ABSTRACT

CLEGHERN, ZACHARY B. Who’s a Good Dog? Fusing Multimodal Environmental and Behavioral Data from IoT Sensors, Analytics, and Machine Learning to Characterize, Predict, and Improve Guide Dog Outcomes. (Under the direction of David L. Roberts.)

Producing guide dogs is resource-intensive, costing at least $50,000 per dog and around 24 months of expert and volunteer labor. Despite great efforts to identify successful dogs early on, many fail, a major loss to often non-profit organizations. Candidate dogs are raised by volunteers for 12 to 16 months, but there exists no objective data about the dog’s experiences during this phase. Remote, real-time objective data collection allows insight into how dogs develop away from guide dog schools and require less labor-intensive effort from canine behavior experts.

To remedy the lack of objective data during raising, we developed the “Smart Collar” system, allowing a sensor-enabled collar to remotely collect behavioral and environmental data that volunteer raisers can passively upload into a centralized cloud-based repository. I used these data to generate higher-level behavioral and environmental metrics to provide insight to guide dog organizations about dogs in raising. Improved selection would reduce costs for guide dog schools and save scarce resources, allowing more guide dogs to be successfully placed with those who need them.
Who's a Good Dog? Fusing Multimodal Environmental and Behavioral Data from IoT Sensors, Analytics, and Machine Learning to Characterize, Predict, and Improve Guide Dog Outcomes

by
Zachary B. Cleghern

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Computer Science

Raleigh, North Carolina
2020

APPROVED BY:

Alper Bozkurt

Margaret Gruen

Tiffany Barnes

David L. Roberts
Chair of Advisory Committee
DEDICATION

To service dogs everywhere.
BIOGRAPHY

Zachary B. Cleghern was born in 1992 and raised in Tennessee Ridge, Tennessee. He attended Tennessee Technological University. There he majored in Computer Science and minored in Mathematics and graduated with a Bachelor of Science degree in 2014. He went on to study at North Carolina State University in the doctoral program with an interest generally in machine learning. Under the direction of Dr. Dave Roberts he researched multiplayer game forecasting and received his Master of Science in Computer Science on the way to his PhD. For his PhD studies, he researched the intersection of canine behavior, wearable computing, and machine learning. He completed his PhD on this topic in 2020.
ACKNOWLEDGEMENTS

I would like to first thank my advisor, Dr. Dave Roberts, for his mentorship and advice over the years, and for challenging me to realize my potential. I would also like to thank the rest of my PhD committee for their advice and feedback: Dr. Alper Bozkurt, Dr. Margaret Gruen, and Dr. Tiffany Barnes.

A special thanks goes to Christa Cody, who always provided critical feedback, pushed me to attain my goals, and along with our cat, Balerion, provided love and support throughout our time in graduate school. Thanks to my family who instilled in me a sense of curiosity and desire to challenge myself and to my friends for their support.

I would also like to thank members of the CATS team, the CIIGAR lab, the iBionicS lab, everyone at Guiding Eyes for the Blind, and all the dogs who have helped our projects along the way: Diesel, Cash, Bruce, Simba, and all the dogs at Guiding Eyes who are or were training to be guide dogs.

This work has been supported by an IBM faculty award and NSF Grant No. 1329738.
TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................ vii

LIST OF FIGURES ..................................................................................................... viii

Chapter 1 Introduction ............................................................................................... 1

Chapter 2 Related Works .......................................................................................... 5
  2.1 Human-Based Methods ....................................................................................... 6
  2.2 Sensor-Based Methods ....................................................................................... 8
  2.3 Environments .................................................................................................... 13
    2.3.1 Types of Environments ............................................................................. 13
    2.3.2 Citizen Science ......................................................................................... 14

Chapter 3 Background ............................................................................................... 15
  3.1 Guide Dog Training ........................................................................................... 15
  3.2 The Smart Collar ............................................................................................... 16

Chapter 4 Smart Collar Data Collection System Design ........................................ 19
  4.1 Design ............................................................................................................... 19
  4.2 Iterations .......................................................................................................... 20
    4.2.1 Major Iteration 1 ....................................................................................... 21
    4.2.2 Major Iteration 2 ....................................................................................... 21
    4.2.3 Major Iteration 3 ....................................................................................... 23
  4.3 Conclusion ......................................................................................................... 24

Chapter 5 Predicting Success with Decision Trees .................................................. 26
  5.1 Abstract ............................................................................................................ 26
  5.2 Introduction ....................................................................................................... 26
  5.3 Data ................................................................................................................ 28
    5.3.1 C-BARQ .................................................................................................... 29
    5.3.2 Behavior Checklist .................................................................................... 29
  5.4 Building the Decision Tree Model .................................................................... 29
  5.5 Results ............................................................................................................. 31
  5.6 Discussion ........................................................................................................ 32
  5.7 Future Work ..................................................................................................... 34

Chapter 6 An IoT and Analytics Platform for Characterizing Adolescent Dogs' Suitability for Guide Work ................................................................. 35
  6.1 Abstract ............................................................................................................ 35
  6.2 Introduction ..................................................................................................... 35
  6.3 Related Works ................................................................................................ 36
  6.4 The IoT and Analytics System ........................................................................ 37
    6.4.1 Mobile App ............................................................................................... 38
    6.4.2 Cloud Back-end ....................................................................................... 39
  6.5 Methods .......................................................................................................... 40
  6.6 Results and Discussion .................................................................................... 41
TABLES

Table 3.1 Sampling Rates and Periods of Each Smart Collar Sensor .......................... 18

Table 5.1 C-BARQ features having statistically significant correlation with guide dog success. ................................................................. 30
Table 5.2 Features with statistically significant correlation with guide dog success. ... 30
Table 5.3 Best hyperparameter set for full feature set ........................................... 31
Table 5.4 Best cross-validation scores and test set accuracy, precision, and recall for each feature subset. Note that “breed” was actually removed as every dog in the data set had the same breed code. ......................................................... 32
Table 5.5 Best cross-validation scores and test set accuracy, precision, and recall for each feature subset. Note that “breed” was actually removed as every dog in the data set had the same breed code. ......................................................... 32
Table 6.1 MW-U Test between Indoor and Outdoor Walking Behaviors. All three tests statistically significant with $p < 0.001$. .................................................. 42
Table 6.2 Variance of IMU Acceleration Magnitude in Three Activities .................. 43
Table 6.3 Levene’s Test between Behavior States ................................................... 43
Table 7.1 Clustering Performance (Silhouette Score and Within-Cluster SSE) DBSCAN and K-Means Algorithms on Collar Location Data ............................. 49
Table 7.2 Comparison of IMU Data, Separated by K-Means Clusters(“C1” and “C2”) 50
Table 7.3 Comparison of Light Data, Separated by K-Means Clusters(“C1” and “C2”) 50
Table 7.4 Comparison of Sound Data, Separated by K-Means Clusters(“C1” and “C2”) 51
Table 7.5 Comparison of Pressure Data, Separated by K-Means Clusters(“C1” and “C2”) 51
Table 8.1 Overall Expert Rating of each High Level Metric ................................. 60
Table 8.2 Min, Max Values for Hyperparameter Optimization for Walking Speed; Best Values for each Feature ......................................................... 64
Table 8.3 Cross-validated Performance of Speed Model compared with Baselines .... 64
Table 8.4 Stats for emotional state categories ...................................................... 66
Table 8.5 Min, Max Values for Hyperparameter Optimization for Emotional State; Best Values for each Feature ......................................................... 67
Table 8.6 Cross-validated Performance of Emotional State models ........................ 68
Table 8.7 Environment Classification Results ...................................................... 70
Table 8.8 Min, Max Values for Hyperparameter Optimization for Sound Origin; Best Values for each Feature ......................................................... 73
Table 8.9 Cross-validated Performance of Sound Origin Model with and without Class-Weighting ................................................................. 74
Table 8.10 Frequency and Duration of Location Visits ........................................... 75
Table 8.11 Error of Familiarity Score with $K = 3$ Clustering Model vs. $K = 5$ Clustering Model ................................................................. 77
# LIST OF FIGURES

| Figure 3.1 | Smart Collar exploded view. | 17 |
| Figure 3.2 | Smart Collar Hardware Summary. | 17 |
| Figure 4.1 | The initial user interface for the mobile app. The interface is a single page. | 21 |
| Figure 4.2 | The user interface for the mobile app during the second major iteration. | 22 |
| Figure 4.3 | The user interface for the mobile app during the third (current) major iteration. | 24 |
| Figure 5.1 | Example portion of decision tree with 3 nodes, each using Gini Impurity splitting criterion. | 28 |
| Figure 6.1 | Mobile app interface for collecting Smart Collar data. Live data is shown to the user. In this example, IMU activity is shown. | 39 |
| Figure 6.2 | An overview of the cloud system architecture hosted on IBM Cloud using Kubernetes. | 40 |
| Figure 6.3 | The dogs wearing the Smart Collar. | 40 |
| Figure 6.4 | IMU acceleration magnitude for each dog in three behaviors: Sitting, Walking, and Stairs. Notable events are labeled with numbers and explained in the text. | 42 |
| Figure 7.1 | Representative example of SSE as $k$ increases for a single dog. | 49 |
| Figure 8.1 | Diagram of the neural network architecture for the Walking Speed metric. | 63 |
| Figure 8.2 | Predicted and actual speed for an example time segment from one of the walks. | 65 |
| Figure 8.3 | Diagram of the neural network architecture for the Emotional State metric. | 67 |
| Figure 8.4 | Precision-Recall curves for each emotional state. From left to right: Inhibited, Inactive/Relaxed, Excited, Activated with Stress. | 68 |
| Figure 8.5 | Plot of sound intensity for a typical bark on the left, no dog noise center, a “quieter” bark on the right. | 71 |
| Figure 8.6 | Diagram of the neural network architecture for the Sound Origin metric. | 73 |
| Figure 8.7 | Precision-Recall curve for optimized neural network model for Sound Origin metric. | 74 |
| Figure 8.8 | Sum of Squared Error as K increases for location data set. | 76 |
| Figure 8.9 | Mean Squared Error of Location Familiarity Model as K Increases. | 77 |
Guide dogs are service dogs with advanced training in order to lead and assist people with visual impairments. The reported costs to breed, raise, train, place and provide lifetime support for a guide dog are estimated to be at least $50,000 for each dog [Ber17; Bli]. In addition to this massive financial investment, each dog requires about two years of raising and training before it can be eventually placed with someone in need of a guide dog. However, many dogs that enter training do not succeed as guide dogs, either failing somewhere in the training process or with the eventual handler. This represents lost resources and lost opportunities to serve more individuals needing these dogs. Organizations expend great effort using behavioral and health data to influence decisions on which individuals are selected as replacement breeders and career sorting young dogs for their estimated potential [Mea17].

*Puppy raisers* volunteer to raise the puppies, teach obedience and social skills, and expose them to a variety of environments and experiences, which is crucial as guide dogs must work effectively in a wide variety of environments. Initially, Guiding Eyes staff visit puppies weekly then later twice per month. Formal assessments occur at 4 and 10 months of age. Finally, the dog takes the In-For-Training (IFT) test at roughly 16 months old. Following this, dogs begin training as a guide dog, enter other training (*e.g.* detection or emotional support), or be released as pets.

Using these formal assessments, I initially investigated how we can predict the eventual outcomes for candidate guide dogs. Outcome here means success or failure at some point in the future, either in training or in the final evaluation. I used already available subjective assessment data in order to build the prediction models. Using data available at the time of the IFT Test, I built decision tree models to classify the binary success outcome. Classification accuracy was low (around 60%) but showed promise in that these assessments are somewhat predictive of final outcome. However, the
input data I used were limited to the results of expert assessments. These are still subjective human generated data as at this time we had no way to collect more objective data about the dog during raising. This study can be found in Chapter 5.

While this phase of puppy development is crucial to a successful outcome of being placed with a permanent handler, it can be difficult for organizations to understand how a particular dog is progressing outside of periodic check-ins. With the high costs of producing a guide dog, a more quantified understanding of each dog's development and an estimate of their future success is a much-needed part of the process. This is a long time span (several months to a year) during which the dog develops considerably, so collecting more-objective data about the dog's behavior and environment may enable organizations with earlier estimates of a dog's suitability. If an organization can accurately predict the future outcome using automated analyses of its behavior and remove them from the program early for a more suitable role (such as a therapy dog or as a pet) this will save scarce resources.

To collect more-objective sensor data, our team has designed and implemented a “Smart Collar” system for dogs and raisers to use. This system makes use of a custom Bluetooth-enabled **Smart Collar** equipped with several sensors. A triaxial Inertial Measurement Unit (IMU) accelerometer senses acceleration of the dog to measure its activity and physical movement. A microphone detects sound levels to measure the level of noise both around and from the dog. A light sensor detects ambient light levels so that we can measure the brightness of the environment. Other environmental sensors include temperature, humidity, and barometric pressure sensors. Some sensors are primarily intended to detect information about the dog’s activities, such as movement or barking. These are intended to reflect the dog’s **behavior**. These behavioral data are collected by the IMU and sound sensors. Other sensors measure factors in the environment that the dog does not directly affect. These **environmental** data sensors include pressure, temperature, and humidity, while the sound sensor can be considered to collect both, as noise levels from the environment and noise from the dog are both detected by the sound sensor.

These data must all be gathered into a database to enable analysis. Instead of attempting to store vast quantities of data on the Smart Collar hardware, we designed the system to gather the data in nearly real time (or as close as possible). I developed a smartphone application to connect to the Smart Collar over a Bluetooth Low Energy (BLE) connection so that the collar can transmit data as it is collected. In addition, the raiser’s phone provides GPS latitude and longitude data when the Smart Collar is connected, which give further insight into the dog’s surroundings. The mobile app stores data temporarily until a Wi-Fi connection is present. When the raiser is connected to Wi-Fi, the app uploads the data to a cloud backend system. We developed this system as a central processing and storage pipeline for multiple types of canine data from Guiding Eyes dogs, but only the collar portions are in the scope of this work. The backend exposes an API with a Flask application that allows the app to initiate data uploads in batches with a unique tattoo identifier. These batches are then submitted as Gearman jobs for asynchronous import into a relational time series database. We host this pipeline on a Kubernetes cluster and architected it in such a way to enable easy horizontal
and vertical scaling.

As a demonstration of the system, I conducted an experiment using two dogs to validate the collected data. Two Labrador retrievers were instructed to exhibit several predefined behaviors in both an indoor and outdoor environment. I used statistical mean and variance tests to show how the resulting data is differentiated between behaviors and environments. This study is described in Chapter 6. We also initially deployed Smart Collar systems to 11 raisers for both general testing and for initial data collection. Their candidate dogs wore the collar while the raisers collected data and also noted what problems arose when using the system. I used these data to perform similar statistical tests and also unsupervised clustering; this time incorporating GPS location data. I found that data were clearly distinguishable between location-based clusters and that the data clustered well based on GPS location, indicating that differences in geographical location do indeed result in differences in collected data, importantly including IMU data, which is considered in the system to represent behavioral data. This study is described in Chapter 7.

With a new version of the Smart Collar ready, I updated the mobile application for ease of use and added additional features. We deployed 20 of these updated collar systems to several testers, the next goal being to scale up to 300 actual raisers. I defined several behavioral and environmental metrics that describe the dog’s behavior and environment at a higher level than the raw collar data. The purpose of this is to enable more advanced analysis, such as automatic reporting that would provide insight to guide dog organizations about dog development, and also for use in prediction of future outcomes using supervised machine learning. I conducted an interview with four guide dog experts about their opinions on these and other potential higher level metrics. Using their feedback, The global pandemic of 2020 caused significant delays and sources of uncertainty in the deployment of Smart Collars and subsequent data collection. However, I completed a study collecting collar data in order to build models that can generate these metrics. I used various machine learning algorithms to estimate the value of the metrics and analyzed their performance.

Specifically, my thesis is that we can utilize on-body sensor and GPS data collected on adolescent candidate guide dogs in an uncontrolled environment over the raising period with their volunteer raiser families to both accurately generate high level behavioral and environmental metrics and to build predictive machine learning models to estimate future success with accuracy higher than subjective expert evaluation alone.

I aimed to address the following research questions:

1. How can we build a system that adequately and non-invasively captures dog behavior, without putting undue burden on the volunteer raisers?

2. How can we collect data in a way that allows us to gain insight into how the dog reacts to stimuli and spends its time during the raising period?

3. How can we derive higher level metrics describing the dog’s current environment and reaction to stimuli?
4. How can the collected objective sensor data be used to model the dog’s behavior and likelihood of success?

Note that this work does not address RQ4. The global pandemic of COVID-19 greatly impacted the ability to collect data and resulted in numerous manufacturing delays and thus this work instead focuses only on the first three research questions.

The rest of this document is outlined as follows. Chapter 2 contains related works and there I discuss similar approaches to understanding canine behavior and collecting various types of data, especially as it applies to service work. In Chapter 3, I provide relevant background information about guide dog raising at Guiding Eyes for the Blind and more in-depth information about the Smart Collar hardware. Chapter 4 describes the design of the data collection system making use of the Smart Collar. Chapter 5 describes a study in which I used decision trees to predict successful outcomes using existing assessment data. An overview of the complete Smart Collar system can be found in Chapter 6, as well as a study using two dogs to test the system. In Chapter 7, I present a study in which I used data collected from raisers as a larger scale test and employed unsupervised learning to extract more information from production data. In Chapter 8, I present an investigation into higher-level metrics and how to generate them from raw data. Finally, Chapter 9 contains a conclusion of the work completed.
Understanding canine behavior, especially as it relates to suitability for service work, is an active, growing area of research. Early prediction of successful service work is particularly important, as the cost of breeding, raising, training, and supporting a single service animal can be highly expensive. For guide dog work in particular, Guiding Eyes for the Blind has estimated that the cost for a single dog is at least $50,000\[Bli]. Researchers have employed both insights from veterinary medicine and innovation in various fields of engineering to better detect, classify, and predict canine behavior\[Maj16; Byr18; Pis14; Mea17; Mea19]. More broadly, the growing field of canine-centered computing\[Fre17] is at the forefront of advancements in better understanding dog behavior with technology. Technological innovation informed by canine behavioral science has a large potential for increasing the understanding of and ability to assess and predict service dog suitability\[Mea19; Byr18; Ber17].

We can characterize approaches to behavioral analysis in canine research along several dimensions. Researchers have used various methods to collect data. Some approaches involve fully subjective human observation and some make use of more objective, electronic sensor-based methods. In the following sections, I discuss both of these approaches, referred to as human-based and sensor-based methods of observation. Human-based methods involve human observation and decision-making and can be further characterized by the level of expertise of the humans involved and the specific methods used. Sensor-based methods make use of various electronic sensors to collect data with less or no input from a human observer, though it should always be noted that this does not fully eliminate human subjectivity. Data can be collected in many types of environments, ranging from an natural home settings to specially-designed evaluation rooms where many variables can be controlled.
For example, the Smart Collar system primarily employs sensor-based approaches for data collection and analysis. The environment for general data collection with this system is very uncontrolled. Raisers use the system in their homes and wherever they may go with their puppy during their typical routine.

### 2.1 Human-Based Methods

A widely used approach to understand canine behavior, including at Guiding Eyes, is the use of human observation without any electronic sensors. Behavioral coding is a data collection technique in this category in which an expert notes when an animal exhibits predefined behaviors. Behavioral and physical tests involve some evaluation process to collect data about the dog. Evaluations may be developed specifically for a particular study or based on standardized metrics such as with the C-BARQ questionnaire.

Bray et al. used behavior coding to evaluate maternal style in the breeding population at a guide dog school. They did this by coding seven variables of interest\[Bra17a\]. These variables were *time in pool*, *vertical nursing per pup*, *lateral nursing per pup*, *ventral nursing per pup*, *contact per pup*, *licking/grooming per pup*, and *orienting out*. This principal component composed of seven variables also had predictive ability of success in guide dog training. By using this measure of maternal style, Bray et al. studied 98 puppies in a guide dog training program\[Bra17b\]. They found a negative association with early intense levels of mothering and success in the guide dog program. They found that puppies whose mothers had a nursing style requiring more effort from the puppy had a higher likelihood of success in the training program. In another study, Tomkins, Thomson, and McGreevy found associations of specific behaviors with future guide dog success, such as frequent panting or licking behaviors during a dog distraction test\[Tom11\]. Behavioral coding is also a practice used at Guiding Eyes to label behavioral data during evaluations such as the eight week puppy test. However, as the raising period largely takes place geographically far away from the school, behavioral coding by experts is not used with the Smart Collar system itself for general data collection.

Behavioral and physical tests are also administered in order to predict future behavior. Weiss and Greenberg administered an 11 item behavioral test on nine dogs in an animal shelter and then gave them obedience and retrieval training\[Wei97\]. They found no correlation between the test and ability to learn retrieval tasks, but that they could predict the “fear/submission” variable from the test. Weiss further developed this study with an experiment using 75 dogs\[Wei02\]. In this updated experiment, Weiss used an updated selection test and then the dogs underwent obedience and retrieval training. They found that some of the selection items did have an association with a variable called ”service success”, comprised of trainer-rated scores: trainability, service dog potential, and attention span. With this metric of success, a subset of the selection test items (sensitivity, opinion, reaction to other dog, stranger, stare, and vertical activity level) accounted for about 36% of the variance in “service success” with $R = 0.603$. Weiss suggested that future studies should also examine cortisol levels should be incorporated into this form of behavioral research. Batt et al. found found
several behavioral tests and physical characteristics, including salivary cortisol, that were predictive of success [Bat08]. In this study, the researchers tested motor laterality, behavioral reactivity, and salivary cortisol concentrations in 43 potential guide dogs across a 14 month period. Saliva testing, temperament and lateralization tests took place at 6 months of age at the guide dog training center. At 14 months of age, these were repeated at the center when dogs returned for training. Finally, a third round of lateralization and saliva tests took place before 20 months of age. This final testing took place after dogs were chosen whether or not to be suitable guide dogs. Here the temperament tests had six components: Social Contact Test, Passive Test, Chase Test, Noise Test, Dog Distraction Test, and Sudden Appearance Test. For example, the Social Contact Test measured a dog’s time taken to approach a human tester in a controlled testing room. These tests are based on tests developed by Svartberg [Sva02] except for the Dog Distraction Test, developed by van der Borg, Netto, and Planta [Bor91]. Lateralization tests consisted of a Tape Test and Kong Test, described in detail in [Bat07]. They found that their results obtained at 14 months of age to be superior to results obtained at 6 months. This was most likely because the dogs had more time to develop as dogs reach social maturity between one and three years [Ove02] and behavioral traits are more stable at this point [Cas99]. This result aligns with the time frame in which Guiding Eyes puppies end the raising phase and are evaluated for advanced training, which is timed such that the puppy’s behavioral traits have had time to develop.

Standardized temperament evaluations are also common in both predicting service dog success and in understanding canine behavior more broadly. Questionnaires provide a repeatable standard set of evaluation metrics that can be used for a variety of purposes, but of course still require a subjective, though usually expert, human evaluator. The C-BARQ [Seg05] is a 100 item questionnaire that is used to evaluate canine temperament. It has been validated for use in dogs [Hsu03; Duf08] and is used by Guiding Eyes for the Blind. Marshall-Pescini et al. used the C-BARQ as part of their data collection process in a study examining the effect of training on dogs’ cognitive abilities [MP08]. To study the effect of training, 118 dogs ranging from 6 months to 10 years of age completed a problem-solving task in an unfamiliar room. They found that dogs with training experience (such as search and rescue or retriever work) were more easily able to obtain food from a testing apparatus. Successful dogs had a significant association with trainability and lower “stranger-fear” scores on the C-BARQ questionnaire.

Standardized questionnaires can also be used to aid non-experts in collecting behavioral data. Serpell and Hsu used the C-BARQ to survey the owners of 1,563 dogs to assess a variable called “trainability” among 11 dog breeds [Ser05]. This study found significant differences in trainability among several breeds without the need for formally-trained experts to physically interact with each dog. The standardized nature of the C-BARQ questionnaire combined with the fact that a specified evaluation area was not needed allowed the survey to reach a much wider population size than would be feasible otherwise, which is a goal of the Smart Collar project which aims to increase the scale at which data collection is possible. This concept is expanded further in Section 2.3 of this chapter which discusses citizen science.
Arata, Momozawa, and Takeuchi [Ara10] assessed that distraction, sensitivity, and docility are important behavioral factors that can predict a dog’s future success based on a questionnaire, and that a dog’s level of distraction predicted its future success with 80.6% accuracy. This study was conducted with candidate guide dogs which had completed three months of advanced guide dog training and had about three months remaining before final evaluation.

Duffy and Serpell used results from the C-BARQ questionnaire at five service dog organizations to predict future success of guide and service dogs [Duf12]. With 11,997 evaluations from 7696 dogs (evaluations took place at both 6 and 12 months of age), they used a generalized linear model to predict a binary success outcome. They found that successful dogs scored better on 27 C-BARQ items at both evaluations. The trait Pulls excessively hard on leash was the most predictive trait in their analysis. Each increase by one point in the C-BARQ score for pulls excessively hard on leash was associated with a 1.4% increase in the probability that a dog would eventually fail the program. Serpell also developed the Behavior Checklist (BCL) for evaluating temperament. The BCL is a scoring system where observers can score up to 52 aspects of behavior such as excitability, various fearful behaviors (such as traffic or noise fear), and body sensitivity. Each item in the Behavior Checklist is scored with ordinal values one through five. Like the C-BARQ questionnaire, the Behavior Checklist is a commonly used scoring system for scoring the behavior of potential guide dogs, including at Guiding Eyes. Both the Behavior Checklist and C-BARQ questionnaires have been validated for their use in canine behavioral evaluation and thus in Chapter 5, I use these two evaluations as data sources from Guiding Eyes to predict success and failure of potential guide dogs.

### 2.2 Sensor-Based Methods

Sensor-based methods of data collection are more objective approaches using electronic sensors, such as physiological or environmental sensors, to record measurements with lessened need for human decision-making in the process. Human interpretation is still be needed, so sensor-based methods aren’t fully objective, but the data collection approach is more objective than the previously discussed human-based methods. Researchers have turned to sensor-based approaches in order to study canine behavior and in particular to predict the suitability of service dogs. Reasons for making use of sensors include reduced (but not eliminated) human bias, precision, consistency, and the fact that these technologies are becoming more and more feasible for use in this research domain. Sensor systems collect data and are equipped either off of the dog’s body (such as a depth sensor) or as a wearable computing device on the dog’s body. In this chapter, I refer to these two forms of sensor systems as “on-body” and “off-body” systems. The Smart Collar system is one such example of an on-body sensor system. With a more objective measurement process that sensors provide, human bias is reduced and there is potential for scalability not possible with human, especially expert, observation. Developments in wireless networks and the internet-of-things (IoT) have also enabled advancement in this field while bringing with it new challenges [Maj16], discussed in this section.
2.2.0.1 Off-Body Sensors

Off-body sensors allow for data collection without any device physically attached to the dog. This can reduce unintended alterations in behavior that could result from attaching sensing equipment to a dog. Using a camera with a depth sensor like the Microsoft Kinect, Pistocchi et al. proposed a solution for classifying a dog’s body parts from a video feed [Pis14]. They used a structural support vector machine to model the dog’s shape from depth sensor data and validated the results using visual comparison. In their study, qualitative analysis of their results by domain experts resulted in 96% of images either correctly classified or partially correctly classified, meaning at least half of the body correctly classified.

In addition to body part detection, the Kinect camera has also been shown to be an effective sensor for the purpose of posture recognition. Mealin, Dominguez, and Roberts used the Microsoft Kinect to extract a model of the dog’s body without human-authored labeling [Mea16]. Using data from the Kinect, they used the model to differentiate between static postures. To evaluate the system, they detected several common dog postures such as standing, sitting, and lying. They correctly classified the standing posture with 70% accuracy, sitting with 69% accuracy, and lying with 94% accuracy.

Berns et al. used functional Magnetic Resonance Imaging (fMRI) to predict successful service dogs at Canine Companions for Independence. They did this by directly analyzing brain scans [Ber17]. While in training, a cohort of 49 dogs learned hand signals corresponding to “reward” and “no reward” and were acclimated to the fMRI machine. Of the 49 dogs, 77% were placed successfully (33 were matched and 4 were selected to be breeding dogs). Using the brain scan images, they developed a generalized linear model to classify success or failure. This resulted in a precision of 0.94 and negative predictive value of 0.67. With 4-fold cross-validation, the model resulted in an area under the ROC curve of 0.80. This was somewhat predictive of future success and performed somewhat better (fewer false negatives) than a more traditional behavior test, but did require dogs to be acclimated to and tolerate the fMRI machine.

Byrne et al. used custom-designed dog toys equipped with electronic sensors to measure potential service dogs’ bite levels and predict service suitability [Byr18]. These sensors consisted of a barometric pressure sensor, accelerometer, gyroscope, and a magnetometer. The barometric pressure sensor allowed for the measurement of bite force, as the air pressure inside the toy increased when bit by a dog. The accelerometer, gyroscope, and magnetometer together make up a 9-axis Inertial Measurement Unit (IMU) for studying motion. In this study, Byrne et al. used a population of 40 dogs who had underwent basic socialization and obedience training, but not advanced service dog training. The successful dogs went on to perform roles such as facility dog, hearing dog, or post-traumatic stress disorder dog. With a logistic model tree for classification, they predicted the eventual outcome (success/failure) of the dogs with an accuracy of 87.5%, precision of 0.872, and recall of 0.875. By their estimates, this level of accuracy could save $70,000 in a cohort size of 40 dogs by identifying dogs that would likely fail the training program. However, this study was conducted in a relatively controlled environment. The experimental setup made use of custom-built hardware.
that collected data locally when the dog interacted with the instrumented toys. The Smart Collar system also employs custom designed hardware but does so remotely.

Off-body sensing methods can minimize alterations to behavior and in some cases are more tolerable for the dogs. These methods have found success in both image classification for posture and behavior recognition and also in predicting behavioral temperament. However, these approaches do require a certain level of control over the environment that could hinder scalability. For example, recording video with depth data requires a room to be setup for data collection with the depth camera installed and adequately placed. Off-body sensing systems are not as well suited for large-scale remote data collection from volunteers.

### 2.2.0.2 On-Body Sensors

On-body sensing systems require physical attachment to the dog’s body, such as electrodes for electrocardiogram (ECG) data, sensor-equipped collars, or other wearable computing devices. This comes with some limitations as the dog does need to be able to tolerate the physical device in order for data collection to take place. However, this method enables the collection of a wide range of data types due to the physical proximity to the dog. Researchers have employed on-body sensor systems both for data collection and also for canine communication \([Jac15; Mea15]\).

Using sensor devices equipped to a dog collar is becoming a common approach to this form of data collection, especially with the use of inertial measurement units (IMUs). Brugarolas et al. developed methods for using an IMU-based system to aid dog trainers and proposed a concept called “canine body area network” or cBAN\([Bru12]\). They optimized placement of the IMUs and found they were able to correctly identify postures from the resulting IMU data. Expanding on this work, Brugarolas et al. used machine learning to classify postures in two dogs\([Bru13b]\). Each dog completed a sequence of five postures several times while wearing the IMU-equipped collar to record its movement data. Using the resulting 3-axis accelerometer and 3-axis gyroscope data, they trained models using random forests, k nearest neighbors, and logistic model tree algorithms. They used a cascade learning approach to first separate postures from the transitions between postures. The combined results were approximately 98% classification accuracy for each machine learning algorithm. To evaluate generalizability, they also trained models on data from one of the dogs and used data from the other dog as a test set. In the worst case, this resulted in about 81% accuracy and in the best case 100% accuracy. In addition to this, Brugarolas et al. also proposed methods for detecting both postures and behavior states using the same collar-mounted IMU system\([Bru13a]\). In this study, they used decision trees to classify static postures and Hidden Markov Models (HMMs) to detect dynamic activities (behaviors) of seven dogs and assessed the reliability of both. The HMMs were trained using maximum likelihood estimation. The HMM model classified behaviors into four states, one of which was “don-dynamic”, which meant the dog was in a static posture. In this case, the algorithm handed off the data to decision trees to classify what static posture the data represented. The HMM correctly classified walking and walking up stairs behaviors with 100% accuracy, walking down a ramp with 100% accuracy (except for one dog which was 92%), and non-dynamic behaviors
(postures) with accuracy above typically about 95%. The two-level decision tree approach for static postures achieved high accuracy using accelerometer data (typically above 95%) and somewhat lower accuracy with only gyroscope data (typically around 70%). This study further validated the collar-mounted IMU method for collecting behavioral data and showed that accurate analysis of behavior is possible in this manner and with this technology.

Den Uijl et al. [Uij17] also used a collar-mounted IMU, in this case a triaxial accelerometer, to detect several behavioral states in dogs. The states they studied were \textit{walk, trot, canter/gallop, sleep, static/inactive, eat, drink,} and \textit{headshake}. They achieved impressive results for most of these states, achieving a precision in the 93-100% range and recall in the 89-98% range. Gerencsér et al. used a 6-axis IMU to classify behavioral patterns[Ger13]. They collected data with a harness-mounted IMU sensor with 24 dogs in an open outdoor environment. They coded seven behaviors from accompanying video data and used these tags as labels for supervised learning. The seven behaviors were \textit{lay, sit, stand, walk, trot, gallop,} and \textit{canter}. They used support vector machines (SVMs) for classification. The SVMs resulted in an accuracy of at least 90% in the best cases in which data from an individual dog was used as both training and test data. Using training and test data from several dogs, they achieved an accuracy of about 80%. Winters et al. combined harness-mounted IMU data with skeletal measurements to improve posture recognition[Win15]. In their approach, the dog wore a harness consisting of a strap around its midsection and a strap attached to a collar, both holding IMU sensors. A video camera recorded the five dogs during the process which was used to manually tag postures and behaviors. A random forest model performed the best at classification for recognizing the postures, with a mean accuracy of 88.55% . The authors found the process of manually tagging data with posture labels to be very time consuming, prompting a need to eliminate this manual tagging of data.

Kim et al. used a wearable microphone device to collect sound data from dogs at home in order to classify their vocalizations [Kim18]. They used a combination of Long Short-Term Memory and Convolutional Neural Network models to classify sound inputs into four categories: barking, growling, howling, and whining and achieved up to 84% accuracy. This is similar to my work in that the authors used sound intensity to classify canine sounds, but in this work they were not attempting to separate dog sounds from environmental noises.

Wearable IMU-based sensing systems are capable of detecting various canine postures and distinguishing between behaviors. This shows promise for the use of IMU sensors to analyze complex facets of canine behavior, which is a goal of my work using collar-based sensing systems. Successful guide dog temperaments involve differences in overall behavior that is detectable by human experts. Therefore, the use of IMU sensors to analyze these behaviors in potential guide dogs is a promising avenue of research. In fact, recent research has begun to use wearable sensors to not only detect specific behaviors but specifically to predict guide dog success as well. Mealin et al. developed an evaluation system to assist domain experts at Guiding Eyes for the Blind in evaluating puppies at the age of eight weeks. This “puppy test” serves as an initial screening for potential guide dogs in which the puppies interact with various objects and stimuli in a room while expert observers
score them using the Behavior Checklist (BCL) [Mea17]. The system Mealin et al. created makes use of a harness worn by the puppies and is equipped with both IMU and electrocardiogram (ECG) sensors. They designed the harness with young puppies in mind to both collect good data and to minimize discomfort for the dog. Puppies that did not tolerate the harness completed the evaluation as normal without sensor data collection. In this system, data is transmitted over Wi-Fi to a laptop workstation running custom software. A camera collects video data that is also recorded and shown on the software’s user interface. In this software, the evaluators can annotate predefined “tasks” and behaviors that comprise the puppy evaluation. These physiological and movement data allow for machine learning algorithms to assist the trainers at guide dog schools with puppy selection. Mealin et al. employed deep learning on ECG data collected by this evaluation system to predict what BCL scores a puppy would receive [Mea19]. They chose to predict BCL scores since final outcome data for any given puppy are not available for up to two years and BCL scores are a major part of the screening process. Thus, BCL scores are a proxy for future success. Mealin et al. focused on 29 BCL score categories over 11 tasks in the puppy evaluation. Using a combination of Long Short-Term Memory (LSTM) layers and Convolutional Neural Networks (CNNs), their model predicted the 29 scores with around 92% accuracy (with a possible five outcomes for each BCL). They also found that any of the tasks were predictive of any of the BCL scores, so the important information that the model extracted from the data was not necessarily tied to any particular task of the puppy test, showing evidence of generalizability. In other words, they found that the machine learning model was able to extract information from any particular task that was predictive of a broad set of categories of behaviors and the specific task of the puppy test was less important.

Wearable sensors have recently become a topic of interest for more objective data collection while dogs are with puppy raisers. Volunteers raise the puppies for typically more than a year during a crucial period of development [Lue11], so there is a great opportunity for studying puppy behavior now that the technology to do so has become cheaper and readily available. However, this also comes with new challenges. One of these new challenges is that of asking volunteer raisers to assist with data collection. Zamansky and van der Linden surveyed current puppy raisers about their attitudes toward activity trackers worn by dogs [Zam18]. While they found that the majority of participants had positive things to say about the value a tracker would provide, some did have concerns about privacy, data sharing, and comfort of the dog. For example, some raisers did have concerns about the use of location data, as the location data of the dog at a particular time also locates the raiser. Ultimately, the study found that the majority of raisers would unconditionally accept activity trackers as part of the program if the guide dog school wanted to use them, but privacy is a serious facet of this research and must be considered in the design of the technology and protocol. Additionally, van der Linden et al. investigated the privacy aspects of commercially-available activity trackers [Lin19] and found several privacy concerns in this domain that should be noted. First, they found a mismatch between product marketing and what data the trackers typically collect. Some, for example, are marketed as fitness trackers with no mention of location tracking, but collect it anyway through the use of mobile application. They also found that many products do not offer clarity on what data, if
any is considered private. Finally, they found that in many cases it is unclear what data is stored and especially what data is inferred from collected data. On the other hand, custom-built sensor systems do not have this issue, since researchers can design them to collect exactly the data they are interested in and thus custom-built systems have a much higher potential for data transparency. Researchers must however take care to design these systems to ensure raiser privacy. The Smart Collar system is no exception and privacy is especially important since the system involves the collection of GPS location data.

2.3 Environments

A dog's behavior is related to its environment, and thus sometimes their behavior in one situation may not be indicative of another. This is illustrated in the case of trying to predict problematic behaviors in shelter dogs as studied by Patronek and Bradley, where they found that eliciting biting and warning behaviors in the chaotic, stressful environment of a shelter is a poor test for problem behaviors when placed with an adoptive family[Pat16]. They instead suggested interaction with the dogs in a normal manner that reflects its typical activities.

A dog's environment greatly affects its development[Lue11] and has been studied in the context of guide dog suitability as well[Har16]. Researchers study canine behavior, particularly in attempt to predict suitability for service work, in various environment types, ranging from controlled experimental settings to owner's homes in a completely uncontrolled large-scale approach. Each of these has benefits and drawbacks. The dog's environment itself is also a topic of interest for researchers interested in service dog behavior. Burrows et al. studied service dogs for children with autism spectrum disorder and used observations and interviews to investigate what factors in the home environment affects a dog's behavior and welfare[Bur08]. Serpell and Duffy analyzed C-BARQ data from 978 potential guide dogs at a guide dog school and found several environmental factors that influence C-BARQ results relevant to guide dog success[Ser16]. Their findings show that households with experienced raisers and households with at least one other dog were both associated with successful behavioral traits at 12 months of age.

2.3.1 Types of Environments

Analyzing dog behavior, especially for suitability for service work, must also consider the environment in which the data collection takes place. Controlled environments such as evaluation rooms offer the ability to control environmental variables external to the dog and also allow for more standardized evaluations. Controlled environments allow for experts to administer specific tests[Bat08; Ber17] and also allow for easier incorporation of technology into the process, such as the use of a controlled room designated and tested for use with a Kinect camera, or highly customized sensing devices without internet connectivity[Byr18]. This comes with limitations as well. For example, to evaluate canine brain scans when shown reward/no-reward signals, Berns et al. first had to acclimate dogs to the fMRI environment[Ber17]. While it was useful to generate these data, this
process required extra time and resources and could have also been biased (acclimating the dogs to the fMRI was in itself a form of training).

Researchers have also studied canine behavior in more natural settings such as observation at home. This can be accomplished by the dog's owner (or raiser in the case of guide dogs). In fact, Batt et al. used questionnaire data (a modified form of the C-BARQ) filled out by raisers to predict eventual success in a guide dog program with logistic regression and achieved a sensitivity of 0.68 and specificity of 0.63. In general, standardized questionnaires can assist non-experts with data collection. Collecting sensor data in the home is now possible in part due to reduced technology costs, but has so far been mostly limited to consumer products such as Whistle\cite{Yas15}. Väätäjä studied the rise in wearable computing devices for dogs and found they are driven in part by pet owners' need to monitor their dog's well-being\cite{Vää18}. Zamansky et al. found that commercial activity trackers have been beneficial for owners' relationship with their dog, motivating them to ensure proper levels of exercise\cite{Zam19}. Using such trackers to directly capture data for researchers to analyze and predict behavior is, however, relatively unexplored, though Zamansky and van der Linden interviewed several guide dog raisers about their perceptions about using such devices\cite{Zam18}. They found perceptions to be positive but identified privacy concerns about data collection and handling. Utilizing wearable sensors for understanding the behavior and environment of potential guide dogs would allow unique insight into the development of successful guide dogs through observation in a natural home environment. As such, the Smart Collar system is intended for use in this natural type of environment. In this manner, we can remotely collect data as the dog naturally interacts with its raiser.

### 2.3.2 Citizen Science

Placing no restrictions on the environment at all, citizen science allows for data collection at very large scales. Researchers have recently used citizen science, collecting data from members of the public, typically without formal scientific training, in the domain of canine behavior. Researchers gather dog behavior data from sample sizes that are normally too large to be practical in a lab setting. Hecht and Rice investigated the benefits and technical issues faced by the use of citizen science in doing canine research \cite{Hec15}. Data quality is, however, an issue that needs to be addressed since it can be difficult to ensure that the observer follows correct procedure. An example of citizen science for canine research is Dognition \cite{Ste15}. Dognition is a project with the goal of gathering behavioral data from dog owners at a large scale. Dog owners fill out a questionnaire and perform various cognitive exercises and then report these results to Dognition using a website. Hare found that this method is useful for collecting large amounts of data from citizen scientists and found little evidence that the citizens manipulated results \cite{Ste15}. This result has less of an impact on the Smart Collar project as data collection is mostly passive through the use of the collar and mobile application, but it does bode well for the evolution of the system to include more raiser input in the data collection process.
3.1 Guide Dog Training

The process of training dogs as service animals is highly resource intensive, both in terms of time and financial resources. Some estimates show the cost to raise and train a service dog to be $20,000 to $50,000 [Ber17], or more, especially in the case of guide work [Bli]. Guiding Eyes for the Blind is a large, nonprofit guide dog organization in the United States. Experts with decades of experience breed and train dogs from birth to placement. For the first few weeks of their lives, puppies are whelped and socialized with their litter. At about 7.5 to 8 weeks, Guiding Eyes conducts the first formal assessment known as the "puppy test". In this test, the puppies enter an evaluation room and complete a series of tasks and react to various stimuli. The evaluators score the puppies and some are screened out of the guide dog program for other roles. About 11% of puppies tested are provided to other organizations that train service dogs, 16% are adopted as pets, and 73% pass and remain in the program.

The 73% of dogs that pass the puppy test enter the next phase of guide dog training: the raising phase. For the next several months, volunteer raiser families (puppy raisers) raise the puppies in their homes. This takes place in several locations across the United States. Raisers socialize the puppies further, expose them to many new environments and people, and teach them basic obedience skills (they do not teach the puppies any guide-specific skills at this point). Raisers take the dogs with them when they run errands, take them to meet other dogs, and expose them to many new types of environments as these experiences shape their future behavioral traits and will be a crucial aspect of being a guide dog. During this time, raisers report on their dog’s progress and experts from the school carry out periodic subjective evaluations on each dog. Initially, Guiding Eyes staff visit
puppies weekly but later this is reduced to twice per month. Formal assessments, conducted by Guiding Eyes staff, occur at 4 and 10 months of age. At the age of one year to 18 months, the dog returns to Guiding Eyes and takes the In-For-Training (IFT) test. Following this, dogs begin advanced training as a guide dog, enter other training (e.g., detection or emotional support), or are released as pets (no dogs at Guiding Eyes are sent to animal shelters). Currently, Guiding Eyes has a plethora of data, including some objective data, during both the puppy’s first few weeks of life and also while it is at the school for formal training, but no source of objective data about the dog’s experiences during raising. This is a long time span (several months to a year) during which the dog develops considerably.

3.2 The Smart Collar

The Smart Collar system we used in this project is a custom wearable computing Internet-of-Things device that uses Bluetooth Low Energy to communicate sensor data [Wil20]. We did not find an off-the-shelf product that both collects the relevant data types we are interested in and has an open design for easy interfacing. The system is housed inside a 3D-printed casing with loops to attach a nylon collar strap.

The hardware design centers around a BLE113 System on Chip (SoC) (BLE113-A-M256K, Silicon Labs, Austin, TX, USA) based on the CC2541 BLE microcontroller (CC2541, Texas Instruments, Dallas, TX, USA). Sensors on the custom printed circuit board (PCB) record several types of data. Sensors detect environmental factors such as ambient light, ambient temperature, barometric pressure, and relative humidity, as well as ambient sound levels and acceleration with an Inertial Measurement Unit (IMU). Importantly, this device only senses ambient sound levels without any frequency information. In other words, we cannot reconstruct audio or record the words that a human says. Recording this information would introduce significant privacy risks. The IMU is a triaxial accelerometer, so the collar can detect acceleration in 3-dimensional space but not how the collar is oriented. A summary of the hardware is shown in Figure 3.2.

Figure 3.1 shows an exploded view of the collar. Air can travel inside the housing through the acoustic air vent which enables measurement of environmental data, but otherwise the casing provides splash-proof protection from the environment and standard wear and tear (though it is not fully waterproof). The 4-layer PCB is 43 mm x 37 mm and the complete encased system is 48 mm x 43 mm x 12 mm which contains attachment supports for the collar nylon webbing.

The firmware is capable of over-the-air updates using a bootloader and external memory on an SD card provides data storage for when the device loses Bluetooth connection to the mobile application.

Table 3.1 shows sensor sampling rates on the collar. As dog behaviors can happen very quickly and require a finer resolution in the data stream, audio levels and IMU acceleration have high sampling rates. The rest of the sensors sample much more slowly and record environmental conditions. These environmental conditions are unlikely to change nearly as frequently as the other sensor types. As
the sensors are housed in a water-resistant plastic packaging, some sensors do not accurately reflect conditions outside the collar (especially barometric pressure), but this error is consistent and our aim is to detect patterns and analyze differences across time and between dogs.
<table>
<thead>
<tr>
<th>Data Type</th>
<th>Sample Rate (Hz)</th>
<th>Sample Period (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMU</td>
<td>15</td>
<td>67</td>
</tr>
<tr>
<td>Sound Level</td>
<td>50</td>
<td>20</td>
</tr>
<tr>
<td>Ambient Light</td>
<td>0.2</td>
<td>5000</td>
</tr>
<tr>
<td>Pressure</td>
<td>0.1</td>
<td>10000</td>
</tr>
<tr>
<td>Temperature</td>
<td>0.01667</td>
<td>60000</td>
</tr>
<tr>
<td>Humidity</td>
<td>0.01667</td>
<td>60000</td>
</tr>
</tbody>
</table>
This chapter outlines the design of the system and an overview of how the mobile application works to collect smart collar data. Then, I describe the development of the system as new features have been added incrementally and general design principles that resulted from this iterative process.

4.1 Design

The need for a mobile application arose from hardware limitations in the Smart Collar. The ability to upload data over a Wi-Fi connection would have been infeasible given the collar design and cost constraints. Additionally, a mobile application provides a visual interface for users to interact with and monitor the collar.

Several considerations informed the app design. Firstly, the app should facilitate communications with the Smart Collar in a manner that does not distract the raiser from the dog as much as possible. Raisers should also be able to use the system to successfully collect data without technical expertise. Finally, the app (and the rest of the Smart Collar system) should not intrude on the raiser’s privacy and all data should be handled with care to minimize privacy risks.

I chose iOS as the first platform for the mobile application because the population of raisers initially available predominantly owned Apple iPhones. For ease of cross-platform development, I chose the Xamarin Forms framework.

Each Smart Collar device is essentially interchangeable with any other barring a unique ID. The app uses this unique ID to pair with the raiser’s own Smart Collar. Once paired, the app remembers
that collar ID so that two raisers with their dogs nearby will not have any problems connecting to
the individual collar. Once data collection begins, the raiser does not need to interact with the app. Minimizing impact on the raiser and dog's routine is important so that the data collected will not be skewed by the presence of the system. The app passively collects data until the raiser disconnects or turns off the collar.

In addition to the sensor data collected by the collar itself, the app collects latitude and longitude information using the phone's built-in GPS. However, this is only active when currently connected to the collar. We are only interested in data about the dog and unnecessarily collecting GPS data would increase privacy risks to the raiser. GPS data is recorded whenever the device moves 5 meters or more from the previous measurement. Position between recorded measurements can be interpolated. GPS data will be used to differentiate between types of environments the dog is exposed to. For example, a dog with typical behavior in rural areas or parks may experience inordinate stress in urban areas, impacting behavior as detected by the collar.

The app stores these data until it can be uploaded to our backend system. Uploads take place in batches when a Wi-Fi network is available and stop if the Wi-Fi connection is lost. This prevents us from inadvertently impacting a user's mobile data plan. Along with each batch of data, the app includes the dog's unique tattoo (as entered by the raiser) so that the backend can identify which dog the data belong to. This is the only identifier we have for entity resolution and allows us to match sensor data to dogs, but we cannot match data to a human raiser. In fact, since Guiding Eyes can only match dogs with people but does not have the sensor data, no one involved can match collar data with individual people.

The collar also stores data when the connection is broken using the SD card. The app stores these data points separately along with the timestamps associated with disconnect and reconnect events. The reason we need to store the timestamp of these events is that the timestamps on stored memory data correspond to when the data is transmitted, not stored (the collar does not have the capability to store the current time). The disconnect/reconnect pairs are then used to reconstruct as closely as possible the original timestamp of the stored data.

Finally, the mobile app also facilitates Over-The-Air (OTA) firmware updates of the Smart Collar hardware. An API endpoint on the backend system contains the version number of the latest collar firmware available. If an update is needed, the app downloads the firmware binary and the user can update the collar firmware. The user interface displays a progress bar that shows the user how many blocks of data have been transmitted as a proportion of how many total blocks of data the firmware image contains.

4.2 Iterations

While I developed the data collection system incrementally as we explored how the system would work, the evolution of the design can be segmented into iterations representing major changes in the functionality and user interface.
4.2.1 Major Iteration 1

The first major iteration had basic functionality and acted as an early proof of concept for collecting Smart Collar data. The user interface consisted of a single page that the user interacted with. Figure 4.1 shows a screenshot of this interface. The user entered the dog’s name which was hard coded into the collar’s firmware. The system used this name to find and pair the collar with the app. We used an MQTT [SC13] connection to IBM Cloud for data transfer. The cloud backend consisted of a Node-Red application [Bla14] connected to a CouchDB [And10] NoSQL database. Using the MQTT connection, the app uploaded data points one at a time.

![Figure 4.1 The initial user interface for the mobile app. The interface is a single page.](image)

This design came with several limitations. Collars were not interchangeable without flashing the firmware, which was cumbersome and required a member of the team to be in physical proximity to the device, limiting who could feasibly act as a tester for the system. The MQTT connection did not allow for large batches of data to be uploaded at a time, so the overall upload speed was slow.

4.2.2 Major Iteration 2

The second iteration underwent major changes to collar communication, cloud connectivity, and the user interface. In this iteration, the collar firmware no longer had dogs’ names hard coded and instead used a universal device name for initial pairing. When no device was paired, the app scanned for nearby devices with the universal name for all Smart Collars. The app saved a Bluetooth address unique to the collar device to pair. By doing this, a user could pair with any collar and connect to their own collar when in the vicinity of multiple Smart Collars (though they would need to have only one collar nearby to pair to the correct collar initially).

The MQTT connection was replaced with an HTTP interface we developed as part of a larger cloud-based canine data processing platform. This allowed much larger amounts of data to be
uploaded and more quickly and for automated preprocessing to be easily integrated into the system. Additionally, data were now stored in a relational time series database, which is a better fit for the types of data being collected in this system.

I redesigned the user interface and broke out the single page into a tabbed page interface consisting of four tabs. Screenshots of these four tabbed pages are shown in Figure 4.2. These four tabs were a Status page, an Info page, a Collar page, and a Cloud page. The Status page displayed the current status of the Bluetooth connection and how many data points were stored on the phone. The Info page displayed the version number and its release date as well as a button to open a web browser to a web page where the user could check for app updates. The Collar page displayed information and controls pertaining to the Smart Collar. Here the user could connect and disconnect the collar, pair or forget the collar, and also showed the current connectivity status. The Cloud page contained information regarding the cloud connection. Here the user entered the identifying tattoo of the dog. This information identified the data being uploaded as belonging to the dog with that particular tattoo.

![Figure 4.2 The user interface for the mobile app during the second major iteration.](image)

Though there were many improvements, this design also came with several limitations. The Status page was lacking and did not provide much information. Having four tabs may have been difficult to navigate and some of the information presented would not have been clear to many users. For example, it may not make sense to a user why the dog tattoo belongs in the Cloud page. In addition, incorrect tattoos were not validated by the system. If a user erroneously entered a typo in the tattoo, all data uploaded while the typo was in place would be uploaded with that incorrect tattoo identifier. Input validation was needed to prevent this.

Finally, some testers noted that it was difficult to know if the system was working as intended. They mentioned that they were not sure that they were both receiving data from the collar correctly and that the app was successfully uploading data. I implemented new features in the user interface
in Iteration 3 to remedy this issue.

### 4.2.3 Major Iteration 3

The third major iteration of the design represents the app in its current state. Several interface changes have taken place and new features have been implemented. Figure 4.3 displays the interface of the third iteration. The Status page displays live collar data as it is received from the Smart Collar. For sound levels and IMU, plots of these data streams show the user a visual representation of the live data. This feature allows the user to verify that the collar is sending data to the smartphone and allows for an informal check that data are reasonable. Tattoo validation ensures that we only collect data from dogs with tattoos that are active in the data collection program and that no data will be incorrectly labeled with an incorrect identifier. Validation takes place server-side and the list of active tattoos can be updated remotely at any time. Tattoo entry is now located on the Collar page, as it makes more sense for the user who is not concerned with the technical details behind the tattoo validation process. The Cloud page now displays some statistics for the user to quickly verify that their system is working as intended. The total number of data points uploaded that day and the last seven days are displayed as well as the total number of data points uploaded overall. The number of data points remaining on the phone to be uploaded is also shown. This readout allows the user to know whether or not their system is uploading correctly. There is no longer an Info page. Instead, version information is shown with the Cloud page. Finally, the fourth page is a new screen called the Events page. This page allows the user to log notable predefined events.

This iteration also introduces three major features to the system: stored memory in the Smart Collar, Over-The-Air updates of collar firmware, and event logging.

#### 4.2.3.1 Stored Memory

The newer Smart Collars have an SD card to temporarily store a buffer of data when the collar is active but the connection is broken, such as when the dog exits the range for a BLE connection. When the connection is re-established, the data are transferred to the phone with a flag denoting that these are “stored memory” data. The app records when disconnects and reconnects occur and uploads the stored memory data and reconnect events to the cloud backend. Later, the original timestamps can be reconstructed using these disconnect and reconnect timestamps.

#### 4.2.3.2 Over-The-Air Updates

The app now also facilitates Over-The-Air (OTA) firmware updates for the hardware. The collar transfers its firmware version number at the beginning of the connection. The app checks this number against the latest firmware version number by connecting to an endpoint in the cloud API. The user is prompted for the update and if they choose to update now, the app initiates the update protocol and a progress bar visually indicates how many bytes have been transmitted.
4.2.3.3 Event Logging

Finally, the app now allows for logging of notable events through the interface. The production version of the app currently has one type of event to test the functionality and user experience of logging. This event is called “Exercise Walk” and corresponds to a raiser taking their dog for a routine walk. There are three buttons (though each button is not always available to the user): “Start”, “End”, and “Cancel”. At first, only the “Start” button is visible. Once the walk begins, this button disappears and “End” and “Cancel” buttons appear. Canceling an exercise walk is made available in case the user accidentally begins a walk when they do not mean to. Once the user presses the “End” button, the walk is considered complete and the starting and ending timestamps are logged in the app’s local database. These timestamp data points are uploaded as normal along with all other data collected by the system. A screenshot of the events page with a walk in progress is shown in Figure 4.3.

Figure 4.3 The user interface for the mobile app during the third (current) major iteration.

4.3 Conclusion

The design of the system underwent numerous changes, both in the user interface and underlying functionality. These changes have been based on internal design decisions and user feedback. Overall, the design evolved to be more open with regard to information presented to the user. Early on, the interface was kept simple and minimal to not distract the user, but this turned out to be counterproductive. Although most of what this system does happens passively with no user input, it is still vital to present status information so that the user can be sure the process is operating smoothly. Another important principle to follow is that the system should minimize distractions so that raisers can focus on their dog. A visual plot of the more frequent data types allows for quick inspection of the incoming data.
Privacy has also been a major concern in the design of the app and collar system. Collecting data from many people in their daily routines introduces privacy risks. This is a major reason why the Smart Collar does not collect sound data beyond simple decibel levels. Even though the smartphone itself is used to gather location data, GPS data points are only collected when the app and Smart Collar maintain an active Bluetooth connection. That way, the raiser only tracks their GPS coordinates when they are actively collecting data from the dog. Entity resolution is accomplished with only the dog’s identifying tattoo. No information identifying the human raiser is uploaded with the collar data.
5.1 Abstract

Training guide dogs is a time and resource intensive process that requires a copious amount of skilled professional and volunteer labor. Even among the best programs, many dogs are released from their training programs. The highest cost in producing guide dogs occurs during the professional training and placement. Selecting dogs for training is a crucial task and guide dog schools can benefit from both an increase in accuracy of their selection and the speed at which dogs can be screened out of the program. We present a method using decision trees to predict the future success or failure of dogs in a guide dog program based on existing data sources from a guide dog school. We achieved 60.6% accuracy on predictions in the test set compared to a failure rate of about 50% at this stage in dog training. Decision trees are easily interpretable; thus evaluators can benefit not only from the model’s prediction but can also examine which features are used to determine success or failure.

5.2 Introduction

Guiding Eyes for the Blind (Guiding Eyes) maintains detailed records of results for assessments from every dog in the program. After puppy raisers house the potential guide dogs in their homes and teach basic obedience and social skills, the dogs return to Guiding Eyes for (potentially) the next phase of training. The next assessment, the In-For-Training (IFT) test, is performed at a Guiding Eyes school when the puppy is about 16 months old, though the specific age varies based on the
individual dog. Following the IFT Test, dogs may enter into training as a guide dog, enter training for other work (detection or emotional support), or be released as pets. The IFT Test results in a large data corpus that provides an opportunity to employ machine learning to assist in decisions about which dogs to invest training resources in. This is especially important at this phase in training as the majority of the cost of training a guide dog comes after this point.

Machine learning can identify subtle, otherwise difficult to discover patterns present in data. By incorporating data from both successful and unsuccessful dogs (“success” here means the dog was successfully placed with a handler and/or became a successful breeding dog), machine learning methods can detect patterns that are reflective of dog performance in ways that can be difficult for humans, even experts, to pick up on. Then, the results can be used to better understand early determinants of future success.

We chose decision trees due to their interpretability, which can be used to find features that are highly predictive of training outcome. The induction of decision trees[Qui86] for classification is a standard method[Wei90] in the realm of supervised machine learning. In this case the dependent variable, or “class”, is a label indicating whether a dog is successful in the program or not. We make predictions with decision trees by examining a set of features, or attributes, about a specific dog in the data set. A decision tree is a graph structure like a flowchart consisting of nodes, which represent a test on a condition, and connections between the nodes called edges, which represent the result of the test. By starting at the root node, or first decision, one can follow the edges to arrive at a leaf node, containing the prediction produced by the tree. The nodes between the root and leaf are internal nodes, which represent tests on specific attributes. For example, an integer attribute $X$ and a node in the tree might test “$X \leq 0.$” When traversing the tree at this node, the expression is evaluated using the attribute $X$ of the current data sample (e.g. the data associated with a single dog) and the edge corresponding to the result is selected.

In machine learning, an algorithm, such as CART[Bre84] (Classification and Regression Trees—what we used), constructs decision trees by analyzing a training set of data samples. Induction of the tree happens incrementally beginning with the root node. Using some measure of impurity or information gain, such as Gini Impurity[Fay92], the features are analyzed for the best possible split according to the measure. A feature and threshold value are chosen that splits the training data accordingly. Then each split is examined, creating new nodes until a stopping condition is met, such as maximum tree depth. Maximum tree depth is a number specified outside of the learning process, a hyperparameter. Other hyperparameters in decision tree learning are splitting criterion (how to measure each feature to select the next split), minimum samples in a split to add a new node, and maximum number of features to consider for a single tree.

Using the induced tree, we can classify a data sample by starting at the root node and testing the features at each node, following the associated path down the tree until a leaf node is reached. Once this happens, we have our prediction of the sample’s class. Because each node explicitly states the attribute and threshold used in the test, this process is easily interpretable by humans allowing non-experts to use decision trees without computer interaction and also understand which features
were relevant in the decision. By viewing the nodes, we can observe which features provide the most information about how to classify any sample. See Figure 5.1 for an example portion of a decision tree with 3 nodes shown. The first feature tested is an item from the Behavior Checklist involving the dog's confidence; the number of samples indicates how many data samples were in the portion of the data that this node is concerned with.

![Figure 5.1 Example portion of decision tree with 3 nodes, each using Gini Impurity splitting criterion.](image)

The goal of this study is to show how machine learning, using existing behavioral data, can be employed to predict which dogs are more likely to succeed as guide dogs. Our objectives were to use decision tree learning with existing data collected by Guiding Eyes, including the Behavioral Checklist and the C-BARQ test, and then to test the developed model using an unseen data set. If successful, our model would allow early predictors of success to be identified and used to increase accuracy of early career sorting and identifying which dogs are most appropriate for training. Ultimately, this would benefit the guide dog schools, the people they assist, and the dogs themselves. It is worth noting that our efforts are not the first to specifically address the problem of predicting success of guide dogs with advanced analytics and machine learning. Yim et al. [Yim17] also used decision trees to predict guide dog success, but we are extending and improving on this work by restricting the time period to enable earlier predictions than previously available; ultimately this allows guide dog schools to filter out candidate dogs earlier in life and conserve valuable resources. In other words, all data used to make predictions in this study are available at the time of the dog's In-For-Training Test.

### 5.3 Data

Data were available for 1561 Labrador retrievers (765 males; 796 females) produced by Guiding Eyes over several years. Of these, 789 (50.5%) were successful in the guide dog program. Our definition of “success” here means that the dog was successfully placed with a person who is blind or visually impaired or possessed exceptional qualities and was selected as a breeding dog. The data used for the machine learning task included basic information about the dog (sex and breed) and the results of two common forms of evaluations, described here.
5.3.1 C-BARQ

The C-BARQ[Seg05] is a 101 item questionnaire used to evaluate a dog’s behavior and temperament. This questionnaire was developed by Serpell and Hsu, and has been validated for use in dogs[Hsu03]. It is commonly used in canine research [MP08; Ser05], and comprises data that many guide dogs schools, in addition to Guiding Eyes, already collect[Duf12]. The data we used from this questionnaire was collected when dogs were 12 months of age.

5.3.2 Behavior Checklist

Also developed by Serpell, the Behavior Checklist is a scoring system where observers can score up to 52 aspects of behavior such as excitability, fearful behavior (such as traffic or noise fear), emotional response to potentially stressful situations, and body sensitivity. Like the C-BARQ questionnaire, the Behavior Checklist is a commonly used scoring system for scoring the behavior of potential guide dogs, which makes the test a good source of data if we want models that are useful for any guide dog school. The Behavior Checklist we utilized was collected at the In-For-Training evaluation.

Combined with the basic information about the dog (sex and a categorical breed code), these information sources produced 197 individual features available for a model to use. We tested these features for correlation with a successful outcome. The purpose of this test was to derive a restricted set of significant features for testing. Twenty-seven features had statistically significant ($p < 0.05$) correlation with final success. These significant correlations in the 12 month C-BARQ and Behavior Checklist data are shown in Tables 5.1 and 5.2, respectively. In the C-BARQ questionnaire, the questions referred to in Table 5.1 are as follows:

- Q10: [Displays aggressive behavior] when approached directly by an unfamiliar adult while being walked/exercised on a leash.
- Q11: [Displays aggressive behavior] when approached directly by an unfamiliar child while being walked/exercised on a leash.
- Q12: [Displays aggressive behavior] toward unfamiliar persons approaching the dog while s/he is in your car (at the gas station, for example).
- Q52: [Displays fearful behavior] when having his/her feet towed by a member of the household.
- Q85: Nervous or frightened on stairs.

5.4 Building the Decision Tree Model

Decision trees are relatively easy to understand by humans, which make them a suitable model for our prediction problem in which professionals evaluating dogs may be interested in why a model
chooses a particular prediction. We used the Python library scikit-learn[Ped11], specifically the CART algorithm[6]. We did not require much preprocessing of our data aside from reformatting some information to be usable by the scikit-learn library. For efficiency of the model building process, we removed any features whose values were equal for every dog in the training set. An example of a feature removed in this way was dog breed, as it turned out that only Labrador Retrievers were in this particular data set. To evaluate the quality of the learned models, we divided the data into training and test sets, with a split of 80% in the training set and 20% in the test set. The test set was only used once the hyperparameters were optimized and we had trained the final model. In addition to a hyperparameter search, we also evaluated several sets of features to find the best feature set to use for prediction.

The hyperparameters to the induction algorithm that we optimized were maximum depth, splitting criterion (in scikit-learn, either the Gini Impurity or Information Gain[Fay92]), and minimum
number of samples on which to split an internal node and also on a leaf node. To find the best set of values, we used a simple grid search[Ber12]. This is a straightforward method for finding sets of hyperparameter values that exhaustively searches the combinations of values in a search space by applying the learning algorithm and comparing the results on validation sets. Although more optimal (and faster) approaches exist[Ber12], grid search was sufficient for our small search space. The process did not take more than 5 minutes for any of the chosen feature sets. For each set of possible values, the grid search algorithm further divides the training set into training and validation sets using 3-fold cross validation[Koh95], which calculates a mean accuracy over the three validation sets. We selected the hyperparameters that scored the highest mean accuracy. Finally, we then used those hyperparameter values to fit a tree model on the entire training set. We evaluated the quality of the resulting decision tree on the previously unused test set using three metrics: accuracy, precision, and recall. Precision is the fraction of true positives (successful dogs that were also predicted to be successful) divided by the true and false positives (the number of times the model said the dog would be successful). Recall is the fraction of true positives divided by the number of successful dogs in the test set.

5.5 Results

The grid search found the best set of hyperparameters for the full feature set as shown in Table 5.3. We also tested the effectiveness of just the set of statistically significant features previously shown in Tables 5.1 and 5.2.

Table 5.3 Best hyperparameter set for full feature set

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Depth</td>
<td>5</td>
</tr>
<tr>
<td>Splitting Criterion</td>
<td>Information Gain</td>
</tr>
<tr>
<td>Min Samples (leaf)</td>
<td>6</td>
</tr>
<tr>
<td>Min Samples (internal)</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 5.4 shows the metrics obtained for several subsets of features. Allowing the learning algorithm to choose only from the set of features across all data sources whose correlation with success was statistically significant ($p < 0.05$) resulted in only slightly worse accuracy than the full feature set, but actually outperformed in terms of precision and recall. The full feature set achieved a 60.6% accuracy on unseen data, but since dogs at this point in time have a 50% failure rate this shows promise for the decision tree learning approach. Several subsets achieved recall scores of almost 70%. This indicates the decision tree models were effective at identifying the successful dogs
(though false positives brought down the overall accuracy).

Table 5.4 Best cross-validation scores and test set accuracy, precision, and recall for each feature subset. Note that “breed” was actually removed as every dog in the data set had the same breed code.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Mean Validation Acc</th>
<th>Test Set Acc</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Dog Info (sex, breed)</td>
<td>0.5152</td>
<td>0.4856</td>
<td>0.5161</td>
<td>0.4819</td>
</tr>
<tr>
<td>BCL</td>
<td>0.638</td>
<td><strong>0.608</strong></td>
<td>0.62</td>
<td><strong>0.698</strong></td>
</tr>
<tr>
<td>C-BARQ</td>
<td>0.5521</td>
<td>0.5144</td>
<td>0.5507</td>
<td>0.4578</td>
</tr>
<tr>
<td>Info, BCL, C-BARQ</td>
<td>0.635</td>
<td>0.606</td>
<td>0.620</td>
<td>0.684</td>
</tr>
<tr>
<td>Statistically Significant Features</td>
<td><strong>0.648</strong></td>
<td>0.596</td>
<td><strong>0.698</strong></td>
<td><strong>0.698</strong></td>
</tr>
</tbody>
</table>

By using a decision tree regressor instead of a classifier, we were also able to learn decision trees that output a number between 0 and 1 and can be interpreted as a probabilistic prediction. To evaluate the quality of these predictions, we calculated the mean absolute value of the error, where error is the difference between the predicted output and the actual class. We treated success as the real number 1.0 and failure as 0. The results of this are shown in Table 5.5.

Table 5.5 Best cross-validation scores and test set accuracy, precision, and recall for each feature subset. Note that “breed” was actually removed as every dog in the data set had the same breed code.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Mean Absolute Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Dog Info (sex, breed)</td>
<td>0.50</td>
</tr>
<tr>
<td>BCL</td>
<td><strong>0.434</strong></td>
</tr>
<tr>
<td>C-BARQ</td>
<td>0.488</td>
</tr>
<tr>
<td>Info, BCL, C-BARQ</td>
<td>0.472</td>
</tr>
<tr>
<td>Statistically Significant Features</td>
<td>0.464</td>
</tr>
</tbody>
</table>

5.6 Discussion

Using decision tree learning, we were able to successfully predict the future success or failure of potential guide dogs with 60.8% accuracy, 62% precision, and 68.4% recall on the full feature set. Since the eventual failure rate of dogs at this time point (the IFT test) is close to 50%, this is suggestive of an opportunity for cost savings for guide dog programs, especially if this accuracy score can
be improved upon. Incorporating larger data sets from various guide dog schools could improve this accuracy. By examining sources of behavioral data that already exist, decision tree classifiers can extract patterns for guide dog schools to utilize in evaluating dogs for training. Including only features with significant correlations to success code was much faster (about 4 minutes and 4 seconds for the full feature set; 1 minute and 9.5 seconds for the significant features), but this is unlikely to be an issue in this domain.

Understanding what influences the ability of dogs to work in various roles, behaviorally or otherwise, has been a significant and often multidisciplinary area of research. There are studies that analyze what factors can tell us about the future behavior of dogs, such as the work by Weiss\cite{Wei02} in which behavioral tests were used to aid in the selection of shelter dogs of service dog training. The behavioral tests include assessments of reaction to touch, walking, fetching, and interacting with strangers. They then used regression analysis to predict the dog’s performance over 5 weeks of training. Slabbert and Odendaal found in a two year longitudinal study several behavioral tests whose correlation with future success as police dogs was statistically significant. Some of these tests were able to be administered as young as 8 weeks of age\cite{Sla99}. Byrne et al.\cite{Byr18} used instrumented dog toys with sensors to develop a logistic model tree for classifying the eventual outcome of dogs in a service dog organization (Canine Companions for Independence) and were able to achieve an accuracy of 87.5%, which by their calculations could save $70,000 by identifying dogs that will likely fail the program.

Arata, Momozawa, and Takeuchi\cite{Ara10} studied behavioral factors that may predict success or failure in guide dog programs with the use of a questionnaire. They were able to assess that distraction, sensitivity, and docility are important behavioral factors that can predict a dog’s future success. They found that distraction could predict a dog’s success with 80.6% accuracy. Batt et al.\cite{Bat08} evaluated the ability of several behavioral tests to predict guide dog success and found several behavioral tests and physical characteristics that were predictive of success. Their results obtained at 14 months of age were more accurate than their results obtained at either 6 months or when the dog completed training. Using the C-BARQ questionnaire, Duffy and Serpell found that tests administered at 6 and 12 months of age were useful for predicting success in guide and service dog programs, but their predictive value varied across organization\cite{Duf12}.

A benefit of our approach is that we used data sets that are already commonly collected and did not implement our own tests. Our method is also able to provide a “line of reasoning” for its predictions by producing a path from root node to leaf node so that evaluators can know why a prediction was made. Probabilistic predictions may also prove useful by identifying dogs who may have high (but not exceptionally high) probability of success. This is useful for at least two reasons. Guide dog schools will be ultimately using their own discretion on the decision whether to train each dog and a probabilistic machine learning prediction may be more helpful in that regard. A second reason is these dogs may have specific issues that need to be addressed and their success could hinge on those issues.
5.7 Future Work

There are many avenues to explore in using machine learning to predict the future capabilities of guide dogs. Furthermore, Guiding Eyes collects a variety of text-based data associated with each dog at certain steps in its development such as training notes and comments recorded for various tests. Convolutional neural networks (CNNs) are well-suited to incorporating this data into a machine learning model alongside numeric features. Neural networks may be able to improve the accuracy but in contrast with decision trees can be very difficult to extract meaning from. In addition, IBM Watson[Hi12] services for natural language processing may also be able to augment machine learning prediction of dog outcomes and also to tackle similar problems, such as optimal pairings of puppies with raisers.
6.1 Abstract

Evaluating potential guide dogs is crucial for guide dog schools as raising and training is an expensive process. During adolescence, volunteers raise dogs in training away from guide dog schools and expose them to a variety of stimuli and teach them obedience skills. However, no objective data exists about the dog's behavior and environment during this period, usually lasting several months to a year. We developed an Internet-of-Things sensor-equipped collar to quantify aspects of the dogs' behavior and their environments during this stage. Raisers collect data from the collar using a smartphone app which in turn uploads data to a central processing pipeline. We present an overview of the system and an evaluation showing how we can learn meaningful information about a dog's environment and physical activities while away from the school for months on end, ideally to help predict which dogs will be successful in training.

6.2 Introduction

To collect objective data, we have designed and implemented a system comprised of a custom Bluetooth-enabled Internet of Things Smart Collar and accompanying mobile application that
collects data with on-board sensors. Data is transmitted into a cloud-hosted processing and storage pipeline. Supervised machine learning will be possible after several dogs have worn the collar and succeeded or failed the program. The system is usable by non-experts for data collection and scalable to hundreds of dogs at once. The Smart Collar’s sensors consist of IMU (accelerometer), ambient light levels, sound levels (decibel level only), temperature, humidity, and barometric pressure, but the raiser’s smartphone also provides GPS data to extract location information, which will differentiate between common locations the dog visits, such as home, parks, and grocery stores. To protect privacy, we do not collect information identifying raisers and instead use a unique tattoo identifier for entity resolution so that no one can match people to collar data.

We present a demonstration using collar data to extract meaningful information about behavior from raw sensor data. We observed two dogs wearing the Smart Collar in four behavioral states and two environment types and analyzed statistical differences in these states. We also discuss the deployment of this system in partnership with Guiding Eyes for the Blind, a U.S. guide dog school. The system is used to better understand the development of potential guide dogs when they are away from the school so they can better select dogs for formal training.

The work described in this chapter was published in the proceedings of the Sixth International Conference on Animal-Computer Interaction in November 2019. The study described here was carried out under Institutional Animal Care and Use Committee (IACUC) approval.

6.3 Related Works

Predicting canine behavior, specifically for future suitability in a service role, is studied from various perspectives including expert observation, questionnaires, and sensor systems. Studies involving expert observation have evaluated maternal style and early life experiences as important traits to consider [Bra17a; Bra17b]. Since early experiences shape future suitability, producing service animals may benefit from behavioral prediction. Furthermore, handler questionnaires are often used as a source of data for behavioral analysis. Batt et al. [Bat08] found several behavioral tests and physical characteristics that were predictive of success. Their results obtained at 14 months of age were the most accurate among the tested ages.

The C-BARQ [Seg05] is a 101 item questionnaire for evaluating a dog’s behavior and temperament, validated for use in dogs [Hsu03]. Commonly used in canine research [MP08; Ser05], it comprises data that many guide dogs schools, in addition to Guiding Eyes, already collect. Using the C-BARQ, Duffy and Serpell found that tests administered at 6 and 12 months of age were useful for predicting success in guide and service dog programs, but their predictive value varied across organization [Duf12]. For example, each increase by one point in the C-BARQ score for “pulls excessively hard on leash” was associated with a 1.4% increase in the probability that a dog would fail the program. Also developed by Serpell, the Behavior Checklist (BCL) involving observers scoring up to 52 aspects of behavior such as excitability, fearful behavior (such as noise fear), and body sensitivity. Like the C-BARQ questionnaire, the Behavior Checklist is a commonly used scoring system for scoring the
behavior of potential guide dogs. Arata, Momozawa, and Takeuchi [Ara10] assessed that distraction, sensitivity, and docility are important behavioral factors that can predict a dog’s future success based on the BCL, and that distraction predicted success with 80.6% accuracy. However, this took place in the formal training process, about three months before final evaluation for guide work.

Another method of predicting service dog suitability is to use more objective data provided by sensors. Byrne et al. [Byr18] used instrumented dog toys with sensors to develop a logistic model tree to classify outcomes in a service dog organization (Canine Companions for Independence) and were able to achieve accuracy of 87.5%. They calculated that this could save $70,000 by identifying dogs that will likely fail the program. This work, however, targeted dogs entering advanced training; we, however, focus on dogs earlier in the process in an uncontrolled setting. A study by Den et al. [Uij17] used a collar-mounted triaxial accelerometer to detect several behavioral states in dogs (walk, trot, canter/gallop, sleep, static/inactive, eat, drink, and headshake). They achieved impressive results, with precision in the 93-100% range and recall in the 89-98% range. However, this was conducted by experts in a controlled environment. Objective data collection systems have been used alongside expert behavioral annotators [Mea17]. In contrast with the IoT system in this work, the system in [Mea17] collects data using a laptop with a vest worn during a controlled evaluation and does not sense environmental data.

Citizen science, collecting data from the public without formal scientific training, has been recently used to study canine behavior and collect sample sizes impractically large in a lab setting. For example, in Dognition [Ste15], dog owners fill out a questionnaire and perform various cognitive exercises and report results on a website. Hare found that this method is useful for gathering large amounts of data from citizen scientists and found little evidence that the citizens manipulated results [Ste15]. Hecht and Rice investigated benefits and technical issues, such as data quality, faced by citizen science for canine research [Hec15]. We similarly employ volunteer puppy raisers who need no formal training to collect objective data in an uncontrolled environment. Van der Linden and al describe privacy concerns in animal-based wearable computing, such as clarity to the user about what is collected, what is considered personal information, and how it is handled[Lin19]. Zamanksy and van der Linden found that raisers are accepting of data sharing with guide dog schools[Zam18], but that we should not ignore privacy. In our case, we only identify data using dogs’ unique tattoo identifiers.

6.4 The IoT and Analytics System

To collect accurate sensor data from as many raisers as possible, we need a system that requires minimal effort from the volunteer raisers to minimize impact on their normal routines. The system consists of a Bluetooth-enabled IoT “Smart Collar” with various sensors, a simple mobile application to collect data from the collar, and a cloud back-end to receive and process the data for analysis. Using this, we can collect measurements from hundreds of raisers located anywhere. We refined the software for reliability and usability using our own observations and feedback from the raisers.
themselves to enable larger scales. I did not design or develop the Smart Collar hardware itself; more detailed information on this component of the system can be found in Chapter 3.

6.4.1 Mobile App

We developed a mobile app on iOS to collect and upload data to the back-end system with a simple interface (see Fig. 6.1). Each collar is essentially interchangeable with any other barring a unique ID. The app uses this unique ID to pair with the raiser’s Smart Collar. Once paired, the app remembers that collar ID so that two raisers with their dogs will not have any problems connecting to their individual collar. Once data collection begins, the raiser does not need to interact with the app. Minimizing impact on the raiser and dog’s routine is important so that the data collected will not be skewed by the presence of the system. The app passively collects data until the raiser disconnects or turns off the collar.

In addition to the sensor data collected by the collar itself, the app collects latitude and longitude information using the phone’s built-in GPS. However, this is only active when currently connected to the collar. We are only interested in data about the dog and unnecessarily collecting GPS data would increase privacy risks to the raiser. GPS data is recorded whenever the device moves 5 meters or more from the previous measurement. Position between recorded measurements can be interpolated. GPS data will be used in this system to differentiate between types of environments the dog is exposed to and to characterize its familiarity with various locations. For example, a dog with typical behavior in rural areas or parks may experience inordinate stress in urban areas, impacting its behavior.

The app stores these data until it can be uploaded to our backend system. Uploads take place in batches when a Wi-Fi network is available and stop if the Wi-Fi connection is lost. This prevents us from inadvertently impacting a user’s mobile data plan. Along with each batch of data, the app includes the dog’s unique tattoo (as entered by the raiser) so that the backend can identify which dog the data belong to. This is the only identifier we have for entity resolution and allows us to match sensor data to dogs, but we cannot match data to a human raiser. In fact, since Guiding Eyes can only match dogs with people but does not have the sensor data, no one involved can match collar data with individual people.

The collar also stores data when the connection is broken using the SD card. The app stores these data points separately along with the timestamps associated with disconnect and reconnect events. The reason we need to store the timestamp of these events is that the timestamps on stored memory data correspond to when the data is transmitted, not stored (the collar does not have the capability to store the current time). The disconnect/reconnect pairs are then used to reconstruct as closely as possible the original timestamp of the stored data.

Finally, the mobile app also facilitates Over-The-Air (OTA) firmware updates of the Smart Collar hardware. An API endpoint on the backend system contains the version number of the latest collar firmware available. If an update is needed, the app downloads the firmware binary and the user is prompted for an update. If they accept, the app initiates the OTA update protocol with the collar. A
progress bar visually indicates how many bytes have been transmitted.

Figure 6.1 Mobile app interface for collecting Smart Collar data. Live data is shown to the user. In this example, IMU activity is shown.

6.4.2 Cloud Back-end

For receiving data, we developed a cloud-hosted system for collecting and processing various canine data. This work focuses on the collar data portion, although it is used to collect puppy test data as well, as described by Mealin et al.[Mea17; Mea19]. The system architecture is designed as a cluster of containers hosted on a Kubernetes cluster[Goo14] on IBM Cloud. Figure 6.2 shows an overview of the architecture. A container running a Python Flask application exposes an API for communication from the mobile app. Batches of data are uploaded over a secure TLS connection. If the tattoo ID submitted with the batch matches a list of active tattoos, the batch is accepted (otherwise an error is returned to the app). These batches are then submitted to a Gearman[Fit09] job server for processing and ultimately stored in a relational time series database[Tim15].

Once a batch of data is imported, we use Gearman to schedule preprocessing, which varies by data type. Each of these Gearman workers exist as Kubernetes pods and read from and write to the time series database. This decomposes the preprocessing into isolated pieces allowing for further scalability. The only data source that is not imported in this manner is data provided manually by Guiding Eyes, which is updated only every 3 months. This data consists of historical data on each dog, scores for assessments such as the BCL and C-BARQ, and training notes (unstructured text data).
Figure 6.2 An overview of the cloud system architecture hosted on IBM Cloud using Kubernetes.

Figure 6.3 The dogs wearing the Smart Collar.

6.5 Methods

As a proof of concept in using the system to answer questions about behavior, we performed tests with two Labrador retrievers. Labrador retrievers are a typical guide dog breed, especially at Guiding Eyes. The subjects of this study were well-trained dogs but not part of any service dog program. Each dog wore the Smart Collar, shown in Figure 6.3, while their owner gave them cues to exhibit each of the four predefined behaviors. We gave the dogs treats for completing the tasks. All procedures were completed under IACUC approval.

Each behavior lasted at least one minute to ensure an adequate quantity of data for statistical analysis. For the Sitting behavior, the dog sits indoors. During this behavior, we expect to see very little IMU activity in the data stream. For the Walking Indoors behavior, the dog walks at heel in a flat indoor environment. Activity should be visible in the IMU data now that the dog is moving and should contrast with the Walking Outdoors behavior in terms of environmental sensors: the indoor environment was an air-conditioned home during the afternoon. For the Walking Outdoors behavior, the dog walks at heel in an outdoor environment, the goal being to differentiate between this behavior and the Walking Indoors behavior mentioned previously. The outdoor environment was a warm, sunny day at about 27°C. For the Stairs behavior, the dog walks with the handler from the top of a set of stairs to the bottom of the stairs, then returns to the top. This repeats until about a minute of data has been collected.

The goal of our pilot study was to investigate whether data collected by our system contains
useful information about the behavior of dogs and their environment. Prediction is a larger goal but outside the scope of this particular study. Instead, we show here that some low-level patterns can be extracted from the data to provide evidence for feasibility of machine learning for this data set. For statistical analysis, we used the Mann-Whitney U (MW-U) test [Het10], which does not require normal distributions, to compare the means between data sets and the Levene’s Test [Bro74] to compare the variances between data sets. For environmental data, we tested the differences in mean, aggregating data for each dog. We tested the differences in variance for IMU data sets to examine differences in activities.

We first used a simple moving average filter (window size 670 ms or 10 samples) over each IMU signal. The approximate magnitude of acceleration has units of \( \frac{1}{1000} \) of a g (9.807 \( m/s^2 \)) or “millig”. We subtracted this by 1000 to approximate subtracting out the gravity component (1 g). This is just an approximation because the IMU is rarely pointing one of its axes directly toward the ground. We calculated overall magnitude of activity from the triaxial accelerometer data using the equation 

\[ mag = \sqrt{x^2 + y^2 + z^2} - 1000. \]

6.6 Results and Discussion

This section discusses the results of applying statistical tests to measure the differences in both the four activities and the two types of environments in the experiment.

6.6.0.1 Distinguishing between Environments

Table 6.1 shows the result of the MW-U test for a few data types. As expected, light was much more intense outdoors \((p < 0.01)\) and the Sun dominates the signal. Indoors and outdoors also resulted in pressure differences \((p < 0.01)\). Sound levels, measured in a windowed voltage difference convertible to decibels, were higher for the activities recorded inside. Temperature and humidity were not tested—the low sampling rate would result in a very small sample size. Despite this, we can report that for the few samples that were produced there were clear differences in these two sensor types between the environments.

There is a stark contrast between sensor readings of the indoor and outdoor environments; in fact the data here are linearly separable. This demonstrates that we can distinguish between various environments, though more nuanced differences will of course not be as simple as an obvious jump in temperature. These data can be further augmented in at least two ways: using an external service to extract location information from GPS coordinates and allowing raisers to log location visits in the app.

6.6.0.2 Distinguishing between Dog Activities

Once we identify environmental patterns in the data, we can examine how a dog reacts to environments. The IMU sensor records physical movement of the dog. The sound sensor detects
Table 6.1 MW-U Test between Indoor and Outdoor Walking Behaviors. All three tests statistically significant with $p < 0.001$.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>State</th>
<th>Mean(SD)</th>
<th>U-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light</td>
<td>Walk-Indoor, n=498</td>
<td>0.0799(0.111)</td>
<td>0.0 ($p &lt; 0.01$)</td>
</tr>
<tr>
<td></td>
<td>Walk-Outdoor, n=145</td>
<td>0.822(0.002)</td>
<td></td>
</tr>
<tr>
<td>Pressure</td>
<td>Walk-Indoor, n=66</td>
<td>354(121)</td>
<td>378 ($p &lt; 0.01$)</td>
</tr>
<tr>
<td></td>
<td>Walk-Outdoor, n=34</td>
<td>139(35.3)</td>
<td></td>
</tr>
<tr>
<td>Sound</td>
<td>Walk-Indoor, n=120225</td>
<td>0.0916(0.134)</td>
<td>4.68*10^9 ($p &lt; 0.01$)</td>
</tr>
<tr>
<td></td>
<td>Walk-Outdoor, n=89344</td>
<td>0.0780(0.007)</td>
<td></td>
</tr>
</tbody>
</table>

sound levels; some of this the dog contributes (barking and movement) and the rest originates in its environment. Here we focus on the dog’s movement (IMU).

Figure 6.4 IMU acceleration magnitude for each dog in three behaviors: Sitting, Walking, and Stairs. Notable events are labeled with numbers and explained in the text.

Figure 6.4 shows activity plots for Dog 1 and Dog 2 during the indoor behaviors (sitting, walking, and ascending/descending stairs). In each plot, notable events are labeled with numbers. In Figure 6.4 Dog 1, the label “1” shows when the dog was released from its sitting position and jumped at the end of the session. The rest of the labels in the Stairs behavior show each time Dog 1 reached the bottom or top of the stairs. This is also shown for Dog 2, except the label “1” here denotes when Dog 2 was initially cued to sit from a standing position, “2” and “3” denote events in which Dog 2 jumped on a bed, and the rest show each time Dog 2 reached the bottom or top of the stairs. The Walk behavior is noisier than the Sit state for both dogs as both produced frequent small accelerations while walking. The difference is more pronounced for Dog 2 in Figure 6.4, though this dog produced smaller jumps in acceleration while traversing the stairs. Finally, the Stairs behavior differs from
the others, with frequent large changes in acceleration magnitude and the largest acceleration magnitude overall.

**Table 6.2** Variance of IMU Acceleration Magnitude in Three Activities

<table>
<thead>
<tr>
<th>State</th>
<th>Dog 1 Var (mG²)</th>
<th>Dog 2 Var (mG²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sit</td>
<td>1952</td>
<td>2775</td>
</tr>
<tr>
<td>Walk</td>
<td>73854</td>
<td>10016</td>
</tr>
<tr>
<td>Stairs</td>
<td>1508887</td>
<td>49714</td>
</tr>
</tbody>
</table>

Table 6.3 shows the Levene's Test results. For each dog, each of the three indoor behaviors have significantly different variances ($p < 0.05$), as Figure 6.4 suggests. Variance clearly increases between the Sit, Walk, and Stairs behaviors, shown in Table 6.2. The dogs' acceleration variance was different but changed consistently with activity. The fact that the variances of these distributions is different does not necessarily pose a problem for creating predictive models. For any dog, we are interested in discovering patterns of behavior over time.

**Table 6.3** Levene's Test between Behavior States.

<table>
<thead>
<tr>
<th>States tested</th>
<th>Test Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dog 1 Sit vs. Stairs</td>
<td>773</td>
<td>1.9E-162</td>
</tr>
<tr>
<td>Dog 1 Sit vs. Walk</td>
<td>919</td>
<td>9.1E-199</td>
</tr>
<tr>
<td>Dog 1 Walk vs. Stairs</td>
<td>253</td>
<td>8.8E-57</td>
</tr>
<tr>
<td>Dog 2 Sit vs. Walk</td>
<td>737</td>
<td>2.1E-153</td>
</tr>
<tr>
<td>Dog 2 Walk vs. Stairs</td>
<td>511</td>
<td>1.4E-108</td>
</tr>
<tr>
<td>Dog 2 Walk vs. Stairs</td>
<td>135</td>
<td>6.5E-31</td>
</tr>
</tbody>
</table>

Our results show the differences in *patterns* of data. As we collect more data, we will use machine learning to ultimately to screen out unsuccessful dogs earlier. Building prediction models using real outcomes to assess guide dog suitability will save time and financial resources and improve the welfare of dogs who will ultimately not succeed, finding them more suitable roles.

### 6.7 Conclusion

There is no objective data regarding behavioral development of candidate guide dogs in the months-long raising stage. To analyze these dogs’ environments and behavioral responses, we developed a scalable Smart Collar system for collecting data from raisers. This is deployed with volunteer
puppy raisers for the dog to wear and is designed to preserve user privacy. We presented results of a controlled demonstration to show how we can extract meaningful information about the dog’s environment and behavioral response to that environment. With statistical tests, we saw a clear difference in the environmental data streams between two typical settings: indoors and outdoors during the day. In the first test, we tested the means of the environmental data. In the second, we found significant differences in the variances of the IMU data for three behaviors (separated by dog). This shows that meaningful information exists within the data, but prediction tasks will come later as the data set grows. Using outcomes as labels, we will train a model for prediction of future dogs based on collar data. Predictive models, especially those suited to time series data such as Long Short-Term Memory (LSTM) neural networks may be effective at prediction tasks with these data. Deep learning approaches such as recurrent neural networks will need to learn low-level features in the data set, possibly such as the patterns we analyzed: basic differences in environment type and overall activity level. Extracting higher level features and examining data at differing resolutions will improve the effectiveness of our models.

Raisers expose the dogs to a multitude of new environments and new people, preparing them for guide work. To more easily analyze these and other potentially important events, we will in an upcoming version of the application provide functionality to tag portions of data with some predefined label e.g. “at home”, “met a new person”, or “at the park”. This will provide contextual information to data sets and allow us to create better models of a dog’s behavioral responses and development over time. However, this will not come without challenges. The system should require minimal real-time input from raisers. Thus, annotations would need to be reported on their own time (low-level labels like barking events would be infeasible). We are currently in the process of finalizing the system and scaling up to our next goal of hundreds of raisers. We designed the system for this and larger scales using modern cloud technologies. Once we have more data, we can then build predictive models to assist guide dog schools with drawing conclusions about hundreds of dogs who may be located anywhere at a given time. This will enable guide dog schools, largely nonprofits who have scarce resources, to gather crucial information and assistance in finding which dogs show the most promise for further training with the goal of increasing the number of successful handler-dog pairs.
7 ANALYZING LIVE SENSOR DATA FROM RAISERS

7.1 Abstract

Training successful guide dogs is time and resource intensive, requiring copious professional and volunteer labor. Even among the best programs, dogs are released with attrition rates commonly at 50%. Increasing success rates enables non-profits to meet growing demand for dogs and optimize resources. Selecting dogs for training is a crucial task; guide dog schools can benefit from both better selection accuracy and earlier prediction. We present a system aimed at improving analysis and selection of which dogs merit investment of resources using custom sensing hardware and a cloud-hosted data processing platform. To improve behavioral analysis at early stages, this chapter presents an overview of an IoT-enabled “Smart Collar” system to gather data from puppies while being raised by volunteers prior to training. Our goal is to identify both puppies at risk and environmental influences on success as guide dogs. This chapter demonstrates the initial results of using live collar data from volunteer raisers for statistical analysis.

7.2 Introduction

The reported costs to breed, raise, train, place and provide lifetime support for a service dog generally costs at least $50,000 per dog [Ber17; Bli]. Unfortunately, many dogs that enter training do not succeed as guide dogs, representing lost resources and lost opportunities to serve more individuals needing these dogs. Organizations expend great effort using behavioral and health data to
influence decisions on which individuals are selected as replacement breeders and career sorting young dogs for their estimated potential [Mea17]. Behavioral assessments are conducted at several points of development including neonatal, puppy-raising and training phases, and again following placement. Guiding Eyes For the Blind, a partner in this research program, scores observations from most assessments using the Behavior Checklist. While some associations with success have been found, particularly at older ages, the accuracy of prediction at puppy testing is about 35% because behavior is shaped over time by the interplay of the environmental experiences and genetics. We aim to improve the selection accuracy of candidate guide dogs using wearable sensing technologies integrated into existing practices by Guiding Eyes, employing technologies such as Kubernetes [Goo14] and IBM Cloud to enable this at scale.

Guiding Eyes for the Blind is a large, innovative not-for-profit guide dog organization in the United States. Guiding Eyes maintains records of the previously mentioned assessments from all of their dogs. In the program, socialization begins at age 1 week and continues as the puppies mature. Observations of socialization events are recorded by staff and a formal puppy test is conducted at around 7.5 weeks of age. About 11% of puppies tested are provided to other organizations that train service dogs, 16% are adopted as pets and the remaining 73% are raised as potential guide dogs in the Guiding Eyes program. These 73% are raised by volunteer puppy raisers, who raise them in their homes, teach basic obedience and social skills and provide exposure to various environments and experiences. In this phase, we employ sensor-equipped “Smart Collars” to collect behavioral and environmental data to analyze how the dog’s behavior is shaped over this crucial period.

This chapter describes work in analyzing live collar data from actual raisers using an early version of the Smart Collar system. This work was submitted to a special issue of the International Journal of Cloud Computing (IJCC) on June 2, 2019. It has been approved and is pending publication. The study described here was carried out under Institutional Animal Care and Use Committee (IACUC) approval.

7.3 Related Works

Predicting canine behavior is an active area of research. Prediction of suitability in a service role is often approached from a few different perspectives: expert observation, questionnaires, sensor systems, and more recently, large-scale data collection via citizen science.

7.3.1 Expert Observation

Expert observation and analysis are a straightforward approach to studying canine behavior and suitability for service work. Studies have evaluated maternal style and early life experiences as important traits to consider [Bra17a; Bra17b], demonstrating how early experiences shape future behavior and thus suitability in service roles. Batt et al. [Bat08] found several behavioral tests and physical characteristics that were predictive of success. They found that prediction at 14 months of age was more accurate than their results obtained at 6 months as the dogs had more time to develop,
which of course comes with the trade off that the prediction happens later.

### 7.3.2 Questionnaires

Handler questionnaires are often used as a source of data for behavioral analysis. The C-BARQ [Seg05] is a 101 item questionnaire for evaluating behavior and temperament, which has been validated for canine use [Hsu03]. The C-BARQ is commonly used in canine research [MP08; Ser05] and many guide dog schools employ it, including Guiding Eyes for the Blind. Using the C-BARQ questionnaire, Duffy and Serpell found that tests administered at 6 and 12 months of age were useful for predicting success in guide and service dog programs, but their predictive value varied across organization [Duf12]. For example, each increase by one point in the C-BARQ score for “pulls excessively hard on leash” was associated with a 1.4% increase in the probability that a dog would fail the program. Also developed by Serpell, the Behavior Checklist (BCL) is a scoring system where observers can score up to 52 aspects of behavior such as excitability, fearful behavior (such as traffic or noise fear), and body sensitivity. Like the C-BARQ questionnaire, the Behavior Checklist is a commonly used scoring system for scoring the behavior of potential guide dogs. Arata, Momozawa, and Takeuchi [Ara10] assessed that distraction, sensitivity, and docility are important behavioral factors that can predict a dog’s future success based on the BCL, and that distraction predicted success with 80.6% accuracy. However, this took place in the formal training process, about three months before final evaluation for guide work.

### 7.3.3 Sensor Systems

More objective data from sources such as physiological or environmental sensors are another source to analyze for behavioral prediction. Researchers have used on-body sensors and off-body sensors for canine behavior. Byrne et al. [Byr18] used instrumented dog toys with sensors to develop a logistic model tree for classifying the eventual outcome of dogs in a service dog organization (Canine Companions for Independence) and were able to achieve an accuracy of 87.5%, which by their estimates could save $70,000 by identifying dogs that will likely fail the training program. Den et al. [Uij17] used a collar-mounted triaxial accelerometer to detect several behavioral states in dogs. The states they studied were walk, trot, canter/gallop, sleep, static/inactive, eat, drink, and headshake. They achieved impressive results for most of these states, achieving a precision in the 93-100% range and recall in the 89-98% range. However, it should be noted that this experiment was conducted by experts in a controlled environment. Other studies have also used inertial sensors to analyze behavior [Ger13] and also incorporated machine learning with inertial sensors to estimate both posture and behaviors [Bru13a]. Researchers have used on-body wearable sensors to estimate dog posture [Bru12; Bru13b; Win15], while others have used an off-body approach with devices like the Microsoft Kinect [Pis14; Mea16].
7.3.4 Citizen Science

Researchers have also recently used citizen science, collecting data from members of the public, typically without formal scientific training, in the domain of canine behavior. Researchers gather dog behavior data from sample sizes that are normally too large to be practical in a lab setting. Hecht and Rice investigated the benefits and technical issues faced by the use of citizen science in doing canine research [Hec15]. Data quality is, however, an issue that needs to be addressed. An example of citizen science for canine research is Dognition [Ste15]. Dognition is a project with the goal of gathering behavioral data from dog owners at a large scale. Dog owners fill out a questionnaire and perform various cognitive exercises and then report these results to Dognition using a website. Hare found that this method is useful for collecting large amounts of data from citizen scientists and found little evidence that the citizens manipulated results [Ste15].

7.4 Collar Data Analysis

For analysis of data from the Smart Collar systems, we used live data from 7 of the raisers who were assisting us in testing the system. As we do not have final outcomes yet for dogs wearing the Smart Collars, we cannot predict them using supervised machine learning. However, the data contains patterns that can be analyzed using statistics and unsupervised learning. As an example, GPS data, in the form of latitude and longitude values, allow us to pinpoint geographically portions of the dog behavior and environmental data. This is useful because puppies in the raising phase are often introduced to a variety of environments, and their response to these environments may provide insight into their behavior in terms of service suitability. Similar geographic locations should be considered together, so here we used clustering on location data to group segments of the data stream into clusters using the scikit-learn library [Ped11] and Watson Studio on IBM Cloud. We chose two methods for this, K-Means, a simple but fast centroid-based clustering algorithm, and DBSCAN (Density-Based Spatial Clustering of Applications with Noise). First, for each dog in the available data, we computed the K-Means clusters for various values of $k$. With this, we can estimate the optimal number for $k$ visually by plotting Sum of Squared Error as $k$ increases. All of the data sets behaved similarly, with an optimal value of 2. A representative plot is shown in Figure 7.1.

To compare the two algorithms, we used the Silhouette Coefficient [Rou87]. The Silhouette Coefficient compares the mean within-cluster distance of a sample with the distance to the nearest cluster (that the sample is not a part of). It is calculated as $\frac{b-a}{\max(a,b)}$ where $a$ is the mean within-cluster distance for a point and $b$ is the mean nearest-cluster distance for that point. As K-Means computes centroids for predefined number of clusters, we can also compute the within-cluster Sum of Squared Error. Results for these metrics is shown in Table 7.1.

For most data sets, the two algorithms performed comparably in terms of the Silhouette Coefficient (with K-Means usually somewhat better), though K-Means strongly outperformed on Dog 3. However, as these data sets grow, $k = 2$ may not always be appropriate for segmenting this location data. It may be true that with more complete data, DBSCAN will outperform the simple K-Means
Figure 7.1 Representative example of SSE as $k$ increases for a single dog.

Table 7.1 Clustering Performance (Silhouette Score and Within-Cluster SSE) DBSCAN and K-Means Algorithms on Collar Location Data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60820</td>
<td>0.02480</td>
<td>1</td>
<td>0.997</td>
</tr>
<tr>
<td>2</td>
<td>18785</td>
<td>0.4150</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>3</td>
<td>52554</td>
<td>113.8</td>
<td>0.913</td>
<td>0.519</td>
</tr>
<tr>
<td>4</td>
<td>56996</td>
<td>0.7988</td>
<td>0.994</td>
<td>0.976</td>
</tr>
<tr>
<td>5</td>
<td>5357</td>
<td>0.29486</td>
<td>0.992</td>
<td>0.992</td>
</tr>
<tr>
<td>6</td>
<td>56</td>
<td>1.710E-07</td>
<td>0.979</td>
<td>0.979</td>
</tr>
<tr>
<td>7</td>
<td>23188</td>
<td>6.680</td>
<td>0.972</td>
<td>0.950</td>
</tr>
</tbody>
</table>

algorithm on this task.

With these results and a known optimal number of clusters ($k = 2$) for this data set, we can use the clusters computed by K-Means to analyze the dog’s behavioral and environmental data during these time segments. As these data come from a testing phase of the Smart Collar systems, sufficient data were not necessarily available for all sensor types for all of the dogs. Only available data sets are shown.

We examined how each sensor’s resulting data sets differed between the two clusters in terms of both the mean and the variance. This was segmented according to the timestamps associated with GPS data. For segments of sensor data that took place within timestamps in which GPS data was in a particular cluster, the segment was associated with that cluster.

Starting with IMU, we tested the overall magnitude of acceleration, using the calculation $mag = \sqrt{x^2 + y^2 + z^2}$. A motivation for combining the 3 dimensions of acceleration is that the IMU sensor in the Smart Collar only detects the 3 dimensional linear acceleration, without a gyroscope for rotational acceleration. Because of this, we cannot orient the data in a consistent way— the X axis for one portion of data may not correspond to the same direction as the X axis for another portion, even
for the same dog (due to variation in collar placement). For this combined IMU data, we investigated if the mean and variance in acceleration detected by the Smart Collar was significantly different between the clustered locations. For testing if the mean of each data set are different, we used the Mann-Whitney U Test, while the Levene’s Test compared the variance of two variables. In Table 7.2, results of the Mann-Whitney U Test and Levene’s Test are shown for IMU data for 2 dogs. This shows a significant difference in the variances between the IMU data sets for both dogs that had IMU data available. The means of the activity levels was significantly different between clusters for only one of the two dogs.

Table 7.2 Comparison of IMU Data, Separated by K-Means Clusters (“C1” and “C2”)

<table>
<thead>
<tr>
<th>Dog</th>
<th>C1 n</th>
<th>C2 n</th>
<th>C1 Mean(SD)</th>
<th>C2 Mean(SD)</th>
<th>MW-U Test</th>
<th>Lev Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>167258</td>
<td>125484</td>
<td>1.0051(0.2974)</td>
<td>1.0331(0.1673)</td>
<td>8.194E10(p&lt;0.01)</td>
<td>13250.06(p&lt;0.01)</td>
</tr>
<tr>
<td>3</td>
<td>20608</td>
<td>10267</td>
<td>0.9947(0.1027)</td>
<td>1.001(0.1454)</td>
<td>1.063E8(p=0.4096)</td>
<td>454.131(p&lt;0.01)</td>
</tr>
</tbody>
</table>

We also examined the differences in environmental data for these particular dogs. We used the same clusters produced by the K-Means algorithm on GPS data. Light data can be interpreted as the overall brightness of the current environment. The results of the two statistical tests are shown in Table 7.3. All but one of the dogs which had light data available for this test had a significant difference in overall light levels and significant difference in light level variances between the two location clusters. For Dog 5 these two locations had similar light levels. However, note that the number of data points available for Dog 5 was much lower than the rest of the dogs, especially for Cluster 1. More data from this dog could have resulted in more distinct differences in the light data.

Table 7.3 Comparison of Light Data, Separated by K-Means Clusters (“C1” and “C2”)

<table>
<thead>
<tr>
<th>Dog</th>
<th>C1 n</th>
<th>C2 n</th>
<th>C1 Mean(SD)</th>
<th>C2 Mean(SD)</th>
<th>MW-U Test</th>
<th>Lev Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6643</td>
<td>142</td>
<td>0.051(0.070)</td>
<td>0.017(0.013)</td>
<td>653338.0(p&lt;0.01)</td>
<td>63.156(p&lt;0.01)</td>
</tr>
<tr>
<td>3</td>
<td>9367</td>
<td>5115</td>
<td>-0.180(1.407)</td>
<td>0.492(1.474)</td>
<td>18289870.5(p&lt;0.01)</td>
<td>30.279(p&lt;0.01)</td>
</tr>
<tr>
<td>4</td>
<td>3208</td>
<td>891</td>
<td>0.056(0.082)</td>
<td>0.014(0.014)</td>
<td>1778123.0(p&lt;0.01)</td>
<td>318.99(p&lt;0.01)</td>
</tr>
<tr>
<td>5</td>
<td>308</td>
<td>510</td>
<td>0.345(0.33)</td>
<td>0.325(0.31)</td>
<td>81250.5(p=0.408)</td>
<td>2.333(p=0.127)</td>
</tr>
</tbody>
</table>

The results of the tests for sound data, or ambient "loudness", are presented in table 7.4. All but one dog, this time Dog 4, showed significant differences in sound data between the two location clusters. In this case, the number of data points was sufficient and would not explain the similarity in locations. It could be that the environments were geographically dispersed but physically similar, or that the clusters for this dog were not meaningful (demonstrating a limitation of simple K-Means
Clustering is just one approach to segment this large spatial and temporal data set. As more data becomes available, especially as raisers take their dogs to more locations, K-Means may be less and less effective and density-based clustering may be more appropriate. To analyze behavior with respect to specific environment types, external services could also be employed to tag portions of data with one of several location types, and these situations that dogs are interacting in could be further divided by environmental conditions during data collection using the sensor data we are collecting. In addition, we intend to allow raisers to tag important events in the data stream such as “Meeting a new person” or “Meeting a new dog” for in-depth analysis.

### 7.5 Conclusion and Future Work

The cloud data collection and processing platform running on IBM Cloud gathers data from various phases of the training program at Guiding Eyes for the Blind. In this paper, we discussed initial results using unsupervised learning and statistical analysis of Smart Collar data. All data were obtained using the full end-to-end system running in production. Results show promise in predicting guide dog outcomes from objective sensor data, which can assist experts with objective analysis with the goal to save precious resources and provide more guide dogs to potential handlers.

During the puppy raising phase, we are in the process of collecting data and preparing to scale
to 300 active raisers, providing a rich data set from which to develop predictive models. Future work includes prediction based on these aforementioned data sources using real outcomes and improvements to the overall system and data pipeline. In addition, as we scale the system, we will incorporate more forms of data input. We will implement a feature in the mobile app to allow raisers to manually tag events such as “Meeting a new person” or “Meeting a new dog”. As a professional guide dog will have to perform under all kinds of conditions, ground truth to augment the data with these types of events will be useful for more in-depth analytical tasks. However, care must be taken to not cause the raiser to be distracted from the dog or to impact their typical routine while they are performing these interactions with the app. Therefore, we will likely allow raisers to retroactively tag these time segments with approximate time stamps. Finally, once real outcomes are obtained for a significant number of dogs, we can build models using these real outcomes as labels for machine learning from each data source.
8.1 Abstract

Volunteer puppy raisers expose potential guide dogs to new stimuli and teach obedience skills away from guide dog schools. Producing successful guide dogs is quite expensive and many do not succeed. Identifying and screening successful dogs early is vital for the success of guide dog programs. As there is little objective data about puppies’ experiences with raisers, there is great potential in providing analytical insight during this important developmental period. We present a collar-based sensor system to remotely collect various types of sensor data to derive higher-level analytics describing dogs' behaviors and their environments. We interviewed guide dog experts to refine and develop metrics derivable from the sensor data. We used the collar system to collect training data and generate machine learning models to capture these metrics in order to provide guide dog schools insight about puppies being raised by volunteers away from schools remotely and at wide scale.
8.2 Introduction

Potential guide dogs are typically raised by volunteer raisers in their homes away from the guide dog experts and trainers at guide dog organizations. The process of raising and training dogs for this role is quite expensive, with some estimates ranging $20,000 to $50,000 [Ber17] or more. Unfortunately, many dogs that enter training do not succeed as guide dogs for various reasons. Great effort has been put into identifying successful dogs early in life based on both genetics and behavior and candidate dogs are screened out of programs at several stages.

Guiding Eyes for the Blind, a large, nonprofit U.S. guide dog school, maintains assessments records of all dogs, providing a rich corpus of historical data. The 73% of puppies that pass the first evaluation at about 8 weeks of age enter the raising phase in which volunteer puppy raisers raise the puppies in their homes. Raisers house and prepare them for training. However, there is very little objective data about the puppies’ experiences while with the raisers, which has been a topic of interest for researchers who want to better understand how this critical stage shapes dogs’ suitability for guide work. Because of the difficulty guide dog organizations face in understanding the day to day details of the experiences of hundreds of puppies during the months-long raising stage, there is great potential in providing sensor-based data and analytics describing the dog’s experiences while raising. A sensor-equipped Smart Collar system [Cle19] for use with dogs in raising has enabled remote sensing of dogs’ behaviors and environments through sensors to detect factors such as movement, noise levels, and environmental data. I used this system for data collection in this study.

I have defined several metrics which describe a dog’s behavior and environment at a higher level of abstraction than raw sensor data to allow for better understanding of the experiences of the candidate guide dogs in the raising period. I conducted an interview with four guide dog experts to refine these analytics and develop new metrics based on their goals in understanding dogs in raising. I then used the Smart Collar system to collect data in controlled situations to generate raw collar data with labeled values for each metric. Next, I built machine learning models to capture these metrics and validate the accuracy of each model. In this chapter I describe the process of interviewing, data collection, model building, and evaluation for higher level metrics from the collar data. The final outcome will then be to develop an automated reporting system to generate the values for these behavioral and environmental metrics, giving guide dog schools insights about puppies in raising in as close to real time as possible.

8.3 Related Works

Understanding what factors influence canine service suitability is an active and growing area of research. Recording aspects of canine behavior, either through human observation or sensor data, takes many forms. Behavioral coding and standardized questionnaires are common forms of human observation, though electronic sensors are growing in use.

Behavioral coding is a technique used to record when an animal exhibits certain predefined
behaviors. Bray et al. coded seven variables relating to maternal style at a guide dog school [Bra17a]. The principal component, called *maternal style* had predictive ability in training. Using this measure of maternal style, they studied 98 puppies in a guide dog training program [Bra17b] and found a negative association of early intense levels of mothering with success in the program. This suggests that their more abstract measure of maternal style had predictive ability. Puppies whose mothers had a nursing style requiring more effort from the puppy had a higher success rate.

Standardized temperament evaluations are also commonly used tools to understand canine behavior. Questionnaires provide a standard, but still subjective, set of evaluation metrics that can be used for a variety of purposes, including prediction of service success. This provides consistency across individual studies and applications. The C-BARQ [Seg05] is a 100 item questionnaire for evaluating canine temperament, validated for use in dogs [Duf08; Hsu03], and is used by Guiding Eyes. The C-BARQ has been used to study the effect of training on dogs’ cognitive abilities [MP08]. Duffy and Serpell used the C-BARQ questionnaire to predict future success of guide and service dogs [Duf12]. With 11,997 evaluations from 7,696 dogs, they used a generalized linear model to predict a binary success outcome. Successful dogs scored better on 27 C-BARQ items at both evaluations. *Pulls excessively hard on leash* was the most predictive trait in this analysis. Serpell also developed the Behavior Checklist (BCL), a scoring system where observers score 52 behavioral categories such as excitability, fearful behavior (such as fear of traffic or noise), and body sensitivity. Like the C-BARQ questionnaire, the BCL is commonly used for scoring the behavior of potential guide dogs.

Arata, Momozawa, and Takeuchi [Ara10] found that distraction, sensitivity, and docility are important factors that can predict a dog’s future success based on questionnaires and that distraction predicted eventual success with 80.6% accuracy. This provided more evidence that understanding specific metrics of dog behavior can allow experts to better predict which dogs will be successful. Standardized questionnaires can also be used to aid non-experts in collecting behavioral data. Serpell and Hsu used the C-BARQ to survey the owners of 1,563 dogs to assess “trainability” among 11 breeds [Ser05]. They found significant differences in trainability among several breeds without the need for formally-trained experts to physically interact with the dogs.

Researchers have also used sensor-based data collection methods to characterize and predict behavior in dogs. Sensor-based approaches allow more objective observation of the dogs’ behaviors and surrounding environments. Sensors allow for automated posture and behavior detection as well as prediction of service suitability for a variety of service roles.

Depth sensing cameras such as the Microsoft Kinect have been used for posture and behavior detection. Pistocchi et al. proposed a solution for classifying a dog’s body parts from a video feed [Pis14]. They used a structural support vector machine to model the dog’s shape from depth sensor data and validated the results via visual comparison. Qualitative analysis of their results by domain experts resulted in 96% of images either fully or partially (at least half of the body) correctly classified. Using Kinect video and depth data, Mealin, Dominguez, and Roberts extracted a model of a dog’s body without human-authored labeling [Mea16] which could differentiate between static postures. They detected several common postures such as *standing*, *sitting*, and *lying*, correctly classifying
the standing posture with 70% accuracy, sitting with 69% accuracy, and lying with 94% accuracy.

Collar-mounted sensing devices are becoming a common approach to this form of data collection, especially inertial measurement units (IMUs) for accelerometer data. Brugarolas et al. used machine learning (random forests, K nearest neighbors, and logistic model trees) to classify postures in two dogs [Bru13b]. The results were approximately 98% classification accuracy for each algorithm. Brugarolas et al. also proposed methods for detecting both postures and behavior states using the same collar-mounted IMU system [Bru13a]. They used decision trees to classify static postures and Hidden Markov Models (HMMs) to detect dynamic activities (behaviors) of seven dogs and assessed the reliability of both. The HMM correctly classified walking and walking up stairs behaviors with 100% accuracy, walking down a ramp with 100% accuracy (except for one dog which was 92%), and non-dynamic behaviors (postures) with accuracy above 94% for all but one dog in the study, with the last dog resulting in 77.8%. This study further validated the ability of collar-mounted IMUs to model behavior. Den et al. [Uij17] also used a collar-mounted triaxial IMU to detect various behavioral states in dogs. The states they studied were walk, trot, canter/gallop, sleep, static/inactive, eat, drink, and headshake. They achieved a precision in the 93-100% range and recall in the 89-98% range.

Gerencsér et al. used a 6-axis harness-mounted IMU to classify behavioral patterns [Ger13] of 24 dogs in an open outdoor environment. They coded seven behaviors as labels for supervised learning with support vector machines as the learning algorithm. The seven behaviors were lay, sit, stand, walk, trot, gallop, and canter. The SVMs resulted in an accuracy of about 80%. These studies show how the detection and measurement of behavioral factors is achievable with collar-mounted triaxial IMU data.

Wearable sensors have recently become a topic of interest for data collection while dogs are with volunteer puppy raisers. Volunteers raise the puppies for typically more than a year during a crucial period of development [Lue11], so there is a great opportunity for studying puppy behavior with wearable computing devices. However, this also comes with new challenges. Zamansky and van der Linden surveyed puppy raisers about their attitudes toward activity trackers [Zam18]. They found that the majority of participants had positive attitudes toward the benefits of activity trackers, though some participants had concerns about privacy and comfort of the dog. Some raisers voiced concerns about location data. Ultimately, the study found that the majority of raisers would unconditionally accept activity trackers if the guide dog school wanted to use them. The findings do support that privacy must be considered in the design of both the technology and protocol. Van der Linden et al. investigated the privacy aspects of commercially-available activity trackers [Lin19] and found several privacy concerns, such as a mismatch between product marketing and what data the trackers actually collect. For example, some collars are marketed as fitness trackers but collect location data without mentioning this to the consumer. There is also little clarity on what data, if any, is considered private. They found in many cases that it is unclear what data are actually stored and what is inferred from collected data.

Environment greatly affects development [Lue11], especially in terms of guide dog suitability [Har16]. Burrows et al. studied service dogs for children with autism spectrum disorder and used
observations and interviews to investigate what factors in the home environment affect a dog's behavior and welfare [Bur08]. Serpell and Duffy analyzed C-BARQ data from 978 potential guide dogs at a guide dog school and found several environmental factors that influence C-BARQ results relevant to guide dog success [Ser16]. Their findings show that households with experienced raisers and households with at least one other dog were both associated with successful behavioral traits at 12 months of age. Researchers have recently used citizen science, collecting data from members of the public, typically without formal scientific training, in the domain of canine behavior. Citizen science allows for data collection at very large scales in unrestricted environments. Hecht and Rice investigated both the benefits and technical issues in applying citizen science to canine research [Hec15]. It can be difficult to ensure that the observer follows correct procedure, leading to data quality issues. Dognition is a citizen science project [Ste15] with the goal of gathering behavioral data from dog owners at a large scale. Dog owners perform cognitive exercises with their dog and report results using a questionnaire on a website. Hare found this approach to be useful for collecting large-scale data from citizen scientists and found that there was little evidence of manipulated results [Ste15].

8.4 Methods

The motivation for creating more abstract metrics from raw sensor data is to provide a more complete picture of dogs' experiences in raising. With these and potential future metrics automatically generated from the data, a wide variety of statistical questions can be answered and reported to guide dog schools. To enable automated reporting and analytics from sensor data, we first defined six potential metrics to describe a dog's current behavior and environment at a higher level than individual raw sensor data. Next, we interviewed guide dog experts at Guiding Eyes for the Blind about their opinions regarding what metrics would be useful in aiding them to help dogs succeed in the months-long raising process. The goal of this process was to create a collection of derived features from Smart Collar data to assist guide dog organizations with understanding vast amounts of sensor data. Reporting such analytics at a higher level than raw sensor data will help schools like Guiding Eyes understand how each of their dogs with raiser families are developing and how they can assist raisers to increase the likelihood of a dog achieving success in the guide dog program. This will also help to contextualize the output of future predictive models.

We conducted a semi-structured interview of four guide dog training experts at Guiding Eyes about their opinions on behavioral and environmental metrics derived from sensor data during the raising period. We first defined six metrics: Average Speed, Current Behavioral State, Current Environment, Sound Origin, Location Familiarity, and Reaction to New Environment. Average speed describes the average speed of the dog as it moves around in its environment, based on IMU and GPS data. We proposed the average speed metric to quantify how fast each dog typically moves. This would also enable analysis of its behavior with respect to speed and what factors affect speed. The current behavioral state describes the dog's overall behavior as a categorical variable with the
proposed values *active* and *inactive*. We proposed *current behavioral state* to capture the overall behavior of the dog in a single variable so that the dog's general behavior would be easier to assess at a given time. The *current environment* describes the dog's surroundings, specifically whether or not the dog is *outdoors* or *indoors*. Although we expected it may be useful for guide dog schools to know how much time the dog spends outside, we also chose *current environment* to provide context to other data sources, such as raw accelerometer data or other derived metrics. *Sound origin* classifies the origin of the current sound signal as primarily contributed by the dog wearing the collar or from its environment. Barking tendencies in various situations may impact suitability as a guide dog. Thus, we chose the *sound origin* metric to better understand when the dog itself is the source of noise. The *location familiarity* metric describes how much a particular place has been visited by the dog. Since guide dogs must work in a variety of environments and guide dog schools encourage raisers to expose the dogs to a multitude of places, we defined *location familiarity* to give context to the collar data and help guide dog organizations understand how the dogs in raising behave in familiar and unfamiliar places. *Reaction to new environment* describes how the dog's behavior changes with respect to new places. *Reaction to new environment* specifically captures when a dog is in a new place. We chose this metric to summarize the dog's typical *current behavioral state* in unfamiliar places. For categorical metrics, I refer to the possible values for the metric as *states*.

### 8.4.1 Interview Methodology

For this study, the interviewees were four experts at the guide dog school serving in various roles relevant to puppy raising. The interviewees have 37, 20, 19, and 1.5 years of experience in this domain and serve in the following roles in the organization: Senior Director- Genetics and Breeding, Director of Canine Program Development, Puppy Program Manager, and Genetic Data Analyst. These experts were already familiar with the collar system and the raw data types that the sensors can collect. The process consisted of two phases. The first phase took place before the interview in which the experts reviewed the six proposed metrics and collectively agreed to an assessment of each metric. Here, they provided their responses in the form of written feedback. Due to the relatively small number of items (six) to assess, we did not restrict their overall rating with any particular scoring system. They also had the opportunity to propose additional analytical metrics that they might find to be useful in bolstering their understanding of the progress of dogs in the raising process. The second phase was the actual interview, which we conducted and recorded over a video teleconference. During the interview, we discussed each metric, both the metrics we proposed and those added by the experts, and the reasons for each rating. The interview itself lasted for one hour.

### 8.4.2 Data Collection and Model Building

Before I describe the interview results and the metrics in detail, I will first discuss how the raw sensor data was collected and used. The overall process is similar across the metrics. In subsequent sections I will present the details and unique characteristics of each of the analytics. To build a model that can generate these metrics on live collar data, labeled data is needed to serve as ground truth for
each metric. For most metrics, we collected data using dogs belonging to members of our team that are not currently in a guide dog raising program due to the greater ease of data collection and coding. This group of three dogs consists of Labrador retrievers of various ages. For current behavioral state, I worked with staff at the Guiding Eyes for the Blind who have access to dogs known to have varying temperaments with respect to the behavioral states. For all metrics, dogs wore the collar while going through typical daily routines to gather data such as during walks and being indoors and outdoors. Any dogs that would not tolerate the collar would not be required to wear it and these dogs would not be used in the study, though this was not an issue.

Each metric is different and thus generating them require somewhat different processes, but the general protocol follows a similar pattern of data collection and labeling, model building, and evaluation. I used a standardized protocol for data collection while the dog and handler use the collar system to capture a variety of situations to increase the robustness of the eventual model. I also defined protocols for labeling the data that incorporates accompanying video when needed. Using those labeled data, I used various machine learning algorithms to build a model that estimates the value of that particular metric. The labeled data enable us to evaluate the accuracy of the model. Additionally, for metrics that make use of multiple sensor types, there are gaps in data between samples as a consequence of differing sampling rate. When necessary, I used interpolation to fill in these gaps once data sets were joined. Once a model has been fully validated, we can deploy it in production, providing Guiding Eyes for the Blind with remote, real-time metrics about their dogs in raising. For each metric, we employed a variety of models and heuristics, including straightforward machine learning algorithms such as decision trees and neural networks. For these models, we developed and tested them using Scikit-Learn [Ped11] and Tensorflow [Aba15]. While each metric requires different techniques, each approach can be evaluated similarly. I also employed hyperparameter optimization where appropriate in order to systematically arrive at an optimal model.

8.4.3 Ethics

All procedures in this study have been completed under IACUC approval. The use of the collar system has also been approved by the Institutional Review Board at North Carolina State University. Any dogs that did not tolerate either wearing the smart collar or participating in data collection were not required to and removed from the study.

8.5 Generating Collar Data Analytics

The experts assessed the proposed metrics in terms of overall importance. Table 8.1 summarizes the overall importance ratings given for each metric, including the new metrics suggested by the experts. In addition to the proposed metrics, the experts suggested two additional metrics: walking speed when on an exercise walk and stairs incidents. Although they were interested to know about the frequency and variety of stairs exposure in the puppies in the raising program, they found stairs...
Table 8.1 Overall Expert Rating of each High Level Metric

<table>
<thead>
<tr>
<th>Metric Name</th>
<th>Suggested?</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Speed</td>
<td>No</td>
<td>Not Sure</td>
</tr>
<tr>
<td>Behavioral State</td>
<td>No</td>
<td>High if Practical</td>
</tr>
<tr>
<td>Current Environment</td>
<td>No</td>
<td>Not Sure</td>
</tr>
<tr>
<td>Sound Origin</td>
<td>No</td>
<td>Not Sure</td>
</tr>
<tr>
<td>Location Familiarity</td>
<td>No</td>
<td>High</td>
</tr>
<tr>
<td>Reaction to New Env.</td>
<td>No</td>
<td>High</td>
</tr>
<tr>
<td>Walking Speed</td>
<td>Yes</td>
<td>High</td>
</tr>
<tr>
<td>Stairs Incidents</td>
<td>Yes</td>
<td>Low</td>
</tr>
</tbody>
</table>

incidents to be less important as an objective metric at this time, rating it as “low” importance.

The following subsections summarize for each metric findings from the interview in more detail and outline the process of data collection, model building, and evaluation. Overall, the experts agreed with the usefulness of producing higher-level metrics from the raw collar data. It should be noted that the experts pointed out that at this stage, there are still many open questions about what types of analytics are truly useful while we explore various metrics and thus are interested in exploration of a wide variety of such metrics. Their interest in these metrics tended to focus on whether or not objective metrics are associated with future measures when evaluating adult dogs, such as if walking speed correlates to working speed as an adult, and whether or not guidance or intervention is needed with puppies in raising to increase the chances of successful outcomes. However, they stressed the importance of the exploratory nature at this phase, warning against using these data to be “prescriptive when what we really need to do is collect [information]” (Interviewee 2).

Overall, the data collection process was similar across the metrics. For all but the behavioral state metric, our team collected training data with their own dogs. Two handlers collected data from three dogs, all Labrador retrievers. These dogs were well-trained, but not associated with any service dog organization and not trained as service dogs. For the behavioral state metric, guide dog experts at Guiding Eyes for the Blind collected data from several dogs in the raising program. Additionally, the experts clarified that behavioral state should be more accurately called emotional state. Thus, further references to behavioral state will refer to the metric instead as emotional state.

8.5.1 Walking Speed

The average speed metric was given an importance rating of “not sure”. They expressed concern about sources of skew in the data such as when the dog is confined and large distance between the raiser and dog (GPS coordinates are provided by the raiser’s phone and are used alongside IMU to
estimate speed). However, the experts suggested a focused estimate of walking speed specifically when the dog is out on an exercise walk, as coded by the raiser through the mobile app. They noted that this would be most useful for puppies older than eight months due to more consistent frequency of exercise walks. This metric would be useful for them in determining if and how walking speed in the raising program is associated with working speed during advanced training. The experts explained that they already know that puppies that walk faster in the program do not always walk faster in training and once placed, but this metric would allow for a data-driven understanding of this relationship. These data could also be analyzed for heritability. The experts noted that the guide dog school can also learn how tolerant dogs are in maintaining their walking speed in various weather conditions such as heat and humidity. They said it “might be useful to know a dog really slows down a lot in humidity” (Interviewee 1), but expressed that correlating walking speed to humidity is uncertain and a lower priority. Due to the concerns mentioned in the interview and the desire for a more focused estimate of speed specifically on exercise walks, we focused on the new walking speed metric as opposed to the more general average speed metric. This was captured with a feature in the mobile app for logging special events. The experts said that it would also be interesting in terms of understanding how raisers affect the walking speed of dogs. They mentioned the difficulty in understanding “the handler influence on the dog speed on any … walks” (Interviewee 2).

8.5.1.1 Data Collection

To collect training data for this metric, the handlers logged the walk events in the mobile application at the beginning and end of routine exercise walks. Separately, they also recorded their walking segments to provide an outside measure of speed. The handlers obtained this outside measure of speed using a popular exercise app called Strava, which is used by runners and cyclists to accurately record their speed. To provide additional training data, the handlers also collected samples in which the dog is stationary in addition to walking to increase the range of speeds in the training data. The data sets collected consisted of 13 exercise walks completed by the three dogs owned by the handlers. This resulted in 543820 total samples. The raw input data for walking speed were the IMU accelerometer data and GPS latitude and longitude data.

8.5.1.2 Building a Model

Estimating walking speed is a regression task. The method to estimate speed centers on the integration of acceleration to derive velocity. However, the accelerometer data necessarily includes gravitational acceleration which must be accounted for. Simply subtracting the gravity component is imperfect as the collar is not typically pointing an entire axis toward the ground. Additionally, error in the three acceleration measurements compound the overall error quickly in the integrated velocity. Fortunately, we have another measure of position to assist in estimating speed. GPS coordinates have been used in prior research using sensor fusion alongside IMU data to more accurately estimate velocity [Car06]. Other work has used known heuristics such as stride length to improve neural network speed estimation from triaxial accelerometers in humans [Son07]. We can use cer-
tain heuristics to further inform the model, such as reasonable top speed estimates for a labrador retriever. I used neural networks to combine various derived features in order to accomplish this task. I experimented with both feed-forward neural networks and recurrent neural networks in the form of Long Short-term Memory (LSTM) networks and Gated Recurrent Unit (GRU) networks.

Input data from the system consisted of IMU accelerometer data and GPS coordinate data. I preprocessed these to extract some additional information from the raw data. First, whenever there was missing data for a feature due to the join, I used linear interpolation to fill in the gaps. I calculated the total acceleration \( (total\_acc) \) using the same calculation as in previously:

\[
\text{total\_acc} = \sqrt{x^2 + y^2 + z^2}
\]

I calculated an estimate for velocity based on the acceleration. To get an estimate of this, first we need to account for the gravity component of acceleration, which is difficult without gyroscopic IMUs. For a given walk session, I found the sum total of acceleration for each axis. This produces a vector that, since gravity produces constant acceleration in one direction, should be dominated by gravity. With a unit vector with the same direction as this sum, I generated an estimated gravity vector with which to subtract from each IMU acceleration sample. Then, I integrated the acceleration samples over time to produce an estimate of velocity. I used the resulting velocities \( (vx, vy, \text{ and } vz) \) along with the total velocity magnitude \( (vel) \) as extracted features.

GPS coordinates with timestamps can also be used to generate estimates of speed. Using known quantities such as the approximate radius of the earth where we are collecting data, we can extract distances on the surface of the earth from latitude and longitude coordinates. Thus, we can calculate an estimate of speed \( (gps\_speed) \) from these GPS samples. Note that the smartphone handles GPS data in this system, which is why GPS data are only used as an estimate of the dog’s true speed.

I used \( vel \) and \( gps\_speed \), \( total\_acc \), the three accelerometer data inputs \( (ax, ay, \text{ and } az) \), and the three axes of velocity estimates \( (vx, vy, \text{ and } vz) \) as input features. To generate an estimate of the dog’s walking speed from the inputs, I used a recurrent neural network architecture to capture the time-dependent nature of the data set. I used Tensorflow[Aba15] for this learning task.

To thoroughly test the performance of models on unseen data, I used K fold cross validation with \( K = 3 \). In this case, \( K \) is relatively low due to the long time in which it takes to train recurrent neural networks. With \( K = 3 \), there are three different validation sets taken from the data set, with the rest of the data used for training. Any particular session of data collection always belonged fully to the training set or to the validation set. For any particular set of hyperparemeters, I tested their performance by calculating the cross-validated mean \( loss \) of the models produced by K fold cross validation. The measure of \( loss \) I used was the mean squared error of the difference in model predictions and actual speed values.

After some informal experimentation, I found that Long Short-Term Memory (LSTM) neural networks outperformed Gated Recurrent Unit (GRU) networks, that 10 “steps” (number of samples to consider before making a prediction) were sufficient for learning, and that 120 epochs (number of times the model trains on the entire training data set) and a batch size of 64 were sufficient for this task. I used the ADADELTA optimization algorithm[Zei12] found in Tensorflow. The general
The structure of the neural networks followed this pattern:

1. LSTM layer (input)
2. Dropout layer
3. 1 or 2 Fully Connected Layers (with dropout after first)
4. Linear Layer (reduces output shape to 1 speed estimate)

I optimized five hyperparameters with a simple genetic algorithm. This genetic algorithm used an independent crossover, so an offspring took each hyperparameter randomly from one of the parents. A mutation step altered one random hyperparameter for every offspring each generation. Otherwise, the process followed a classic genetic algorithm with 10 generations, 10 members of the population each generation, and 4 parents each generation. This process did not drastically improve the accuracy of the model, possibly due to the relatively small search space. I chose the best values for the hyperparameters based on the “fittest” set at the end of the process, with some manual tuning. The results of the hyperparameter optimization process are shown in Table 8.2. The “learning rate” was allowed to range from 0.1 to 1.0, though the ADADELTA optimizer actually adjusts the learning rate as training proceeds. “Rho” acts as a decay rate for the optimization algorithm and the optimal value was found to be not far from ADADELTA’s default value in Tensorflow of 0.95. There were up to three layers in the networks, with the second dense layer being optional. This layer turned out to be unnecessary as the best results were achieved with a single layer of LSTM units followed by a single dense layer (then the linear output layer). Finally, each layer also used L2 regularization and was followed by a dropout layer during training, with dropout equal to 0.5.

![Diagram of the neural network architecture for the Walking Speed metric](image.png)

**Figure 8.1** Diagram of the neural network architecture for the *Walking Speed* metric

The cross validation performance of the trained neural network models, averaged across the three cross validation folds, are compared with the baseline metrics in Figure 8.3. Each model’s Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are shown. The units of the RMSE are in terms of miles per hour. The mean training error for the neural network models in terms of MSE after 120 epochs was 0.5638 (RMSE 0.7508). The worst performing method of estimating speed was actually the GPS speed estimate, likely because of the level of precision
Table 8.2 Min, Max Values for Hyperparameter Optimization for Walking Speed; Best Values for each Feature

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Min</th>
<th>Max</th>
<th>Best Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning Rate</td>
<td>0.1</td>
<td>1.0</td>
<td>0.33</td>
</tr>
<tr>
<td>Rho (ADADELTA)</td>
<td>0.85</td>
<td>0.999</td>
<td>0.952</td>
</tr>
<tr>
<td>LSTM Units</td>
<td>4</td>
<td>200</td>
<td>79</td>
</tr>
<tr>
<td>Dense Layer 1 Units</td>
<td>1</td>
<td>300</td>
<td>9</td>
</tr>
<tr>
<td>Dense Layer 2 Units</td>
<td>0</td>
<td>150</td>
<td>0</td>
</tr>
</tbody>
</table>

available. This method by itself performed worse than the baseline. The estimate calculated from the accelerometer data alone performed better, but only slightly. However, combining these estimates in a temporal neural network model performed much better, with MSE less than half of that of the two individual estimates and also the mean baseline predictor, with an MSE of 0.422. This does leave much room for improvement, but is certainly a better model for speed than the two estimates alone. Estimating speed from these triaxial accelerometer data is a difficult task, but the inclusion of both GPS estimated speed IMU estimated speed in a recurrent neural network appear to increase our ability to estimate the dog’s walking speed.

To further characterize the performance of the model, a plot of the actual speed and predicted speed for an example time segment is shown in Figure 8.2. This exercise walk was not included in any of the training sets (the model was trained on the rest of the exercise walks). For this particular walk, the mean squared error was 0.294 and the standard deviation of the error was 0.257, meaning that most errors fell fairly close the mean. However, there were some outliers, with the maximum error being 2.942 miles per hour. The predicted speeds were typically noisier and the model had a narrower range, causing it to miss large spikes in the recorded speed. These were errors were in both directions, so in terms of characterizing the dog’s walking speed this is not necessarily a large problem.

Table 8.3 Cross-validated Performance of Speed Model compared with Baselines

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean Val Error (MSE)</th>
<th>Mean Val Error (RMSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Speed Baseline</td>
<td>1.181</td>
<td>1.087</td>
</tr>
<tr>
<td>Velocity Estimate</td>
<td>1.160</td>
<td>1.077</td>
</tr>
<tr>
<td>GPS Speed Estimate</td>
<td>1.331</td>
<td>1.154</td>
</tr>
<tr>
<td>Neural Network</td>
<td><strong>0.442</strong></td>
<td><strong>0.665</strong></td>
</tr>
</tbody>
</table>
8.5.2 Emotional State

Given a segment of collar data, this metric classifies the dog’s overall behavioral state in terms of its activity level. The experts rated the importance of emotional state as “high if practical”. They suggested that instead of two states, active and inactive, four states would be more useful in characterizing the state:

- excited
- activated with stress
- inactive/relaxed
- inhibited

The split into four states is because “inactive could be both inhibited and relaxed—doesn’t necessarily mean that they are still.” (Interviewee 1) and when the dog is not still, active could actually be “excited or activated with stress” (Interviewee 1). They noted the expected difficulty in coding and modeling these states, stating that it is “hard to pull apart” (Interviewee 2). The four states used by the data collectors are described in the Behavior Checklist definitions as used by Guiding Eyes for the Blind and other organizations scoring behavior. When a dog is activated when exposed to potentially stressful situations, it is exhibiting signs of stress by becoming more active such as faster movements which may include jerky body movements, displacement activities such as scratching or head shaking. Activated dogs are typically less responsive to the handler, exhibit body language of lips and ears pulled back. However, in the excited state, it is exhibiting signs of increased energy and arousal levels without observed stress signals. A dog that is inhibited is showing signs of stress and demonstrating “shutting down” behaviors such as avoidance and withdrawal. An inactive or relaxed dog exhibits smooth movements, relaxed body language and is able to remain responsive.
8.5.2.1 Data Collection

To create a model that classifies emotional state, manually labeled segments of data in each state were needed. Canine behavioral experts provided these labels by collecting collar data and coding them in the app with the refined set of states: excited, activated with stress, inactive/relaxed, and inhibited. This was accomplished at the guide dog school with several dogs known by the experts to be prone to the various states, with the handlers noting times in the app when the dogs are exhibiting these states. I worked with the guide dog experts at Guiding Eyes who developed a protocol (see Appendix A) in order to elicit the four states from candidate guide dogs in raising.

In total, Guiding Eyes professionals collected data for 48 dogs. Table 8.4 summarizes the data collected in terms of the four emotional state events.

Table 8.4 Stats for emotional state categories

<table>
<thead>
<tr>
<th>State</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inactive/Relaxed</td>
<td>86</td>
</tr>
<tr>
<td>Excited</td>
<td>134</td>
</tr>
<tr>
<td>Activated with Stress</td>
<td>46</td>
</tr>
<tr>
<td>Inhibited</td>
<td>55</td>
</tr>
<tr>
<td>Total</td>
<td>321</td>
</tr>
</tbody>
</table>

8.5.2.2 Building a Model

To build this model, I used IMU and sound data as input features, as well as the engineered features used for the sound origin metric. One example of data for the machine learning task consisted of a number of consecutive IMU and sound samples, corresponding to a small segment of time. The output label was a categorical variable with four values– the emotional state as coded by the experts. The models had a similar structure as the models used to estimate walking speed:

1. LSTM layer (input)
2. Dropout layer
3. 1 or 2 Fully Connected Layers (with dropout after first)
4. Sigmoid Layer (reduces output shape to 4, outputs probabilities for each state)

Due to the extreme imbalance of the class distribution, I duplicated the data for the activated with stress and inhibited emotional states for the training set so that all four states had roughly equal
representation before applying K fold cross validation. The validation sets were not modified in this way in order to retain the true distribution of states. I again used a genetic search to optimize hyperparameters and the results for this process are shown in Table 8.5. A visual diagram of the neural network model is shown in Figure 8.3.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Min</th>
<th>Max</th>
<th>Best Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning Rate</td>
<td>0.1</td>
<td>1.0</td>
<td>0.71</td>
</tr>
<tr>
<td>Rho (ADADELTA)</td>
<td>0.85</td>
<td>0.999</td>
<td>0.989</td>
</tr>
<tr>
<td>LSTM Units</td>
<td>4</td>
<td>200</td>
<td>146</td>
</tr>
<tr>
<td>Dense Layer 1 Units</td>
<td>1</td>
<td>300</td>
<td>37</td>
</tr>
<tr>
<td>Dense Layer 2 Units</td>
<td>0</td>
<td>150</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 8.5 Min, Max Values for Hyperparameter Optimization for Emotional State; Best Values for each Feature

Figure 8.3 Diagram of the neural network architecture for the Emotional State metric.

Using the optimized hyperparameters, the final K fold cross validation results are shown in Table 8.6. The best performing model, using unweighted classes, only performed marginally better than a simple majority classifier, indicating that it was very difficult to separate the emotional states accurately. The last row, LSTM, Combined States, uses the same approach as the unweighted class model, but combines the Inactive/Relaxed and Inhibited states into one state as well as Excited and Activated with Stress into one state. This binary classifier performed the best. This suggests that it was quite difficult to separate the more “inactive” states correctly into Inactive/Relaxed and Inhibited, and it was also difficult to properly separate Excited and Activated with Stress.

Examining each state individually, we can analyze the performance in terms of the precision and recall of each state. The curves for Inhibited and Activated with Stress further show the difficulty in correctly detecting these states. The sharp decline in precision when lowering the prediction threshold for Inactive/Relaxed shows that high precision was not really possible without an unreal-
Table 8.6 Cross-validated Performance of Emotional State models

<table>
<thead>
<tr>
<th>Model</th>
<th>Overall Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Classifier Baseline</td>
<td>0.427</td>
</tr>
<tr>
<td>LSTM, Unweighted</td>
<td><strong>0.468</strong></td>
</tr>
<tr>
<td>LSTM, Half-Weighted</td>
<td>0.437</td>
</tr>
<tr>
<td>LSTM, Weighted</td>
<td>0.389</td>
</tr>
<tr>
<td>LSTM, Combined States</td>
<td><strong>0.651</strong></td>
</tr>
</tbody>
</table>

Logistically high probability threshold. However, there was not much loss of precision after this even though gains in recall were possible with lower and lower prediction thresholds. There was a larger area under the curve for Excited, indicating the best performance.

![Figure 8.4](image.png)

Figure 8.4 Precision-Recall curves for each emotional state. From left to right: Inhibited, Inactive/Relaxed, Excited, Activated with Stress.

It was difficult in general for a machine learning model to directly classify data into four emotional states. However, detection of specific stress indicators may be a promising avenue for investigation.

Head shaking has been a successful target of machine learning detection for collar-mounted triaxial IMU devices [Uij17], and in general detection of simpler behaviors have been explored with some success using triaxial IMU sensors [Kum18]. Scratching behaviors and body shaking may also be useful for detection. These behaviors are more specific and may produce a more consistent type of signal in the sensor data that a machine learning model can more easily learn and generalize from. If we can better detect stress indicators, then it is more likely that we can accurately identify stressful emotional states in the dogs.

8.5.3 Current Environment (Indoors, Outdoors)

The experts assessed the importance of the current environment metric as “not sure”. They mentioned that it would be helpful to know how much socialization is occurring both indoors and outdoors. In fact, they noted that if we can also refine the location by type, “such as a busy area, like a town setting that's pretty valuable” (Interviewee 1). A previous study [Cle19] presented in Chapter 6
has shown that statistical tests easily show a difference in indoor and outdoor environments through the collar's environmental sensors. However, classification of indoor and outdoor states in a variety of weather conditions requires more than testing differences in sensor data.

### 8.5.3.1 Data Collection

To generate labeled data, we collected collar data while the handlers and their dogs were both indoors and outdoors. Data were collected during the summer on several different days to increase the variety of environmental data. Handlers collected data with their dogs at home while their dogs were both indoors and outdoors. Handlers began data collection sessions in either environment (outdoors or indoors) and spent approximately 15 minutes with their dog in that environment. Then, they went with their dog to the other environment and spent 15 minutes there, producing data sets of similar sizes. Handlers noted the time at which they brought the dog indoors or outdoors to provide timestamp labels for the resulting data sets. In total, there were 9 data collection sessions from 3 dogs.

### 8.5.3.2 Building a Model

For input data, I used raw data from the light, pressure, temperature, and humidity sensors. Thus, each complete sample consisted of five features. Data were linearly interpolated where necessary due to differing sampling frequencies of the sensors. The total size of the resulting data set was 56095 samples.

I tested the performance of both a logistic regression model and decision tree model in generating this metric. The decision tree learning approach used the default parameters available for the DecisionTreeClassifier class in Scikit-Learn\[Ped11\]. To vary the data used for training, I used K fold cross validation, with $K=10$. In other words, 10 different subsets of the entire data set were used as test sets, and each time the rest of the data was used for training. Because the nature of the data collection process naturally used discrete sessions to collect data, all of the data for any given session was always kept completely in the training set or completely in the test set. I cordoned off the data sets in this way to prevent mixing of test and training information.

Logistic regression fits a linear model according to the training data in order to induce a binary classifier. I also used the Scikit-Learn library for this model. I compared both the decision tree and logistic regression approach with a baseline majority classifier. The majority classifier simply takes the training set data and always predicts the class that appears most frequently in the training data.

The results for both approaches using K fold cross validation are shown in Table 8.7. Majority classifiers and logistic regression scored the same across 10 folds. This suggests that a logistic regression model was unable to extract a meaningful pattern from the training data and instead guessed the majority class. However, decision trees performed exceptionally well without much tuning, with only a small number of mis-classified samples in the test set. This suggests that decision trees are a reliable method for generating the current environment metric. However, more data are needed from a wide variety of outdoor environments to strengthen this claim.
Table 8.7 Environment Classification Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean Val Accuracy</th>
<th>Mean Val Precision</th>
<th>Mean Val Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Classifier</td>
<td>0.615</td>
<td>0.629</td>
<td>0.408</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.615</td>
<td>0.629</td>
<td>0.408</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
</tbody>
</table>

8.5.4 Sound Origin (Dog, Environment)

Given a segment of collar data, sound origin classifies the sound signal when noise is present as primarily coming from the dog vs. coming from the environment. For this metric, the experts said that frequency of vocalizations (e.g. barking, whining) may be interesting to them, especially if paired with other data, such as vocalization while on outdoor socialization outings, which may correlate to dogs with traits such as excitability or fearfulness.

The experts noted the possibility of inadvertently detecting other dogs’ barks as coming from the dog wearing the collar and that this would diminish the quality of evaluations using these analytics. Understanding the difference is important—“If it barks a lot? That’s different than if they’re walking by dogs that are barking at them.” (Interviewee 1). They mentioned this as an issue specifically because puppies in raising regularly attend training classes with their raisers and come into close contact with each other. One of the experts noted that “we have dogs that bark a lot in classes and I think it’s a really undesirable guide dog behavior” and that it could “tell us something about the dog’s level of excitability and ability to self modulate” (Interviewee 1). They also pointed out that since barking is part of their Behavior Checklist (BCL), objective measures of barking would be insightful to compare with “what we marked down for them versus reality” (Interviewee 2).

8.5.4.1 Data Collection

To collect training data, we collected collar data as the dogs barked and whined while the handler coded the timestamp of the events. For this study, we focused on barking and whining (none of the dogs howled or growled and panting was ignored). This process produced labels so that machine learning models can learn to detect these noise events. We also included training and test data using incidents of other dogs barking nearby the collar to address the potential issue of false positives from other dogs’ noises, a potential challenge noted by Interviewee 1. In total, we collected 5 sessions of data, where each session consisted of roughly 15 bark or whine events.

8.5.4.2 Building a Model

To build this model, I used IMU and sound data as input features. I used IMU data as an input in attempt to capture the small movements in the dog’s body produced by or alongside the noisy behavior. One sample of data consisted of a number of consecutive IMU and sound samples and
the output label was a binary variable corresponding to whether or not the data contained a bark event. Thus, this was framed as a classification problem.

In general, classification of sound events is a well-studied problem, including the classification of dog noises such as bark events [San18]. However, sound classification tends to focus on the use of full sound information (both intensity and frequency) to classify sound events. In our system, the Smart Collar only detects sound intensity. This task (classifying canine sounds using only sound intensity) has been studied before by Kim et al., motivated to only use sound intensity levels due to its resource efficiency [Kim18]. They used sound data from dog vocalizations to classify them into four categories: barking, growling, howling, and whining. In this study, the interest is instead in differentiating ambient environmental noise and dog vocalizations using the Smart Collar, with a focus on barking and whining.

Raw sound data from the Smart Collar is not immediately interpretable as a sound level but instead as a voltage reading from the sensor. Sound analysis requires a preprocessing step in which the data is converted by, over short time windows, calculating the maximum voltage difference in samples over the windows. I calculated this for three window sizes (50ms, 100ms, and 150ms) to use as input features. In addition, I used the calculation for total IMU acceleration (total_acc) as in the Walking Speed metric. For each time stamp associated with the dog noise, I assigned a variable called “dog noise” with the value of 1 to every sample in the two second window before and after the recorded timestamp to ensure that the entire event was counted. Every sample without a noise event was assigned the value 0 to the “dog noise” variable, meaning “not dog noise”. Then each “dog noise” and “not dog noise” segment was separated. Finally, I also augmented the data due to the relatively low number of noise events in the data set compared to the background. The augmentation I used was, for each segment in the training set, I generated a set of random numbers from the normal distribution in the range -0.001 to 0.001 and added the noise terms to the sound segment. This resulted in double the number of segments in the training set. This process was always completed after data had been split into training and test sets so that, for example, an augmented and original version of a noise segment did not end up in both sets.

Figure 8.5 Plot of sound intensity for a typical bark on the left, no dog noise center, a “quieter” bark on the right.
Visually, barks and whines in the data set are easily seen as large peaks, so I extracted new features to capture that for a model to easily detect. I used the find_peaks function in the SciPy library [Vir20] to find the peaks of the sound signal with two thresholds (0.05 and 0.02). I then created a rolling calculation of, for any given timestamp, how many peaks were in the last 200ms and the average height of those peaks. This resulted in four extracted features: recent_peaks_05, recent_peaks_02, recent_peak_avg_05, and recent_peak_avg_02.

For machine learning models, I again used LSTM architectures. The models had a similar structure as the models used to estimate walking speed:

1. LSTM layer (input)
2. Dropout layer
3. 1 or 2 Fully Connected Layers (with dropout after first)
4. Sigmoid Layer (reduces output shape to 1 with value to range (0,1) for classification)

The notable difference here is that the final output layer is a Sigmoid layer with one unit. As before, the single neuron unit reduces the output shape to a single value and the use of a Sigmoid function restricts the output between 0 and 1, which represent each class ("not dog noise" and "dog noise"). The output therefore represents the estimated probability of a given segment of data representing part of a noise event.

After some initial experimentation, I used the same genetic algorithm used for Walking Speed to find the best set of hyperparameters for the machine learning model. The minimum and maximum values for the hyperparameters as well as the best values found are shown in Table 8.8. The fitness score I used in this process was the F-Measure (or “F₁ Score”), defined as \( \frac{2 \times p \times r}{p + r} \), where \( p \) is the precision and \( r \) is the recall of the model. The purpose of this is that we do not necessarily want to optimize the accuracy measure, due to the fact that the classes of the data set are very imbalanced. A model could achieve very accuracy by simply predicting “not dog noise” for every sample. Using the F-Measure allows us to optimize both the precision and recall measures with equal importance.

The genetic search found an architecture that was larger overall than the architecture found for Walking Speed, with 44 LSTM units and two dense layers with 93 and 78 units each. A diagram of the neural network architecture is shown in Figure 8.6.

Based on the optimized model architecture, I tested the performance on validation sets in K-fold cross validation (with \( K = 5 \) folds). Additionally, I tested three variations of class weighting since the classes were quite imbalanced. This allows the model to weight the loss function according to the class label and thus we can define an “importance” to that class, in this case assigning a much higher weight to that class. To weight the classes, I assigned the weight of \( \frac{1}{pos} \times \text{total} \) to the positive ("dog noise") class and \( \frac{1}{neg} \times \text{total} \) to the negative ("not dog noise") class, where \( pos \) and \( neg \) are the total number of examples in each class of the training set, and \( \text{total} \) is just \( pos + neg \). This resulted in much higher weights for the positive class. As a balanced approach between no weights and the full weighting method, I also included a model that used the above formula with the weights brought
Figure 8.6 Diagram of the neural network architecture for the *Sound Origin* metric

Table 8.8 Min, Max Values for Hyperparameter Optimization for Sound Origin; Best Values for each Feature

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Min</th>
<th>Max</th>
<th>Best Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning Rate</td>
<td>0.1</td>
<td>1.0</td>
<td><strong>0.74</strong></td>
</tr>
<tr>
<td><em>Rho (ADADELTA)</em></td>
<td>0.85</td>
<td>0.999</td>
<td><strong>0.895</strong></td>
</tr>
<tr>
<td>LSTM Units</td>
<td>4</td>
<td>200</td>
<td><strong>44</strong></td>
</tr>
<tr>
<td>Dense Layer 1 Units</td>
<td>1</td>
<td>300</td>
<td><strong>93</strong></td>
</tr>
<tr>
<td>Dense Layer 2 Units</td>
<td>0</td>
<td>150</td>
<td><strong>78</strong></td>
</tr>
</tbody>
</table>

closer to equal. These models used the weighting method with the negative weight multiplied by two and the positive weight divided by two.

The results for all three approaches are shown in Table 8.9. No approach outperformed the majority classifier in terms of accuracy due to the highly imbalanced classes. However, the majority classifier of course achieved no recall or precision (precision and thus the F-Measure are undefined if there are no predictions of the positive class). The approach that performed the best in terms of accuracy were the unweighted models. The weighted class approach achieved the highest precision but the unweighted approach achieved the highest recall. In fact, as the weighting increased, the precision increased while the recall decreased. The unweighted approach achieved the highest F-Measure score, suggesting that it was the best trade-off between precision and recall.

Examining the relationship between the precision and recall of the task further, Figure 8.7 displays the precision-recall (P-R) curve of the unweighted model. Since a classification model like this outputs the predicted *probability* of a sample being in a particular class, the PR curve shows how the precision and recall values change as the threshold varies between 0 and 1 on the validation set. As Figure 8.7 shows, the area under the curve is low, indicating that the model had trouble learning to accurately separate the dog’s sounds from the background, though high precision and high recall are possible with threshold values near 1 and 0, respectively. While some events were clearly distinguishable from background noise, some were quieter and did not produce as prominent of a peak in the sound signal. Additionally, some of the ambient noise occasionally produced large peaks in the sound signal that may have resembled a bark or a whine to the model. However, the
curve does show at what threshold the model performs better, which is where the point on the curve is closer to the top right-hand corner of the graph. Very high threshold values perform poorly, losing precision very quickly while gaining very little or no recall. Threshold values close to 0.5 perform the best according to this P-R curve.

Figure 8.7 Precision-Recall curve for optimized neural network model for Sound Origin metric

Table 8.9 Cross-validated Performance of Sound Origin Model with and without Class-Weighting

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Classifier Baseline</td>
<td>0.995</td>
<td>n/a</td>
<td>0.000</td>
<td>n/a</td>
</tr>
<tr>
<td>Neural Network, Unweighted</td>
<td>0.939</td>
<td>0.247</td>
<td>0.338</td>
<td>0.285</td>
</tr>
<tr>
<td>Neural Network, Half-Weighted</td>
<td>0.890</td>
<td>0.347</td>
<td>0.164</td>
<td>0.223</td>
</tr>
<tr>
<td>Neural Network, Weighted</td>
<td>0.832</td>
<td><strong>0.457</strong></td>
<td>0.127</td>
<td>0.199</td>
</tr>
</tbody>
</table>

8.5.5 Location Familiarity

Location familiarity describes how much a place has been visited by the dog. For example, if the dog has never been to a location before, the location familiarity corresponding with that set of collar data would be 0. This value is incremented by one every time the dog returns to that particular location (defined by clustering of GPS data). The experts assessed location familiarity as having “high” importance. They noted specifically that this measure could help answer important questions such as if there is a level of new locations visited that is associated with increasing dogs’ confidence, if more locations visited in general is better, and if there is an optimal level for more inhibited and more activated dogs. They noted that the dog experiencing new situations is “really invaluable to
“genetics” (Interviewee 1) especially if we go further and track “frequency, duration, and variety” (Interviewee 1) of new places. The experts said that generally they see better outcomes with raisers that frequently expose the dogs to new places and situations, but there are other factors at play. “So we don’t really know the answers, I’ve been dying for years to learn this” (Interviewee 1). We discussed privacy concerns arising from collecting GPS data. They mentioned that it is important for raisers to know when the system does and does not collect GPS data, pointing out that only collecting these data while there is an active Bluetooth connection to the collar “makes it a lot more reasonable” (Interviewee 3).

8.5.5.1 Data Collection

To generate data for building a model for this metric, only GPS data is required as input. Therefore, I generated data by physically taking the collar system to a set of predefined locations multiple times, noting the timestamps for when the collar system is active at each location. This data collection process was affected by the COVID-19 pandemic of 2020 in that the requirement to visit public locations was more difficult. Thus, I didn’t ask the data collectors to collect data for this metric. Unsupervised clustering has been used in a previous study to separate distinct locations using this type of data. I made use of unsupervised clustering to model locations and then used a simple algorithm to count locations (clusters) visited. Once locations are clustered, this algorithm separates unique visits to these locations by segmenting the timestamps of the data by location cluster. Each unique visit increments the location familiarity score for that place by one.

Data sets spanned several days involving trips to various locations (5 locations in total). I noted the timestamp of my arrival and departure to define the beginning and ending of the “visit” to that location. Number of visits per location and duration of visits varied. Statistics for number of visits and mean duration of each visit are shown in Table 8.10. Note that there is also in this data set a location 0. This denotes that the system was not in any of the 5 predefined locations during this time; here the system was in transit between locations and therefore location 0 is not a true location itself, but it may be interesting to examine how the algorithm performs in separating this from the “true” locations.

Table 8.10 Frequency and Duration of Location Visits

<table>
<thead>
<tr>
<th>Location</th>
<th>Number of Visits</th>
<th>Total Duration (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>85</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>98</td>
</tr>
</tbody>
</table>
8.5.5.2 Building a Model

To cluster the data by location, I used each recording of latitude and longitude as one sample of data. In total, there were 1281 samples. However, we first need to remove data that we can be confident that we are not interested in. To do this, I set a threshold for estimated speed to determine when the system is actually at a stationary location instead of traveling by vehicle. I calculated the estimated $\text{gps\_speed}$ as was done for the Walking Speed metric and removed all samples where the estimated speed (moving average over 15 seconds) was above 10 meters per second.

Next, I used K-Means clustering and found the optimal value for $K$ using the “elbow method”, as done previously in Chapter 7. The results, in terms of mean squared error (MSE), are shown in Table 8.8. This resulted in an optimal value of $K = 3$, though I explored other values of K for this metric as well.

![Figure 8.8 Sum of Squared Error as K increases for location data set.](image)

I used these clusters to define the model’s definitions of locations. The model assigned each sample a location number based on which modeled cluster it belonged to. This enabled the calculation of a location familiarity score. The first time a particular location was visited in the data set, the location familiarity score for that group of data was assigned to be 1. Then, each subsequent unique visit to that location increased the score by 1.

To evaluate the accuracy of this metric, I compared the generated location familiarity with the ground truth location visit data. To evaluate how well the model performed, I used the mean squared error between the true familiarity and the estimated familiarity. Performance metrics by location and aggregated over the whole data set are shown in Table 8.11 using both the optimal value of K according to the process described above ($K = 3$) and also the true number of locations ($K = 5$). Both models correctly calculated the familiarity of locations 2,3, and 4, which were visited once. The $K = 3$ model scored better on location 0 (in transit) but otherwise the $K = 5$ model performed the best, including on the data set as a whole.
A plot of errors as the value of K increases is shown in Figure 8.9. The minimum error is achieved by models based on both 4 and 5 clusters. The algorithm was unable to minimize the error with only $K = 3$. While the resulting location familiarity score was reasonable, the best possible model was found using the known number of locations in the data set. A more flexible method of generating clusters is needed to deploy this with live collar systems so that no labeled data is needed.

![Figure 8.9 Mean Squared Error of Location Familiarity Model as K Increases.](image)

<table>
<thead>
<tr>
<th>Location</th>
<th>Mean Squared Error $K=3$</th>
<th>Mean Squared Error $K=5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4.959</td>
<td>5.184</td>
</tr>
<tr>
<td>1</td>
<td>0.150</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>4</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>1.901</td>
<td>1.000</td>
</tr>
<tr>
<td>All</td>
<td>2.000</td>
<td>1.635</td>
</tr>
</tbody>
</table>

### 8.5.6 Reaction to New Environment

Reaction to new environment describes how a dog reacts when it is in an environment it has not yet been exposed to. This metric received a rating of “high” importance. The experts noted this would be
important especially in terms of identifying excitable puppies in the program. They again expressed the difficulty in assessing a dog's behavioral state, such as in differentiating between a dog that is still and relaxed and a dog that is inhibited with stress. This is a derived metric that would be calculated from the current emotional state and location familiarity metrics and thus the accuracy depends on the accuracy of these two other models. Because this metric depends on the two aforementioned metrics, the input for this model can be considered to be the underlying input for these two metrics, which are IMU, sound, and GPS data.

Collection of training data for reaction to new environment itself is much more difficult due to the need for video with which to code behaviors specifically when dogs are in new environments. In normal circumstances, data collection at an adequate scale would be a challenge, but this is exacerbated by the ongoing COVID-19 pandemic. Data collection and experimentation for the reaction to new environment metric is left as future work.

8.6 Conclusion

The Smart Collar system collects from any given dog a large amount of raw sensor data. These raw sensor data alone may be difficult for guide dog organizations to interpret without context. To provide more insight into the progress of dogs in raising, I defined several higher-level metrics that describe more abstract aspects of their behaviors and environments. I conducted an interview with four guide dog experts at a guide dog school (Guiding Eyes for the Blind) to understand if these and other potential metrics would be useful to understand how any particular dog is progressing.

The interview provided a wealth of information about what guide dog experts need from wearable sensing devices in order to understand the progress of their dogs in raising. While the concept of using Smart Collar sensor data to extract more abstract measures of the dog's behavior and environment is still exploratory, the experts found several of the metrics to be useful indicators of characteristics that they are interested in. For example, the dog's walking speed during routine exercise walks may correlate with the dog's adult working speed as a guide dog, which can be studied if such a metric is accurately generated using data from raisers. I facilitated the collection of training data for five of the potential metrics. Using the data, I derived machine learning models to output the higher-level metrics from the raw collar data. I found that for some of the metrics, high accuracy was easily achieved, while the performance of the models on some of the metrics was much lower. Larger data sets will likely alleviate this, and more data from a variety of dogs and situations will increase the robustness of these models.

Future work should include expansion of the set of metrics explored, such as with the metrics that I discussed with Guiding Eyes but did not gather data for: reaction to new environment and stairs incidents. Future research should also work closely with guide dog experts to define what metrics would provide the best insight into the dog's development. Existing metric models can be improved and made more robust with more data and a wider variety of data. The metrics that make use of deep learning models will likely benefit in terms of accuracy from both additional training.
data and further exploration of the model architectures.
This chapter discusses my conclusions and summarizes my research contributions. I then describe avenues for future work to build on the completed research.

### 9.1 Conclusion

Producing successful guide dogs is quite expensive and the number of dogs who actually succeed in training is relatively low. During the months-long raising period of guide dog training, candidate guide dogs develop greatly and there is great potential in capturing and analyzing quantifiable data about the puppies' environments and behaviors in this phase. The Smart Collar system was developed to address this gap of data.

I validated the data produced by the Smart Collar system to ensure that the raw data was useful for analysis. This consisted of statistical tests to confirm that differences could be seen in the data between two environments and three behaviors. After our initial deployment, I used the resulting data for further analysis, this type investigating if there were statistical differences in each sensor data type when separating the data by geographical location. I used unsupervised clustering techniques to separate by location and compared the data between the clusters. I found statistically significant differences in the means and standard deviations for several of the data sets that I analyzed. This was also the first study in which the Smart Collar system was used with external beta testers. Because of this, we obtained feedback about actual raisers regarding their experiences in using the system.

Next, I defined several higher-level metrics to give further context to the raw sensor data. The benefit of this is to enable automated reporting for each dog so that Guiding Eyes can better understand how each dog in the program is progressing. I interviewed four experts on guide dog
training at Guiding Eyes about their opinions on the metrics. They provided me with feedback and new potential metrics to investigate. After incorporating the feedback, I conducted a study in which dog owners in our team and Guiding Eyes staff both collected and labeled data to enable machine learning models to generate five of these metrics. Accuracy of the models varied between the metrics, but overall showed promising results and supported the case for deriving higher-level metrics in order to summarize and report on Smart Collar data. Accurate summaries and detection of important behaviors would allow Guiding Eyes and other guide dog organizations to have up to date remote observation of dogs when in critical periods of behavioral development.

As a reminder, the research questions my work sought to address are:

1. How can we build a system that adequately and non-invasively captures dog behavior, without putting undue burden on the volunteer raisers?
2. How can we collect data in a way that allows us to gain insight into how the dog reacts to stimuli and spends its time during the raising period?
3. How can we derive higher level metrics describing the dog’s current environment and reaction to stimuli?
4. How can the collected objective sensor data be used to model the dog’s behavior and likelihood of success?

The COVID-19 pandemic of 2020 greatly impacted my work and created delays in manufacturing of Smart Collar devices and in our ability to collect data in the Guiding Eyes raiser program. Thus, RQ4 is left as future work, discussed in the next section. Otherwise, the aforementioned research questions were addressed by various research contributions over the course of my work, listed here with the corresponding research questions:

- The design of a mobile application that collects data from a wearable IoT device (RQ1, RQ2)
- Design considerations to minimize undue burden on volunteer raisers that need to concentrate on their dog (RQ1)
- Metrics derived from raw sensor data that are useful to guide dog organizations in remotely understanding the behavior and environment of the dogs (RQ2, RQ3)

### 9.2 Future Work

There are several avenues for next steps in this work. In this section I describe some potential future work to build on my contributions and improve the system.
9.2.1 Software Features

In Chapter 8 I used two versions of the mobile app. In both, the user is able to log the beginning and end of special events such as exercise walks to give further context to collected data. In the special version for collecting emotional state data, the user can log each of the four predefined emotional states. In the future, more types of events should be included in this feature. There are likely different ways in which these data should be reported by the user. Long duration events, such as an exercise walk, should allow the user to tag the start and end of the events so that the event data consists of a time duration. Shorter events, or events that may require the raiser's attention, such as meeting a new dog, may be best suited for only logging the event as a single timestamp, but future research will have to investigate this, possibly with an A/B test of the feature.

Additionally, quality of life features for the mobile application will improve the user's experience, such as more transparent information about how much data is currently stored on the collar as well as the smartphone, a cleaner user interface, and notification when the collar becomes disconnected. Finally, an Android version of the app will allow a larger segment of the raiser population to participate in the data collection process. Much of the code is reusable across platforms as a result of being developed with Xamarin.

9.2.2 Metrics and Reporting

The study described in Chapter 8 is a starting point for analytics based on the Smart Collar data. Next steps should focus on three areas: new metrics, more robust modeling of existing metrics, and an automated reporting platform.

The set of metrics I explored can be expanded, including with the metrics that I discussed with Guiding Eyes but did not gather data for: reaction to new environment and stairs incidents. Future research should work closely with guide dog experts to define what metrics would provide the best insight into the dog's development.

Next, existing metric models can be improved and made more robust. For example, the current environment metric should include cold weather data in order to increase the accuracy of the model in a variety of weather environments. Additional inputs such as the date of a recording could also be used to improve the accuracy in various weather conditions. The metrics that make use of deep learning models will likely benefit in terms of accuracy from both additional training data and further exploration of the model architectures, with more training data being a priority.

Finally, the generated metrics can be displayed to guide dog training experts through a reporting platform. With raw data augmented by generated metrics, we can answer more analytical questions about dogs in raising, such as how various analytical measures correlate with each other and with other measures that Guiding Eyes is interested in. Ideally, such a reporting platform would have a visual interface that would give technical and non-technical users alike a more complete picture of the progress of a candidate guide dog and if interventions are needed to improve its chances of success.
9.2.3 Prediction of Outcomes

Prediction of future outcomes based on Smart Collar data is another potential future goal for analytics with the Smart Collar system, as I did not address this directly (RQ4). The proportion of successful dogs in the In-For-Training test at the end of raising is about 50%. Identifying successful and unsuccessful dogs while still in raising is an open problem.

Using the derived metrics as extracted features in the prediction process may improve the ability to predict success and failure, but there are a number of challenges. Any given dog produces a massive amount of Smart Collar data in a variety of situations. How to separate these data into machine learning examples is one such challenge. Another is that of which features to use. It may be that some combination of raw data and derived features such as summaries or generated metrics are most predictive. Additionally, this problem will require data collection from a large number of dogs. The system is designed with scale in mind, but scalability will need to be tested and likely refined, and coordination with hundreds of raisers will be another challenge. However, using sensor data to increase the percent of successful dogs by screening out dogs that will not succeed will greatly benefit guide dog schools in terms of monetary and time resources, and will help those unsuccessful dogs in finding more suitable roles (such as other service work or as pets).
BIBLIOGRAPHY


Timescale. TimescaleDB. 2015.


Described here is the data collection protocol used to collect and code training data for the emotional state metric (found in Chapter 8). Helen West primarily developed the protocol after my discussions with Guiding Eyes for the Blind about the requirements needed to collect training data for the metric.

1. Expose to novel object (owl statue) – Dog to experience novel object suddenly. Allow dog to react without interaction from handler with a loose leash.

2. Walk over upside down bread crate – Allow dog to walk over bread crate, if dog avoids ask dog again with hand gesture across the top of the crate, if dog refuses use food motivator

3. Sit for high value food reward (e.g.pupperoni) – Handler asks dog to sit 3 times and rewards with higher value food.

4. Sit on bench/chair and allow dog to settle without interaction for 5 minutes – handler sits on bench/chair with dog on shorter leash and does not interact with it for 5 minutes. Ensure there are no items to scavenge near this area

5. Someone other than handler walks up to dog to say hi and pet – helper approaches dog with happy upbeat voice. Do not touch dog when overly excited. Handler to use minimal interaction with the dog.

6. Expose to metal can which is shaken to make noise – Dog approaches helper, when dog is 6ft
away handler shakes can 5 times. Allow dog to react without interaction from handler with a loose leash.

7. Handler conducts physical exam in the manner of a vet exam – handler examines ears, eyes, teeth, paws and tail then gives dog a 5 second hug.

8. Sit on bench/chair and allow dog to settle without interaction for 5 minutes - handler sits on bench/chair with dog on shorter leash and does not interact with it for 5 minutes. Ensure there are no items to scavenge near this area.

9. Someone other than handler conducts physical exam in the manner of a vet exam – helper examines ears, eyes, teeth, paws and tail then gives dog a 5 second hug.

10. Handler plays with dog with favorite toy – handler engages the dog to play with a toy that the dog likes to play with for one minute.

11. Exposure to another dog, first at a distance then moving closer – helper brings 2nd dog into view 20ft away, allow reaction without handler interaction. Helper brings 2nd dog to 10 ft away, allow reaction without handler interaction. Move 2nd dog away after a maximum of 20 seconds.

12. Sit on bench/chair and allow dog to settle without interaction for 5 minutes - handler sits on bench/chair with dog on shorter leash and does not interact with it for 5 minutes. Ensure there are no items to scavenge near this area.