A Study of Accuracy and Usability of AI-based Waste Classification within a Mobile Application

Yash Narayan[1], June Flora[2], Chad Zanocco[2], Jeremy Irvin[2], Ram Rajagopal[2]

Abstract*—**Every year, solid waste generates over 1.5 billion metric tons of CO2 equivalent greenhouse gases, making it one of the major causes of global warming. While significant effort has been expended in educating the public about the benefits of recycling, errors arising from human confusion at the point of disposal result in both missed opportunities for recycling as well as expensive contamination that can cause entire recycling bins to end up in a landfill. In this work, we demonstrate the efficacy of deep learning, available at the point of disposal within a mobile app, for accurate and instantaneous waste classification. Our application, ***DeepWaste, achieves an overall accuracy of 92.0% with an F-1 score of 92.1%. We show the effectiveness of our model by benchmarking it against human waste classification performance where DeepWaste outperforms the average human accuracy by nearly 50% while performing in real-time. We also provide results from a System Usability Scale (SUS) test where the DeepWaste mobile app achieves a score of 81/100, earning an A grade and adjective rank of "Excellent" in the SUS scale.

I. INTRODUCTION

Climate change is described as one of the greatest challenges of the 21st century [1]. Across numerous scientific studies, it has been shown that greenhouse gas (GHG) emissions caused by human activities have warmed the climate at a an unprecedented rate since the Industrial Revolution. The Intergovernmental Panel on Climate Change (IPCC), a UN body for assessing science related to climate change, projects that to avoid catastrophic and irreversible climate change, global warming needs to be limited to 1.5°C rise from pre-industrial levels. The latest IPCC report update released in August 2021 states that climate change is happening at an even faster rate than previously understood, and that currently under all scenarios of carbon emissions, the threshold of 1.5°C is very likely to be exceeded latest by 2040, and potentially as early as by the end of this decade unless drastic emission reductions measures are enacted soon.

Accurate waste disposal plays an important role in reducing GHG emissions. Every year, the world generates over 2 billion tons of solid waste [2]. This waste generates over 1.5 billion metric tons of CO2 equivalent greenhouse gases [2], contributing nearly as much to climate change as all the cars on the U.S. roads. Solid waste also includes significant embodied GHG emissions. For example, most of the GHG emissions associated with paper occur before it becomes waste [3]. Therefore, encouraging waste minimization through recycling programs can have significant up-stream GHG minimization benefits. In the U.S., even though 75% of this waste is capable of being recycled, only 34% is actually recycled [4]. Further, 91% of plastic isn't recycled [5] and only about 5% of food and other organic waste is composted [6].

Despite massive investment to educate the public about accurate waste disposal, efforts so far have been only moderately successful. People are often confused by what they can recycle, or compost. Signs and boards found near waste bins are difficult to understand and are often incomplete. Furthermore, disposal of waste varies based on the local recycling facilities' capabilities, and therefore rules for disposal are subject to change on a county-by-county basis.

Errors in waste disposal are not only missed opportunities to recycle or compost, but also lead to the contamination of recycling and compost bins. Often, an entire bin can end up at a landfill due to a single error leading to contamination of the whole bin. According to industry estimates, human confusion in the identification and correct disposal of waste into our waste bins results in nearly 25% of recyclables getting contaminated [7], diverting materials that could be recycled into landfills. When a recyclable or compostable material ends up in the landfill, it releases methane, a greenhouse gas that is twenty one times more potent than CO2 in contributing to global warming over a 100-year timescale.

In this work, we present DeepWaste, the first mobile application targeted at the problem of erroneous waste disposal, available right at the point of disposal. DeepWaste leverages recent breakthroughs in convolution neural networks (CNNs) for image-recognition tasks [8] and the availability of increased computational power in everyday cell phones, to provide a novel approach for waste identification that is fast, low-cost, and accurate for anyone, anywhere. In this paper, we present experimental results showing the accuracy of the DeepWaste model which provides a significant improvement over previously published research. In addition, we benchmark the performance of DeepWaste against the accuracy and speed of human waste classification on a unique and diverse set of previously unseen waste images and show that DeepWaste significantly outperforms average human accuracy by over 50% and is also significantly faster. Finally, we evaluate the human-computer-interface efficacy of our application by running a statistical user evaluation with a random group of DeepWaste users to calculate System Usability Scale (SUS) score where the DeepWaste mobile application achieves a score of 81/100, earning the grade of A and adjective rank of "Excellent" within the scale.

The rest of the paper is organized as follows: Section 2 covers the related work in applying AI methods to the problem of waste classification. Section 3 describes the methodology used in this work. Section 3.1 discusses the deep learning

[1] The Nueva School, San Mateo, CA
[2] Stanford University, Department of Civil and Environmental Engineering

model used for DeepWaste and discusses various techniques applied to achieve the high accuracy.Section 3.2 describes the methodology used to benchmark human accuracy and Section 3.3 describes the setup for calculating the SUS usability score. Section 4 covers the results for the DeepWaste model. Section 4.1 provides accuracy metrics including precision, recall, F-1 scores, Receiver Operating Curve and Confusion Matrix across the test set. Section 4.2 provides the results of comparing DeepWaste against human classification and shows that DeepWaste outperforms average human accuracy. Section 4.3 provides the results for the SUS usability test and some feedback from users on the application. Section 5 provides a summary and direction for future work.

II. LITERATURE REVIEW

The topic of applying machine learning (ML) or artificial intelligence (AI) for waste classification has recently begun to garner considerable research interest. One of the earliest attempts to use AI for waste classification was Thung et al [13] and Awe et al [14]. Their model, TrashNet, used R-CNN technique to classify waste into three categories: recycle, landfill, and paper. The algorithm achieved accuracy of 68%, demonstrating the promise of using artificial intelligence for the problem of waste classification. Another attempt to use deep learning was RecycleNet proposed by Bircanoglu et al [15]. The authors experimented with several well-known CNN architectures and achieved the highest accuracy of 95% using a DenseNet model. However, the authors noted the performance of this model was too slow to be used in a real-time classification context. They proposed a new architecture, RecycleNet, which achieves an overall accuracy of 81%. More recently, Adedji et al [33] describe a model in which a CNN based ResNet-50 model is used for classification except that the last layer is replaced with a mult-class Support Vector Machine based classification to achieve overall accuracy of 87%. Another recent is described by Sai et al [38] which is similar to [15] where a DenseNet model achieves 92% overall accuracy. While the model performs well for Cardboard (98%) and Paper (90%), it has low accuracy for Glass (82%), Metals (80%), Plastics (83%), and Trash (68%), and is not suitable to be used within a mobile device due to the large model size. A novel hybrid approach for image detection was proposed by Chu et al [34] where they proposed a new hardware to use not only the images but other sensory inputs such as sound and smell to improve waste classification. While they report an overall accuracy of 98%, the results were demonstrated on a small data set of 50 images and the images had to be placed in a very specific way for their specialized hardware to correctly capture the sensory inputs.

There are several other attempts to use AI within a 'smart bin' or an industrial grade binning system located at a recycling plant [9], [10], [11], and [12]. This approach requires expensive hardware that costs thousands of dollars, thus deterring their wide-spread adoption. In addition, as these smart bin based hardware solutions are typically deployed at a recycling center, it is too late to prevent bin contamination that happens at the point of disposal. Our approach with DeepWaste is novel as it uses widely available smartphones, and therefore has the potential for large-scale adoption at little or no cost. DeepWaste's ubiquitous and easy access also allows its use right at the time of disposal to prevent errors, thus lowering the probability of bin contamination.

There is some previous work related to waste classification in embedded systems, but it has generally required specialized hardware built for this purpose, and none of these approaches in the literature describe a system for classification that is available on a smartphone. In [16] authors describe using SSD-MobileNet embedded on a Raspberry Pi 4 to classify plastic bottles, glass bottles, and metal cans achieving accuracies of 95%, 82% and 86% respectively. Another attempt to create a system available at the point of disposal was Auto-Trash, which required a Raspberry Pi 4 based attachment to the garbage disposal can. No accuracy results were published on this. In [35], Mittal et al describe a smartphone app, SpotGarbage, however the purpose of this app is not waste classification. The app is designed as a way for citizens to take pictures and easily geo-tag and report garbage for civic purposes. In [36], authors describe ThanosNet, a system to collect images along with metadata fields such as location and traffic. However, no results of waste classification are reported. More recently, White et al [39] describe WasteNet, a smart bin for waste classification in a non peer-reviewed arXiv submission where they claim to achieve 97% accuracy. However, no details of the model are provided and the proposed methodology uses specialized hardware in the smart bin context. On searching through AppStore, authors were able to find only one other application that uses AI for waste classification, Waste Classifier [40]. This application provides classification into various categories of recycling, paper and organic waste. However no references were provided on the modeling approach used or accuracy of the results on any benchmark data set for this application.

In addition to the high accuracy and widespread availability, DeepWaste is the first model in the literature, to the best of our knowledge, that considers compost as a new category for waste classification. Misclassification of compostable material is particularly harmful for the environment because when compostable material such as food scraps and green waste are diverted into a landfill, it is compacted down and covered. This removes the oxygen and causes it to break down in an anaerobic process. Eventually, this releases methane, a greenhouse gas that is 25 times more potent than carbon dioxide in warming the earth over a 100-year timescale (and more than 80 times on a 20-year timescale) [17].

Another important contribution of this research is to establish a baseline accuracy for human waste classification. To the best of our knowledge, our work provides the first quantitative study comparing DeepWaste to overall human accuracy, by conducting a blind accuracy benchmark and comparing AI-based classification to humans. We also perform a SUS usability survey for the DeepWaste mobile app to demonstrate its applicability as a practical tool for everyday

III. METHODS

Classifying waste using AI is a challenging problem for a number of reasons. First, whether a waste item is recyclable, compostable or trash depends on the properties of the material

which can be hard to detect simply from the image. Second, materials can come in any shape or size such as a broken bottle or a crumpled can; any image processing technique needs to be able to deal with this variation. The two primary Research Questions (RQ) we sought to study were the following:

RQ1: Can we utilize the recent advances in artificial intelligence for image recognition to develop a model to classify waste items with high accuracy and generalizability?

RQ2: Can we then embed such a model into a smartphone via an app and by using the increased computational power of these everyday devices provide an accurate, low-cost, fast and universally accessible classifier at the point of disposal?

Figure 1 summarizes the pipeline to create the DeepWaste mobile application.

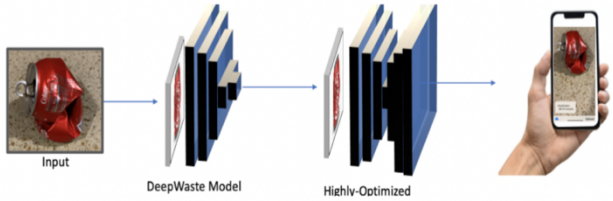


Figure 1: DeepWaste mobile application pipeline

Data Collection: Since there was no publicly available dataset, a significant part of this research was to first collect a dataset from scratch by collecting images from our local neighborhood. Towards this goal, we implemented user-based model training capabilities so that users can easily take a picture, label it, and upload it to the cloud on Amazon Web Service (AWS) for on-demand model training. In total, we manually collected 1218 images at various lightings and angles: 396 images compostable item(s), 427 images recyclable item(s), and 395 images landfill item(s). *Figure 2* shows some sample images from our dataset. To ensure each image in the dataset was labeled accurately, we verified the correct classification through our local waste provider.

Figure 2: *Sample images from DeepWaste dataset*.

Data Augmentation: On this dataset, prior to training the model, we apply aggressive data augmentation to each input image. Data augmentation is used to improve the generalizability and accuracy of our model by matching real-world noise of consumer images; while each waste item can be identified by its unique properties, such as its shape and size, each user image item may vary in terms of size, illumination, blur, and background depending on how a given individual takes the image from their smartphone. To address this challenge, each input image was rotated with an angle randomly selected among 0, 90, 180, and 270 degrees and cropped, sheared, and blurred.

Model: Utilizing this dataset, we develop our deep learning neural network model called DeepWaste. Neural Networks are inspired by how the brain works and consists of a series of processing nodes organized as a hierarchy of layers, each simple in its operation, but collectively the overall network is able to implicitly 'learn' complex relationships between inputs and outputs that are not easy to model by other more traditional computational methods. In this study, we select a specific type of neural network architecture called a convolutional neural network (CNN) that has been highly successful in image classification and recognition problems in a wide range of settings. CNNs achieve this performance by exploiting features from local structures in an image and aggregating the local features to make a prediction on the full image.

We experimented with a number of different state-of-the-art CNN architectures including InceptionV3 [18] , Inception ResnetV2 [19], MobileNet [20], PNASnet [21], and Resnet50 [22]. Based on the results, we selected Resnet50 as our model's backbone. The Resnet50 network is a variant of the ResNet architecture and contains 48 Convolution layers along with 1 MaxPool and 1 Average Pool layer. Our full dataset is used to train the model and then a balanced test set consisting of 100 real-world images (34 trash, 33 recycle, 33 compost), entirely new to the model, are used to evaluate the performance of the model. Then advantage of this architecture over other types of CNN architectures is it allows for training of very deep networks by introducing a residual block into the network that prevents the gradient from 'vanishing' or 'exploding' during training [23]. To the ResNet50 network, we added an additional 'attention layer' [24] before the final FC layer. This attention layer is traditionally used in sequence-to-sequence tasks, but here we use the attention mechanism to identify the most discriminative features that distinguish between images, and make the algorithm pay more attention to these more discriminative local regions of an image, thereby improving the classification accuracy of the model in complex scenes. We use this approach because of the large variance in the same subcategory of images in our dataset and the small variance among different subcategories.

Training: As part of our preprocessing step during training, each of the images in the dataset are resized to 224 by 224 pixels to fit the input of our model. All of our dataset is used to train the model and then a test set consisting of 100 real-world images (34 trash, 33 recycle, 33 compost), entirely new to the model, are used to evaluate the performance of the model. The training process consists of iteratively updating the parameters to decrease the prediction error. The prediction error is computed by comparing the network's prediction to the actual classifications from the dataset.

Typically, to train such deep CNN networks, huge data sets with millions of images are required to get acceptable accuracy. A small dataset, such as in our case, could cause a network to overfit, or not be able to sufficiently generalize. To overcome this challenge, we utilize a technique called transfer learning [25] to initialize our models from weights pre-trained on the 2014 ImageNet Large Scale Visual

Recognition Challenge dataset, which consists of around 1.3 million images and 1000 object classes. Transfer learning leverages the previously learned low-level features such as lines, edges, and curves. Since these low-level features are common to any image classification task, transfer learning requires significantly less data to achieve high accuracy. After initializing our model with pre-trained weights, we freeze the hidden layers of our model and add a final fully connected (FC) layer to our CNN to speed up model training.

This model is then trained using a differential learning rate for 40 epochs with early termination if convergence is detected sooner. A differential learning rate allows different parts of the network to train at different rates, speeding up the overall training process. We use a dropout rate of 0.2 in the final FC layer. Dropout is a regularization technique that drops out a percentage of the layer from activating and thus reduces the overfitting of our neural network model. The final layer of the network feeds into a SoftMax layer that takes the output of the network, and for a given input image assigns a probability between 0 and 1 of that image belonging to a particular category. The model selects the category with the highest probability as the classification label and provides the user the probability value. All layers of the network are trained using the Adam optimizer [26] with a learning rate of 1e-3. Adam is an extension of the optimization algorithm called stochastic gradient descent (SGD) that iteratively updates the weights of a network based on the training loss to improve the overall model performance. The final network has a total of 23, 516, 228 learnable parameters. We use an open-source deep learning framework called PyTorch to develop all of our models.

3.1: Evaluating DeepWaste Accuracy against Human Classification accuracy

In addition to benchmarking the accuracy, precision, recall, and F-1 scores for DeepWaste on a test set, we also establish a baseline for average human classification accuracy and benchmark DeepWaste against human classification by testing both humans and DeepWaste on a previously unseen data set by the algorithm. To the best of our knowledge, this is the first such quantitative study establishing the baseline for human waste classification accuracy. As part of our benchmark, our goal was to investigate the following research questions.

RQ3: Can we establish a benchmark for the accuracy and speed of human-based waste classification on a diverse set of real-world waste images?

RQ4: Can an AI-based model for waste classification outperform human-based classification with high confidence?

RQ5: What items do humans commonly get confused by and how does accuracy vary between the three classes: trash, recycle and compost?

To ensure that a robust comparison was established between humans and DeepWaste, our images for the survey comparison consisted of the same test set images used to evaluate the performance of the model. This was a balanced set of 100 real-world images, roughly evenly split between trash, recycling, and compost classes (34 trash, 33 recycle, 33 compost), none of which were included in the dataset to train the DeepWaste model. To establish a benchmark for human-based waste classification, we restricted our sample from cities across ZIP codes in the San Francisco Bay Area. This was done since regionally recycling rules can vary significantly based on facilities' capabilities, and therefore humans' responses on the survey would vary based on respondents' regional recycling or composting rules. While restricting the sample to the SF Bay Area significantly reduced local variations in rules, there still existed some small differences between counties that needed to be accounted for (e.g., black colored plastic needed to be thrown into the trash in some Bay Area counties while in other Bay Area counties it could be recycled). To ensure that our survey accounted for such local variations, all images were carefully verified with each local municipalities' rules; items where local variations in rules affected the classification were removed from the survey until we achieved a survey set that worked for all Bay Area counties.

Our survey was developed using QualtricsTM, and participants were sampled using a convenience sample through Amazon Mechanical Turk (AMC), a crowd-sourcing platform to perform on-demand tasks. Our survey was administered from May 2021 to June 2021, resulting in 73 valid responses. The goal of the survey was to benchmark human-based accuracy of waste classification and measure convergence and divergence between humans' classification on specific items. Respondents' answers were completely anonymous and no identifiable information, such as their name, age, email address, phone number etc., was collected. Prior to beginning the survey, respondents were provided brief instructions, and then once started, respondents were asked to classify each of the 100 images in the survey into either recycling, compost, or trash. Associated with each image was a title of the item as well as important information that would affect the respondent's response such as whether the item was clean, unused, or dirty. Respondents were not allowed to return back to or skip questions and the order of images presented to each respondent was randomized. While respondents had unlimited time to complete the survey, we recorded the time the respondent took to finish the survey.

3.2: Evaluating DeepWaste User Experience

In addition to accuracy testing, we also conducted a usability study of our mobile application by running a statistical user experience evaluation. We conducted a blind System Usability Scale (SUS) test with a random group of DeepWaste users. A SUS test is a quick and straightforward 10 item Likert scale that assesses the subjective opinion from a user regarding the usability of a particular system [27]. SUS is one of the most widely used usability tests which has shown across a variety of studies that it can reliably measure

the perceived usability of a system by sampling a relatively small number of users as in our case [28]. Additionally, as compared to other types of statistical user experience evaluations, a SUS test is simple and quick, and only asks for users' reactions after using the system (e.g. if they found the system complex or easy-to-use) instead of asking for the user to assess specific features of the system (e.g. the visual appearance and organization of the system).

RQ6: Can we assess the ease-of-use of our application, how confident users feel in our application, and how likely they are to use our application in the future?

We developed our survey using Qualtrics to provide instructions and record responses for the SUS test. In total, we had 20 user responses. Participants were randomly sampled using a convenience sample through AMC and the sample was restricted to individuals who lived in California and had an iOS device. Participants were also individuals who had never used the DeepWaste mobile app previously. Prior to completing the SUS test, participants were asked to download DeepWaste from the App Store and use it for at least 2-3 minutes by classifying items near them. Once finished using the app, participants were asked to immediately evaluate the app by filling out the SUS test. The test consisted of 10 statements that covered a variety of aspects of system usability, such as the complexity of the app and if they would use the app frequently, and asked users to choose a value from 0 (strongly disagreeing with the statement) to 5 (strongly agreeing with the statement) on a Likert scale. We adapted the traditional SUS test slightly by adding a question after participants had already finished completing the scale for them to provide written feedback about their experience using DeepWaste. This was done to check if users' feedback corresponded to their SUS score. To calculate the SUS score, we used the standard SUS calculation (For items 1,3,5,7, and 9, we subtracted the users' scale position by 1. For items 2,4,6,8, and 10, the score is 5 minus the original scale position. After applying this procedure, the item's score contribution ranges from 0 to 4. We then summed the converted score contribution and multiplied the score by 2.5. The final SUS score ranges from 0 to 100). While the final SUS score ranges from 0 to 100, it is not a percentage. To interpret our score, we translate our final SUS score to a percentile score from [29], and an "adjective rank" and "letter grade" from [30]

IV. RESULTS

4.1: Can our AI model achieve high accuracy for waste classification? (RQ1)

Out of the various CNNs we benchmarked on the dataset in the DeepWaste, Resnet50 showed the best accuracy and convergence on the test set in terms of average precision. We then added an additional layer to the Resnet50 backbone architecture called an attention correction layer which improved classification performance by an additional 5%. The final DeepWaste model achieves an average precision score (arithmetic mean of the precision and recall score) of 0.920 and an F1 score (harmonic mean of the precision and recall score) of 0.921 (Table 1).

Accuracy	InceptionV3	MobileNet	PNASNet	Resnet50	DeepWaste
Trash	0.771	0.751	0.722	0.761	0.927
Recycle	0.891	0.949	0.864	0.924	0.921
Compost	0.806	0.873	0.841	0.882	0.919
Overall	**0.84**	**0.842**	0.852	**0.811**	**0.920**

Table 1: *DeepWaste performance on test set*

We measured the accuracy of our classifier using a variant of a traditional Receiver Operating Characteristic Area Under the Curve (AUC-ROC) designed for multiclass classification. The ROC-AUC score is an evaluation metric for classifiers that consider the false positive rate (proportion of the positive class incorrectly classified by the model) and true positive rate (proportion of the positive class that got correctly classified). A classifier with an AUC = 1 is able to perfectly distinguish between all the Positive and the Negative class points; an AUC = 0.5 is a classifier that cannot distinguish between Positive and Negative class points, i.e., the classifier is predicting randomly. Figure 4 shows the ROC curves from our best performing model. DeepWaste achieves a high AUC on all three classes: an AUC = 0.963 for the recycling class, an AUC = 0.965 for the trash class, and AUC = 0.905 for the compost class. We also generate a confusion matrix of DeepWaste results on the test set to visualize how well our model predicted against the actual annotations of compost, trash, and recycle (Figure 4).

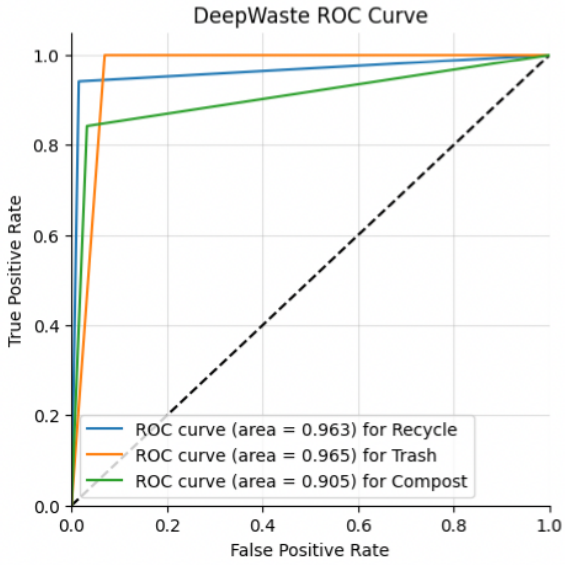


Figure 3: *ROC curves for DeepWaste model on test set*

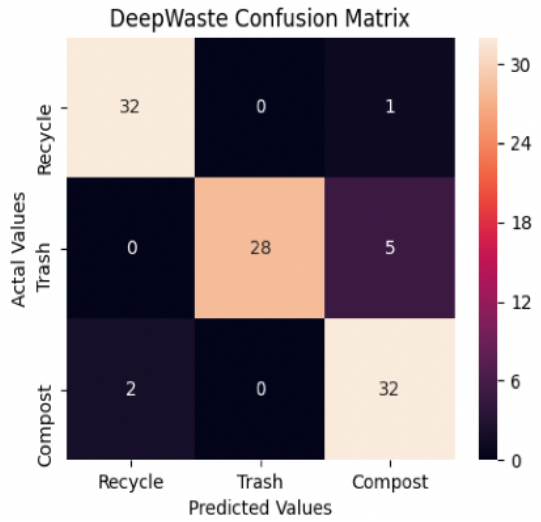


Figure 4: *confusion matrix of DeepWaste model on test set*

4.2: Can we embed an optimized DeepWaste model into a smartphone via an app? (RQ2)}

DeepWaste was then optimized and subsequently deployed inside of a mobile application using Apple CoreML. CoreML optimizes on-device performance by leveraging the CPU, GPU, and Neural Engine while minimizing its memory footprint and power consumption. DeepWaste model is running strictly on the user's mobile device, therefore removing the need for internet connection and sharing data. We designed an easy-to-use mobile app interface for users using Swift 4 programming language and Xcode environment. ***Figure 5*** shows the DeepWaste app classifying commonly confused items in real-life. A user can simply point their phone camera to any piece of waste and get instantaneous feedback, with an average prediction time of around 100ms. DeepWaste is able to correctly identify items with high accuracy, even when the shape has been deformed such as a crushed soda can, orange peels, an apple core, crumpled paper, and a plastic bag. Note that the plastic bag in ***Figure 5*** is classified as trash because plastic bags, films, and wraps cannot be recycled in your curbside recycling bin; they must be dropped-off to a special retail store that can collect plastic grocery bags for recycling. Throwing this plastic bag into the recycling bin has the potential of contaminating the entire bin. The DeepWaste model can be retrained and personalized to account for different local rules.

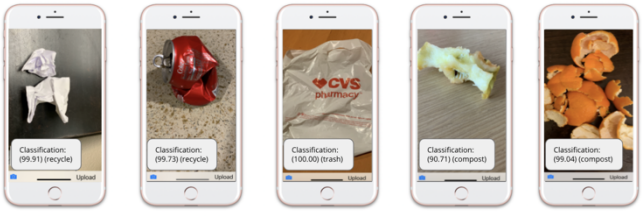


Figure 5: *DeepWaste app classification output*

4.3: Can we establish a benchmark for human-based waste classification? (RQ3)

We randomly sampled 73 respondents in the Bay Area. To ensure that a robust comparison was established between humans and DeepWaste, we asked them to classify the same set of 100 waste items (34 trash, 33 recycle, 33 compost) from our test set. The mean human accuracy achieved on the benchmark was 61.36% (95% Confidence Interval between 57.7% to 65%), with the minimum score achieved on our benchmark = 28% and the maximum score achieved on our benchmark = 89%, with a deviation of 16.06 (Table 2).

While respondents had unlimited time to complete the survey, we recorded the total time it took them to complete the survey from start to finish. Table 3 shows the mean time individuals took to complete the survey as well as the maximum and minimum time spent to complete the survey. It is important to note that there was no causal relationship found between the time taken by the user and their accuracy score. Someone who has taken longer will not necessarily have higher accuracy as evidenced by the fact that the individual who took the minimum amount of time had 74% accuracy, while the individual who took the maximum amount of time had 63% accuracy.

Sample Size: N = 73
Mean Score: 61.36% (95% CI 57.7% to 65%)
Standard Deviation: 16.06
Margin of Error: 3.68
Maximum Human Score: 89%
Minimum Human Score: 28%

Table 2. Benchmark Results from Human Accuracy on Survey Images

Sample Size: N = 73
Mean Time Spent: 10 minutes 47 seconds (95% CI 57.7% to 65%)
Minimum Time Spent: 3 minutes 9 seconds
Maximum Time Spent: 55 minutes 18 seconds
DeepWaste Algorithm Time Spent: 10 seconds

Table 3. Human Time Spent on Survey Images

4.4: Can DeepWaste outperform human-based classification? (RQ4)

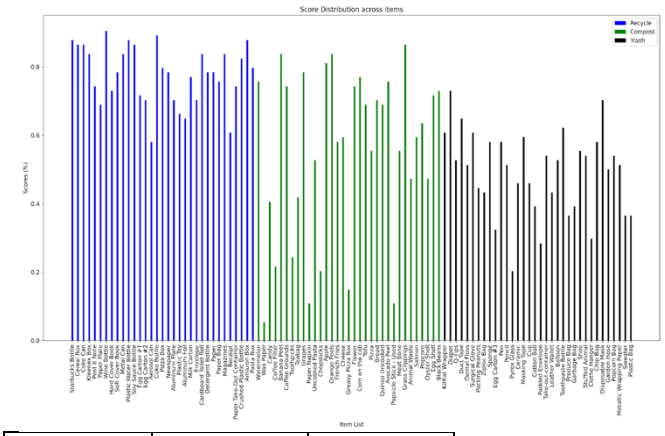
Comparing the human classification accuracy and speed against DeepWaste, the current DeepWaste model outperforms the average human, best performing human, and worst performing human, with a high confidence margin, on each accuracy metric including the average accuracy, F1 Score, and Precision and Recall Score. (Table 4).

	DeepWaste	Average Human	Maximum Human	Minimum Human
Average Accuracy	0.92	0.61	0.89	0.28
F1 Score	0.921	0.627	0.889	0.22
Precision Score	0.927	0.677	0.90	0.187
Recall Score	0.919	0.641	0.889	0.279

Table 5: DeepWaste vs Maximum Human Score vs Minimum Human Score vs Average Human Score

4.5: What items do humans commonly confuse? (RQ5)}

Figure 7 shows the average human accuracy for each of the 100 items. The average human accuracy of classification for each category (recycle, trash, and compost) is shown in comparison with DeepWaste accuracy in Table 5.



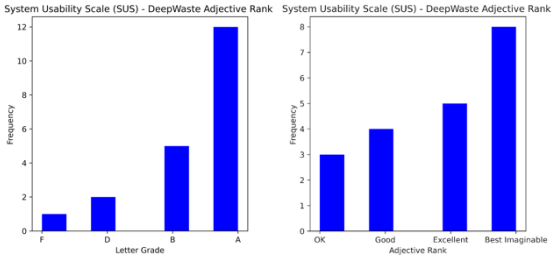
	DeepWaste	Average Human
Recycle	0.921	0.778
Trash	0.927	0.492
Compost	0.919	0.555

Table 5: DeepWaste vs Average Human Accuracy in each category

Overall, based on the results the human accuracy is highest on average for Recycle and lowest for Trash. This disparity between human accuracy on the recycling and trash classes indicate that many users are taking waste items and incorrectly classifying them as recyclables. For example, the sweater (which is trash) about ~50% of humans classified it as recyclable and ~18% of humans classified it as Compost, while only ~32% of humans classified it correctly as Trash. False positives for recycling is a particularly important point as this indicates potential contamination of the entire recyclable bin. Another interesting observation is that DeepWaste was able to differentiate between items that look similar but have different material properties better than the average human. For example, the survey included three different types of egg cartons: one that is plastic (recyclable), another that is cardboard (recyclable), and third that is styrofoam (trash). DeepWaste correctly classified all three of these while the majority of individuals selected recyclable for all three egg cartons.

4.6: How do users feel while using our application? (RQ6)

In the usability survey of a group of 20 users, randomly selected using AWS Mechanical Turk service, DeepWaste achieved an SUS score of 81.00 which translates to around the top 10% of scores as shown in Figure 8. As an adjective rank, DeepWaste's SUS score translates to "Excellent" and as a grade the SUS score is equal to an A. The histograms in Figure 9 below summarizes the "Letter Grade" and Adjective Rank distribution based on the SUS test results from each user. Figure 10 shows the average response for each of the 10 questions along with standard deviation for each.



V. CONCLUSION AND FUTURE WORK

In this work, we have developed DeepWaste, an easy-to-use mobile application that utilizes highly-optimized deep learning techniques to provide fast, accurate, and low-cost waste classification. DeepWaste is one of the first mobile waste classification solutions that is universally accessible at the point of disposal, for anyone with a smartphone, to mitigate climate impact. DeepWaste achieves overall accuracy of 92.0% and an F-1 score of 92.1%, making it one of the highest accuracy results reported in the literature. We also perform a robust comparison between DeepWaste and human accuracy on a diverse set of waste images, establishing the first such benchmark to measure average human accuracy on waste classification. DeepWaste outperforms the mean human classification accuracy by nearly 50% while being able to classify these images in near real-time and significantly faster than humans taking the survey. Finally, we also show the human-computer efficacy of DeepWaste by conducting a SUS usability study in which DeepWaste achieves a score in the top 10 percentile. The DeepWaste mobile app is available to the general public on the App Store and is being piloted at several school and university campuses. We hope our work can reduce the amount of incorrect waste disposal, and over time raise more awareness around the impacts of waste on our climate. If DeepWaste can even reduce erroneous waste disposal by 1%, it will be equivalent to removing over 6.5 million gasoline-burning vehicles from the road, making this a promising application of AI to tackle challenge of climate change.

VI. REFERENCES

[1] Intergovernmental Panel on Climate Change, "Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change" (2012)

[2] S. Kazam, L. Yao, P. Bhada-Tata, F. Van Woerden, "What a Waste 2.0: A Global Snapshot of Solid Waste Management to 2050," World Bank, pp. 3-5 (2018)

[3] Containers, Packaging, and Non-Durable Good. "Documentation for greenhouse gas emission and energy factors used in the waste reduction model (WARM)." (2016).

[4] US Environmental Protection Agency (EPA). "National Overview: Facts and Figures on Materials, Wastes and Recycling." (2019).
[5] Parke, L (National Geographic). "A Whopping 91 Percent of Plastic Isn't Recycled."
[6] Buzby, Jean, Claudia Fabiano, and Jeanine Bentley. "USDA and EPA estimation methods for food loss and waste in the United States." The Economics of Food Loss in the Produce Industry. Routledge, 2019. 77-89.
[7] Koerth, M (FiveThirtyEight). "The Era Of Easy Recycling May Be Coming To An End." (2019).
[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, Vol. 521, pp. 436, 2015
[9] I. Salimi, B. S. Bayu Dewantara, I. K. Wibowo, "Visual-based trash detection and classification system for smart trash bin robot,"
International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC), Bali, Indonesia, pp. 378-383, doi: 10.1109/KCIC.2018.8628499 2018.
[10] D. Vinodha, J. Sangeetha, B. Cynthia Sherin, M. Renukadevi, "Smart Garbage System with Garbage Separation Using Object Detection," International Journal of Research in Engineering, Science and Management 2020.
[11] D. Ziouzios, M. Dasygenis, "A Smart Recycling Bin for Waste Classification," Panhellenic Conference on Electronics & Telecommunications (PACET), pp. 1-4, doi: 10.1109/PACET48583.2019.8956270, 2019.
[12] Jacobsen, Rune Møberg, et al. "Waste Wizard: Exploring Waste Sorting using AI in Public Spaces." Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society. 2020.
[13] G. Thung, M. Yang, "Classification of Trash for Recyclability Status," Stanford CS229 Project Report, 2016.
[14] O. Awe, R. Mengitsu, V. Sreedhar, "Final Report: Smart Trash Net: Waste Localization and Classification," Stanford CS229 Project Report, 2016.
[15] C. Bircanoglu, M. Atay, F. Be¸ser, Ö. Genç and M. A. Kızrak, "RecycleNet: Intelligent Waste Sorting Using Deep Neural Networks," Innovations in Intelligent Systems and Applications (INISTA), Thessaloniki, pp. 1-7, doi: 10.1109/INISTA.2018.8466276, 2018
[16] Thokrairak, Sorawit, Kittiya Thibuy, and Prajaks Jitngernmadan. "Valuable Waste Classification Modeling based on SSD-MobileNet." 2020-5th International Conference on Information Technology (InCIT). IEEE, 2020.
[17] United States Environmental Protection Agency (EPA), "Overview of Greenhouse Gases," 2018. URL https://www.epa.gov/ghgemissions/overview-greenhouse-gases
[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826 (2016)
[19] C. Szegedy, S. Ioffe, V. Vanhoucke, A.Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," arXiv preprint arXiv:1602.07261 (2016)
[20] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, Adam, H, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861 (2017)
[21] C. Liu, B. Zoph, J. Shlens, W. Hua, L. Li, L. Fei-Fei, A. Yuille, J. Huang, K. Murphy, "Progressive neural architecture search," arXiv preprint arXiv:1712.00559 (2017)
[22] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, (2016)
[23] Hochreiter, Sepp. "The vanishing gradient problem during learning recurrent neural nets and problem solutions." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 6.02 (1998): 107-116.
[24] Wang, Fei, et al. "Residual Attention Network for Image Classification." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 6450–58. DOI.org (Crossref), doi:10.1109/CVPR.2017.683.
[25] Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." IEEE Transactions on knowledge and data engineering 22.10 (2009): 1345-1359.
[26] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).
[27] Brooke, J. SUS: A 'quick and dirty' usability scale. In Usability Evaluation in Industry; Jordan, P.W., Thomas, B., McClelland, I.L., Weerdmeester, B., Eds.; Taylor and Francis Group, CRC Press: Cleveland, OH, USA, 1996; Chapter 21; pp. 189–194.
[28] Tullis, Thomas S., and Jacqueline N. Stetson. "A comparison of questionnaires for assessing website usability." Usability professional association conference. Vol. 1. 2004.
[29] Percentile rankings of SUS scores from "A Practical Guide to the System Usability Scale: Background, Benchmarks, & Best Practices," by J. Sauro, 2011, Denver, CO: Measuring Usability LLC. Reprinted with permission.
[30] Bangor, Aaron, Philip Kortum, and James Miller. "Determining what individual SUS scores mean: Adding an adjective rating scale." Journal of usability studies 4.3 (2009): 114-123.
[31] Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv preprint arXiv:1312.6034 (2013).
[32] Brooke, J. SUS: A retrospective. J. Usability Stud. 2013, 8, 29–40.
[EPA2006b]
[33] Adedji, Olugboja, and Wang Zenghui, "Intelligent Waste Classification Using Deep Learning Neural Networks", 2nd International Conference on Sustainable Materials Processing and Manufacturing (2019)
[34] Chu, Yinghao, Huang, Chen, Xiaodan, Xie, Tan, Bohai, Kamal, Shyam , and Xiong, Xiaogang, "Multilayer Hybrid Deep-Learning Method for Waste Classification and Recycling", Hindawi Computational Intelligence and Neuroscience Volume 2018
[35] J. Donovan, "No Title," Auto-trash sorts garbage automatically at the techcrunch disrupt hackathon, 2016. [Online]. Available: https://techcrunch.com/2016/09/13/auto-trash-sortsgarbage-\%0Aautomatically-at-the-techcrunch-disrupthackathon/.
[36] G. Mittal, K. B. Yagnik, M. Garg, and N. C. Krishnan, "SpotGarbage: Smartphone app to detect garbage using deep learning," UbiComp 2016 - Proc. 2016 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput., 2016.
[37] Sun, Alan, and Xiao, Harry, "ThanosNet: A Novel Trash Classification Method Using MetaData", IEEE International Conference on Big Data (2020)
[38] Sai Susanth G1 , Jenila Livingston, Agnel Livingston, "Garbage Waste Segregation Using Deep Learning Techniques", IOP Conf. Series: Materials Science and Engineering (2021)
[39] Gary White, Christian Cabrera, Andrei Palade, Fan Li, Siobhan Clarke., "WasteNet: Waste Classification at the Edge for Smart Bins", https://arxiv.org/pdf/2006.05873.pdf (2020)
[40] "Waste Classifier", available on the App Store for iPhone