighbors” is greater). For this project, an “n_neighbors” value of 40 will be utilized.

Results and Discussion:

After running tests on all of the algorithms, the results shown in the figure below represent the F1 accuracy scores of each respective algorithm.

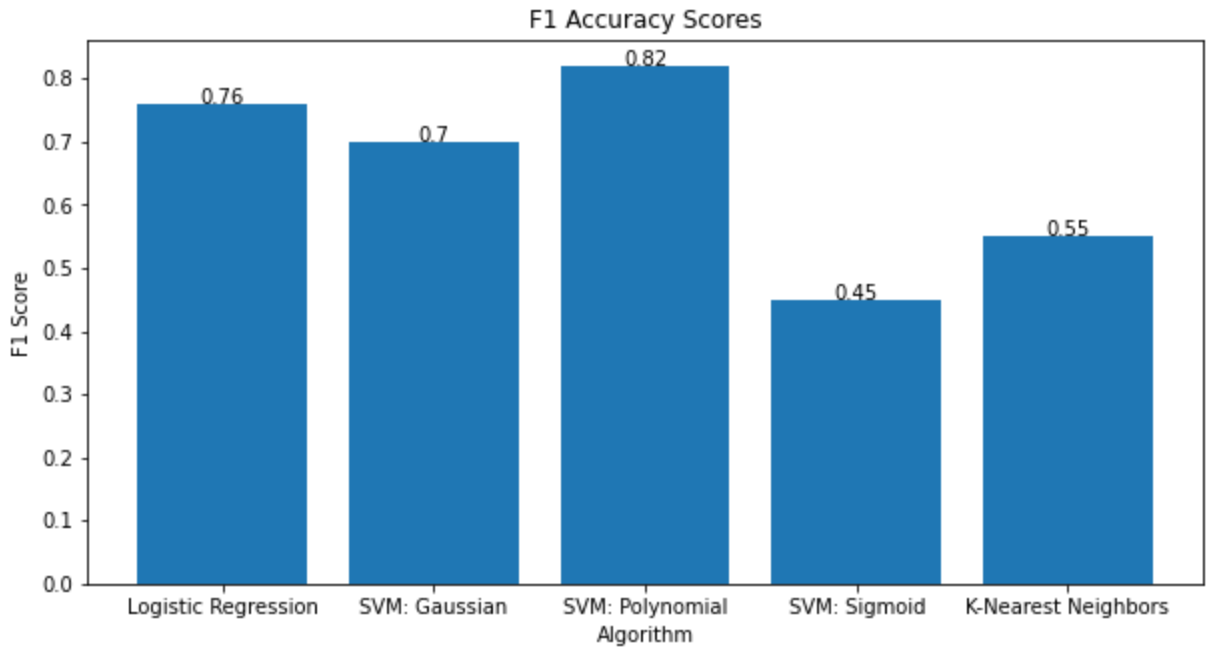

Figure 3: F1 Accuracy Scores of Tested Machine Learning Algorithms

While Logistic Regression, SVM: Gaussian, SVM: Polynomial all had respective F1 accuracies above 70% (with SVM: Polynomial having the highest at 82%) and would be feasible to deploy to consumer products just from this standpoint, the SVM: Sigmoid and K-Nearest Neighbors algorithms had much lower values that would be far from acceptable regardless of the their performance in other aspects of consideration such as time and memory management (it should be considered that at accuracies this low, there is still a considerable amount of manual sorting that must be done, quickly driving overall costs up). While not exactly clear what the cause of these low scores are, we hypothesize that they have to do with the limitations of the analyzed dataset, given that it only has 100 different objects in it and this could be an insufficient number for proper training of a high level algorithm.

For the three remaining algorithms, the factors of time and memory should now be considered given that the accuracy scores are relatively similar. In terms of the time that it took for each algorithm to run on the testing set of data, Logistic Regression finished first at 2.33

seconds, followed by SVM: Polynomial at 304.207 seconds, and finally SVM: Gaussian at 868.254 seconds. Keep in mind that this is the time required to identify 800 different images and that the time it would take for each algorithm to identify a couple of objects is a few seconds at most.

Given the ways that Logistic Regression and SVM models function as outlined in the Methodology/Models section, both are relatively memory efficient after training has been completed and either one should qualify to run on decently cheap computer hardware in extreme cases (a RaspberryPi, for example).

With the time and memory efficiency of each algorithm in mind, it seems reasonable to eliminate the SVM: Gaussian algorithm from consideration given that it both takes considerably longer than the other algorithms and has a lower overall F1 accuracy score, leaving the decision to fall between Logistic Regression and SVM: Polynomial for being used in a potential consumer product. While Logistic Regression is incredibly fast in categorizing samples and does so at a respectable accuracy, the drawbacks of the algorithm with its tendency to overfit and the already respectable speed of SVM: Polynomial (less than a half a second per object analyzed) at a slightly higher degree of accuracy, it can be concluded the SVM: Polynomial is the best machine learning algorithm for our purposes.

Conclusion:

In this paper, it was determined that Support Vector Machine utilizing the Polynomial Kernel best fit a consumer recycling can powered by AI for its good balance of accuracy, memory allocation, required compute power, and time for analyzing a subject. Furthermore, this conclusion serves as a proof of reasonable possibility, or rather that if given enough time and

effort, there is indeed a path to both efficiently and effectively integrate AI plastic sorting systems on the consumer end. And once again, while that path forward may not seem like the most beneficial for recycling companies in the short term, it ultimately drives in more profit in the long run and also helps to address the issues of plastic waste that currently plague our world.

While this research offers a good starting point for recycling groups to build on, potential further developments of this research include exploring more powerful machine learning algorithms like neural networks and expanding the scope of the database used for training and testing (using more than just 100 objects). By making these changes to the experimental setup, there is not an insignificant chance that F1 accuracy scores will increase by a good margin while maintaining similar hardware and time limitations/factors.

Acknowledgements:

I would like to thank Polish researchers Janusz Bobulski and Jacek Piatkowski for granting me access to the WaDaBa Image Database, as the data that they collected formed the basis of the evidence for my project. Additionally, I would like to thank Eric Bradford and the Inspirit AI team for providing resources and help throughout my journey in this research project. None of this would have been possible without their guidance.

References:

Bobulski, Janusz, and Jacek Piatkowski. "Plastic Waste DataBase of Images – WaDaBa." *WaDaBa Image Database*, Czestochowa U of Technology, wadaba.pcz.pl/#home. Accessed 12 Oct. 2022.

Chariot Energy. "How Long Does It Take for Plastic to Decompose?" *Chariot Energy*, 10 Feb. 2021, chariotenergy.com/blog/how-long-until-plastic-decomposes/. Accessed 13 Oct. 2022.

EPA. "Plastics: Material-Specific Data." *Facts and Figures about Materials, Waste and Recycling*. *U.S. Environmental Protection Agency*, United States Environmental Protection Agency, 19 Sept. 2022, www.epa.gov/facts-and-figures-about-materials-waste-and-recycling/plastics-material-specific-data. Accessed 12 Oct. 2022.

Morgan, Blake. "Why Is It So Hard To Recycle?" *Forbes*, 21 Apr. 2021, www.forbes.com/sites/blakemorgan/2021/04/21/why-is-it-so-hard-to-recycle/?sh=4fdfed603b77. Accessed 12 Oct. 2022.

Parker, Laura. "A Whopping 91 Percent of Plastic Isn't Recycled." *National Geographic*, 20 May 2022, education.nationalgeographic.org/resource/whopping-91-percent-plastic-isnt-recycled. Accessed 12 Oct. 2022.

PET Resin Association. "An Introduction to PET." *PET Resin Association*, 2015, www.petresin.org/news_introtoPET.asp#:~:text=PET%20is%20completely%20recyclable%2C%20and,States%20each%20year%20for%20recycling. Accessed 13 Oct. 2022.