APS360

Project Final Report

**Predicting Pet Adoption Speed using ANN models**

April 9, 2020

Word Count: 2374

| Prerna Anand | 1003841435 |
| --- | --- |
| Sagar Chadha | 1004204274 |
| Neevi Shah | 1004191340 |
| Daman Sharma | 1004179998 |

**Table of Contents**

[**Introduction**](#_coo2ve8mmcmj) **1**

[**Illustrations**](#_npmp2oyov675) **1**

[**Background & Related Work**](#_upw2ncku416o) **2**

[**Data Processing**](#_ay25x2iaz3nu) **2**

[**Architecture**](#_kiicromhqbwc) **4**

[**Baseline Model**](#_lw90j2uh17lc) **6**

[**Quantitative Results and Discussion**](#_xvfxpm4qao30) **7**

[**Qualitative Results and Discussion**](#_xmk98fvh08ra) **10**

[**Evaluate Model on New Data**](#_o64f3vrsvys1) **12**

[**Ethical Considerations**](#_osqt2utf0wqc) **12**

[**Project Difficulty Reflection**](#_1d3u9vxge7u3) **13**

[**Link to Code Files**](#_h1q73yapqfke) **13**

[**References**](#_3nglzt1jrnzi) **14**

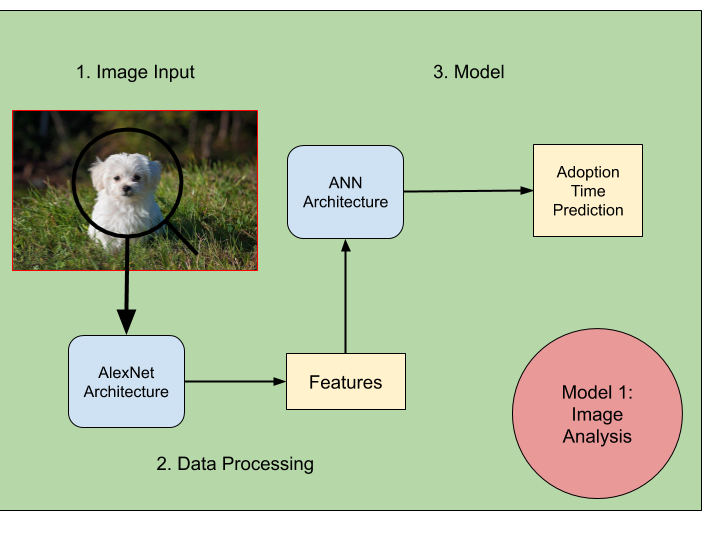
# 

# Introduction

The project is to predict the adoption speed of a pet based on feature and image data provided by the PetFinder.my Kaggle competition [1]. The project is vital because it can provide insights to adoption shelters on which types of pets will get adopted faster as well as how to prioritize the allocation of resources for each animal accordingly. In turn, this can also reduce the chance of the euthanization of these animals. Machine learning is a reasonable approach for this project because the project involves finding relations between data points of each animal and making a classification decision. The project can also benefit from general features of an ANN such as the ability to detect all possible combinations among the data points.

# Illustrations

The project consists of three models that were created and tested separately. Figure 1 describes the model that performs an analysis of pet image features to give an adoption speed. Figure 2 shows the model that takes the features provided from a csv file and sentiment data to give an adoption speed. Figure 3 is of the joint training model which is a combination of the two previous models with the outputs of each model being averaged to get the adoption speed.

*Figure* *1: Image Analysis Model Layout [1] Figure* *2: Feature Analysis Model Layout [1]*

  
*Figure* *3: Joint Training Model Layout [1]*

# Background & Related Work

A study had been conducted by members of the School of Veterinary Medicine at the University of California to determine the key factors that led to an animal’s adoption or euthanization. After gathering data based on 4813 dogs and 3301 cats impounded at Sacramento County, the members of the study constructed multiple models using logistic regression. The results of the final model were presented in the form of a ratio (comparison between the baseline value for each pet characteristic) with a 95% confidence interval for each characteristic [2].

The Kaggle competition issued by PetFinder.my to determine adoption speeds resulted in many submissions. The winner solution consists of four models along with stacking, but with little focus on hyperparameter tuning. The output of the model was a classification of the time taken for an animal to be adopted [1].

The inspiration behind using joint training for the third model stems from a research paper. In the paper, it is stated that the multi-task learning techniques on models such as learning tasks jointly have resulted on average in improved generalization performance and decreased overfitting [3]. This idea was applicable to the models because the first two models were able to be trained jointly together, benefitting from this finding.

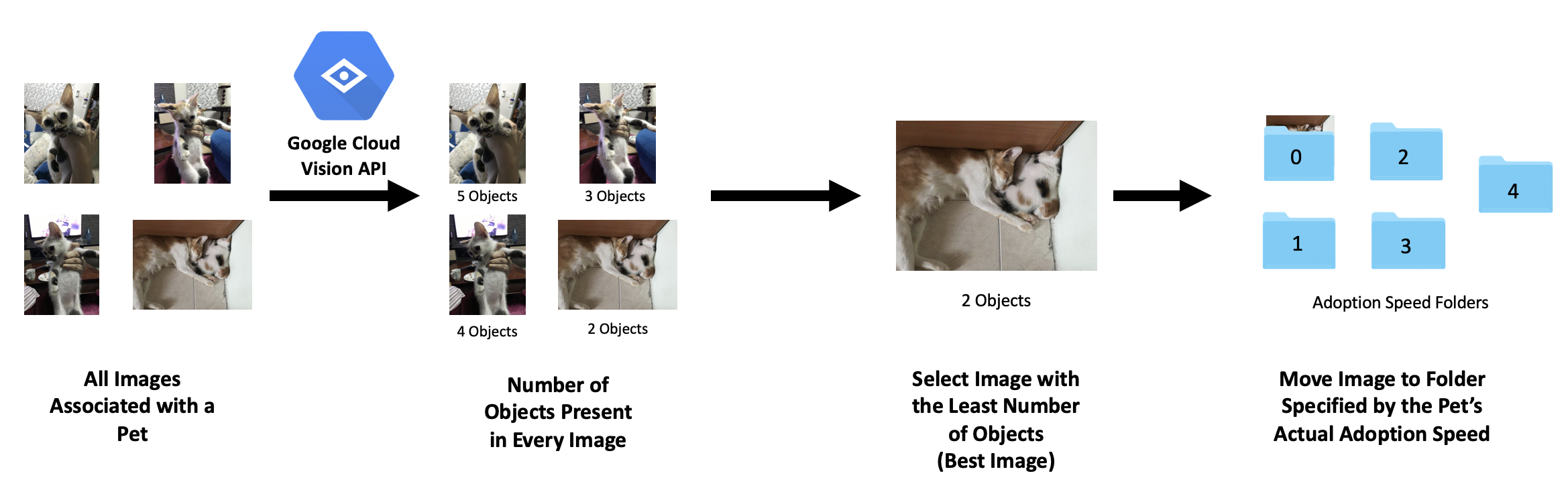
# Data Processing

To prepare the data in a more useful form for the ANN for the csv files, several data preprocessing steps needed to be performed**.**

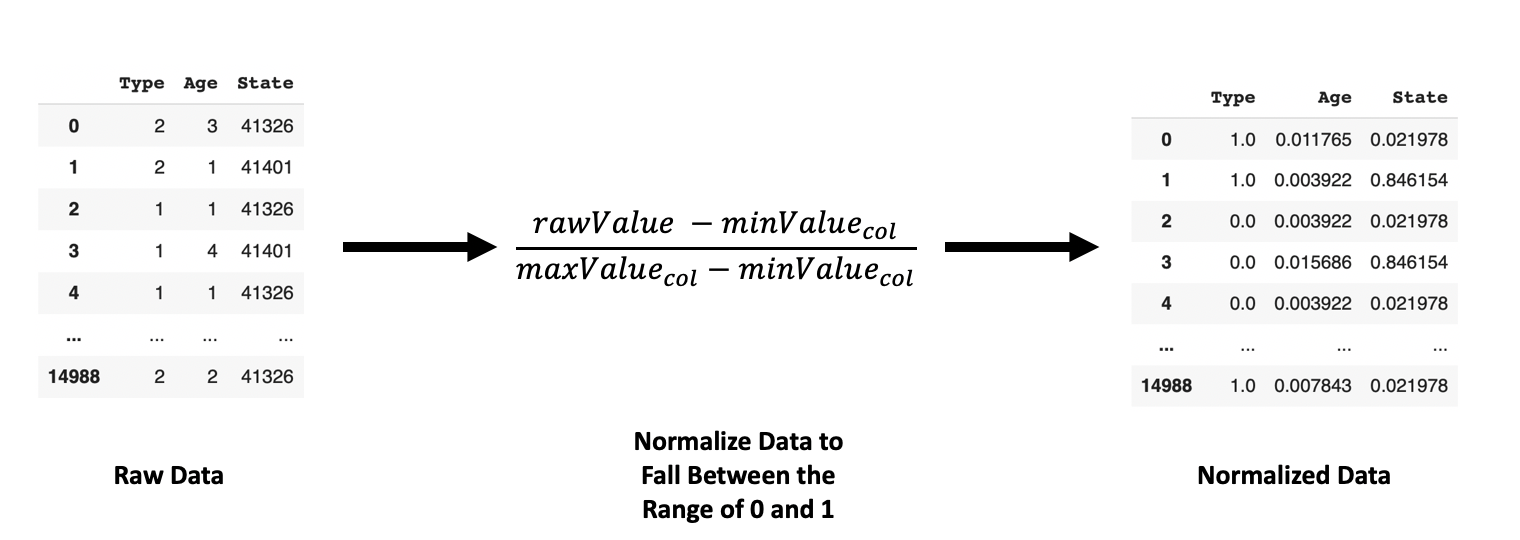
1. To compute the image sentiment value, labels from every image metadata of a pet was analyzed by Google’s Natural Language API that returned a value between -1 (negative sentiment) to 1 (positive sentiment). The pet description sentiment data was also provided for every pet in a separate sentiment JSON file within the same sentiment range. Thus, the overall image sentiment value and pet description sentiment value were appended to every pet in the pet dataset.
2. Weights were added to the age and breed of the pet in the dataset to mimic desirable attributes in a pet. A binary weight of 0 or 1 for age was added if the age of the pet is less than 12 months. A binary weight for breed was added if the pet is a desirable breed based on popular breeds [4][5].
3. Unrelated features that did not contribute to the prediction of the adoption speed were dropped from the pet dataset.
4. All data was normalized to be in the range of 0 and 1 because most attributes had different magnitudes of number ranges.
5. The training data was balanced to ensure that both classes (0 and 1) have the same amount of data. Before data balancing, class 0 had 5228 pets and class 1 had 5267 pets. Pets in class 0 were duplicated to ensure both classes have 5267 pets.

In addition, data preprocessing steps were performed on images for the ANN for images.

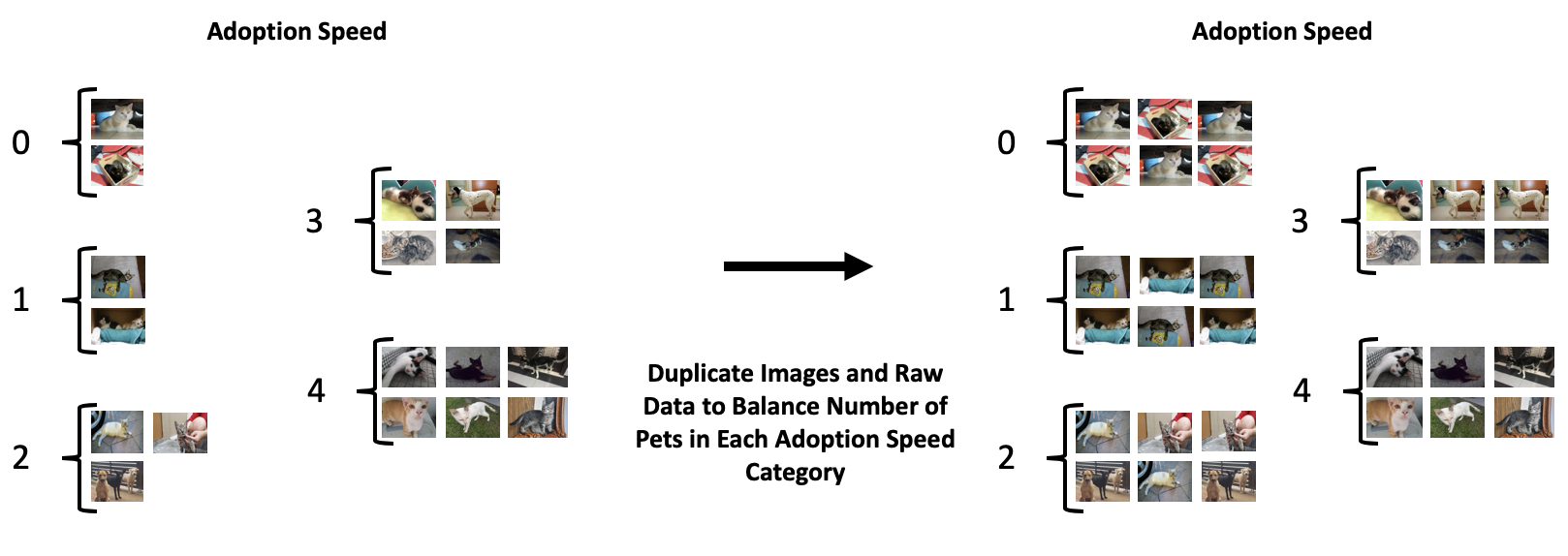
1. All images of a pet were inputted to the Google Cloud Vision API, which returns the number of objects present per image. The image with the least amount of objects other than the pet was selected and was cropped and resized to the pet. This reduced the images in the training dataset from 58312 to 12781 (reduction of 78.1%).
2. Images were organized into folders depending on the adoption speed class of the pet.
3. Tensors for each image were created by inputting the image through AlexNet and saved to a folder depending on the adoption speed class of the pet.



*Figure 4: Image Reduction and Organization [1]*



*Figure 5: Data Normalization [1]*



*Figure 6: Data Balancing [1]*

Table 1 presents the data dictionary for the features in the .csv files, given by Kaggle.

*Table 1: Data Dictionary [1]*

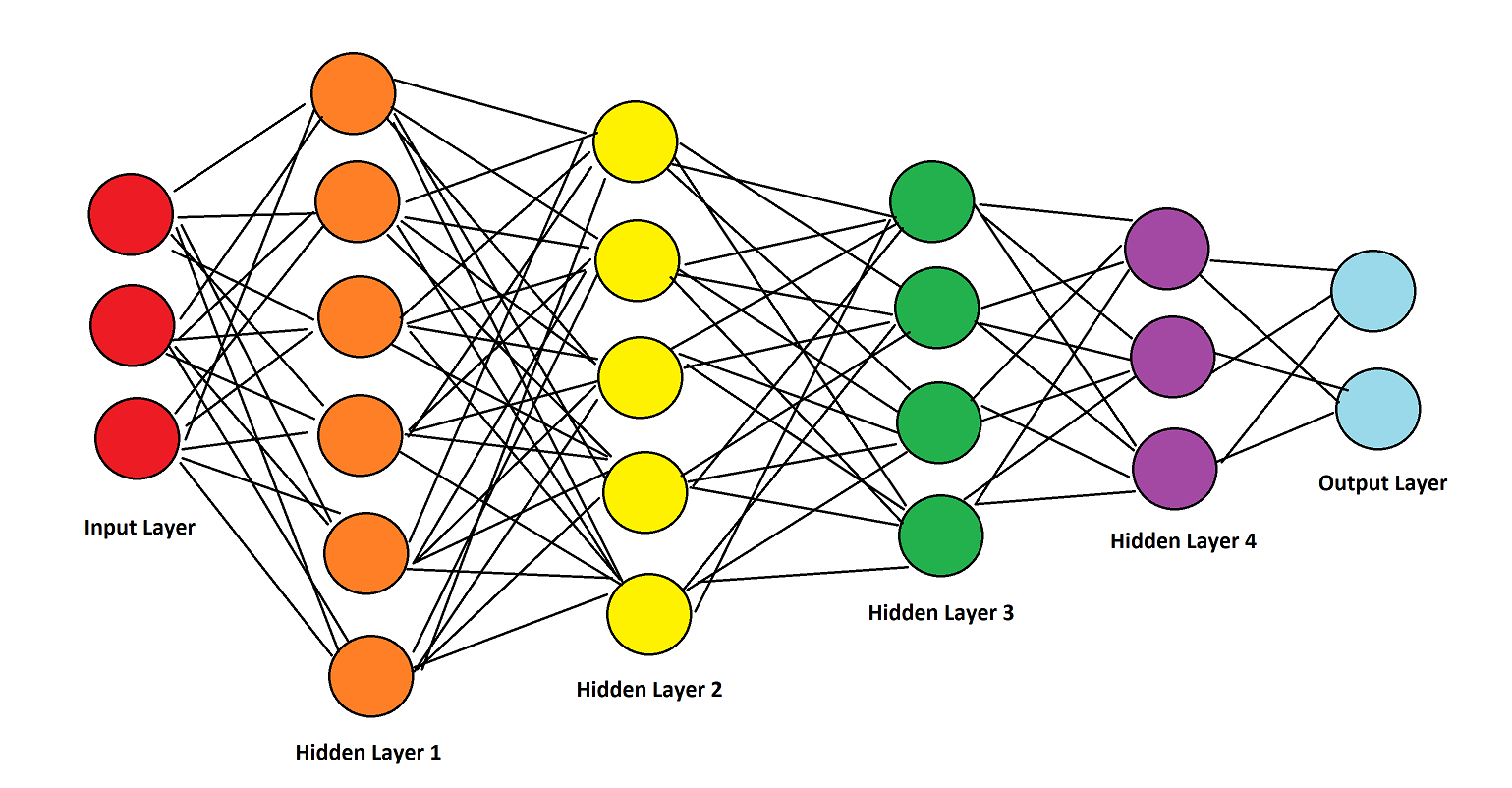
| PetID - Unique hash ID of pet profile |
| --- |
| AdoptionSpeed - Categorical speed of adoption. Lower is faster. This is the value to predict. The value is determined by how quickly, if at all, a pet is adopted. The values are determined in the following way:  0 - Pet was adopted between 0 to 30 days (1st month) after listing.  1 - Pet was adopted after 30 days (1 month) after listing. |
| Type - Type of animal *(1 = Dog, 2 = Cat)* |
| Name - Name of pet *(Empty if not named)* |
| Age - Age of pet when listed, in months |
| Breed1 - Primary breed of pet |
| Breed2 - Secondary breed of pet, if pet is of mixed breed |
| Gender - Gender of pet *(1 = Male, 2 = Female, 3 = Mixed, if profile represents group of pets)* |
| Color1 - Color 1 of pet | Color2 - Color 2 of pet | Color3 - Color 3 of pet |
| MaturitySize - Size at maturity *(1 = Small, 2 = Medium, 3 = Large, 4 = Extra Large, 0 = Not Specified)* |
| FurLength - Fur length *(1 = Short, 2 = Medium, 3 = Long, 0 = Not Specified)* |
| Vaccinated - Pet has been vaccinated *(1 = Yes, 2 = No, 3 = Not Sure)* |
| Dewormed - Pet has been dewormed *(1 = Yes, 2 = No, 3 = Not Sure)* |
| Sterilized - Pet has been spayed / neutered *(1 = Yes, 2 = No, 3 = Not Sure)* |
| Health - Health Condition *(1 = Healthy, 2 = Minor Injury, 3 = Serious Injury, 0 = Not Specified)* |
| Quantity - Number of pets represented in profile |
| Fee - Adoption fee *(0 = Free)* |
| State - State location in Malaysia |
| RescuerID - Unique hash ID of rescuer |
| VideoAmt - Total uploaded videos for this pet |
| PhotoAmt - Total uploaded photos for this pet |
| Description - Profile write-up for this pet. |

# Architecture

There are three models created for this project. An ANN to classify AlexNet features and an ANN to account for csv files. The third model was a joint training model incorporating both ANN models.

**ANN architecture for images (Figure 7):**

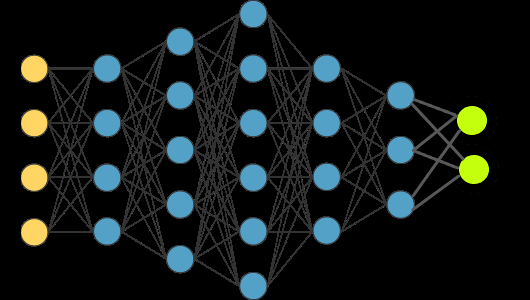
The ANN for images includes 4 hidden layers. The number of parameters are 9216\*5000 + 5000 + 5000\*1000 + 1000 + 1000\*200 + 200\*2 + 2 = **51286402**, and ReLU is used as the activation function.

**

*Figure 7: ANN architecture for images*

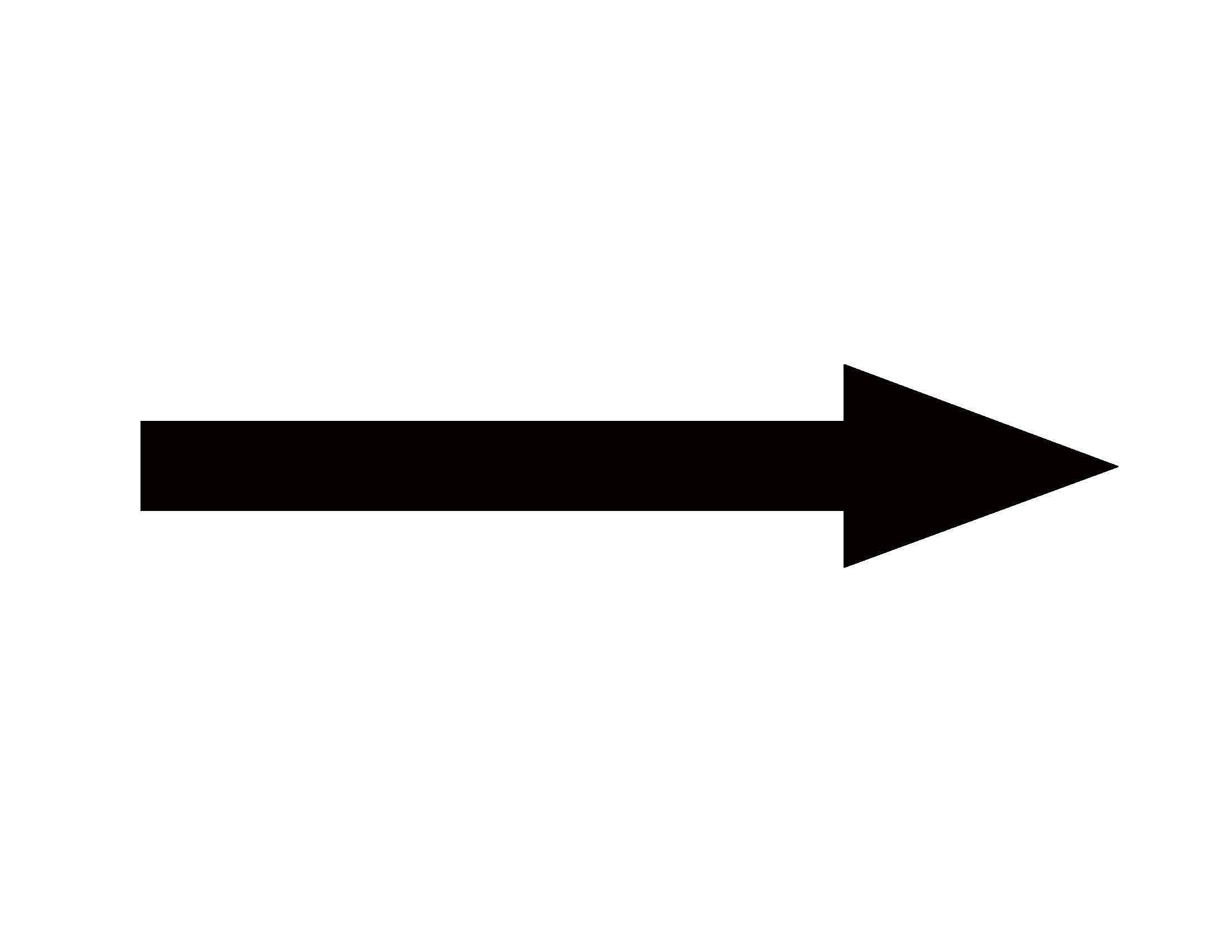
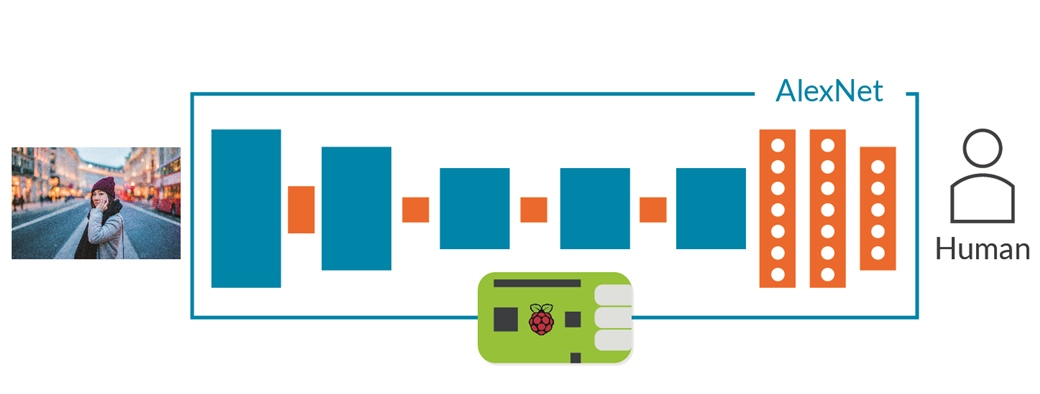
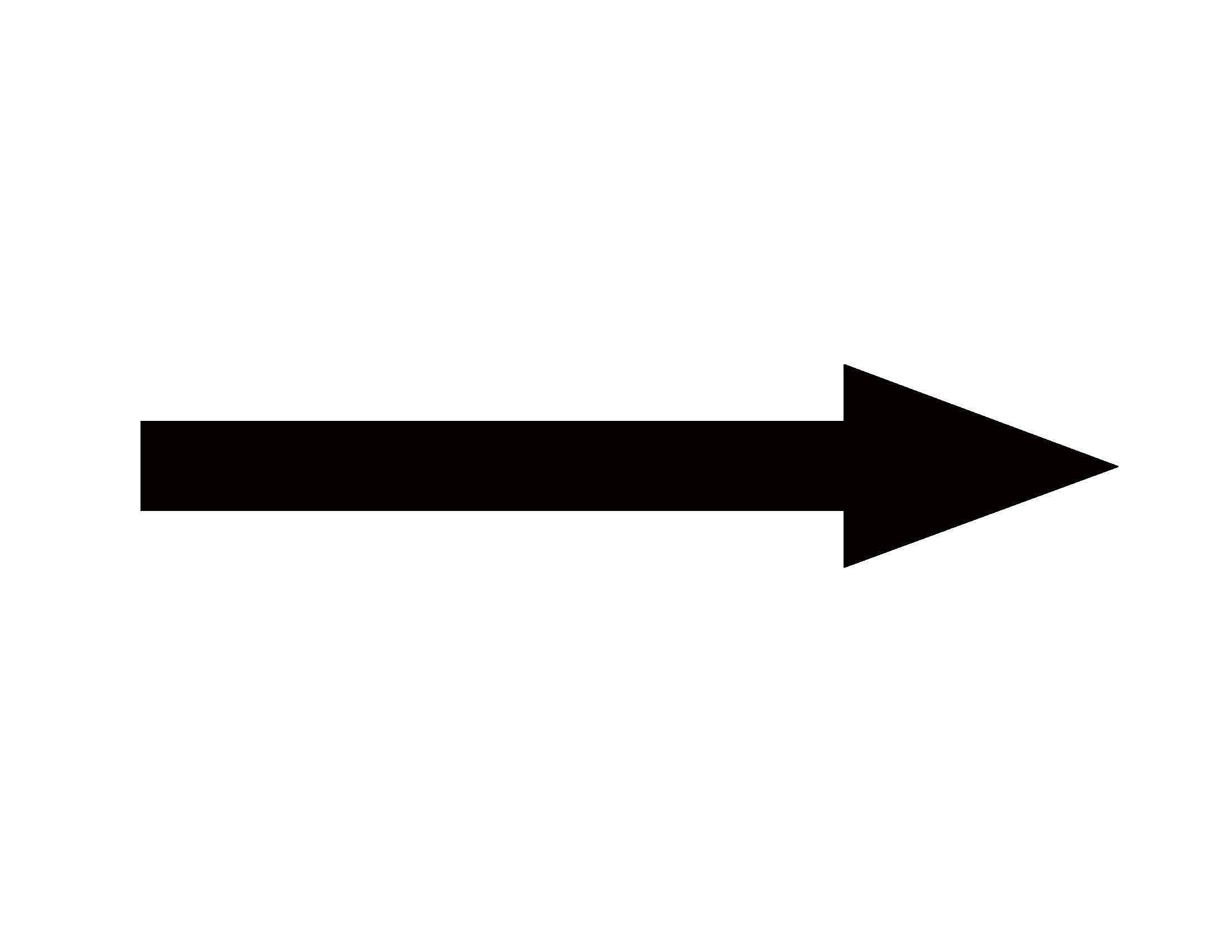
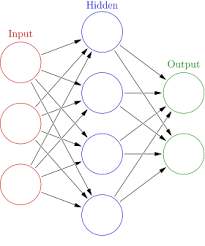
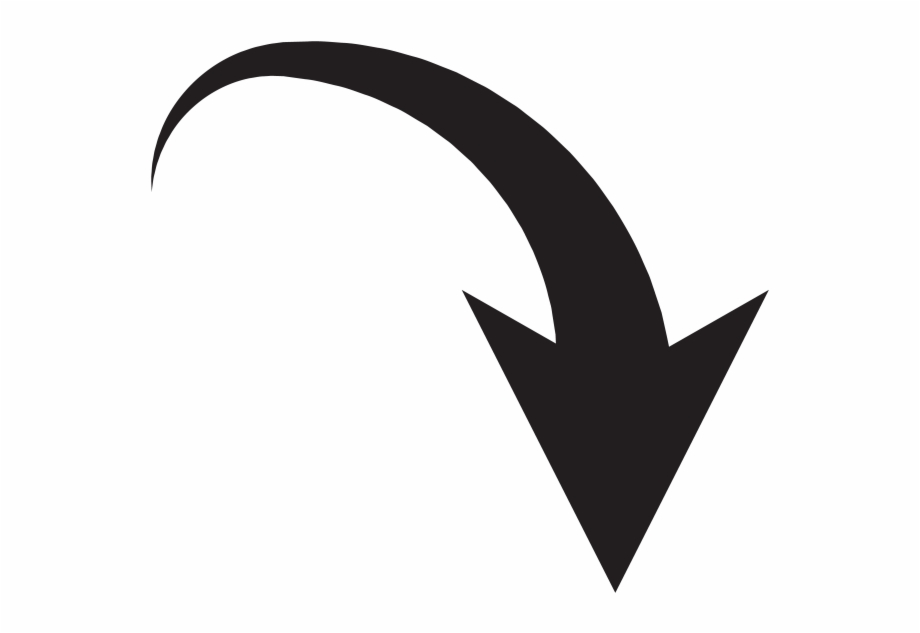
**ANN architecture for csv files (Figure 8):**

The ANN for csv files includes 5 hidden layers. The number of parameters are 18\*500 + 500 + 500\*200 +200 + 200\*100 + 100 + 100\*50 + 50 + 50\*2 = **134950**, and ReLU is used as the activation function.

**

*Figure 8: ANN architecture for csv files*

For joint training (Figure 9), both models were trained together which is aggregating and updating their losses together. The models give one set of predictions together.

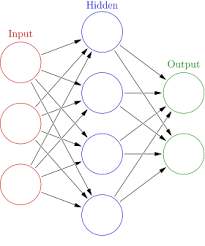
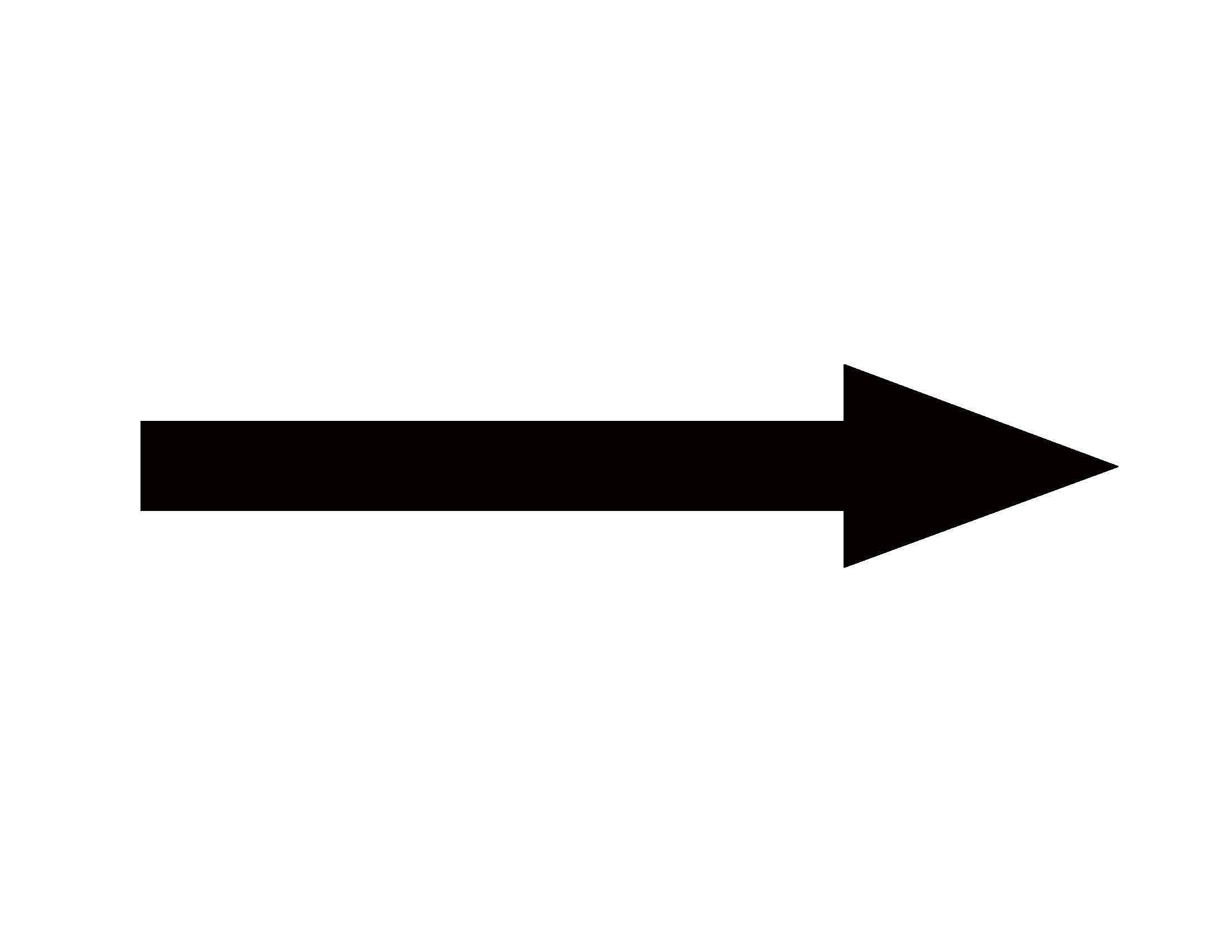
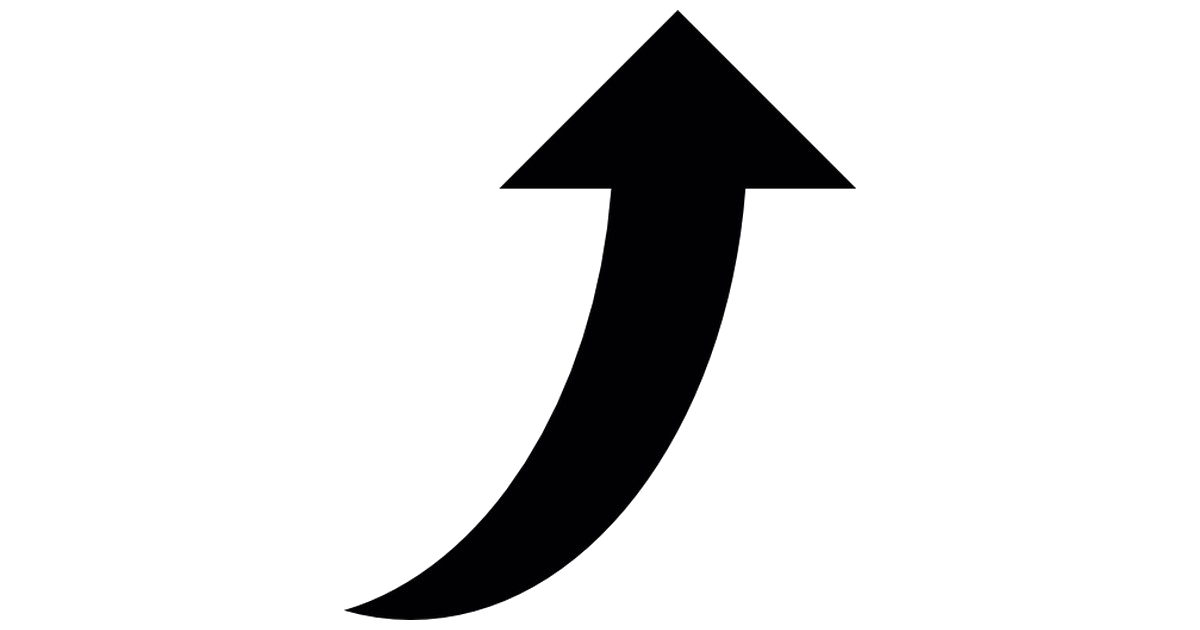
**

*Pet Image AlexNet to extract features ANN*



*Adoption speed*

*prediction*

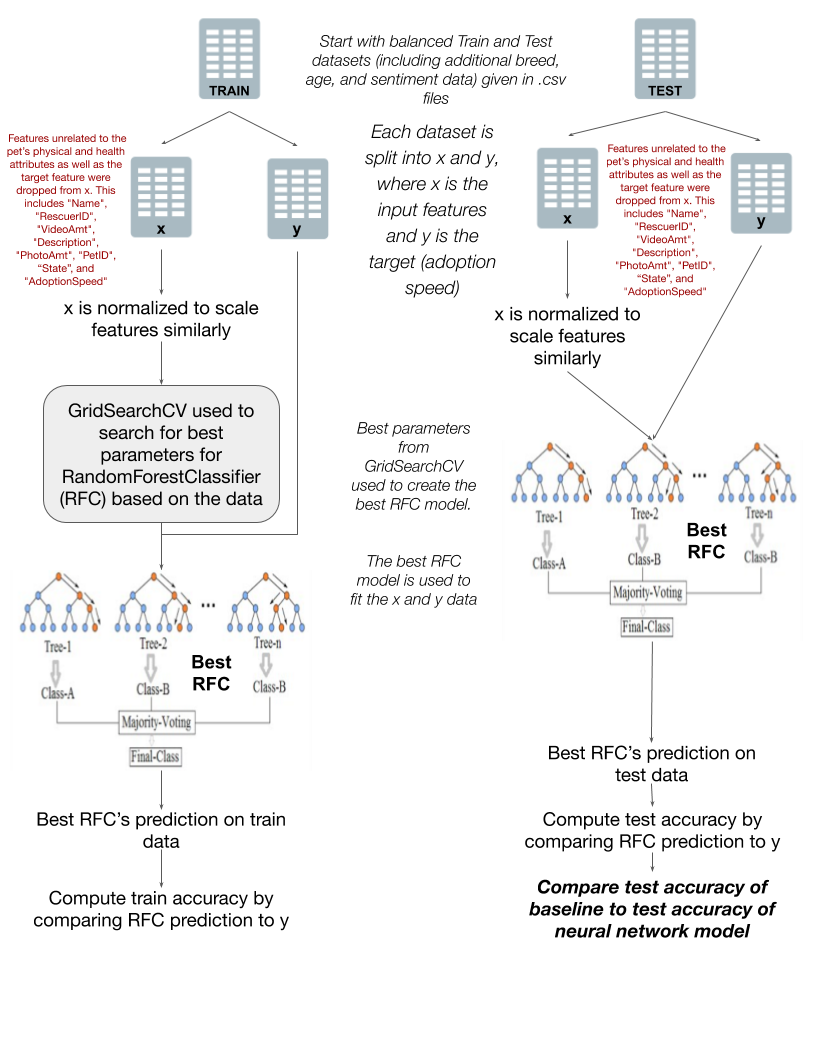
 

*Data from .csv files ANN*

*Figure 9: Joint Training Model*

# Baseline Model

A random forest classifier was used to classify the adoption speed according to the classification system (Table 1) because it is able to aggregate many decision trees together to limit overfitting as well as error due to bias [6]. It also works well on multi-class problems with different types of data [7]. Figure 10 depicts the modelling process from input to output. The baseline model’s test accuracy was 65.5%.

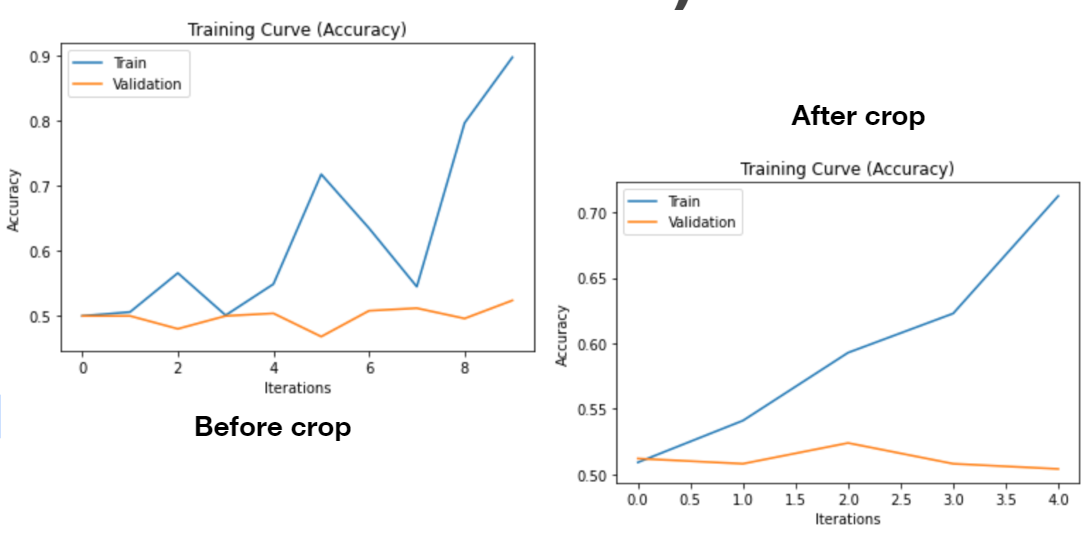


*Figure 10: Modelling process for baseline Random Forest Classifier model [8][9]*

# Quantitative Results and Discussion

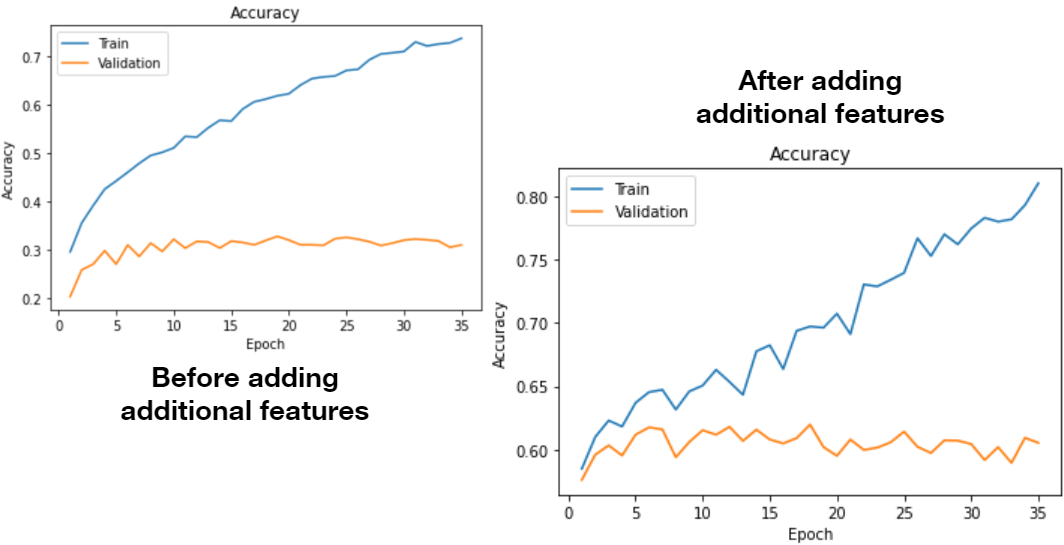
In terms of quantitative results, the team compared the performance of the three models using training curves to illustrate the change in accuracy and determine how the models were fitting.

For the image ANN, the training curve initially looked very erratic, and there was overfitting with the training set (Figure 11). After the images were cropped to get rid of background noise, and epochs were reduced to prevent overfitting, the curve became less erratic (Figure 11).



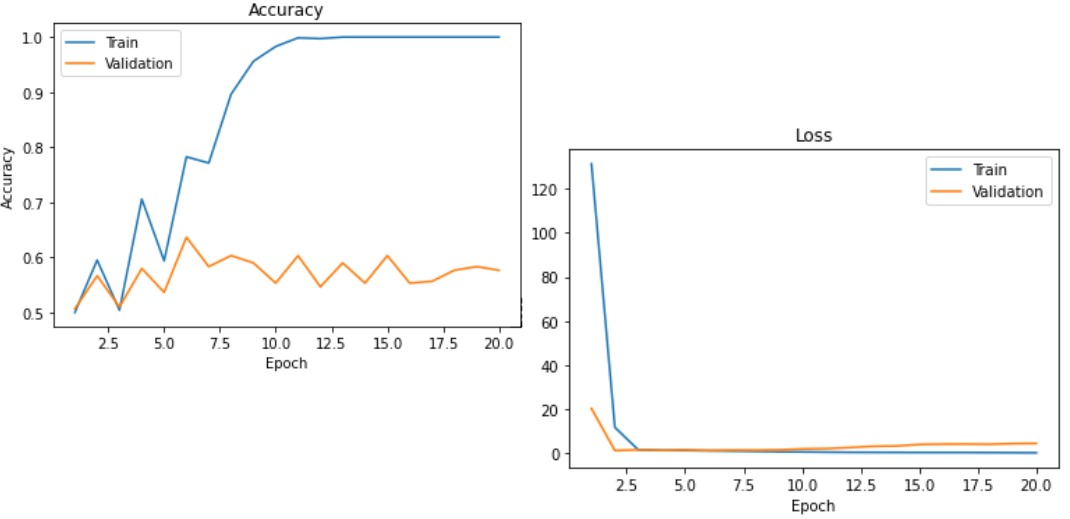
*Figure 11: Training Curves for the AlexNet + ANN Model*

For the csv ANN, after adding additional sentiment, breed, and age features, training accuracy increased. The curve on the right in Figure 12 shows that training accuracy is increasing whereas validation accuracy is plateauing. However, both training and validation curves for accuracy are not extremely far apart compared to the curve on the left, which is a good sign.

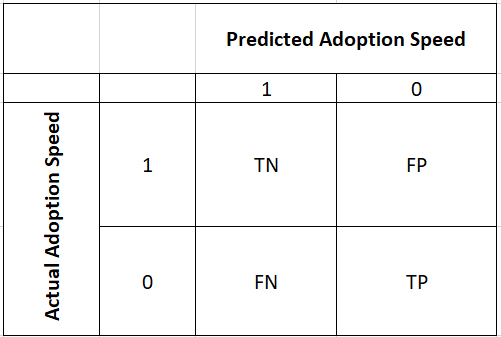


*Figure 12: Training Curves for the csv ANN Model*

For the joint training model, the loss is aggregated from both the image and csv ANN models, so the scale of the loss graph is much higher than usually seen in training curves (Figure 13). However, the training and validation losses are almost convergent by the end, which is good. This was also a smaller dataset, so training accuracy was able to hit 100%, whereas similar to the csv ANN, validation accuracy is plateauing around 60% (Figure 13).

*Figure 13: Training Curves for the Joint Training Model*

The performance of the baseline model was compared to the csv ANN model (since it had the highest accuracy of all models) on the actual test data labels using a confusion matrix to identify the amount of true predictions vs. misclassifications. For the following confusion matrices, 0 is when a pet is adopted within 1 month (positive) and 1 is when a pet is adopted after 1 month (negative), as shown in Figure 14.



*Figure 14: Confusion Matrix Format*

The csv ANN (Table 2) had a precision of 69.11%, which means 69.11% of our results are relevant. 32.20% are misclassified overall (). The model was better at detecting positives (71.62%) compared to true negatives (63.43%).

*Table 2: Confusion Matrix For csv ANN Predictions Compared To Actual Labels*

| True Negative = 1176 | False Positive = 678 |
| --- | --- |
| False Negative = 601 | True Positive = 1517 |
| Sensitivity = 71.62% = proportion of true positives correctly identified  Specificity = 63.43% = proportion of true negatives correctly identified | |

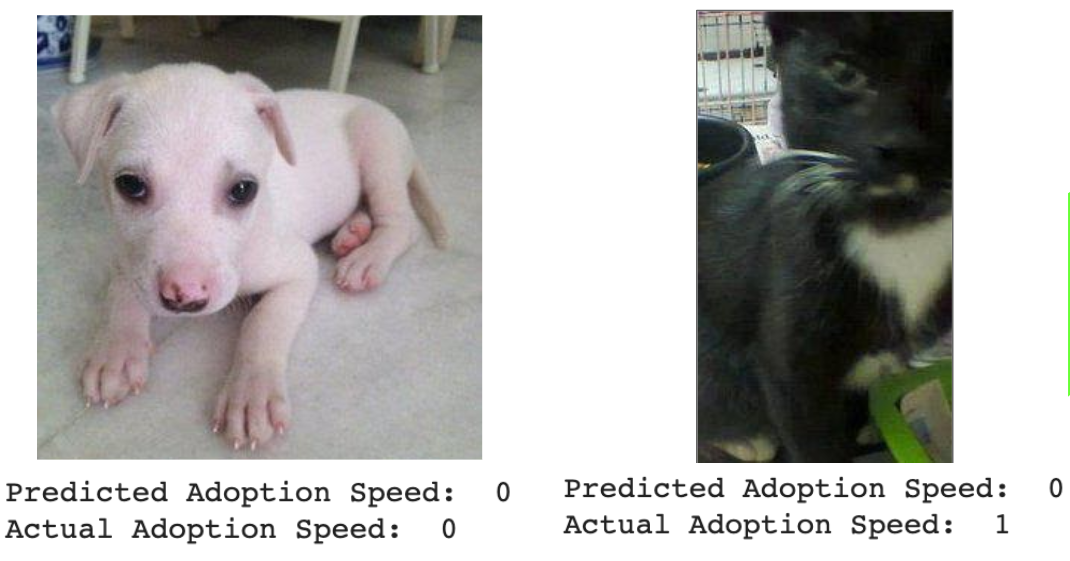
The baseline model (Table 3) had a precision of 72.09%, slightly higher than the csv ANN. 34.52% are misclassified overall (false positive + false negatives). It had a 10% lower sensitivity compared to the csv ANN, and a 10% higher specificity rate. Thus, the baseline model was better at detecting negatives (74.45%) compared to the csv ANN (63.43%).

*Table 3: Confusion Matrix For Random Forest Baseline Model Predictions Compared To Actual Labels*

| True Negative = 1382 | False Positive = 472 |
| --- | --- |
| False Negative = 899 | True Positive = 1219 |
| Sensitivity = 57.55% = proportion of true positives correctly identified  Specificity = 74.45% = proportion of true negatives correctly identified | |

# Qualitative Results and Discussion

Sample predictions for the image ANN model are shown in Figure 15 below. The training accuracy (Figure 11) for this model was high, but the validation accuracy was lower and plateauing because of the input data. Quite a few images had issues, such as bad quality, dark, or they did not show the entire body/face of the pet.



*Figure 15: Image ANN Predictions (good input and correct prediction on the left, bad input and incorrect prediction on the right)*

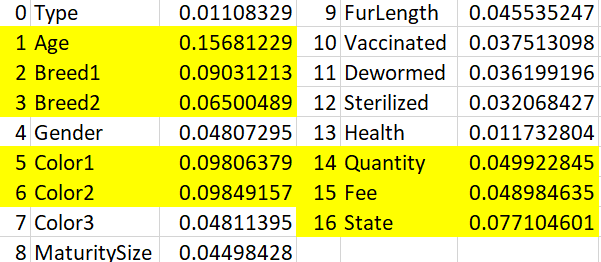
For the csv ANN, the training accuracy (Figure 12) is 20% higher than the validation accuracy. Similar to the image ANN, it was hypothesized that this was due to the input data. There were discrepancies in the way unspecified values were labelled by the adoption agency in the data dictionary (Table 4). For instance, unspecified “FurLength” is marked with **0**, whereas unspecified “Vaccinated” is marked with a **3**. Mixed groups of pets were given a Gender value of 3, rather than 1 (male) or 2 (female). Despite intuitively seeming very important, the inconsistent numbering structure of unspecified values led to those features being less helpful for the model, and thus, are not included in the top eight important features of the baseline model, which outputs feature importances (Figure 17).

*Table 4: Subset of Data Dictionary, Discrepancies Highlighted in Yellow*

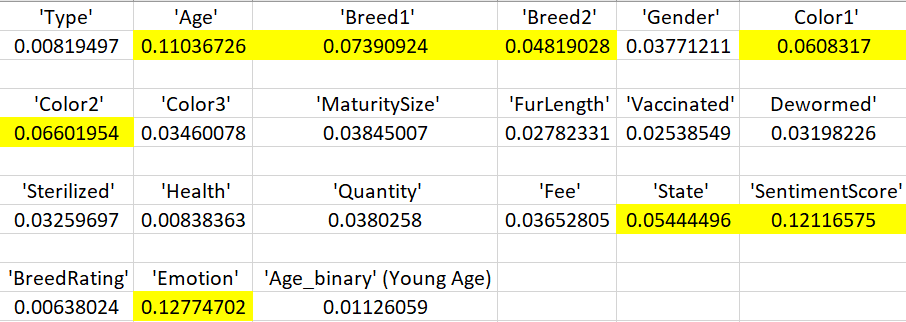
| MaturitySize - Size at maturity *(1 = Small, 2 = Medium, 3 = Large, 4 = Extra Large,*  *0 = Not Specified)* |
| --- |
| FurLength - Fur length *(1 = Short, 2 = Medium, 3 = Long, 0 = Not Specified)* |
| Vaccinated - Pet has been vaccinated *(1 = Yes, 2 = No, 3 = Not Sure)* |
| Dewormed - Pet has been dewormed *(1 = Yes, 2 = No, 3 = Not Sure)* |
| Sterilized - Pet has been spayed / neutered *(1 = Yes, 2 = No, 3 = Not Sure)* |
| Health - Health Condition *(1 = Healthy, 2 = Minor Injury, 3 = Serious Injury, 0 = Not Specified)* |

Based on the issues with the input data, throughout the project, the focus has been on increasing accuracy of the models, such as by cropping images (Figure 11) or adding additional data as described in the [Data Processing](#_ay25x2iaz3nu) section.

The way additional data was added was strategic. As mentioned previously, the baseline model was a random forest model, so it outputted feature importances denoting how useful each feature was towards making the prediction. For instance, breed and age are in the top three features for the random forest model, and are intuitively important too (Figure 16). Based on this, a top 10 breeds feature was added to denote popular breeds and a young age feature was created to denote pets less than or equal to 12 months [4][5]. Sentiment data was also added for a more human aspect. The feature importances was run again and Emotion and Sentiment Score are seen as highly important, showing that adding more data did help the csv ANN model (Figure 17).



*Figure 16: R*a*ndom Forest Classifier Feature Importances, Top 8 Highlighted*

**

*Figure 17: R*a*ndom Forest Classifier Feature Importances After More Data, Top 8 Highlighted*

# Evaluate Model on New Data

The test dataset provided 3972 pet details in a different csv file from the training dataset. There was a different folder for testing images as well. All data processing for testing and training data was done separately in different files/folders to make sure there was no duplication of data.

The testing csv file and images were not used for any training purposes or hyperparameter tuning. Testing was only done using pre-trained models; a function was run on the pre-trained model using all testing data to get the accuracy. The highest testing accuracy was found to be 67.8% for the csv ANN based on the predictions of the winning team because the competition does not provide the adoption speed of testing data.

In order to evaluate the model on new data that is not already provided, values for all of the features in Table 1 would need to be provided.

# Ethical Considerations

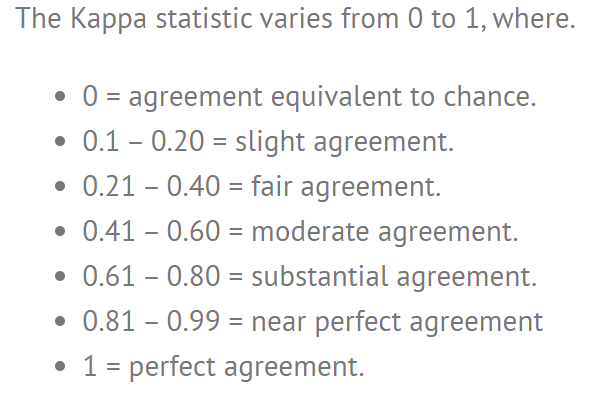
In terms of the ethics landscape, harmfulness and fairness are the main issues that apply to this project.

There are no current standards on pets giving consent because that would be hard to enforce - the pets are owned by the Malaysian adoption agency who posted the dataset on Kaggle. In the dataset, there are pet images as well as text-based feature data. Sensitive information relating to the images have already been masked for privacy. The original dataset provided is a largely unbalanced dataset as it has more dogs than cats, and more females compared to males. There are also groups of pets in the dataset, which may be a combination of animals and genders. Accordingly, the prediction may be less accurate for such unbalanced data. However, the team did duplicate data entries to get an equal number of data for the two adoption speeds. Pet adoption preferences and pet types available for adoption may vary worldwide - this is a limitation of the model as it is based on the Malaysian dataset provided only.

Additionally, the project’s goal is to help identify features that speed up pet adoption to be applied to improve other pet’s chances. However, there is a possibility that this information may be used maliciously to euthanize pets that have a low adoption speed or stimulate breeding of pets that specifically meet these standards.

# Project Difficulty Reflection

Overall, this was a difficult project due to the variety and quality of data that was provided by the competition. Due to this, the team had to determine which pet characteristics in the pet dataset and pet features in the images were the most important and had the greatest influence on determining the adoption speed. Furthermore, there were inconsistencies in the handling of data, which included low quality images, and different methods to handle missing data for pets. Finally, the criteria when adopting pets is highly subjective for humans, which also adds a layer of difficulty when developing models for this problem. This is also evident with other teams on Kaggle as the Cohen Kappa value of the winning team is 0.453, which is only considered “moderate agreement” with the actual values [10].



*Figure 18: Kappa Statistic Explanation [10]*

Even after implementing different models to utilize as much data as we can, including training two models jointly, the team noticed that the increase in complexity of the models did not result in an increase in accuracy. Thus, the team finally opted to use the ANN for CSV data model because it resulted in the highest training accuracy of 67.8% when compared to the winning team’s results on Kaggle.

# Link to Code Files

<https://drive.google.com/drive/folders/1M6EDVhL5DX1YPY0VDaEW5w4n4ATN_4-j?usp=sharing>

# References

[1] Kaggle.com. (2020). PetFinder.my Adoption Prediction | Kaggle. [online] Available at: https://www.kaggle.com/c/petfinder-adoption-prediction/data [Accessed 22 Feb. 2020].

[2] Lepper, M., Kass, P. and Hart, L. (2002). Prediction of Adoption Versus Euthanasia Among Dogs and Cats in a California Animal Shelter. Journal of Applied Animal Welfare Science, 5(1), pp.29-42.

[3] Papers.nips.cc. 2020. [online] Available at: https://papers.nips.cc/paper/7296-supervised-autoencoders-improving-generalization-performance-with-unsupervised-regularizers.pdf [Accessed 8 April 2020].

[4] CatTime. 2020. *The 25 Most Popular Cat Breeds In The World [PICTURES] - Cattime*. [online] Available at: https://cattime.com/lifestyle/781-the-25-most-popular-cat-breeds [Accessed 8 April 2020].

[5] Dogtime. 2020. *10 Most Popular Dog Breeds In America - Dogtime*. [online] Available at: https://dogtime.com/dog-health/general/4333-most-popular-dog-breeds-list [Accessed 8 April 2020].

[6] Medium. (2020). Understanding Random Forest. [online] Available at: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 [Accessed 22 Feb. 2020].

[7] The Kernel Trip. (2020). Random forest vs SVM. [online] Available at: https://www.thekerneltrip.com/statistics/random-forest-vs-svm/ [Accessed 22 Feb. 2020].

[8] "Data Icon" Wikipedia, 2011. [Online]. Available: https://commons.wikimedia.org/wiki/File:AWS\_Simple\_Icons\_Non-Service\_Specific\_Corporate\_Data\_Center.svg. [Accessed: March 15, 2020].

[9] "Random Forest Simplified Diagram" Wikipedia, 2017. [Online]. Available: https://commons.wikimedia.org/wiki/File:Random\_forest\_diagram\_complete.png. [Accessed: March 15, 2020].

[10] Statistics How To. 2020. *Cohen's Kappa Statistic - Statistics How To*. [online] Available at: https://www.statisticshowto.com/cohens-kappa-statistic [Accessed 8 April 2020].



*Figure* *1: Image Analysis Model Layout [1] Figure* *2: Feature Analysis Model Layout [1]*