***DOG BREED IDENTIFICATION VIA APPLIED MACHINE LEARNING***

A COMPUTER VISION APPROACH TO DOG BREED IDENTIFICATION

WESTERN GOVERNORS UNIVERSITY

Lydia Strough | SID#002452624

**TABLE OF CONTENTS**

[A. PROJECT PROPOSAL FOR BUSINESS EXECUTIVES 3](#_Toc135142425)

[A.1. LETTER OF TRANSMITTAL 3](#_Toc135142426)

[A.2. PROJECT RECOMMENDATION 4](#_Toc135142427)

[A.2.a. PROBLEM SUMMARY 4](#_Toc135142428)

[A.2.b. APPLICATION BENEFITS 4](#_Toc135142429)

[A.2.c. APPLICATION DESCRIPTION 5](#_Toc135142430)

[A.2.d. DATA DESCRIPTION 5](#_Toc135142431)

[A.2.e. OBJECTIVES AND HYPOTHESIS 5](#_Toc135142432)

[A.2.f. METHODOLOGY 5](#_Toc135142433)

[A.2.g. FUNDING REQUIREMENTS 6](#_Toc135142434)

[A.2.h. DATA PRECAUTIONS 6](#_Toc135142435)

[A.2.i. DEVELOPER’S EXPERTISE 7](#_Toc135142436)

[B. PROJECT PROPOSAL 8](#_Toc135142437)

[B.1. PROBLEM STATEMENT 8](#_Toc135142438)

[B.2. CUSTOMER SUMMARY 8](#_Toc135142439)

[B.3. EXISTING SYSTEM ANALYSIS 8](#_Toc135142440)

[B.4. DATA 11](#_Toc135142441)

[B.5. PROJECT METHODOLOGY 11](#_Toc135142442)

[B.6. PROJECT OUTCOMES 12](#_Toc135142443)

[B.7. IMPLEMENTATION PLAN 12](#_Toc135142444)

[B.8. EVALUATION PLAN 13](#_Toc135142445)

[B.9. RESOURCES AND COSTS 13](#_Toc135142446)

[B.10. TIMELINE AND MILESTONES 14](#_Toc135142447)

[C. APPLICATION 17](#_Toc135142448)

[D. POST-IMPLEMENTATION REPORT 18](#_Toc135142449)

[D.1. ORGANIZATION VISION 18](#_Toc135142450)

[D.2. DATASET 18](#_Toc135142451)

[D.3. DATA PRODUCT CODE 19](#_Toc135142452)

[D.4. OBJECTIVE AND HYPOTHESIS VERIFICATION 23](#_Toc135142453)

[D.5. EFFECTIVE VISUALIZATION AND REPORTING 23](#_Toc135142454)

[D.6. ACCURACY ANALYSIS 24](#_Toc135142455)

[D.7. APPLICATION TESTING 25](#_Toc135142456)

[D.8. APPLICATION FILES 26](#_Toc135142457)

[D.9. USER GUIDE 27](#_Toc135142458)

[D.10. SUMMATION OF LEARNING EXPERIENCES 28](#_Toc135142459)

[E. REFERENCES 29](#_Toc135142460)

# A. PROJECT PROPOSAL FOR BUSINESS EXECUTIVES

## A.1. LETTER OF TRANSMITTAL

April 03, 2023  
  
Janelle Dixon, CEO  
Animal Humane Society  
1411 Main St NW  
Coon Rapids, MN 55448

Ms. Dixon,

As I am sure you are aware, the Animal Humane Society (AHS) relies on its capable veterinary technicians and shelter staff to identify all companion animal intakes by breed/species. Last year, in 2022, the AHS admitted 5,193 dogs – 40% of the organization’s companion animal intake (2022 Annual Report, n.d.). However capable the organization’s shelter staff may be, with more than 200 dog breeds recognized by the American Kennel Club, assigning the correct dog breed to a dog purely based on visuals can be difficult and time-consuming. Understandably, it is not unheard of for these dogs to be mislabeled. “-you’ll see dogs in our shelter whose breed description doesn’t seem to match their appearance” (How breed is determined at AHS and other animal shelters, n.d.).

To minimize the risk of human error, the AHS should incorporate machine learning (ML) into their dog breed identification process. Specifically, I propose the implementation of a dog breed image recognition application to aid AHS vet techs in classifying new shelter dog admissions. The incorporation of machine learning technologies into business practices has been known to come with several benefits. ML has been known to improve efficiency, aid decision-making, save costs, and increase customer satisfaction (GeekforGeeks, 2021). With the implementation of this ML tool, AHS vet techs will be able to classify dogs with improved accuracy and efficiency - thus minimizing the risk of human error and saving the organization valuable resources.

The proposed product is a user-friendly web application with the ability to quickly and accurately classify a dog by breed with a single image. Developing, installing, and maintaining this product will cost the organization approximately $83,070, with a yearly 15% maintenance fee of $9,585. The developer in question for the project has a strong computer science and project management background, with over two years of software development experience and several certifications including, CompTIA Project+ and ITIL 4 Foundation, to validate their expertise.

Thank you for your time and consideration. I look forward to hearing from you. Feel free to reach out anytime if you have any questions or concerns.

Sincerely,   
Lydia Strough

## A.2. PROJECT RECOMMENDATION

### A.2.a. PROBLEM SUMMARY

In 2022, 40% of the Animal Humane Society (AHS) companion animal intake were dogs - totaling 5,193 admissions. All dogs admitted to the AHS must be identified by breed. Unfortunately, with more than 200 dog breeds recognized by the American Kennel Club, assigning the correct dog breed to every dog purely based on visuals can be difficult. Understandably, it is not unheard of for these dogs to be mislabeled. A University of Florida study showed that different shelter staffers who evaluated the same dogs at the same time had only a *moderate* level of agreement among themselves. These results raise concern as shelter workers and veterinarians are expected to determine the breeds of dogs in their facilities daily (University of Florida, n.d.).

To minimize the risk of human error, the AHS should incorporate machine learning (ML) into their dog breed identification process - specifically, a dog breed image recognition application. The incorporation of ML technologies into business practices has been known to come with several benefits. ML has been known to improve efficiency, aid decision-making, save costs, and increase customer satisfaction (GeekforGeeks, 2021). The proposed product will be a user-friendly web application with the ability to quickly and accurately classify a dog by breed with a single image. Providing AHS vet techs with this ML tool will help them identify dogs with increased accuracy, thus decreasing the number of mislabeled dogs housed within the AHS, and saving the organization valuable resources.

### A.2.b. APPLICATION BENEFITS

In recent years, machine learning (ML) has become increasingly significant. Large companies like Netflix, Amazon, YouTube, Facebook, Google, MIT, and more have implemented ML into their business practices (Reilly, 2021). In healthcare settings, ML-enabled tools are used to analyze medical reports and images. These tools are associated with improved patient diagnoses, cost savings, and more (Ramakrishnan, 2023). In a 2022 study, it was shown that ML-enabled tools performed with equal or better accuracy than experienced medical professionals in diagnosing patients. The study concluded with the following results: the model performed with an accuracy of 99.3%, compared to the hand orthopedic surgeons with accuracy ratings of 97.3%, 94.7%, and 96.7%. “The accuracy of the (model) was better than that of the hand orthopedic surgeons” (Suzuki, 2022). It is reasonable to assume that an ML-enabled tool could perform similarly when asked to classify a breed of dog. In fact, there are numerous studies involving ML classification models and species identification.

In 2020 researchers conducted a study that provided a new approach to the identification of Amur tigers. Rather than manually identifying the tigers, they utilized deep (machine) learning algorithms to automatically identify them. The study aimed to identify individual tigers by the body stripes of the tigers, as the stripes of the tiger on both the left and right sides of their body are inconsistent. Images of each tiger were separated into two categories, “left” and “right” to distinguish the different sides. The experiment resulted in an *excellent* recognition accuracy of 90.48% (left) and 93.50% (right) (Shi, 2020).

In summary, there is a positive correlation between the implementation of machine learning into business practices. ML-enabled tools can be used in species identification, and it is also evident that they can outperform experienced medical professionals in their field. It is safe to say that if implemented,ML technology could be an invaluable asset for the AHS as it could provide an accurate, real-time solution for discriminating dog breeds, saving the organization valuable time and other resources.

### A.2.c. APPLICATION DESCRIPTION

The final product of the proposed project will be a user-friendly dog breed image classification tool, in the form of a web application, compatible with various operating systems and browsers. The user will select the “upload image” icon and be prompted to upload an image of the dog in question. The application will run the image through the ML model and output the top five most likely dog breeds with their associated prediction percentages. With the implementation of this ML tool, AHS vet techs will be able to classify dogs with improved accuracy and efficiency - thus minimizing the risk of human error, decreasing the number of mislabeled dogs housed within the AHS, and saving the organization valuable time and other resources.

### A.2.d. DATA DESCRIPTION

The proposed project will extract data from a renowned machine learning and data science community, Kaggle.com. Utilizing this *free*, pre-existing dataset will save the AHS money and time, rather than expending extensible company resources to create data for the project. The dataset in question is the “70 Dog Breeds-Image Data Set”, 225.78 MB of dog images (9,346 in total). A CSV file accompanies the dataset, containing the image’s associated metadata, including image paths and their dog breed classification. Although utilizing a pre-existing dataset saves the company time and money because they do not have to create a dataset from scratch, the data will need to undergo thorough exploration to ensure that it meets all project requirements.

### A.2.e. OBJECTIVES AND HYPOTHESIS

The final product of the proposed project will be a user-friendly dog breed image classification tool, in the form of a web application. It is hypothesized that the ML-enabled tool will be able to identify a breed of dog in real-time with an accuracy rating of 90% or higher.

### A.2.f. METHODOLOGY

It has been determined that the waterfall methodology is the best development strategy for this project. With its predictability, this method will provide the team with a clear project scope, schedule, and cost estimation - all aspects that will aid the team in staying on track. Although the project will struggle more with changes than its competitors, the waterfall method will save time, money, and other resources that can be allocated to unforeseen changes when the time arises. In general, the waterfall method’s strengths will lead to faster and cheaper project completion.

A rough outline of the project methodology and each phase can be found below:

1. **Initiation (Pre-project setup):** The project is chartered. The core team is assembled.
2. **Planning:** Plans are outlined for the management of all aspects of the project.
3. **Execution:** Actual work of the project is completed. Objectives and requirements are met.
4. **Deployment:** Deploy the product to a small subset of users.
5. **Monitor and Control:** Project progress and performance are measured. Changes are proposed where necessary to meet existing or shifting goals.
6. **Closing:** Confirm objectives were met. Deliver the final product to the customer.

### A.2.g. FUNDING REQUIREMENTS

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost ($ USD)** |
| 240 work hours | Project management team (4 members) | 40,320 |
| 120 work hours | IT team (4 members) | 13,920 |
| 80 work hours | Development team (3 members) | 5,160 |
| Hardware | Personal computers, Desktop, etc. | 4,500 |
| Software | Google colab, Gardio.app | 0 |
| Hosting service | HuggingFace.co “public space” | 0 |
| Dataset | Kaggle dataset (“70 Dog Breeds-Image Data Set”) | 0 |
| 15% buffer |  | 9,585 |
| Annual maintenece (15%) | Quality assurance, performance measuring, etc. | 9, 585 |
|  | **Total** | 83,070 |

### A.2.h. DATA PRECAUTIONS

The proposed project will not be handling any sensitive data throughout project production. However, precautions will be taken to ensure access to the application model and other sensitive organizational data regarding the project is controlled. This can be accomplished by securing company devices with strong passwords and the most up-to-date software, operating systems, and malware protection. During the project’s planning phase, the team will also produce a risk management plan (RMP) to help mitigate potential risks and incidents that could impact the project. The RMP will include backup and recovery plans, etc.

### A.2.i. DEVELOPER’S EXPERTISE

The project will be placed in the hands of an ambitious, detail-oriented computer science graduate with strong interpersonal skills and a passion for technology. This person has a solid computer science and project management background, with over two years of software development experience and several certifications including, CompTIA Project+ and ITIL 4 Foundation, to validate their expertise. Coupled with their relentless work ethic, I am confident that they can produce the desired project’s output accurately and efficiently.

# B. PROJECT PROPOSAL

## B.1. PROBLEM STATEMENT

This proposal describes a machine learning (ML) project that will aid the Animal Humane Society (AHS) with dog breed identification. Details on the project including benefits, design, data set descriptions, and more can be found below.

## B.2. CUSTOMER SUMMARY

In 2022 the AHS admitted 5,193 dogs – 40% of the organization’s companion animal intake (2022 Annual Report, n.d.). All dogs admitted to the AHS must be identified by breed. The AHS relies on its vet techs to label these dogs. With more than 200 dog breeds recognized by the American Kennel Club, assigning the correct dog breed to a dog purely based on visuals can be difficult. Understandably, it is not unheard of for these dogs to be mislabeled.

“-you’ll see dogs in our shelter whose breed description doesn’t seem to match their appearance” (How breed is determined at AHS and other animal shelters, n.d.).

To minimize the risk of human error, the AHS should incorporate ML into their dog breed identification process - specifically, a dog breed image recognition application. The incorporation of ML technologies into business practices has been known to come with several benefits. ML has been known to improve efficiency, aid decision-making, save costs, and increase customer satisfaction (GeekforGeeks, 2021). The proposed product will be a user-friendly web application with the ability to quickly and accurately classify a dog by breed with a single image. Providing AHS vet techs with this ML-enabled tool will help them identify dogs with increased accuracy, thus decreasing the number of mislabeled dogs housed within the AHS, and saving the organization valuable time and resources.

## B.3. EXISTING SYSTEM ANALYSIS

The AHS’s dog classification process is as follows: If available, the AHS shelter staff looks at the dog’s rabies vaccination certificate. This legal document, like an I.D. or passport, includes all the dog’s pertinent information including date of birth, sex, breed, and fur color or markings. If unavailable, which is not uncommon, the AHS turns to their vet techs to identify these dogs. The vet tech will analyze the dog’s weight, paw size, markings, ears, head shape, coat texture, and coloring to identify the dogs to the best of their ability (How breed is determined at AHS and other animal shelters, n.d.). Unfortunately, with more than 200 dog breeds recognized by the American Kennel Club, assigning the correct dog breed to every dog purely based on visuals can be difficult. Understandably, it is not unheard of for these dogs to be mislabeled. A University of Florida study showed that different shelter staffers who evaluated the same dogs at the same time had only a *moderate* level of agreement among themselves. These results raise concern as shelter workers and veterinarians are expected to determine the breeds of dogs in their facilities daily (University of Florida, n.d.).

In recent years, machine learning (ML) has become increasingly significant. Large companies like Netflix, Amazon, YouTube, Facebook, Google, MIT, and more have implemented ML into their business practices (Reilly, 2021). In healthcare settings, ML-enabled tools are used to analyze medical reports and images via neural networks and deep learning. These tools are associated with improved patient diagnoses, cost savings, and more (Ramakrishnan, 2023). Below I will discuss several articles whereby similar machine learning technologies were implemented. I will showcase the effectiveness of these ML-enabled tools, providing insight into the many possibilities for innovation within the AHS – specifically their dog breed identification process.

A 2022 study, *Detecting Distal Radial Fractures from Wrist Radiographs Using a Deep Convolutional Neural Network with an Accuracy Comparable to Hand Orthopedic Surgeons,* evaluated the ability of convolutional neural networks (CNN) to diagnose distal radius fractures (DRF) using frontal and lateral wrist radiographs. The model was then compared to the performance of three hand orthopedic surgeons with seven, eight, and ten years of experience. The study involved 503 patients with DRF and 289 patients without fractures. No multiple images, more than one anteroposterior (AP), and one lateral view were obtained from a single patient. The CNN classification model\* was constructed, trained, and validated under the following conditions:

Programming language: Python, version 3.6.7  
Deep Learning Framework: Keras (version 2.2.4) and TensorFlow (version 1.14.0)  
CPU: Core i7-9750 (Intel, Santa Clara, CA)  
GPU: GeForce RTX 2060  
RAM: 16 GB  
\*Separate models were constructed for AP and lateral radiographic views.

The hand orthopedic surgeons interpreted the same radiographs used for evaluating the CNN model (the validation dataset). It was noted that the images were in jpeg format - different conditions than those used in a clinical setting. The clinicians were also blind to the mechanism of injury and the patient’s age. This was all done to ensure consistent conditions between the model and clinicians. The study concluded with the following results: the CNN model performed with an accuracy of 99.3%, compared to the hand orthopedic surgeons with accuracy ratings of 97.3%, 94.7%, and 96.7%.

“The accuracy of the CNN was better than that of the hand orthopedic surgeons” (Suzuki, 2022).

The accuracy of the CNN was equal to or better than that of the three orthopedic surgeons under the same conditions. Based on the results of this study, it is reasonable to assume that another CNN classification model trained on a dog breed image dataset could perform similarly when asked to identify a dog breed.

There are numerous studies involving ML classification models and species identification. For example, in 2020 Chunmei Shi and other researchers conducted a study, *Amur tiger stripes: individual identification based on deep convolutional neural network*, that provided a new approach to the identification of Amur tigers. Rather than manually identifying the tigers, they utilized deep CNN algorithms to automatically identify them. The researchers obtained experimental data from 40 Amur tigers located in Guaipo Tiger Park, China – approximately 200 images per tiger, a total of 8,277 images. The study aimed to identify individual tigers by the body stripes of the tigers, as the stripes of the tiger on both the left and right sides of their body are inconsistent. Images of each tiger were separated into two categories, “left” and “right” to distinguish the different sides. Most of the images were cropped automatically with Python code, and a small percentage were cut manually preserving as much of the body stripes as possible. The images were then divided randomly into training and test sets (5:1 ratio). The deep CNN model was constructed with the deep learning framework, Keras. The experiment consisted of 500 iterations (epochs), each iteration returning an accuracy value and loss function. The experiment resulted in an *excellent* recognition accuracy of 90.48% (left) and 93.50% (right). Concluding the study, the researchers compared their CNN model to other popular CNN models including AlexNet, LeNet, ResNet34 VGG13, and ZF\_Net. It was found that ResNet34 performed the best with an accuracy of 93.25% (left) and 93.31% (right). However, the run time for Chunmei’s study was superior - 60.26 (left) and 61.43 (right), compared to ResNet34 with runtimes of 75.52 (left) and 75.67 (right).

*A Computer Vision-Based Approach for Tick Identification Using Deep Learning Models* illustrates yet another example of an ML classification model being used for species identification. Chu-Yuan Luo and other researchers set out to develop a computer vision method to identify common tick species. They accomplished this through modified CNN-based models. A 12,000 tick image dataset that differentiated between the different developmental stages of three tick species (A. americanum, D. variabilis, and I. scapularis), as well as their sex, feeding status, and host were, used in this study. All images were resized to 224 x 224 pixels to reduce memory usage. In the experiment, 90% of the dataset was used for training and validation purposes, while the other 10% was used for testing. The training dataset was divided into ten subsets containing an equal number of ticks of each species. The deep learning models were constructed from five well-known pre-trained (ImageNet dataset, 1000 categories with 1.2 million images) models (VGG16, ResNet50, InceptionV3, DenseNet121, and MobileNetV2). All experiments were written in Python (version 3.8.5) and conducted in Keras (version 2.4.0) and TensorFlow (version 2.4.1) deep learning frameworks. The study concluded with the following results: The Inception-V3 model achieved the best rating with an accuracy of 99.5% on the test set. The other models, ResNet50, VGG16, DenseNet121, and MobileNetV2, achieved accuracies of 99.42%, 99.37%, 99.2%, and 98.73%. All models used within the study were successful in providing over a 98% recognition rate – indicating that DCNN-based classification approaches are effective for tick species identification.

“These results demonstrate that a computer vision system is a potential alternative tool to help in prescreening ticks for identification… and, as such, could be a valuable resource for health professionals” (Luo, 2022).

In summary, ML models can be used in species identification, and it is also evident that deep CNN-based models can outperform experienced medical professionals in their field. It is safe to say that if implemented,ML technology could be an invaluable asset for the AHS as it could provide an accurate and real-time solution for discriminating dog breeds. To minimize the risk of human error, the AHS should incorporate machine learning into their dog breed identification process. Providing AHS vet techs with an ML-enabled tool will help them identify dogs with increased accuracy, thus decreasing the number of mislabeled dogs housed within the AHS, and saving the organization valuable time and other resources.

## B.4. DATA

The proposed project will extract data from Kaggle.com. The dataset in question is the “70 Dog Breeds-Image Data Set”, 225.78 MB of dog images (9,346 in total) and their associated metadata (a CSV file containing the image paths, the image’s dog breed classification, and more). Project development will follow the data mining methodology, SEMMA. An overview of the data mining methodology is illustrated below.

1. **Sample:** “70 Dog Breeds-Image Data Set” will be extracted from Kaggle.com. A sample of the data, representing the full dataset (several images for multiple breeds), is taken out.
2. **Explore:** The data is explored, noting all variables, and searching for outliers. The data is then visually checked for trends and groupings, including data distribution bias.
3. **Modify:** The data is manipulated by removing missing values, deleting duplicate images, and reaffirming that all dog breeds contain the same number of images, to ensure a lack of bias within the data.
4. **Model:** Post exploration and modification, the CNN model is constructed. This model is trained and validated with the “70 Dog Breeds-Image Data Set”.
5. **Assess:** The accuracy of the trained model is assessed - the model is tested using the test data from the “70 Dog Breeds-Image Data Set”.

## B.5. PROJECT METHODOLOGY

It has been determined that the waterfall methodology is the best development strategy for this project. With its predictability, this method will provide the team with a clear project scope, schedule, and cost estimation - all aspects that will aid the team in staying on track. Although the project will struggle more with changes than its competitors, the waterfall method will save time, money, and other resources that can be allocated to unforeseen changes when the time arises. In general, the waterfall method’s strengths will lead to faster and cheaper project completion.

A rough outline of the project methodology and each phase can be found below:

1. **Initiation (Pre-project setup):** The project is chartered. The core team is assembled.
2. **Planning:** Plans are outlined for the management of all aspects of the project.
3. **Execution:** Actual work of the project is completed. Objectives and requirements are met.
   1. Data collection
   2. Data exploration
   3. Data manipulation and refinement
   4. Model construction
   5. Model training and validation
   6. Model testing
4. **Deployment:** Deploy product to a small subset of users
5. **Monitor and Control:** Project progress and performance are measured. Changes are proposed where necessary to meet existing or shifting goals.
   1. Quality assurance
   2. Performance measurement
   3. Risks and issues
   4. Change control
6. **Closing:** Confirm objectives were met. Deliver the final product to the customer.

## B.6. PROJECT OUTCOMES

The final product of the proposed project will be a user-friendly ML-enabled dog breed image classification tool, permanently hosted on a web application. This product will be accompanied by a user guide that will include any necessary installation instructions to ensure customer ease of use.

## B.7. IMPLEMENTATION PLAN

The proposed project will implement supervised learning, via the convolutional neural network algorithm. Convolutional neural networks are powerful tools used for image classification tasks. Specifically designed for processing images and finding patterns within them, CNNs are a popular choice for image classification projects (GeekforGeeks, 2023).

The CNNs algorithm advantages include, but are not limited to:

* The ability to handle large amounts of data and achieve high accuracy.
* The ability to detect patterns and features in images, videos, and audio signals.

(GeekforGeeks, 2023)

A 2022 study was conducted that evaluated the ability of CNNs to diagnose distal radius fractures (DRF) using frontal and lateral wrist radiographs. The CNN model was implemented using the Keras and TensorFlow frameworks. The model was then compared to the performance of three experienced orthopedic surgeons. The accuracy of the CNN model was equal to or better than that of the three orthopedic surgeons (Suzuki, 2022). In summary, this ML technology has proven to outperform licensed professionals in their fields. If implemented, this tool could be an invaluable asset for the AHS.

The proposed project will be built in the following environment:

PC: MacBook Pro  
Chip: Apple M1  
Memory: 8 GB  
MacOS: Ventura 13.3.1  
IDE: Google Colab  
Programming Language: Python version 3.10  
API: Gradio.app  
Framework: TensorFlow, Keras

The application will then be permanently hosted in a HuggingFace.co “public space” post-project completion, and will be accessible via a website “link”. Hosting the ML tool in a public space will increase user reach and improve customer ease of use. Post-project completion, the product will be distributed to a small subset of users, whereby the application will be monitored and controlled before its full release to the customer. After the product has received approval from the customer that it meets all project requirements, the project will receive annual maintenance to ensure product quality, model accuracy, and customer satisfaction.

## B.8. EVALUATION PLAN

Post-project completion, the following success criteria will be used to evaluate whether or not the project has achieved all desired outputs and functions.

|  |  |
| --- | --- |
| **Objective** | **Success Criteria** |
| Ease of use | User satisfaction survey scores an 80% rating or higher. |
| Performance | The model identifies dogs with a 90% accuracy rating or higher. |

## B.9. RESOURCES AND COSTS

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost ($ USD)** |
| 240 work hours | Project management team (4 members) | 40,320 |
| 120 work hours | IT team (4 members) | 13,920 |
| 80 work hours | Development team (3 members) | 5,160 |
| Hardware | Personal computers, Desktop, etc. | 4,500 |
| Software | Google colab, Gardio.app | 0 |
| Hosting service | HuggingFace.co “public space” | 0 |
| Dataset | Kaggle dataset (“70 Dog Breeds-Image Data Set”) | 0 |
| 15% buffer |  | 9,585 |
| Annual maintenece (15%) | Quality assurance, performance measuring, etc. | 9, 585 |
|  | **Total** | 83,070 |

## B.10. TIMELINE AND MILESTONES

|  |  |  |  |
| --- | --- | --- | --- |
| **Sprint** | **Start** | **End** | **Tasks** |
| 1 | **Date**  05/01/2023  05/04/2023  05/11/2025 | **Date**  05/04/2023  05/11/2023  05/12/2023 | **Initiation (Pre-project setup):**  Validate project – business case document  Write project charter  Get project charter approved |
| 2 | **Date**  05/17/2023  05/17/2023  05/22/2023  05/22/2023  05/22/2023  05/22/2023  06/02/2023  06/02/2023  06/02/2023  06/12/2023  06/13/2023 | **Date**  05/17/2023  05/19/2023  06/02/2023  06/02/2023  06/02/2023  06/02/2023  06/09/2023  06/09/2023  06/09/2023  06/13/2023  06/16/2023 | **Planning:**  Kick-off meeting  Scope planning  Project schedule  Communication planning  HR planning  Procurement planning  Cost estimating & budgeting  Quality planning  Risk planning  Project management plan sign-off  Transition plan |
| 3 | **Date**  06/19/2023  06/20/2023  06/22/2023  07/03/2023  6/20/2023  07/18/2023  07/26/2023 | **Date**  06/19/2023  06/21/2023  06/30/2023  07/17/2023  07/17/2023  07/25/2023  08/04/2023 | **Execution:**  Project kick-off meeting  Data collection  Data exploration  Data manipulation & refinement  Model construction  Model training & validation  Model testing |
| 4 | **Date**  08/07/2023 | **Date**  08/08/2023 | **Deployment:**  Deploy model to small subset of users |
| 5 | **Date**  08/10/2023  08/10/2023  08/10/2023  08/10/2023 | **Date**  ongoing  ongoing  ongoing  ongoing | **Monitor and Control:**  Quality assurance  Performance measuring  Risks and issues  Change control |
| 6 | **Date**  08/24/2023  08/28/2023  08/28/2023  09/06/2023  09/11/2023 | **Date**  08/25/2023  08/31/2023  09/05/2023  09/08/2023  09/15/2023 | **Closing:**  Obtain sign-off and acceptance from organization  Conduct post-mortem analysis  Conduct post-project reviews  Lessons learned document  Project close report |

# C. APPLICATION

Below you will find links to the running application (permanently hosted on a hugging face public space), as well as its associated Google Colab Notebook that includes all project code and post-project reports.

Notebook:  
<https://colab.research.google.com/drive/1w-fX8ZXZvoPuOLjNbp_GlfSMSuOpXvmk?usp=sharing>

Application:  
<https://huggingface.co/spaces/Lydia-Strough/What-Pup>

# D. POST-IMPLEMENTATION REPORT

## D.1. ORGANIZATION VISION

All dogs admitted to the Animal Humane Society (AHS) must be identified by breed. The AHS relies on its vet techs to label these dogs. With more than 200 dog breeds recognized by the American Kennel Club, assigning the precise breed to a dog purely based on visuals can be difficult. Understandably, it is not unheard of for these dogs to be mislabeled. To minimize the risk of human error, the AHS commissioned a machine-learning-enabled dog breed image recognition application to aid their vet techs and shelter staff with dog breed identification. The delivered product is a user-friendly web application that quickly and accurately classifies a dog by breed with a single image. This ML-enabled tool, trained on 70 dog breeds, provides a real-time solution to dog classification, aiding shelter staff in identifying dogs with increased accuracy (decreasing the number of mislabeled dogs within the AHS) and efficiency, saving the organization valuable time and other resources.

## D.2. DATASET

The project could not have been completed without the development of the machine learning model, which required a large image dataset to train, validate, and test it. The Kaggle.com dataset used, “70 Dog Breeds-Image Data Set”, consisted of 225.78 MB of dog images (9,346 in total) along with their associated metadata (a CSV file containing the image paths, the image’s dog breed classification, and more). This dataset was extracted, then following the SEMMA data mining methodology, the dataset was sampled, explored, modified, assessed, and prepared for model usage before application. Preprocessing of the data went as follows.

The data was explored, noting all classes (dog breeds) within the dataset (70 in total). The dataset, already organized into training, validation, and testing directories was evaluated – by counting the number of total images within each directory. It is noted that each of these directories was organized by dog breed (Eg., train/Afghan/image.jpg). Several images from varying breeds were then visually checked, looking for trends and noting the format of each image (224x224x3, jpg format). Finally, the distribution of the data was assessed – looking for bias within the training data. It was concluded that there was bias within the data, as some dog breeds had nearly 200 images, while others contained under 70.

Post-data exploration, the training data directory was copied and then manipulated, as to remove the data distribution bias. It was determined that the “American Hairless” folder within the training directory contained the least number of images, totaling 65. This number became the new maximum number of images to be housed within each (training) dog breed folder. A new training directory was created, containing copies of the first 65 images from every dog breed folder. This ensured no data distribution bias within the training data. The new non-bias training data was then preprocessed for model usage – the training data was split 5:1 into a new training and validation dataset (80% training and 20% validation). This new training data was then augmented (rescaled, flipped, zoomed, etc.) to limit the chances that the machine learning model would overfit. All images were formatted to 224x224 to limit memory usage and ensure uniformity.

## D.3. DATA PRODUCT CODE

The final product was developed and tested in Google Colab, then permanently hosted on a hugging face public space for user ease of use and accessibility. Following the SEMMA, data mining methodology, the dataset was sampled, explored, modified, assessed, and prepared for model usage before application. An overview of the project development and data mining process is illustrated below.

1. **Sample:** “70 Dog Breeds-Image Data Set” was extracted from Kaggle.com. A sample of the data, representing the full dataset (several images from multiple breeds including a beagle, dalmatian, and a pug) was taken out for exploration. An example sample can be seen below.

A picture containing dog breed, text, pet, dog

Description automatically generated

1. **Explore:** The data was explored, noting all classes (dog breeds) within the dataset (70 in total). The dataset, already organized into training, validation, and testing directories was evaluated – by counting the number of total images within each directory. It is noted that each of these directories was organized by dog breed (Eg., train/Afghan/image.jpg). Several images from varying breeds were then visually checked, looking for trends and noting the format of each image (224x224x3, jpg format). Finally, the distribution of the data was assessed – looking for bias within the training data. A bar chart was plotted, for easy visualization of the data distribution. The bar chart produced during this phase of the data mining process can be seen below. It was concluded that there was bias within the data, as some dog breeds had nearly 200 images, while others contained under 70. It was determined that the “American Hairless” folder within the training directory contained the least number of images, totaling 65.

A picture containing screenshot, text, plot, line

Description automatically generated

1. **Modify:** Post-data exploration, the training data directory was copied and then manipulated, as to remove the data distribution bias. It was determined that the “American Hairless” folder within the training directory contained the least number of images, totaling 65. This number became the new maximum number of images to be housed within each training class (dog breed) folder. A new training directory was created, containing copies of the first 65 images from every dog breed folder. This ensured no data distribution bias within the training data. Another bar chart was produced showcasing the new distribution of the training data. This bar chart can be found below. The new non-bias training data was then preprocessed for model usage – the training data was split 5:1 into a new training and validation dataset (80% training and 20% validation). This new training data was then augmented (rescaled, flipped, zoomed, etc.) to limit the chances that the machine learning model would overfit. All images were formatted to 224x224 to limit memory usage and ensure uniformity.

A picture containing text, screenshot, rectangle, line

Description automatically generated

1. **Model:** Post-data exploration and modification, the machine learning model was constructed. This model was trained and validated with the pre-processed data for 100 iterations (epochs). Analysis of the model data was expedited via model accuracy and loss plots. These diagrams indicated the presence of overfitting during model development. These figures lead to the incorporation of a dropout layer as well as additional convolutional 2D layers and dense layers into the model’s architecture to create a “better fitted” model. The final model’s accuracy and loss plots can be seen below.

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated with low confidence

1. **Assess:** Post-model construction and training, the accuracy of the trained model was assessed - the model was tested using the raw test data from the “70 Dog Breeds-Image Data Set”. This data is comprised of 700 images – 70 dog breeds, 10 images each. This data was used to verify the accuracy of the trained model and is what was used to create the trained model’s confusion matrix, which can be found below. This confusion matrix did not lead to any modifications to the model. This figure simply illustrates the model’s areas of “confusion” while predicting a dog image's breed. As the data used in this project’s build was smaller than anticipated (post-bias removal), this chart could not be used to its fullest potential. However, it is theorized that future projects that contain larger datasets, and more variables to help the model predict the breed of dog, including the dog’s size, age in months, and more, will benefit from this chart. It will indicate patterns and trends, where the model is “confused”, leading to further model accuracy improvement and more.

A picture containing text, screenshot, line, plot

Description automatically generated

## D.4. OBJECTIVE AND HYPOTHESIS VERIFICATION

To minimize the risk of human error, the AHS commissioned a machine-learning-enabled dog breed image recognition application to aid their vet techs and shelter staff with dog breed identification. The delivered product is a user-friendly web application that quickly and accurately classifies a dog by breed with a single image. This ML-enabled tool, trained on 70 dog breeds, provides a real-time solution to dog breed classification, aiding shelter staff in identifying dogs with increased accuracy (decreasing the number of mislabeled dogs within the AHS) and efficiency, saving the organization valuable time and other resources. The product delivered meets the project’s objective as the model predicts a breed of dog in real-time. However, it was hypothesized that the ML-enabled tool would be able to identify a breed of dog with an accuracy rating of 90% or higher, but the trained model, in its current state, only has a testing accuracy rating of approximately 45%. It is theorized that this is due to the lack of data used to train the model. The model’s accuracy could be improved by increasing the dataset size, or including other variables for the model to take into account while predicting the dog’s breed, including the dog's size, age in months, etc.

## D.5. EFFECTIVE VISUALIZATION AND REPORTING

All project visuals and reports played a part in the effective development of the dog breed classification application. During data exploration, bar charts were used to discern the distribution of the data – indicating whether or not there was any bias. Without this descriptive visual, leading to the dataset’s modification, the result would have been a biased model. Data analysis was expedited by model accuracy and loss plots. These diagrams indicated the presence of overfitting during model development. These non-descriptive figures lead to the augmentation of the training data (rescaling, flipping, zooming, and shuffling training images), as well as the incorporation of dropout into the model’s architecture to aid the model with overfitting, creating a “better fitted” model. The confusion matrix, another non-descriptive figure, produced post-model testing, was used to indicate the presence of trends (or lack of) within the model’s training “confusion” as the model predicted an image's dog breed. This figure’s lack of patterns indicated to developers that the data used in this project’s build was not sufficient and that future projects could benefit from a larger dataset, and/or more variables to help the model predict the breed of dog, including the dog’s size, age in months, and more.

## D.6. ACCURACY ANALYSIS

The model’s accuracy was assessed using the ‘accuracy’ metric, a commonly used metric, especially in classification tasks. This metric calculates the percentage of correctly classified samples out of the total number of samples. This metric was used to represent the model's overall performance in correctly predicting all classes (dog breeds). A plot illustrating the model’s training history across 100 epochs can be found below. This diagram showcases the model’s changing accuracy with each epoch.

A picture containing text, screenshot, plot, line

Description automatically generated

## D.7. APPLICATION TESTING

Post-model training and validation, the trained model was tested using the raw testing data. This data is comprised of 700 images – 70 dog breeds, 10 images each. This data was used to verify the accuracy of the trained model and is what was used to create the trained model’s confusion matrix, which can be found below. This confusion matrix did not lead to any modifications to the model. This figure simply illustrates the model’s areas of “confusion” while predicting a dog image's breed. As the data used in this project’s build was smaller than anticipated (post-bias removal), this chart could not be used to its fullest potential. However, future projects that contain larger datasets, and more variables to help the model predict the breed of dog, including the dog’s size, age in months, and more, will benefit from this chart. It will indicate patterns and trends, where the model is “confused”, leading to further model accuracy improvement and more.

A picture containing text, screenshot, line, plot

Description automatically generated

## D.8. APPLICATION FILES

Below you will find links to the running application (permanently hosted on a hugging face public space), as well as its associated Google Colab Notebook that includes all project code and post-project reports.

Notebook:  
<https://colab.research.google.com/drive/1w-fX8ZXZvoPuOLjNbp_GlfSMSuOpXvmk?usp=sharing>

Application:  
<https://huggingface.co/spaces/Lydia-Strough/What-Pup>

## D.9. USER GUIDE

1. Click the following link: <https://huggingface.co/spaces/Lydia-Strough/What-Pup>
2. A new page will open, presenting the running application.
3. The “What Pup?” dashboard should appear.

A screenshot of a computer

Description automatically generated with medium confidence

You can now interact with the application! It contains an input (“Drop image here” or “Click to upload”) and an output (Prediction) section. At the bottom of the page, there are example images that you, the user, can use to test the application. You can test its functionality by doing the following:

1. Select a dog image that you would like to identify - an example image or by clicking the input section, whereby the application will ask you to upload an image of your choice.
2. The image you have selected will display in the input section of the application.
3. The output section will then produce a chart displaying the model’s top five dog predictions for the selected image. An example of the functioning application can be seen below.

A screenshot of a computer

Description automatically generated with medium confidence

## D.10. SUMMATION OF LEARNING EXPERIENCES

My previous academic experiences at Western Governors University readied me for this project in many ways. The completion of university courses such as Scripting and Programming, Data Structures and Algorithms, as well as Software I and II laid the foundation for my technical skills throughout this project. Other courses like Technical Communication, Business of IT – Project Management, and Business of IT – Applications, provided me with the necessary soft skills and certifications (CompTIA Project+ and ITIL 4 Foundation) that were necessary for writing this post-implementation report and planning the development of this project.

Additional resources were necessary for this project - I built my project in an unfamiliar IDE and API and utilized unfamiliar libraries and machine learning frameworks. I watched tutorials on YouTube and utilized other resources from the TensorFlow website to gather a greater understanding of the TensorFlow and Keras libraries, as well as Google Colab, Gradio, and Hugging Face. Additionally, my previous course instructors, Mark Denchy and Sidney Rubey, aided me throughout the planning and development process. Each of them helped me to brainstorm and problem-solve throughout this project.

Without my previous academic experiences at WGU, which taught me the value of consistency and persistence, as well as the necessary technical and soft skills, this project would have been more daunting than it was. I took each phase of the project one step at a time, and I am more confident now in my abilities than I was at the beginning of this project. I will take these lessons with me into my future endeavors. I am forever thankful for this opportunity to have learned and grown with WGU. I will continue to challenge myself and advance my technical and soft skills in the future.

# E. REFERENCES

*2022 Annual report*. Animal Humane Society. (n.d.). Retrieved April 12, 2023, from https://www.animalhumanesociety.org/about/2022-annual-report

*DNA studies reveal that shelter workers often mislabel dogs as 'pit bulls' " College of Veterinary Medicine " University of Florida*. UF monogram. (n.d.). Retrieved April 12, 2023, from https://www.vetmed.ufl.edu/2016/02/17/dna-studies-reveal-that-shelter-workers-often-mislabel-dogs-as-pit-bulls/

GeeksforGeeks. (2021, June 22). *Top 10 benefits of machine learning in businesses*. GeeksforGeeks. Retrieved April 14, 2023, from https://www.geeksforgeeks.org/top-10-benefits-of-machine-learning-in-businesses/

GeeksforGeeks. (2023, March 24). *Introduction to convolution neural network*. GeeksforGeeks. Retrieved April 14, 2023, from https://www.geeksforgeeks.org/introduction-convolution-neural-network/

*How breed is determined at AHS and other animal shelters*. Animal Humane Society. (n.d.). Retrieved April 12, 2023, from https://www.animalhumanesociety.org/resource/how-breed-determined-ahs-and-other-animal-shelters

Luo, C.-Y., Pearson, P., Xu, G., & Rich, S. M. (2022). A computer vision-based approach for tick identification using Deep Learning Models. *Insects*, *13*(2), 116. https://doi.org/10.3390/insects13020116

Ramakrishnan, M. (2023, January 24). *How machine learning in healthcare works and why it matters*. Emeritus Online Courses. Retrieved April 24, 2023, from https://emeritus.org/blog/ai-ml-machine-learning-in-healthcare/#:~:text=Machine%20learning%20in%20healthcare%20can,on%20training%20in%20similar%20cases.

Reilly, J. (2021, December 22). *20 companies using machine learning in cool ways*. Akkio. Retrieved April 24, 2023, from https://www.akkio.com/post/10-companies-using-machine-learning-in-cool-ways

SHI, C., LIU, D., CUI, Y., XIE, J., ROBERTS, N. J., & JIANG, G. (2020). Amur tiger stripes: Individual identification based on deep convolutional neural network. *Integrative Zoology*, *15*(6), 461–470. https://doi.org/10.1111/1749-4877.12453

Suzuki, T., Maki, S., Yamazaki, T., Wakita, H., Toguchi, Y., Horii, M., Yamauchi, T., Kawamura, K., Aramomi, M., Sugiyama, H., Matsuura, Y., Yamashita, T., Orita, S., & Ohtori, S. (2021). Detecting distal radial fractures from wrist radiographs using a deep convolutional neural network with an accuracy comparable to hand orthopedic surgeons. *Journal of Digital Imaging*, *35*(1), 39–46. https://doi.org/10.1007/s10278-021-00519-1