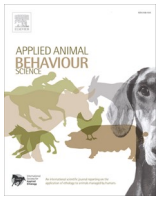
Applied Animal Behaviour Science 287 (2025) 106632

Contents lists available at ScienceDirect

Applied Animal Behaviour Science

journal homepage: www.elsevier.com/locate/applanim



Prediction of assistance dog training outcomes using machine learning and deep learning models

Mohammad Hossein Amirhosseini [a], James A. Serpell [b,*], Emily E. Bray [c,d], Theadora A. Block [c], Laura E.L.C. Douglas [c], Brenda S. Kennedy [c], Katy M. Evans [e,f], Kathleen Freeberg [e], Piya Pettigrew [g]

[a] Department of Computer Science and Digital Technologies, School of Architecture, Computing and Engineering, University of East London, London, UK
[b] School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA, USA
[c] Canine Companions for Independence, Santa Rosa, CA, USA
[d] College of Veterinary Medicine, University of Arizona, Tucson, AZ, USA
[e] The Seeing Eye, Inc., Morristown, NJ, USA
[f] School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonnington, Leics, UK
[g] Dogvatar, Inc., Miami, FL, USA

ARTICLE INFO

Keywords:
Machine learning
Deep learning
C-BARQ
Dog training
Predictive modeling
Behavior assessment
SVM
XGBoost
Working dogs

ABSTRACT

This study investigates the predictive power of machine learning and deep learning models for forecasting training outcomes in assistance dogs, using behavioral survey data (C-BARQ) collected from volunteer puppy-raisers at two developmental stages: 6 months and 12 months. We used data from two assistance dog training organizations–Canine Companions and The Seeing Eye, Inc.– to assess model performance and generalizability across different training contexts. Six models, including traditional machine learning approaches (SVM, Random Forest, Decision Tree, and XGBoost) and deep learning architectures (MLP and CNN), were trained and evaluated on C-BARQ behavioral scores using metrics such as accuracy, F1 Score, precision, and recall. Results indicate that Support Vector Machine (SVM) and XGBoost consistently delivered the highest prediction accuracy, with SVM achieving up to 80 % accuracy in the Canine Companions dataset and 71 % in the Seeing Eye dataset. Although deep learning models like CNN showed moderate accuracy, traditional machine learning models excelled, particularly in structured, tabular data where feature separability is essential. Models trained on 12-month data generally yielded higher predictive accuracy than those trained on 6-month data, highlighting the value of extended behavioral observations. This research underscores the efficacy of traditional machine learning models for early-phase prediction and emphasizes the importance of aligning model selection with dataset characteristics and the stage of behavioral assessment.

1. Introduction

The domestic dog's extraordinary diversity of breeds and types reflects a long history of human selection for behavioral attributes that enhance the ability of these animals to perform specific working roles, ranging from hunting, guarding, herding and detection work to the provision of companionship and social support (Dutrow et al., 2022; Hall et al., 2021; Serpell, 2021). As the global demand for selectively bred and trained working dogs continues to grow, organizations and agencies dedicated to producing these animals are increasingly focused on ways of improving their selection and training (Bray et al., 2021). According to organization estimates, the average cost to produce most trained working dogs ranges from $40,000 to $75,000 USD (Cleghern et al., 2018; Mercato et al., 2022), a significant portion of which can be attributed to the effort and resources invested in the production of dogs that ultimately fail to become successful in their assigned roles. For example, data derived from assistance (guide and service) dog organizations suggest that 50–70 % of dogs are released and re-homed because they are deemed unsuitable for assistance work (Duffy and Serpell, 2012; Harvey et al., 2017). Behavioral issues of one kind or another account for the majority (63–87 %) of these releases (Duffy and Serpell, 2012). Given these high costs and losses, working dog organizations and

* Corresponding author.
E-mail address: serpell@vet.upenn.edu (J.A. Serpell).

https://doi.org/10.1016/j.applanim.2025.106632
Received 5 January 2025; Received in revised form 10 April 2025; Accepted 15 April 2025
Available online 21 April 2025

agencies have a strong interest in being able to use behavioral characteristics to predict an individual dog's likelihood of being successful as early as possible in its development (Bray et al., 2021).

A variety of different methods and approaches have been used to predict future performance in working dogs, including behavioral test batteries administered at various ages. Several of these studies across multiple working dog programs have confirmed that numerous behavioral attributes are correlated with successful program completion and may provide useful indicators of training potential, even as early as eight weeks of age (Arata et al., 2010; Asher et al., 2013; Batt et al., 2008; Brady et al., 2018; Bray et al., 2017b, 2019, 2021; Goddard and Beilharz, 1986; Harvey et al., 2016; Lazarowski et al., 2021; MacLean and Hare, 2018; Sinn et al., 2010; Svobodová et al., 2008; Tiira et al., 2020; Tomkins et al., 2011; Wilsson and Sundgren, 1997).

To enhance early socialization, most working dog organizations tend to foster their puppies with volunteer puppy raisers who care for the dogs from about 7–8 weeks of age. Foster homes are often widely dispersed geographically, thus rendering these dogs relatively inaccessible for direct behavioral observation or testing until they return to their host organizations to be professionally trained at around 14–18 months old. To overcome the difficulty of evaluating dogs during this period, several studies have used surveys or questionnaires to collect behavioral assessments from working dog puppy raisers (Bray et al., 2019; Cleghern et al., 2018; Duffy and Serpell, 2012; Foyer et al., 2014; Serpell and Hsu, 2001). The most widely used and validated survey of this type is the Canine Behavioral Assessment & Research Questionnaire (C-BARQ), developed at the University of Pennsylvania (Hsu and Serpell, 2003; Duffy and Serpell, 2012). Among working dog organizations, C-BARQs are usually completed by puppy raisers serially when their dogs are 6 and 12 months old.

Previous attempts to use C-BARQ scores to predict training outcomes have been moderately successful. One study, conducted in 5 different assistance dog organizations, investigated whether a dog's success (being placed with a handler or selected as a breeder) or release from the program (for behavioral reasons) could be predicted based on its earlier C-BARQ scores. Guide and service dogs that successfully completed their training obtained more favorable scores on 27 out of 36 C-BARQ behaviors, and 'pulling excessively hard on the leash' was the most highly predictive trait for failure. Logistic regression models indicated that the overall C-BARQ evaluations were able to discriminate between successful and unsuccessful dogs significantly above chance levels (areas under the ROC curves 0.64–0.72) (Duffy and Serpell, 2012). A subsequent study within a single service dog organization obtained similar results (AUC of 0.71) but also determined that C-BARQ evaluations were more accurate at identifying dogs with the lowest probabilities of being successful (85–92 % accuracy) compared with the most successful dogs (62–72 % accuracy) (Bray et al., 2019). Both these studies included demographic and contextual covariates such as breed, sex (Duffy and Serpell, 2012), coat color, training region, and year of assessment (Bray et al., 2019) in their models. The primary research objective of the current study was to determine whether AI models (machine and deep learning) can improve on these earlier attempts to predict training outcomes in assistance dogs using only data derived from C-BARQ assessments.

To date, few studies have implemented predictive AI modelling approaches to forecast outcomes in dogs, relying instead on traditional statistical analyses (but see Panthirana and Balalle, 2024). The current study seeks to address this gap by introducing machine learning and deep learning approaches to predict training outcomes in assistance dogs based on C-BARQ data. The predictive ability of the developed models is evaluated at two distinct stages of puppy raising—when the dogs are 6 months and 12 months old—to determine if training outcomes can be accurately forecasted relatively early in development. This research also compares the effectiveness of AI models on datasets from two distinct working dog organizations that breed and train different categories of assistance dogs (service vs. guide) to assess the generalizability of the approach. While previous studies have established correlations between C-BARQ scores and training success in working dogs using traditional statistical approaches, our goal here was to determine whether AI models could enhance predictive accuracy and provide a more automated, flexible, and scalable solution for the early identification of successful assistance dog candidates.

2. Materials and methods

2.1. Data collection

Datasets from two distinct assistance dog training organizations were used to enable a comparative analysis and to assess the generalizability of AI-based predictive approaches across different sources. Additionally, by analysing data collected at different life stages (6 and 12 months), we evaluated the feasibility of predicting training outcomes earlier in the puppy-raising process, offering insights into the model's reliability in detecting training potential at earlier developmental phases.

The first dataset was provided by Canine Companions, a U.S.-based nonprofit organization that breeds, trains, and places service dogs free of charge with individuals with disabilities–including adults, children, and veterans–as well as facility dogs for professionals in healthcare, criminal justice, and educational settings. This dataset contains demographic information for each dog, such as breed, age, and sex, along with raw behavioral scores from the comprehensive 100-item Canine Behavioral Assessment & Research Questionnaire (Supplementary Table S1, see Hsu and Serpell, 2003; Duffy and Serpell, 2012). These C-BARQ assessments, completed by puppy raisers, capture quantitative behavioral information from when the dogs are approximately 6 and 12 months old and contain 7627 records from dogs born between the years 2005 and 2023. The outcome variable captures three classes of training results: *Graduate* (indicating that a dog successfully completed training and was assigned a working role), *Released* (indicating that a dog was released from training due to behavioral reasons), and *Breeder* (indicating that a dog enrolled in the breeding program). The *Graduate* category at Canine Companions includes service dogs serving a variety of different roles—e.g., dogs placed with adults with physical or auditory disabilities, children with developmental disabilities, veterans with post-traumatic stress disorder, and facility dogs. Approximately 43 % of the dogs in this sample graduated successfully, 52 % were released for behavioral reasons, and 5 % became breeders. Dogs released for medical reasons were excluded from the dataset since our focus is on predicting behavioral success.

The second dataset was provided by The Seeing Eye, Inc., another US-based nonprofit organization that breeds, trains, and places guide dogs with blind and partially sighted individuals. This dataset includes background details for each dog, including age, sex, and breed, as well as raw behavioral scores on the 100-item Canine Behavioral Assessment & Research Questionnaire (C-BARQ) completed by puppy raisers when the dogs were 6 and 12 months old (Duffy and Serpell, 2012). The dataset contains 7719 records from dogs born between the years 2004 and 2022, and the outcome variable captures the same three classes of training results—*Graduate*, *Released*, and *Breeder*—as in the first dataset. The *Graduate* category at The Seeing Eye includes any dogs assigned to be guide dogs for blind or partially sighted persons aged 16 or older. Approximately 49 % of this sample graduated successfully, 45 % were released for behavioral reasons, and 6 % became breeders. As above, dogs released for medical reasons were excluded from the dataset.

2.2. Data cleaning and feature engineering

All demographic features such as age, sex, and breed were excluded from both datasets to ensure that training outcome predictions were based solely on C-BARQ behavioral features. Analyses were performed separately on the Canine Companions and Seeing Eye datasets, but the same methods were applied to each. Four specific C-BARQ items

(AGG32, AGG33, AGG34, and AGG35) were excluded from both datasets for the 6-month and 12-month age groups due to a high volume of missing data (>20 %) in the Seeing Eye data. These items assess a dog's aggressive behavior toward other dogs in the same household and were frequently left incomplete by puppy raisers who did not have another dog in the household. While the level of missingness for these variables was < 20 % in the Canine Companions data, the same variables were excluded to facilitate comparisons of the two sets of data.

To handle missing values among the remaining variables in both datasets and to ensure that all samples were available for training and testing machine learning models, we applied mean imputation in which each column is replaced with the mean of the observed values within that column, thus preserving the datasets' original sample sizes. This method is effective when data is missing at random, as it provides an unbiased estimate by maintaining the mean of the data. As the target variable (training outcomes) in both datasets was categorical, we employed OneHotEncoder() method to convert the target into a numerical format. This method ensures that categorical labels are represented appropriately without imposing ordinal relationships between the categories. This transformation enables compatibility with machine learning algorithms.

To manage label imbalance in both datasets and mitigate overfitting risks during model training, we employed the Synthetic Minority Oversampling Technique (SMOTE). Unlike traditional methods that replicate minority samples, SMOTE generates synthetic samples by interpolating between existing instances of the minority class. This approach enhances the representation of the minority class without merely duplicating existing data points (Elreedy, Atiya, and Kamalov, 2024). Specifically, for each minority class sample x_i, SMOTE selects one of its K-nearest neighbors x_{nn} and generates a synthetic sample x_{new} positioned between x_i and x_{nn}. This process follows the formula:

$$x_{new} = x_i + \lambda \times (x_{nn} - x_i)$$

where λ is a random value between 0 and 1, creating new points that maintain the distribution of minority class features. By enriching the feature space representation for the minority class, SMOTE enhances class balance, which can be especially beneficial for algorithms sensitive to imbalance. This method helps in reducing model bias toward the majority class, thereby improving classification accuracy and generalizability (Elreedy, Atiya, and Kamalov, 2024).

2.3. Feature selection

Training outcomes in both datasets were determined by the organizations based on the dog's behavior as they progressed through the training programs. In Experiment 1, we utilized the C-BARQ scores recorded by puppy-raisers at 12 months of age to train and test machine learning models and predict training outcomes, allowing for a comparison of prediction results across the two datasets. In Experiment 2, we used the C-BARQ scores recorded by puppy-raisers at 6 months of age to assess whether training outcomes could be reliably predicted at an earlier life stage, again comparing prediction results between datasets based on the 6-month scores.

Principal Components Analysis (PCA) was also applied to both datasets to achieve dimensionality reduction while retaining 95 % of the dataset's variance. PCA achieves this by transforming the original dataset into a series of uncorrelated variables, termed principal components, which prioritize capturing the highest variance. Specifically, PCA calculates the covariance matrix of the dataset X and performs an eigenvalue decomposition to obtain eigenvectors (principal components) and eigenvalues, which indicate the variance captured by each component (Härdle et al., 2024). Selecting the principal components with the highest eigenvalues ensures that they explain a cumulative threshold of variance—here, set at 95 %. Mathematically, this relationship can be expressed as:

$$Variance\ explained = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

where λ_i denotes the eigenvalues, and k represents the number of selected components. By retaining only the leading components necessary to reach this variance threshold, PCA reduces feature count while maintaining the critical dataset structure, aiding in preventing overfitting and enhancing computational efficiency in subsequent model training (Härdle, Simar, and Fengler, 2024).

2.4. Preparing training and testing sets

As Python programming language was used for implementation, the train_test_split() function from the sklearn library was used to split the pre-processed datasets randomly into 80 % training and 20 % testing sets, as recommended.

2.5. Developing machine learning models

2.5.1. Experiment 1: 12-month C-BARQ scores

To predict training outcomes (i.e., *Graduate*, *Released*, and *Breeder*) in the Canine Companions and Seeing Eye datasets based on 12-month C-BARQ scores, four machine learning and two deep learning models were developed and evaluated: (1) Support Vector Machine (SVM), (2) Random Forest, (3) Decision Tree, (4) XGBoost, (5) Multi-Layer Perceptron Neural Network (MLP), and (6) Convolutional Neural Network (CNN). Each model underwent systematic hyperparameter tuning using RandomizedSearchCV to identify the best-performing parameter configurations. This method was selected because it allows for efficient hyperparameter optimization by sampling from a defined distribution rather than exhaustively searching all possible combinations. This approach significantly reduces computational cost while maintaining high-quality model performance (see Supplementary Materials).

The identified optimum values for each model's parameters in the Canine Companions and Seeing Eye datasets are presented in Supplementary Table S2.

Optimal hyperparameter values frequently vary across models when applied to different datasets. As presented in Table S2, tuning the hyperparameters of machine learning models resulted in slightly different optimal values across the two different datasets. This variation is driven by differences in data characteristics, such as feature distributions, relationships, and sample size. Hyperparameter tuning, therefore, aims to adapt models to these unique data features, with results showing distinct optimal values for parameters like 'C' and 'gamma' in the Support Vector Machine model, 'min_samples_split' in the Random Forest model, and 'max_depth', 'min_samples_leaf', and 'min_samples_split' in the Decision Tree model. These adjustments enhance the models' performance by better aligning them with the inherent structures and complexities of each dataset.

In deep learning models, hyperparameters such as learning rate, activation functions, and layer configurations tend to be more stable across different datasets. This is because deep learning models excel at learning hierarchical, complex feature representations directly from data, which can make them inherently more adaptable to diverse data structures. Once hyperparameters are fine-tuned on one set of training data, these models often generalize well, requiring fewer adjustments to perform effectively on similar tasks with different datasets. This adaptability contrasts with traditional machine learning models, where hyperparameters are often more sensitive to changes in dataset characteristics. We confirmed the stability of hyperparameters, achieving optimal predictive performance across both datasets using consistent parameter values.

2.5.2. Experiment 2: 6-month C-BARQ scores

In the second experiment, we used the same machine learning and

deep learning models to predict training outcomes in the Canine Companions and Seeing Eye datasets based on 6-month C-BARQ scores. Similar to experiment 1, the hyperparameter tuning was performed using RandomizedSearchCV to identify the best-performing parameter configurations. The identified optimum values for each model's parameters in this experiment are presented in Supplementary Table S3. The hyperparameter values that were considered in RandomizedSearchCV for each model and parameter are the same as in Experiment 1 and are presented in Supplementary Table S4.

3. Results

3.1. Experiment 1 results: performance of the models based on 12-month C-Barq scores

The performance of the models that were trained based on the 12-month C-BARQ scores was evaluated using 5-Fold Cross Validation. Accuracy (overall proportion of correct predictions), F1 Score (harmonic mean of precision and recall), Precision (proportion of positive predictions that are actually correct), Recall (a.k.a. Sensitivity; the proportion of actual positive cases correctly classified), and Specificity (the proportion of negative cases correctly classified) metrics were generated in this evaluation phase to enable a thorough comparative analysis. Table 4 presents the results for each dataset following the 5-Fold Cross Validation process.

As shown in Table 4, the Support Vector Machine (SVM) and XGBoost models demonstrated superior performance in comparison to the other models, though each model exhibited unique strengths and limitations across the two datasets.

In the Canine Companions dataset, the SVM model achieved the highest overall performance, scoring an accuracy, F1 Score, precision, and recall of 0.80 across the board, indicating balanced performance with consistent predictive strength. The SVM model also achieved an impressive level of specificity (0.90) for the 12-month data. XGBoost closely followed SVM, with slightly lower metrics but still demonstrating solid classification ability at 0.77 accuracy and with a 0.76 F1 Score. Random Forest and neural network models (Multi-Layer Perceptron and CNN) showed moderate performance, with accuracies around 0.74–0.75. The Decision Tree model lagged behind, achieving the lowest accuracy of 0.63, highlighting its limitations in handling the complexity of the data when compared to more sophisticated models like SVM and XGBoost. Our findings suggest that ensemble methods and deep learning architectures may better capture the nuanced patterns in the Canine Companions dataset.

In the Seeing Eye dataset, all models generally exhibited lower performance compared to the Canine Companions dataset. SVM remained the best-performing model, achieving 0.71 across all metrics and a specificity of 0.84, followed by XGBoost with an accuracy of 0.69. Multi-Layer Perceptron achieved moderate accuracy score of 0.67. CNN and Random Forest models had relatively low performance, with accuracy scores of 0.64 and 0.65, respectively. Similar to the experiment on the Canine Companions dataset, the Decision Tree model lagged behind, achieving the lowest accuracy of 0.54, indicating its limitations in handling the complexity of the data when compared to other models. The overall lower performance in the Seeing Eye dataset suggests that the feature set may not capture the behavioral nuances as effectively in this context, or the data distribution might be more challenging for the models.

In summary, SVM and XGBoost consistently performed well across both datasets, with SVM excelling in the Canine Companions dataset and providing stable performance in the Seeing Eye dataset. The results suggest that SVM and XGBoost are particularly effective for this classification task, likely due to their ability to model complex decision boundaries and account for non-linear relationships in the data. The neural networks (Multi-Layer Perceptron and CNN) performed moderately well, indicating potential but perhaps a need for further tuning or additional layers to improve generalization. The Decision Tree model, while interpretable, showed the weakest performance, which suggests that it may struggle with complex, high-dimensional datasets. The findings indicate that ensemble and SVM models are more robust choices for predicting behavioral outcomes based on 12-month C-BARQ scores.

In addition to accuracy-based evaluation metrics, we also assessed the confidence of the best-performing model's predictions using the probability scores generated by the classifier. These scores reflect the model's certainty in its classification decisions and were calculated as the predicted probability for the assigned class using the predict_proba() method. It is important to distinguish between accuracy score and confidence score: while accuracy represents the overall proportion of correct predictions across the test set, confidence scores reflect how certain the model is about individual predictions. A model can have high accuracy overall but still be uncertain about specific classifications—confidence scores provide insight into that nuance.

The average confidence score for the SVM model across all test samples was 80.29 % for the Canine Companions dataset and 70.37 % for the Seeing Eye dataset, indicating a generally high level of certainty in the model's predictions, particularly for the Canine Companions sample. These confidence scores further reinforce SVM's status as the most robust and reliable model in this experiment, providing not only accurate but also consistent and confident predictions.

3.2. Experiment 2 results: performance of the models based on 6-month C-Barq scores

The performance of the models trained using the 6-month C-BARQ scores was evaluated following the same 5-Fold Cross Validation methodology as applied in the previous experiment. Accuracy, F1 Score, Precision, Recall and Specificity were again utilized to enable a detailed comparative analysis. Table 5 presents the results for each model, tested across both the Canine Companions and Seeing Eye datasets.

According to Table 5, the Support Vector Machine (SVM) and XGBoost models once again demonstrated strongest performance, while other models exhibited limitations. Differences between the Canine Companions and Seeing Eye datasets were also evident, reflecting dataset-specific challenges in predictive modelling.

In the Canine Companions dataset, the SVM and XGBoost models both achieved the highest accuracy scores of 0.75, with balanced F1 Scores, precision, and recall, suggesting their robustness in capturing the

Table 4
Experiment 1: 5-Fold Cross Validation Results for Models Based on 12-Month C-BARQ Scores.

ML Model	Canine Companions					Seeing Eye				
	Accuracy	F1 Score	Precision	Recall	Specificity	Accuracy	F1 Score	Precision	Recall	Specificity
SVM	0.80	0.80	0.80	0.80	0.90	0.71	0.71	0.71	0.71	0.84
Random Forest	0.75	0.74	0.74	0.75	0.88	0.65	0.64	0.64	0.65	0.83
Decision Tree	0.63	0.62	0.63	0.62	0.82	0.54	0.53	0.52	0.54	0.77
XGBoost	0.77	0.76	0.76	0.77	0.89	0.69	0.69	0.67	0.69	0.84
Multi-Layer Perceptron Neural Network	0.74	0.74	0.74	0.74	0.87	0.67	0.65	0.66	0.67	0.84
Convolutional Neural Network (CNN)	0.74	0.73	0.74	0.73	0.87	0.64	0.63	0.63	0.64	0.82

Table 5
Experiment 2: 5-Fold Cross Validation Results for Models Based on 6-Month C-BARQ Scores.

ML Model	Canine Companions					Seeing Eye				
	Accuracy	F1 Score	Precision	Recall	Specificity	Accuracy	F1 Score	Precision	Recall	Specificity
SVM	0.75	0.74	0.75	0.75	0.88	0.70	0.69	0.67	0.70	0.85
Random Forest	0.72	0.71	0.71	0.72	0.85	0.65	0.64	0.64	0.65	0.83
Decision Tree	0.61	0.60	0.60	0.61	0.80	0.53	0.52	0.52	0.53	0.77
XGBoost	0.75	0.75	0.75	0.75	0.87	0.67	0.66	0.66	0.67	
Multi-Layer Perceptron Neural Network	0.69	0.68	0.68	0.69	0.84	0.65	0.65	0.64	0.66	0.82
Convolutional Neural Network (CNN)	0.68	0.67	0.67	0.68	0.84	0.63	0.63	0.63	0.64	0.81

patterns associated with behavioral outcomes even at the 6-month stage. SVM also displayed the highest level of specificity (0.88) at this life stage. Random Forest followed with a moderate accuracy of 0.72, indicating a reasonable but slightly lower predictive power than SVM and XGBoost. The Multi-Layer Perceptron Neural Network and CNN performed moderately well, with accuracies of 0.69 and 0.68, respectively, reflecting their potential but a possible need for further tuning. The Decision Tree model, with an accuracy of 0.61, lagged behind the other models, likely due to its relatively simple structure and limited capacity to handle complex, high-dimensional data.

The results for the Seeing Eye dataset presented a slightly different trend, with overall lower performance across most models compared to the Canine Companions dataset. SVM remained the best-performing model with an accuracy of 0.70, followed by XGBoost with an accuracy of 0.67. The specificity of the SVM model (0.85) was also relatively high. Both models demonstrated balanced F1 Scores, precision, and recall, suggesting effective performance, albeit with slightly reduced predictive strength relative to the Canine Companions dataset. The Random Forest and Multi-Layer Perceptron models showed comparable performance, each achieving accuracy scores around 0.65, while CNN achieved 0.63. The Decision Tree model struggled significantly on the Seeing Eye dataset, with an accuracy of only 0.53, indicating potential challenges in adapting to the specific features or data distribution of this dataset.

As in the previous experiment, we also evaluated the average confidence of the best performing model's predictions based on probability outputs. These confidence scores, calculated as the model's predicted probability for the assigned label, offer additional insight into prediction certainty beyond accuracy alone. For the 6-month predictions, the SVM model achieved an average confidence score of 71.04 % on the Canine Companions dataset and 70.63 % on the Seeing Eye dataset. These results suggest that, while the SVM model maintained stable predictive ability across both datasets, its confidence in predictions was somewhat lower compared to the 12-month models, potentially reflecting the increased challenge of making outcome predictions at an earlier behavioral stage.

In summary, the SVM and XGBoost models displayed the most consistent performance across both datasets, highlighting their adaptability and effectiveness in predicting outcomes based on the 6-month C-BARQ scores. The moderate results from Random Forest and neural networks suggest that these models may also be viable options with further optimization. The Decision Tree's lower accuracy and F1 Scores across datasets underscore its limitations for this task, particularly in handling complex or noisy data. Overall, the findings suggest that SVM and XGBoost are the most reliable choices for early behavioral prediction, with consistent performance that aligns with their results in the 12-month C-BARQ analysis, though the predictive challenge appears slightly greater at the 6-month stage.

3.3. Comparison of model performance between 6-month and 12-month C-BARQ scores

Generally, models trained on the 12-month C-BARQ data achieved higher accuracy across both the Canine Companions and Seeing Eye datasets. The progression in performance underscores the importance of temporal data maturity, particularly in structured behavioral datasets where characteristics develop and stabilize over time.

To facilitate a comparative analysis, the prediction accuracy of each model based on 6-month and 12-month C-BARQ scores for the Canine Companions dataset and the Seeing Eye dataset is presented in Table 6.

As presented in Table 6, all models exhibited greater accuracy when trained with 12-month scores from the Canine Companions dataset. The SVM and XGBoost models, which had shown strong results at the 6-month stage, further enhanced their predictive power with accuracies rising from 0.75 to 0.80 for SVM and from 0.75 to 0.77 for XGBoost. These results suggest that these models are particularly effective at capturing the complex relationships in the dataset, and the additional six months of behavioral data provide a clearer signal for outcome prediction. Random Forest also showed improvement, moving from 0.72 at 6 months to 0.75 at 12 months, reflecting its robust adaptability to enriched data. Similarly, both neural network models—Multi-Layer Perceptron (MLP) and CNN—saw accuracy gains, reaching 0.74 with the 12-month data. The Decision Tree model had the lowest improvement, indicating potential limitations in its ability to leverage the added temporal data as effectively as the other models.

In the Seeing Eye dataset, trends were slightly different. While most models still showed a small improvement with the 12-month data, the gains were more modest than those seen in the Canine Companions dataset. SVM, XGBoost, and MLP saw minor increases in accuracy, with SVM rising from 0.70 to 0.71, XGBoost from 0.67 to 0.69, and MLP from 0.65 to 0.67. Interestingly, the Decision Tree model performed better with the 12-month data (0.65) compared to the 6-month data (0.53), suggesting that it may have benefitted more from extended observations in this dataset than in the Canine Companions dataset. Overall, CNN showed limited improvement, with accuracy increasing only marginally from 0.63 to 0.64, indicating it may be less sensitive to temporal depth compared to other models in this dataset.

In addition to improvements in accuracy, models trained on 12-month C-BARQ data also demonstrated higher prediction confidence,

Table 6
Prediction accuracy of each model based on 6-month and 12-month C-Barq scores Across both datasets.

ML Model	Accuracy based on 6-month C-BARQ scores		Accuracy based on 12-month C-BARQ scores	
	Canine Companions Dataset	Seeing Eye Dataset	Canine Companions Dataset	Seeing Eye Dataset
SVM	0.75	0.70	0.80	0.71
Random Forest	0.72	0.65	0.75	0.65
Decision Tree	0.61	0.53	0.63	0.65
XGBoost	0.75	0.67	0.77	0.69
Multi-Layer Perceptron Neural Network	0.69	0.65	0.74	0.67
Convolutional Neural Network (CNN)	0.68	0.63	0.74	0.64

as measured by average probability scores. Specifically, the best performing model (SVM) achieved an average confidence score of 80.29 % for the Canine Companions dataset and 70.37 % for the Seeing Eye dataset at 12 months, compared to 71.04 % and 70.63 % respectively for the 6-month models. This suggests that not only does predictive performance improve with additional data, but the model also becomes more certain in its classifications—highlighting the importance of behavioral data maturity over time.

4. Discussion

The aim of this study was to assess the ability of Machine Learning (ML) and Deep Learning (DL) models to predict training outcomes for young assistance dogs using behavioral data (C-BARQ item scores) collected at approximately 6 and 12 months of age. The results indicate that the best performing ML model can learn to assign dogs correctly to specific career outcomes (*Graduate, Released,* and *Breeder*) with an impressive level of accuracy, reaching up to 80 % using the 12-month data from Canine Companions and 71 % based on the Seeing Eye dataset, and 75 % and 70 %, respectively, using the 6-month data. In the past, researchers have been able to generate individual-level predictions using conventional statistical approaches, such as linear or logistic regression models, that take input data for a single dog – e.g., that dog's 6- or 12-month C-BARQ scores – and generate a specific prediction about that dog's likely training outcome, along with a probability or confidence score. We can do the same thing with our current ML approach, but with several advantages (Bray et al., 2019; Duffy and Serpell, 2012). First, the current approach performs as well or better with less information – it only uses the C-BARQ scores of the dogs, without any additional covariates (e.g., sex, breed, coat color, training year, etc.). ML approaches are also better equipped to handle complex, high-dimensional, and non-linear behavioral data. These findings are especially encouraging given the substantial time gap—often 6–12 months or more—between C-BARQ assessments and the eventual decision to graduate or release a dog, and the fact that the behavioral evaluations are completed by a diversity of puppy raisers with varying levels of experience regarding canine behaviour, dog training, and puppy raising. Furthermore, these surveys represent a very minor time investment on the part of the puppy raisers who spend approximately 10–15 minutes completing the survey.

It is unclear why the models performed better based on behavioral data from Canine Companions compared with The Seeing Eye, particularly in view of the more varied career outcomes for the former. Differences in breed representation between the two organizations may be a factor. The majority of Canine Companions dogs are either Labrador retrievers or Labrador x golden retriever crosses, whereas the Seeing Eye also breeds and trains significant numbers of German shepherds as dog guides. Given observed behavioral differences between German shepherds and the two retriever breeds (Serpell and Duffy, 2014), it is possible that greater variance in breed-related traits impacted the accuracy of the ML models for the Seeing Eye.

Model performance generally improved with data collected at 12 months compared with 6 months. This is likely due to the greater stability and maturity of behavioral traits in the older cohort of dogs, as well as their closer temporal proximity to the period of training and ultimate career determination. These findings also highlight the value of extended data collection periods when developing predictive models for datasets capturing gradual developmental changes in behavior and temperament. The SVM and XGBoost models emerged as the most reliable across both datasets, showcasing their adaptability to both earlier and later observational periods. The Canine Companions dataset exhibited stronger performance gains with additional data than the Seeing Eye dataset, suggesting that dataset-specific factors, such as behavioral consistency or breed representation, may also play a role in model effectiveness over time.

The Support Vector Machine (SVM) model was the top performer across both datasets, particularly excelling with 12-month data. This strong performance underscores SVM's adaptability to structured, tabular data, where its ability to create optimal class boundaries enhances predictive accuracy. SVM's consistent accuracy between 6 and 12-month data highlights its robustness, especially for datasets with well-defined class boundaries, as seen in guide dog training outcomes.

In addition to traditional performance metrics, we also evaluated the confidence of the SVM model's predictions using probability estimates generated by the predict_proba() function. This analysis provided insight into how certain the model was in its classifications, beyond simply being correct or incorrect. The average confidence score, calculated as the predicted probability associated with each predicted label (Graduate, Released or Breeder), was highest for the 12-month Canine Companions dataset (80.29 %), followed by the 12-month Seeing Eye dataset (70.37 %), and the 6-month Canine Companions (71.04 %) and Seeing Eye (70.63 %) datasets. These results suggest that the SVM model was not only more accurate at later timepoints, but also more certain in its predictions—supporting the interpretation that temporal maturity in behavioral traits improves both performance and model confidence. This adds another dimension to our evaluation, indicating that the predictions made at 12 months are both more accurate and more trustworthy.

XGBoost also demonstrated strong performance, with accuracy levels of 77 % for the Canine Companions dataset and 69 % for the Seeing Eye dataset at 12 months. While Random Forest provided slightly lower accuracy than SVM and XGBoost, it proved to be a stable baseline model. On the other hand, the Decision Tree model exhibited modest accuracy (63 % on the Seeing Eye dataset at 12 months), reflecting its limitations in generalizability and sensitivity to noise compared to ensemble approaches.

Deep Learning models, including the Multi-Layer Perceptron (MLP) and Convolutional Neural Network (CNN), exhibited moderate performance, generally trailing behind the ML models. For instance, MLP achieved 74 % accuracy on the Canine Companions dataset and 67 % on the Seeing Eye dataset with 12-month data, indicating that DL models may be less effective on structured, tabular data without spatial or sequential dependencies. The lower performance of DL models compared to ML models suggests that while deep learning is advantageous for complex data structures, traditional ML models like SVM and XGBoost can outperform them when dealing with structured datasets such as C-BARQ. This outcome highlights that traditional ML models, with proper tuning, can better leverage the structured nature of C-BARQ data, focusing on feature separability to deliver accurate predictions.

These findings have significant implications for assistance dog breeding and training programs. Knowing a dog's C-BARQ scores at 6 or 12 months provides valuable insights into training suitability, thereby allowing for informed decisions earlier in the dog's development. Such predictive insights can help to optimize resource allocation; for example, by ensuring that dogs at risk of failing receive appropriate behavioral remediation as early as possible, or by informing the decision to release a dog from the program before it enters the more cost- and resource-intensive phase of professional training. Overall, these results underscore the importance of selecting ML models aligned with dataset characteristics and suggest that ML-driven tools offer promising potential for early decision-making in assistance and other working dog programs. Because they are powered by AI, these ML algorithms can also be fully automated, allowing rapid and relatively inexpensive processing of data on new dogs as they progress through the puppy-raising system.

The present study's reliance solely on C-BARQ data, and the inherent limitations of using a single, though well-validated, phenotypic measure, suggest that incorporating additional sources of behavioral, cognitive, physiological, and demographic information may be valuable in improving the accuracy of predictions. For example, Hare et al. (2024) recently developed a modified version of the C-BARQ for working detection dogs that includes additional behavioral domains, such as distractibility, impulsivity, playfulness, and basophobia (fear of falling),

which may be relevant to the assessment of assistance dogs. Also, while the C-BARQ focuses primarily on problem behaviors that tend to have a negative impact on a dog's working ability, behavioral tests or assessments that capture uncorrelated positive dimensions of working performance are likely to contribute to the predictive accuracy of future ML models (e.g., Asher et al., 2013; Batt et al., 2008; Berns et al., 2016; Brady et al., 2018; Bray et al., 2017a, 2019, 2021; Harvey et al., 2016; Lazarowski et al., 2021; MacLean and Hare, 2018; Sinn et al., 2010; Svobodová et al., 2008; Tiira et al., 2020; Tomkins et al., 2011; Wilsson and Sundgren, 1997). Further work, however, is needed to determine the most informative and reliable test components, and the most appropriate ages for testing.

Additional analyses using combined datasets may offer further insights. However, the primary goal of the present study was to evaluate model performance across two separate canine populations that differ in terms of breed composition and training objectives, as our aim was to support individual working dog organizations that focus on different breeds and prepare dogs for distinct roles. While we initially considered training a model on one dataset and evaluating it on another to assess generalizability, we determined that such an experiment would present significant feasibility concerns due to contextual and demographic differences between the study populations. Despite this limitation, we were able to demonstrate consistency in model performance across both datasets, suggesting that C-BARQ behavioral features are relatively stable and reliable indicators of training outcomes.

In sum, this study provides several meaningful scientific contributions:

- It demonstrates that ML models, particularly SVM and XGBoost, can effectively predict training outcomes in assistance dogs with higher accuracy and greater confidence than previously reported methods.
- It confirms that behavioral assessments conducted at 12 months yield more reliable predictions than those at 6 months, emphasizing the importance of developmental stability in behavioral traits.
- It provides insights into how different ML models perform on structured, tabular data, contributing to the broader field of applied ML in behavioral science.
- It offers a methodological framework for working dog organizations to implement AI-driven decision-making, potentially optimizing resource allocation and training investments.
- It demonstrates that prediction confidence scores can serve as a useful proxy for evaluating model reliability and should be considered alongside accuracy in future working dog assessments.

5. Conclusions

This study offers a comparative analysis of traditional machine learning and deep learning models for predicting training outcomes in working dogs, utilizing C-BARQ scores collected at two key developmental stages—6 months and 12 months—across two assistance dog organizations. Our findings reveal that traditional machine learning models, especially SVM and XGBoost, consistently deliver strong predictive performance across both age stages and datasets, demonstrating their robustness and suitability for structured, tabular data. By contrast, deep learning models like CNN and MLP, while advantageous for high-dimensional and sequential data, did not surpass traditional models in this setting, achieving only moderate accuracy levels.

The comparison between 6-month and 12-month C-BARQ scores illustrates that longer observational periods and/or capturing behavior at a more mature stage in a dog's life leads to improved model accuracy, with all models showing enhanced performance with 12-month data. SVM and XGBoost, however, performed reliably even at the 6-month mark, making them viable for early-stage predictions where timely assessments are essential. This research emphasizes the importance of selecting models that align with the dataset's structure and the specific predictive goals, particularly in contexts like working dog training, where early prediction can have substantial impacts on training investments and resource management.

Future research could focus on enhancing deep learning models for structured data by incorporating advanced feature engineering or custom model architectures, alongside additional sources of behavioral, cognitive, physiological, and demographic data. Such enhancements might further bridge the performance gap between traditional ML and deep learning approaches, offering more nuanced insights into early-stage predictions and assisting in the optimization of training decisions within assistance dog programs. Further work on applying ML models to automate the process of classifying and assigning individual dogs to specific career outcomes would be beneficial, as would additional analysis of the specific C-BARQ variables that are most/least informative as predictors of success and failure.

CRediT authorship contribution statement

Kennedy Brenda S.: Writing – review & editing, Resources, Data curation. **Douglas Laura E.L.C.:** Writing – review & editing, Resources, Data curation. **Block Theadora A.:** Writing – review & editing, Resources, Data curation. **Bray Emily E.:** Writing – review & editing, Resources, Data curation. **Pettigrew Piya:** Funding acquisition. **Freeberg Kathleen:** Writing – review & editing, Resources, Data curation. **Evans Katy M.:** Writing – review & editing, Resources, Data curation. **Serpell James A:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Amirhosseini Mohammad Hossein:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization.

Funding

Funding for this project was provided through Dogvatar Inc.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.applanim.2025.106632.

References

Arata, S., Momozawa, Y., Takeuchi, Y., Mori, Y., 2010. Important behavioral traits for predicting guide dog qualification. J. Vet. Med Sci. 72, 539–545. https://doi.org/10.1292/jvms.09-0512.

Asher, L., Blythe, S., Roberts, R., Toothill, L., Craigon, P.J., Evans, K.M., et al., 2013. A standardized behavior test for potential guide dog puppies: methods and association with subsequent success in guide dog training. J. Vet. Behav. 8, 431–438. https://doi.org/10.1016/j.jveb.2013.08.004.

Batt, L.S., Batt, M.S., Baguley, J.A., McGreevy, P.D., 2008. Factors associated with success in guide dog training. J. Vet. Behav. (3), 143–151. https://doi.org/10.1016/j.jveb.2008.04.003.

Berns, G.S., Brooks, A.M., Spivak, M., Levy, K., 2016. Functional MRI in awake dogs predicts suitability for assistance work. Sci. Rep. 7, 43704. https://doi.org/10.1101/080325.

Brady, K., Cracknell, N., Zulch, H., Mills, D.S., 2018. A systematic review of the reliability and validity of behavioral tests used to assess behavioral characteristics important in working dogs. Front Vet. Sci. 5. https://doi.org/10.3389/fvets.2018.00103.

Bray, E.E., Sammel, M.D., Cheney, D.L., Serpell, J.A., Seyfarth, R.M., 2017b. Effects of maternal investment, temperament, and cognition on guide dog success. Proc. Natl. Acad. Sci. 114 (34), 9128–9133.

Bray, E.E., Sammel, M.D., Seyfarth, R.M., Serpell, J.A., Cheney, D.L., 2017a. Temperament and cognition in a population of adolescent guide dogs. Anim. Cogn. 20, 923–939. https://doi.org/10.1007/s10071-017-1112-8.

Bray, E.E., Levy, K.M., Kennedy, B.S., Duffy, D.L., Serpell, J.A., MacLean, E.L., 2019. Predictive models of assistance dog training outcomes using the canine behavioral assessment and research questionnaire and a standardized temperament evaluation. *Front. Vet. Sci.* b 6, 49. https://doi.org/10.3389/fvets.2019.00049.

Bray, E.E., Otto, C.M., Udell, M.A.R., Hall, N.J., Johnston, A.M., MacLean, E.L., 2021. Enhancing the selection and performance of working dogs. Front Vet. Sci. 12, 8.
Cleghern, Z., Gruen, M., Roberts, D., 2018. Using decision tree learning as an interpretable model for predicting candidate guide dog success. *Measuring Behavior 2018*. Manchester: Manchester Metropolitan University, pp. 252–258.
Duffy, D.L., Serpell, J.A., 2012. Predictive validity of a method for evaluating temperament in young guide and service dogs. Appl. Anim. Behav. Sci. 138, 99–109. https://doi.org/10.1016/j.applanim.2012.02.011.
Dutrow, E., Serpell, J.A., Ostrander, E., 2022. Domestic dog lineages reveal genetic drivers of behavioral diversification. Cell 185. https://doi.org/10.1016/j.cell.2022.11.003.
Elreedy, D., Atiya, A.F., Kamalov, F., 2024. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Mach. Learn 113, 4903–4923. https://doi.org/10.1007/s10994-022-06296-4.
Foyer, P., Bjällerhag, N., Wilsson, E., Jensen, P., 2014. Behaviour and experiences of dogs during the first year of life predict the outcome in a later temperament test. Appl. Anim. Behav. Sci. 155, 93–100.
Goddard, M., Beilharz, R., 1986. Early prediction of adult behaviour in potential guide dogs. Appl. Anim. Behav. Sci. 15, 247–260. https://doi.org/10.1016/0168-1591(86)90095-X.
Hall, N.J., Johnston, A.M., Bray, E.E., Otto, C.M., MacLean, E.L., Udell, M.A.R., 2021. Working dog training for the twenty-first century. Front Vet. Sci. 27, 8.
Härdle, W.K., Simar, L., Fengler, M.R., 2024. Principal Component Analysis. Applied Multivariate Statistical Analysis. Springer, Cham. https://doi.org/10.1007/978-3-031-63833-6_11.
Hare, E., Essler, J.L., Otto, C.M., Ebbecke, D., Serpell, J.A., 2024. Development of a modified C-BARQ for evaluating behavior in working dogs. Front Vet. Sci. 11, 1371630 doi: 10.3389/fvets.2024.1371630.
Harvey, N.D., Craigon, P.J., Sommerville, R., McMillan, C., Green, M., England, G.C., et al., 2016. Test-retest reliability and predictive validity of a juvenile guide dog behavior test. J. Vet. Behav. 11, 65–76. https://doi.org/10.1016/j.jveb.2015.09.005.
Harvey, N.D., Craigon, P.J., Blythe, S.A., England, G.C., Asher, L., 2017. An evidence-based decision assistance model for predicting training outcome in juvenile guide dogs. PLoS ONE 12, e0174261. https://doi.org/10.1371/journal.pone.0174261.
Hsu, Y., Serpell, J.A., 2003. Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. J. Am. Vet. Med Assoc. 223 (9), 1293–1300.
Lazarowski, L., Rogers, B., Krichbaum, S., Haney, P., Smith, J.G., Waggoner, P., 2021. Validation of a behavior test for predicting puppies' suitability as detection dogs. Animals 11 (4), 993. Apr 1.
MacLean, E.L., Hare, B., 2018. Enhanced selection of assistance and explosive detection dogs using cognitive measures. Front Vet. Sci. 5, 1–14. https://doi.org/10.3389/fvets.2018.00236.
Mercato, M., Kenny, J., O'Riordan, R., O'Mahoney, C., O'Flynn, B., Galvin, P., 2022. Assistance dog selection and performance assessment methods using behavioral and physiological tools and devices. Appl. Anim. Behav. Sci. 254, 105691. https://doi.org/10.1016/j.applanim.2022.105691.
Panthirana, A.M.B., Balalle, H., 2024. Predicting adult dog temperament based on puppy behaviors: a machine learning approach for enhancing canine welfare. Open J. Appl. Sci. 14, 3028–3049 (Available at: https://www.scirp.org/journal/ojapps).
Serpell, J.A., 2021. Commensalism or cross-species adoption? A critical review of theories of wolf domestication. Front Vet. Sci. 8, 662370.
Serpell, J.A., Duffy, D.L., 2014. Breeds and their behavior. In: Horowitz, A. (Ed.), In *Domestic Dog Cognition and Behavior*. Springer-Verlag, Berlin, pp. 31–57.
Serpell, J.A., Hsu, Y., 2001. Development and validation of a novel method for evaluatiing behavior and temperament in guide dogs. Appl. Anima. Behav. Sci. 72, 347–364.
Sinn, D.L., Gosling, S.D., Hilliard, S., 2010. Personality and performance in military working dogs: Reliability and predictive validity of behavioral tests. Appl. Anim. Behav. Sci. 127, 51–65. https://doi.org/10.1016/j.applanim.2010.08.007.
Svobodová, I., Vápeník, P., Pinc, L., Bartos, L., 2008. Testing German shepherd puppies to assess their chances of certification. Appl. Anim. Behav. Sci. 113, 139–149.
Tiira, K., Tikkanen, A., Vainio, O., 2020. Inhibitory control – important trait for explosive detection performance in police dogs? Appl. Anim. Behav. Sci. 224, 104942.
Tomkins, L.M., Thomson, P.C., McGreevy, P.D., 2011. Behavioral and physiological predictors of guide dog success. J. Vet. Behav. 6, 178–187. https://doi.org/10.1016/j.jveb.2010.12.002.
Wilsson, E., Sundgren, P.E., 1997. The use of a behaviour test for the selection of dogs for service and breeding, I: method of testing and evaluating test results in the adult dog, demands on different kinds of service dogs, sex and breed differences. Appl. Anim. Behav. Sci. 53 (4), 279–295.