PROPOSAL FOR A MACHINE LEARNING SOLUTION

Alano Peirce

05/14/2022

Table of Contents

A. PROJECT OVERVIEW

At the request of the Wildlife Conservation Association (WCA), I would like to submit this detailed proposal introducing a thorough and robust machine-learning solution that fulfills a pressing need of the organization. The discussions that follow include a description of said organizational need, the benefits of the proposed machine-learning solution, the resources and costs associated with the machine-learning solution, and the technical details of said machine-learning solution.

A.1. Organizational Need

The Wildlife Conservation Association (WCA) is an organization that performs various duties relating to wildlife conservation. To assist in these efforts, motion-activated wildlife cameras have been installed in various locations around national parks for the purpose of documenting the local wildlife population (through the use of pictures and/or video). These images provide an invaluable resource for many different facets of wildlife conservation — for example, they can be used to determine the population count for a specific animal species, to track the behavior and habits of various animal species, and to identify threats facing a specific animal species (such as pollution, disease, overabundant or invasive species, or disturbance of their environment).[1]

As part of the process used to convert these images into useful information, each image must be individually examined and classified according to the species of animal(s) that appear in it. Currently, this task is performed manually by employees, which (considering the thousands of images captured daily by WCA) constitutes an incredibly time-consuming and inefficient approach. In fact, with the recent deployment of hundreds of new cameras, employees are struggling to keep up with the task of classifying all new images within a reasonable timeframe.

To ease the burden on these employees, I propose that a machine-learning (ML) system be developed in order to take over the responsibility of classifying these wildlife images. Such a system would be able to automate the image classification process (as well as the underlying species-identification processes), thus leaving employees free to attend to other needs of the organization. Additionally, the proposed ML system would exponentially improve performance — astonishingly, it would be able to fully process an image within a fraction of a second, thus majorly expediting the classification of said images.

A.2. Context and Background

The Wildlife Conservation Association (WCA) is, as its name suggests, an organization whose mission is to effect the conservation and protection of wildlife and the habitats

they occupy. More specifically, the WCA is a company that contracts with the government in order to provide these conservation services.

In order to accomplish their main objective, WCA needs to collect specific data about the wildlife species that they are tasked with protecting — for example, this can include data regarding sex ratios, population counts, the spread of disease, or the behavior of a particular species.[1]

One trusted method used to collect such data is the use of wildlife cameras. The images captured by these contraptions are highly versatile and can provide many different types of data. Additionally, some of these images can be used to increase public awareness of conservation efforts. However, wildlife cameras also come with a drawback — the images captured must be individually reviewed by a person, who must then decide what to do with each image.

In the early days of WCA, this wasn't a problem — WCA was initially contracted to provide conservation efforts for only one national park, meaning that there were only a small number of wildlife cameras whose images needed to be processed. However, as WCA proved themselves to be extremely competent and effective in the field of wildlife conservation, the government began extending more contracts for additional national parks; WCA gladly accepted these contracts.

A few years have elapsed since then, and WCA now provides conservation efforts to more than fifty national parks covering nearly one million acres of land. Although the company has increased dramatically in size, there aren't enough employees to keep up with the image processing needs necessitated by the thousands of wildlife cameras that WCA now uses.

These events have necessitated a search for a solution that will somehow enable WCA to process the images from all of their wildlife cameras in a timely manner. This is where the machine-learning system that I am proposing comes into play — if developed, this system will provide a robust and highly effective solution to this problem. It will both automate and expedite the image classification process, thus conserving time, money, and resources.

A.3. Outside Works Review

The journal article written by Schneider et al. (2020) details a study that the authors performed regarding the accuracy of machine-learning solutions used for species identification in cases where the training data is limited (i.e. where the data set consists of thousands of images rather than millions). Additionally, they tested the ability of these machine-learning solutions to identify animal species in images whose background locations were not included in the training data set.[2]

In cases with "trained" background locations, the authors found that species with at least 1000 "example" images (a.k.a. images in the training data set) were identified with high

accuracy in new images by the machine-learning solutions — specifically, these species had “a high and stable recall of 0.971 ± 0.0137”.[2] On the other hand, species that had less than 500 “example” images were identified with low accuracy in new images — specifically, these species had a “low and highly variable recall of 0.750 ± 0.329”.[2]

Cases with “untrained” background locations were classified less accurately than those with “trained” background locations — the best-performing machine-learning model accurately classified images with “untrained” background locations only 68.7% of the time.[2]

The discussed journal article is relevant to the current proposal because it explores the accuracy of machine-learning models when used to perform image classification in various different scenarios. Additionally, it specifically focuses on image classification based on species identification of involved animals, which is exactly what the ML solution I present in this proposal aims to accomplish.

From this article, we can discern that larger training data sets are more favorable and produce higher accuracy. Additionally, we can ascertain that all image background locations should be included in the training set when possible, as this leads to higher accuracy of species identification by machine-learning models. Both of these deductions are highly useful and directly applicable to the ML solution that I am proposing.

The journal article written by Tabak et al. (2019) describes a study in which a machine-learning model was designed in order to “automatically classify wildlife species from camera trap images”.[3] This ML model was trained using a data set of 3,367,383 images, and was then tested on wildlife images from locations other than those involved in the training data set — that is, the model was tested against images with “untrained” background locations (as discussed in the previous study by Schneider et al. [2020]), which provided a more rigorous challenge for the ML model.[3]

Specifically, the ML model was tested against images from three separate “untrained” locations in the United States, Canada, and Tanzania. The ML model produced an accuracy rate of 98% for the first test group (i.e. 98% of the images were correctly classified), 82% for the second test group, and 94% for the third test group. Additionally, the study found that “The trained model classified approximately 2,000 images per minute on a laptop computer with 16 gigabytes of RAM”.[3]

The discussed journal article is relevant to the current proposal because it involves the construction and testing of a machine-learning model whose purpose is to classify images of wildlife by species; this proposal advances a similar machine-learning solution.

Additionally, the article demonstrates that higher accuracy rates can be achieved even when classifying images that have “untrained” background locations; this correlates with a high number of training images (3,367,383 images)[3], thus providing further evidence

towards the inference that larger training data sets cause machine-learning models to produce more accurate image classification results. Furthermore, the article demonstrates that machine-learning models can classify images very quickly — a classification rate of 2,000 images per minute translates to an approximate classification rate of 0.03 seconds per image. These findings are extremely useful and relevant to the current proposal because they provide crucial information that may affect the development of the machine-learning solution that I am proposing.

The journal article produced by Gaggiotti et al. (2019) details a study in which the authors create a machine-learning model capable of identifying the species of animals found in images; however, this ML model is also capable of discerning whether or not images contain humans, vehicles, or nothing of interest (meaning no animals, humans, or vehicles). The ML model produced highly favorable results — “accuracies for identifying empty images across projects ranged between 91.2% and 98.0%, whereas accuracies for identifying specific species were between 88.7% and 92.7%”.[4]

Amazingly, the article also reports that after the addition of certain improvements, the machine-learning model was able to accurately classify images at a rate equal to that of citizen scientists.[4] Furthermore, the article states that because of the ML model, “human effort [for the purposes of image classification] was reduced by 43% while maintaining overall accuracy”.[4]

The discussed journal article is relevant to the current proposal because the functionalities of the machine-learning model created in the study aligns very closely with the needs of WCA (and thus aligns closely with the suggested features of the ML solution that I am proposing).

Furthermore, the article provides empirical evidence that machine-learning solutions designed to classify wildlife images do indeed take a large burden off of the people who were previously responsible for this task. It also proves that machine-learning solutions are capable of performing image classification at an accuracy rate equivalent to that of an experienced person. Both of these findings directly relate to the informed assertion that the ML solution presented in this proposal would be highly beneficial to WCA.

A.4. Solution Summary

The machine-learning system that I am proposing will classify wildlife images using the K-Nearest Neighbors (KNN) algorithm. This algorithm relies on a training data set, which will be taken from the vast array of already-classified images in WCA’s database. After being supplied with the training data set, the KNN algorithm will be able to group new wildlife images into existing classification groups by discerning which of the groups in the training data set contains images that are most similar to the image currently being considered.

A.5. Machine Learning Benefits

My proposed machine-learning system will greatly benefit WCA by considerably expediting and automating the process of classifying images taken by wildlife cameras. This will substantially reduce the current strain on WCA's resources.

Additionally, the algorithm used in the ML system is rather straightforward and easy to understand,[5] thus suggesting that current and future employees of WCA can easily be trained to perform maintenance and alterations on the system.

Furthermore, the algorithm utilizes WCA's existing database of already-classified wildlife images to function as its training data set; thus, no additional memory storage space is required to store the (generally extensive) training data.

B. MACHINE LEARNING PROJECT DESIGN

B.1. Scope

The following items are **within the scope** of the proposed project:

- The development of a machine-learning model that can classify images captured by WCA's wildlife cameras
- The development of a software program that will act as a visual interface from which users will be able to utilize the developed machine-learning model
- Compatibility with both Windows and macOS operating systems

The following items are **outside the scope** of the proposed project:

- The ability to analyze and classify videos taken by WCA's wildlife cameras
- The ability to analyze and classify images taken by WCA's thermal-imaging unmanned aerial vehicles (UAVs)
- Compatibility with Linux or Unix operating systems

B.2. Goals, Objectives, and Deliverables

Goals

- To design and create a machine-learning system that will largely automate the classification of the images produced by WCA's wildlife cameras
- To enable the processing of the images produced by WCA's wildlife cameras in a timely manner
- To reduce the substantial workload currently encumbering various WCA employees

Objectives

- The proposed machine-learning system will have an image classification accuracy of at least 85%.
- The proposed machine-learning system will reduce the human effort involved in wildlife image classification by at least 35%.
- The proposed machine-learning system will be able to fully process at least two images per second.

Deliverables

- A machine-learning system capable of classifying images based on the species of animals appearing in said images
- Documentation relating to all aspects of the machine-learning system — this will be extremely useful for modifying and maintaining said system

B.3. Standard Methodology

To direct the design and development of the proposed machine-learning system, the SEMMA methodology will be used.

The steps involved in the SEMMA methodology are delineated as follows:

- **Sample:** During this step, training data and validation data will be selected from an extensive data set.[6] Training data refers to the data used to train the machine-learning model, while validation data refers to the data set used to measure the accuracy of the machine-learning model.

  In our case, these two groups of data will be selected from the already-classified wildlife images stored in WCA's database. The training data set that we select must be accurately representative of the whole data set; it must also be a good size such that a) the training data set is not so small as to provide poor training to the ML model, and b) the training data set is not so large that the ML model exhibits unacceptable image processing times.[7] Additionally, the validation data set that we select must *also* be accurately representative of the whole data set; this is necessary so that the validation data set will provide an accurate gauge for comparison.

- **Explore:** This step involves an exploration of the data — this consists of inspecting the data for expected relationships, unexpected correlations, and outliers/anomalies.[7]

  In our case, we will be utilizing data visualization (the act of representing data by using visuals, such as graphs or charts) to accomplish this step.[6] An example of an "expected relationship" that we might discover is that the past classifications of wildlife images are highly accurate. An example of an "unexpected correlation" that we might uncover is that reptiles are classified incorrectly three times more often than mammals. An example of an "anomaly" we might discover is that an image of a bear somehow found its way into the "corn snake" classification.

- **Modify:** This step involves modifying (i.e. cleaning and potentially transforming) the data. This is done using the insight gained from the previous SEMMA step ("Explore").[6]

In our case, “cleaning” the data might include measures such as moving the previously-discussed misplaced bear image into the correct classification. An example of a “transformation” would be converting the image files from PNG format to JPEG format.

- **Model:** This step involves testing various different machine-learning models on the processed data. The outcomes resulting from these tests are examined, and using this information, a suitable machine-learning model is chosen.[8]

  In our case, since we have already decided on the K-Nearest Neighbors (KNN) algorithm as our machine-learning model, we will instead perform more specific tests — for example, we may experiment with changing different variables involved in the KNN algorithm (such as the value of “K”), which will allow us to fine-tune our chosen machine-learning model.

- **Assess:** In this step, the chosen machine-learning model is evaluated — is this model reliable enough for its intended purpose, and does it meet the requirements of the project?[6]

  In our case, this might involve testing our chosen model on new images of wildlife, and then using the results from this process to determine the model’s suitability for our project.

B.4. Projected Timeline [9]

- **05/26/2022** – The proposal is accepted by WCA after a few days of consideration.
- **06/14/2022** – A technical proof of concept (which demonstrates the technological feasibility of the proposed machine-learning solution) is presented to WCA. Four full days (following this date) are set aside to allow WCA executives to review and approve this proof of concept.
- **07/08/2022** – A detailed and extremely thorough design prototype for the machine-learning solution is submitted for review. Four full days (following this date) are set aside to allow WCA executives to review and approve the design prototype.
- **08/10/2022** – The completed machine-learning solution is presented to WCA.
- **08/21/2022** – The machine-learning solution has been fully integrated with current systems.
- **09/03/2022** – Testing and subsequent fixes to the machine-learning solution are complete; the solution is deployed.

SPRINT SCHEDULE

Sprint	Start	End	Tasks
1	05/08/2022	05/22/2022	Creation of proposal
2	05/27/2022	06/14/2022	Development of a technical proof of concept
3	06/19/2022	07/08/2022	Development of a thorough design prototype for the machine-learning solution
4	07/13/2022	08/10/2022	Development of the machine-learning solution itself
5	08/11/2022	08/21/2022	Integration of the machine-learning solution with current systems
6	08/22/2022	09/03/2022	The solution is extensively tested by Quality Assurance (QA) engineers, and any bugs found are fixed.

B.5. Resources and Costs

Resource	Description	Cost
Lead Machine Learning Engineer (me)	This person will perform the primary duties related to the design and development of the machine-learning model. This person will be paid $40/hr and will be needed for 500 hours of work.	$40 x 500h = $20,000
Secondary Machine Learning Engineer	This person will assist the Lead Machine Learning Engineer in the performance of various duties. This person will be paid $30/hr and will be needed for 300 hours of work.	$30 x 300h = $9,000
Software Designer	This person will design and implement the GUI of the software program from which users will be able to utilize the developed machine-learning model. This person will be paid $30/hr and will be needed for 160 hours of work.	$30 x 160h = $4,800
Software Engineer	This person will design and implement the back-end functionality of the software program from which users will be able to utilize the developed machine-learning model. This person will be paid $35/hr and will be needed for 270 hours of work.	$35 x 270h = $9,450
QA (Quality Assurance) Engineer #1	This person will thoroughly test the software in order to uncover any bugs that may be present in the system. This person will be paid $25/hr and will be needed for 90 hours of work.	$25 x 90h = $2,250

QA (Quality Assurance) Engineer #2	This person will also thoroughly test the software in order to uncover any bugs that may be present in the system. This person will be paid $25/hr and will be needed for 90 hours of work.	$25 x 90h = $2,250
Desktop Computer (x4)	Each of the four team members working on this project will require a desktop computer. This will not cost anything, as WCA already has four desktop computers available.	$0
Windows 11 Pro (x4)	Each of the four desktop computers will require Windows 11 Pro, which is the operating system that will be used to develop this project. This will not cost anything, as Windows 11 Pro is already installed on the four desktop computers.	$0
Oracle VM VirtualBox 6.1.34 (x4)	Each of the four desktop computers will require VirtualBox to be installed (for the purpose of testing the developed software and/or machine-learning model in various environments). This will not cost anything, as WCA already possesses a VirtualBox Enterprise license that allows hundreds of separate users in a business/commercial context.	$0
Windows 11 Pro (x4)	Each of the four computers will need Windows 11 Pro installed on a VM created by their respective VirtualBox programs (so that the software / ML model can be tested in Windows 11). Each instance of Windows 11 Pro costs $150.	$150 x 4 = $600
Windows 10 Pro (x4)	Each of the four computers will need Windows 10 Pro installed on a VM created by their respective VirtualBox programs (so that the software / ML model can be tested in Windows 10). Each instance of Windows 10 Pro costs $200.	$200 x 4 = $800
macOS 12 (macOS Monterey) (x4)	Each of the four computers will need macOS 12 installed on a VM created by their respective VirtualBox programs (so that the software / ML model can be tested in macOS 12). This will not cost anything, as WCA already possesses a volume business license that allows hundreds of separate users to download macOS 12 in a business/commercial context.	$0
macOS 11 (macOS Big Sur) (x4)	Each of the four computers will need macOS 11 installed on a VM created by their respective VirtualBox programs (so that the software / ML model can be tested in macOS 11). This will not cost anything, as WCA already possesses a volume business license that allows hundreds of separate users to download macOS 11 in a business/commercial context.	$0
TensorFlow 2.8.0	This is a free library used to develop machine-	$0

	learning models.	
Keras 2.4.0	This is a free API that provides a Python interface for TensorFlow.	$0
	Total	**$49,150**

B.6. Evaluation Criteria

The following criteria will be used to evaluate the success of the completed machine-learning solution.

Objective	Success Criteria
Ease of Use	At least 75% of test users give the solution a rating of 5 or under on a scale of 1 to 10 (with 1 being “extremely easy to use” and 10 being “extremely difficult to use”)
Algorithm Efficiency	On average, the machine-learning algorithm is able to fully process at least two images per second.
Algorithm Accuracy	The machine-learning algorithm is able to classify images correctly at least 85% of the time.

C. MACHINE LEARNING SOLUTION DESIGN

C.1. Hypothesis

After the completion of development, the proposed machine-learning solution will be able to classify wildlife images with a general accuracy rate that falls between 85% and 90% (inclusive).

This hypothesis will be tested as follows:

1. First, the solution will be used to classify 5000 images.
2. Then, WCA employees will manually check each image to determine whether or not it has been classified correctly. During this process, two tallies will be kept — a) the number of correctly-classified images, and b) the number of incorrectly-classified images.
3. After the employees have completed their inspection of all 5000 images, the percentage of correctly-classified images will be calculated from the two tallies.
4. Next, the same process (steps 1 through 4) will be repeated four more times (for a total of five times in all).
5. Finally, the average of the five accuracy percentages (from the five separate experiments) will be calculated; this average is the final result and represents the accuracy rate of the machine-learning solution.

   This result will be compared against the accuracy rate predicted in the hypothesis (85% to 90%, inclusive); if the result is in the predicted range, then the hypothesis is correct. However, if the result is outside of the predicted range, then the hypothesis is incorrect.

C.2. Selected Algorithm

For this project, we have chosen to use a **supervised learning approach**. This is a broad term that refers to machine-learning techniques that make use of labeled training data (i.e. training data that is labeled with the correct classifications) in order to “teach” the algorithm how to classify data points properly.[10]

The specific algorithm that we have selected for the machine learning aspect of this project is the **K-Nearest Neighbors (KNN) algorithm**. Since this algorithm falls under the umbrella of supervised learning approaches,[5] it relies on a training data set — this will be sampled from the considerable collection of already-classified images in WCA’s database.

In order to classify a new image, the KNN algorithm compares the new image against those in the training data and determines which images in the training data set most closely match the image currently being considered. The classification that houses the most “close match” images will be identified as the classification that the new image belongs in.[5]

The “K” in “KNN” is a numerical value set by the machine learning engineer; it refers to the number of “close match” images that will be considered by the algorithm. For example, if “K” is set to 12 (by the machine learning engineer), then the algorithm will decide where a new image belongs by considering the classifications of the 12 “closest match” images in the training data set. In other words, if 7 of the 12 closest matches are classified as “alligators”, and the other 5 of the 12 closest matches are classified as “crocodiles”, then the new image will be classified as an “alligator” (because, for the KNN algorithm, “the majority wins”).[5]

C.2.a Algorithm Justification

The KNN algorithm was selected for this project largely due to the fact that it is relatively straightforward and easy to understand,[5] even for those who possess no prior knowledge in the field of machine learning. This implies that current and future employees of WCA can easily be taught how to work with, maintain, and potentially even modify the machine-learning system.

As an example of the simplicity of its implementation, the training process for the KNN algorithm hardly takes any work — in fact, it merely involves supplying the algorithm with a previously-classified data set; the algorithm does the rest of the work by itself.[5] In WCA’s case, this will be especially easy because WCA’s database already consists of previously-classified images that will provide prime material for the training data.

Furthermore, KNN makes for a great choice because it “doesn’t make any assumptions about the underlying data distribution”.[5] This suggests that, unlike many other machine-learning algorithms, KNN is able to handle imbalanced training data sets (e.g. a data set that has 500 images classified as “bison” but only 5 images classified as “moose”) without assuming that the data set it is tasked with classifying will be similarly imbalanced.

C.2.a.i. Algorithm Advantage

As previously stated, the KNN algorithm is incredibly simple and easy to understand,[5] which makes it an ideal choice for a company like WCA that has very few employees experienced in machine learning.

C.2.a.ii. Algorithm Limitation

Because the KNN algorithm's method for classifying images is modeled almost exclusively on the data set used to train it, KNN does not have the ability to create new classifications beyond those that exist in the training data. For example, if the training data only consists of two classifications — "alligators" and "crocodiles" —- and the algorithm is confronted with a picture of a raccoon, the algorithm will classify it as either an alligator or a crocodile.

C.3. Tools and Environment

The operating system that will be used to develop the entirety of the machine-learning solution is **Windows 11 Pro**.

However, the solution will be designed to be compatible with both Windows and macOS. To facilitate this feature, **Oracle VM VirtualBox 6.1.34** (a type-2 hypervisor) will be used to create various virtual machines (VMs) on which the solution can be tested.

To enable this VM-based testing, several different operating systems will be required — **Windows 11 Pro**, **Windows 10 Pro**, **macOS 12** (macOS Monterey), and **macOS 11** (macOS Big Sur). These operating systems will be installed on various VMs created by VirtualBox; these VMs will then be used to test the solution in different environments.

The solution will be developed using the **Python** programming language with the help of **TensorFlow 2.8.0** (a free library used to develop machine-learning models). Additionally, **Keras 2.4.0** (a free API) will be used to provide a Python interface for TensorFlow.

The training data will be sampled from WCA's existing database, which is **MySQL Enterprise Edition 8.0**.

C.4. Performance Measurement

The performance of the proposed machine-learning solution will be evaluated based on its conformity (or lack thereof) to the predetermined accuracy and efficiency objectives delineated in Section B.2. In order for the solution's performance to be considered satisfactory, it must meet the predetermined metric for accuracy (the accuracy rate must be at least 85%) as well as the predetermined metric for efficiency (the solution must be able to [on average] fully process at least two images per second).

The accuracy rate of the solution will be measured using the steps detailed in Section C.1., while the efficiency of the solution will be measured using the steps detailed below:

1. First, the solution will be used to classify 5000 images. The time it takes for this to occur will be recorded.

2. Then, step 1 will be repeated four more times (for a total of five times in all).

3. Next, each of the five recorded time values will be converted to seconds and then divided by 5000; the average of the resulting values will then be calculated. This average is the final result and represents the average efficiency of the machine-learning solution.

   This result will be compared against the predetermined metric for efficiency. If the result is equal to or below 0.5 seconds, then it meets the predetermined metric; otherwise, it does not.

D. DESCRIPTION OF DATA SETS

D.1. Data Source

Both the training data set and the validation data set for the machine-learning algorithm will be sampled from WCA's database, which is of type MySQL Enterprise Edition 8.0. This database contains a large collection of wildlife images that have already been classified at a previous point in time; this is exactly what is needed for the machine-learning algorithm's training and validation data sets.

D.2. Data Collection Method

Since WCA's database is up-to-date, the desired data will be collected from said database via **logical extraction** (as compared to physical extraction).[11] This is a simple process that involves querying the database (e.g. using Structured Query Language [SQL]) for the desired data.

The specific logical extraction technique that will be used is **full extraction**. This refers to extracting an entire block of data (versus, for example, extracting only the pieces of data that have been altered).[11] If it is the first time that a set of data is extracted from a source (as is true in this case), full extraction is necessary by default.[12]

D.2.a.i. Data Collection Method Advantage

Full extraction is advantageous because accomplishing it is "relatively uncomplicated when performed with the right data extraction tools";[11] this is because the desired data can be "extracted without [requiring] … additional logical information from the system".[11] Simply put, full extraction is very simple and straightforward to implement.

D.2.a.ii. Data Collection Method Limitation

One limiting factor of full extraction is that it "involves high data transfer volumes, which can put a load on the network".[12] This is especially true in our case, as we will be extracting all classified wildlife images that are currently in the database (which is a sizable amount of data). Fortunately, full extraction is generally a process that only needs to occur once for a particular set of data (i.e. the first time that set of data is extracted).[12]

D.3. Quality and Completeness of Data

One concern regarding the extracted data set is the possible presence of **dirty data**, which refers (in our case) to images that have been erroneously classified. If these errors are left uncorrected in the training data set, the machine-learning algorithm will be incorrectly trained; this will likely result in the algorithm classifying certain new images incorrectly.

In order to get rid of the dirty data, we (the persons working on this project) will meticulously comb through the training data in order to catch and correct any erroneous image classifications; this will be completed before the training data is used to train the machine-learning algorithm. For example, we may discover that there is an image of a cougar in the “elk” classification; this error will be rectified by moving the image to the appropriate classification (the “cougar” classification).

Another concern regarding the extracted data set is the potential for **missing data**. For example, it may be possible that the data extracted from the database does not include images (and thus classifications) of specific animal species that are found within the national parks that WCA serves; if this is the case, then the training data (which will be sampled from the extracted data) will also exhibit this characteristic. This is problematic because, as previously discussed, the KNN algorithm (which is the specific machine-learning algorithm that we will be using) does not have the ability to create classifications beyond the ones that already exist in the training data set. For example, if there is no “hare” classification in the training data and the algorithm is presented with an image of a hare, the algorithm will still assign the image to one of the existing classifications, which will be decided by determining which of the existing classifications most closely matches the new image (e.g. the image of the hare might be placed in the “rabbit” category).

To fill in the gaps caused by missing data, we (the individuals working on this project) will take note of the classifications included in the training data; we will then carefully compare this list against a list of animal species found in the national parks that WCA serves. If any animal species on the latter list are not included as classifications in the training data set, then we will add this classification to the training data set along with images of the indicated animal species (which will be assigned to the newly-added classification for purposes of accurately training the algorithm).

D.4. Precautions for Sensitive Data

Since the images taken by WCA’s wildlife cameras direct the organization’s activities to a large extent, the data set consisting of WCA’s wildlife images is considered to be sensitive and proprietary business information. Because of this, certain precautions must be taken in order to prevent the theft, tampering, or unauthorized access of the data.

One precaution we will take is the application of the principle of least privilege. Essentially, all individuals working on this project will be given the least amount of access possible to the data (but still enough access that each individual can perform their required duties). Additionally, each individual working on this project will be required

to sign a non-compete agreement stating that they will not use any of WCA’s data or information to benefit a competing business organization within the next two years.

Furthermore, all work on the project will be restricted to the desktop computers provided by WCA; the use of personal devices for project-related work will not be allowed (including, for example, individuals texting each other regarding developments in the project). Each of the provided desktop computers will require two-factor authentication in order to log into a user account. Moreover, all communications regarding the project must take place using official WCA email accounts; these accounts will also require two-factor authentication.

The software that will be developed as part of the project (which will act as a visual interface from which users will be able to utilize the developed machine-learning model) will also contain enhanced security features. For example, the utilized training data set will be hidden and inaccessible to the vast majority of users (except for those who need access to it, such as the administrator account). The user accounts that possess elevated privileges will require two-factor authentication.

References

[1] U.S. National Park Service. (2021, November 4). *Wildlife Monitoring and Wildlife Viewing Camera Systems: Frequently Asked Questions*. National Park Service. Retrieved May 14, 2022, from https://www.nps.gov/pore/learn/nature/wildlife_monitoring.htm

[2] Schneider, S., Greenberg, S., Taylor, G. W., & Kremer, S. C. (2020). Three critical factors affecting automated image species recognition performance for camera traps. *Ecology & Evolution (20457758)*, *10*(7), 3503–3517. https://doi.org/10.1002/ece3.6147

[3] Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., Vercauteren, K. C., Snow, N. P., Halseth, J. M., Di Salvo, P. A., Lewis, J. S., White, M. D., Teton, B., Beasley, J. C., Schlichting, P. E., Boughton, R. K., Wight, B., Newkirk, E. S., Ivan, J. S., Odell, E. A., Brook, R. K., & Lukacs, P. M. (2019). Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology & Evolution*, *10*(4), 585–590. https://doi.org/10.1111/2041-210X.13120

[4] Gaggiotti, O., Willi, M., Fortson, L., Pitman, R. T., Cardoso, A. W., Locke, C., Swanson, A., Veldthuis, M., & Boyer, A. (2019). Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology & Evolution*, *10*(1), 80–91. https://doi.org/10.1111/2041-210X.13099

[5] Joby, A. (2021, July 19). *What Is K-Nearest Neighbor? An ML Algorithm to Classify Data*. G2 Learn Hub. Retrieved May 14, 2022, from https://learn.g2.com/k-nearest-neighbor

[6] Data Science Process Alliance. (n.d.). *What is SEMMA?* Retrieved May 14, 2022, from https://www.datascience-pm.com/semma/

[7] SAS Institute Inc. (2017, August 30). *Introduction to SEMMA*. SAS Help Center. Retrieved May 14, 2022, from https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm

[8] Navlani, A., Fandango, A., & Idris, I. (2021). *Python Data Analysis - Third Edition*. Van Haren Publishing. https://subscription.packtpub.com/book/data/9781789955248/2/ch02lvl1sec06/semma

[9] Diceus. (2020, December 29). *Step by step software development: 7 phases to build a product*. Retrieved May 14, 2022, from https://diceus.com/step-step-software-development-7-phases-build-product/

[10] Delua, J. (2021, March 12). *Supervised vs. Unsupervised Learning: What's the Difference?* IBM. Retrieved May 14, 2022, from https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

[11] Kleinings, H. (2022, April 20). *What Is Data Extraction? Techniques, Tools & Examples*. Levity. Retrieved May 14, 2022, from https://levity.ai/blog/what-is-data-extraction

[12] Talend, Inc. (n.d.). *What is Data Extraction? [Tools & Techniques]*. Stitch: A Talend Product. Retrieved May 14, 2022, from https://www.stitchdata.com/resources/what-is-data-extraction