Toward Predicting Infant Developmental Outcomes from Day-Long Inertial Motion Recordings

Naomi T. Fitter, *Member, IEEE*, Rebecca Funke, José Carlos Pulido, Maja J. Matarić, *Fellow, IEEE*, and Beth A. Smith

***Abstract*—As improvements in medicine lower infant mortality rates, more infants with neuromotor challenges survive past birth. The motor, social, and cognitive development of these infants are closely interrelated, and challenges in any of these areas can lead to developmental differences. Thus, analyzing one of these domains - the motion of young infants - can yield insights on developmental progress to help identify individuals who would benefit most from early interventions. In the presented data collection, we gathered day-long inertial motion recordings from N = 12 typically developing (TD) infants and N = 24 infants who were classified as at risk for developmental delays (AR) due to complications at or before birth. As a first research step, we used simple machine learning methods (decision trees, k-nearest neighbors, and support vector machines) to classify infants as TD or AR based on their movement recordings and demographic data. Our next aim was to predict future outcomes for the AR infants using the same simple classifiers trained from the same movement recordings and demographic data. We achieved a 94.4% overall accuracy in classifying infants as TD or AR, and an 89.5% overall accuracy predicting future outcomes for the AR infants. The addition of inertial data was much more important to producing accurate future predictions than identification of current status. This work is an important step toward helping stakeholders to monitor the developmental progress of AR infants and identify infants who may be at the greatest risk for ongoing developmental challenges.**

***Index Terms*—infant motion sensing, early motion interventions, rehabilitation engineering**

I. Introduction

DUE to improvements in obstetric and neonatal medicine, an increasing number of infants with neuromotor challenges survive past birth [1]. Infants with complications at or before birth are classified as being at risk for developmental delay (AR), and some are later diagnosed with cognitive and/or physical developmental delays [2]. Early motor delays are often the initial signs of later developmental impairments [2]. Identifying neuromotor impairment accurately and early enough so that interventions can take place before the developmental delay is pronounced is a current challenge in the field of physical therapy. One approach to early identification is tracking infants' spontaneous movements, which have been shown to be correlated to future motor control [3]. Specifically, researchers propose that neurological deficits could be identified by collecting high-quality spontaneous movement patterns from wearable sensors and kinematic analysis systems [4]–[6].

One goal of our work is to use day-long kinematic data collected from wearable inertial sensors to classify infants as typically developing (TD) or AR. Although these labels are already known for each infant based on the condition of that infant at birth, a classifier that uses inertial motion data to label infants as TD or AR would further establish and support past work that indicates differences in the spontaneous movements of TD and AR infants. Previous non-machine learning efforts to accomplish this task on the same dataset appear in [7]. The main impactful and challenging goal of our work is to predict future outcomes for AR infants. By conducting supervised learning using the inertial motion data of AR infants from our data collection, we can establish whether recorded movement patterns may enable the prediction of future developmental challenges for these AR infants. This forecasting could support strategic interventions for AR infants who are most likely to experience developmental challenges.

This article discusses related work in Section II. A description of the data collection and initial data processing appears in Section III. Section IV outlines the methods and results for the TD/AR classification. Section V provides the methods and results for predicting future outcomes of AR infants. We discuss the implications of this work in Section VI and draw overarching conclusions in Section VII.

II. Related Work

Exploratory motions ranging from batting an overhead mobile to reaching out to a caregiver are essential for the development of young infants. TD infants engage in these types of exploratory movements naturally, learning to control their bodies and interact with their environment [8], [9]. In contrast, AR infants often have neuromotor impairments involving coordination, strength, and proprioception, which can lead to developmental delays [2]. As medical care improves, a larger number of infants who experience complications at or before birth survive; recent estimates determined that 9% of infants in the United States are born at risk [10]. Generally, these infants could benefit from early intervention services

Naomi T. Fitter is with the Collaborative Robots and Intelligent Systems Institute, Oregon State University, Corvallis, OR 91777, USA. naomi.fitter@oregonstate.edu

Rebecca Funke and Maja J. Matarić are with the Interaction Lab, Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA. rfunke@usc.edu,mataric@usc.edu

José Carlos Pulido is with the Planning and Learning Group, Departamento de Informática, Universidad Carlos III de Madrid, Madrid, Spain. jcpulido@inf.uc3m.es

Beth A. Smith is with the Division of Biokinesiology and Physical Therapy and Department of Pediatrics, University of Southern California, Los Angeles, CA 90089, USA. beth.smith@usc.edu

Manuscript received September 16, 2019. Manuscript revised and returned February 6, 2020.

to support their development [10]. However, resources for early interventions are limited, and the standard of care for these infants is often infrequent movement therapy or no intervention until after infancy [11], [12].

The growing number of AR infants leads to emergent needs to understand how the behavior of TD and AR infants differs and identify which AR infants are at the highest risk for later developmental delays. For infants, motor, social, and cognitive development are all closely interrelated [13], so an analysis of infant motion alone leads to insights on general developmental progress. Past work involving short-term (5- to 10-minute) analyses of infant motion have demonstrated differences in motion features of TD and AR infants. Using wearable sensors and kinematic analysis systems, these studies have shown that kinematic variables such as spatiotemporal organization, kicking frequency, and interjoint and interlimb coordination are different between TD infants and infants with myelomeningocele [14], [15], intellectual disabilities [16], Down syndrome [17], as well as infants born preterm [18]. The presented work for early, data-driven identification of infants who are at high risk for developmental delays can improve outcomes for these individuals; early and intense targeted interventions have the potential to improve neurodevelopmental structure and function [19].

Because of the past documented differences between the motion features of TD infants and infants with various neuromuscular challenges, simple machine learning techniques are good candidates for 1) *distinguishing between TD and AR infants* and 2) attempting to *predict developmental outcomes* using infant motion features. Past work has applied machine learning to inertial sensor data to address similar classification problems, such as the detection of gait anomalies in individuals with Parkinson's disease [20], pathological gait [21], and Huntington's disease [22]. In this work, we approach the stated classification goals using relatively simple and interpretable machine learning tools including decision trees (DT) [23], k-nearest neighbors (KNN) [24], and support vector machines (SVM) [25]. We also consider an ensemble approach that combines the above methods as a way to reduce bias and overfitting [26].

Certain existing neonatal assessments have shown promise for early detection of developmental delays via observation of movement, but the present tools in this space largely exhibit poor sensitivity and specificity for predicting neurodevelopmental outcomes. One example is the Alberta Infant Motor Scale (AIMS) [27]; on this assessment, a score of less than five or ten percent is considered a "cutoff" for identifying delay [28], but it is not shown to be a good predictor of later neurodevelopmental outcomes [29]. Bias is a concern with AIMS results [30], as are false positives [31]. One successful counterexample is assessments for the detection of cerebral palsy, which can be identified at approximately 3-4 months chronological or adjusted age using the General Movements Assessment (GMA) [32] with summary estimates of 98% sensitivity (95% confidence interval [CI] 74–100%) and 91% specificity (95% CI 83–93%) [33]. At the same time, the GMA requires a trained observer to rate spontaneous movements in early infancy. Our proposed quantitative approach with accessible, low-cost sensors has potential to more accurately predict future outcome compared to AIMS and to help predict neurodevelopmental outcomes more broadly than the GMA.

III. Data Collection and Processing

To better understand TD and AR infant motion and its implications, we conducted day-long data collections of in-home infant motion using APDM Opal inertial sensors [34]. The participant information, procedures, and data processing steps are outlined in this section.

A. Participants

Data for twelve TD infants were collected in the Portland, OR, metro area and data for 24 AR infants were collected in the Los Angeles, CA, metro area. To be included in the study, TD infants were required to come from singleton, full-term pregnancies. Infants with scores below the 5th percentile on the AIMS assessment were excluded from the TD group. AR infants included in the study were from a broad group meeting the state of California's criteria for being at risk for developmental delay and eligible for state-administered early intervention [35]. AR infants with unstable medical conditions were excluded. The participant group was heterogeneous because our goal was not to predict specific impairments or diagnoses, but rather to study general motion characteristics that might broadly reflect atypical motor control.

During each session, the AIMS score for the infant was recorded. To help us capture longitudinal data including various developmental stages, each infant participated in three data collections between the ages of one and 20 months, with the exception of two AR infants who only completed two sessions. Corrected ages were used for any infant born preterm. Data were collected only from infants who were not yet walking independently; the two AR infants who did not complete a final session were walking independently by the time of their planned third data collections, and thus were excluded from completing a third recording.

B. Procedure

During data collection sessions, we used two APDM Opal inertial sensors [34] to measure tri-axial acceleration and angular velocity at a sampling rate of 20 Hz. The sensors were synchronized to one another throughout the recording using Bluetooth. A researcher affixed one sensor to each of the infant's ankles using custom leg warmers with pockets, as sketched in Fig. 1. Sensor placement was not consistent in orientation (i.e., the axes were not placed in a consistent way relative to the infant's leg), but past work has validated that Opal sensors affixed in this way can accurately record the quantity of infant limb movements without impeding or promoting movement [36], [37].

Inertial data were collected from typical infant behavior over a full day of activity (8-13 hours) in the infant's natural environment, and the sensors were removed at the end of the day. During the data collection, sensor data were stored on the sensors' internal memory. The recordings were downloaded

Fig. 1. A sketch of an infant in this data collection. The sensors were worn in pockets on custom leg warmers, as shown in gray.

following the data collection. Participating families were instructed to go about their normal activities during the data collection. This process was repeated during each of the three data collection sessions, which were spaced approximately two months apart, starting from an infant age of anywhere from one to fifteen months. More detail on infant ages and AIMS scores during each session appears in Table I.

For AR infants, we followed up with the parents when each infant was 24 months old to determine whether the infant was experiencing developmental delays or higher outcomes resemblant of typical development. *Infants with developmental delays/lower outcomes* were defined as those who were undergoing ongoing therapy after the identification of specific delays. *Infants with higher outcomes* were defined as those who were not undergoing any physical or occupational therapy, although they could still be under developmental stimulation or monitoring. We obtained follow-up developmental status information for nineteen of the 24 total AR infants, corresponding to 55 total infant movement recordings. One of the infants without follow-up information passed away before the age of 24 months, and the families of the other four infants could not be successfully reached for follow-up information.

C. Data Preprocessing

The Opal sensors collected raw accelerometer, gyroscope, and magnetometer data from infant movements. The raw triaxial accelerometer and gyroscope data were initially processed using custom MATLAB programs described in more detail in [37]. This previously validated algorithm [37], [38] detected leg movement occurrences using thresholding on the root sum of squares signal from the three filtered accelerometer axes and the gyroscope axes. The duration of movement, average motion acceleration, peak motion acceleration, and type of movement (unilateral, bilateral synchronous, or bilateral asynchronous) were also determined by our MATLAB software.

After processing, the basic dataset contained 23 features that summarized the general demographics and movement characteristics of each infant, as listed below:

- session number
- infant age
- AIMS developmental score
- movement duration (mean and S.D. for right and left)
- movement acceleration (mean and S.D. for right and left)
- average peak acceleration (mean and S.D. for right and left)
- hours of awake time
- unilateral movement rates (right and left)
- bilateral asynchronous movement rates (right and left)
- overall leg movement rates (right and left)
- bilateral synchronous movement rates

For the unilateral, bilateral, and overall leg movement data, the number of movements was normalized by the total awake time for the infant to give the rate of movements per hour. Other than session number, infant age, and AIMS score, all features originated from the processed accelerometer and gyroscope data. Each feature was drawn from the entire length of the recording (day-long time range of 8-13 hours).

Because two AR infants completed only two sessions, the resulting dataset contained 36 observations of day-long TD infant leg movement and 70 AR infant observations. Since the dataset is imbalanced, we used several accuracy metrics in our evaluations, including precision, recall, and F1 score. Results in later sections suggest that the 2:1 imbalance present in our overall dataset does not prevent high classifier performance.

D. Feature Engineering and Extraction

To more completely represent the infants' behaviors, we expanded the above-listed 23-feature set with additional features computed as the ratios of each leg motion rate type divided by the mean movement duration for the same leg. Here, the considered leg motion rate types are the unilateral movement rates (right and left), bilateral asynchronous movement rates (right and left), and overall leg movement rates (right and left) mentioned above.

For each classification problem, we used the Lasso feature and variable selection method [39] to remove redundant or unnecessary features. Specifically, we identified variables in

TABLE I
MEAN (STANDARD DEVIATION; AND RANGE) OF TD AND AR INFANT AGES (IN MONTHS) AND AIMS SCORES FOR EACH SESSION.

	TD		AR	
	Age	AIMS	Age	AIMS
Session 1	4.75 (2.49; 7.0)	18.00 (10.30; 27.00)	5.35 (3.62; 14.5)	17.17 (11.78; 42.00)
Session 2	6.75 (2.49; 7.0)	29.33 (11.75; 38.00)	7.43 (3.59; 14.5)	24.04 (15.81; 49.00)
Session 3	8.75 (2.49; 7.0)	40.50 (11.66; 32.00)	9.04 (3.20; 14.5)	27.55 (12.96; 40.00)
Overall	6.75 (2.93; 11.0)	29.28 (14.36; 48.00)	7.22 (3.75; 18.5)	22.79 (14.12; 49.00)

the Lasso model that corresponded to the minimum cross-validated mean squared error (MSE) and variables in the sparsest model within one standard error of the minimum MSE. Based on the meaningful feature subsets identified by the union of these two techniques, we removed highly correlated and low-variance features to reduce the dimensionality of the feature set from 35 to eight features for each classification task. Lasso's selections were verified using random stumps that generated single-split decision trees on subsets of the data to determine which features were the most influential in dividing the data for maximal information gain.

IV. TD vs. AR Classification

As introduced in Section II, the first goal of this work is to *distinguish between TD and AR infants* using the observations collected from infants in our dataset. This section details our approach to this classification problem.

A. Selected Features

Before training models to distinguish between TD and AR infants, we used Lasso to eliminate redundant and uninformative features, resulting in the following feature set:

- infant age
- AIMS developmental score
- mean movement duration of the right leg
- standard deviation of movement duration of the right leg
- mean left leg average acceleration
- mean right leg average peak acceleration
- mean left leg average peak acceleration
- right leg bilateral asynchronous motion rate

B. Initial Classification Results

During model training and evaluation, we used leave-one-out cross-validation (LOOCV) to create binary classifiers that could label the omitted infant observation as TD or AR. In other words, 106 distinct models were *trained* using 105 observations each (and omitting one observation, which rotated through the 106 total movement observations). The remaining observation was then used as the *test* set for the classifier trained without that data. This approach is commonly used for classification tasks with small datasets [40].

For each considered classifier, we used grid search to tune the hyperparameters. Specifically, we considered 1 to 5 splits (for DT), nearest neighbors (for KNN), and polynomial orders (for SVM). Our ensemble approach computed the majority vote of the top-performing DT, KNN, and SVM models together. Since we completed multiple motion observation sessions with each infant, we determined the overall TD or AR classification for a given infant by taking the majority vote label across all observations for any given classification approach.

In addition to overall accuracy, we used metrics such as precision, recall, and F1 score to evaluate the classifiers. We aimed to generate models with higher true positives (TP) and lower false negatives (FN) because labeling the AR status correctly is most important for supporting healthy infant development. False positives (FP) were less potentially harmful, although we should be mindful of how this type of misclassification might affect resource allocation or lead to other unintended consequences.

The classification accuracy of the tuned models varied from 77.8% for the KNN approach to 88.9% for the SVM approach. The ensemble approach achieved an accuracy of 83.3% with an F1 score of 0.880, falling short of the SVM model alone in some aspects of the classification task. Accordingly, the SVM approach was the top-performing classifier in this first round of analysis. Table II shows the classification accuracy of each approach, along with the corresponding precision, recall, and F1 scores.

Although the SVM classifier was the most accurate overall, the recall values for the KNN and ensemble approaches were higher. This has implications on the treatment of AR infants; if AR infants are misclassified as TD, they might miss out on important early care and interventions. On the other hand, in a scenario with limited resources (also the case in interventions for AR infants) false positive classification of TD infants could lead to stigma and wasted resources. Thus, the choice of classifier depends on the objectives of researchers, care providers, and families. For now, since the TD vs. AR classification of infants is informed by the circumstances of birth, we conclude that the SVM model is the top-performing option in this case.

C. Effects of Considering Longitudinal Data

In the above analyses, we considered the majority vote label across all infant observations to determine each classifier's prediction for each infant. With the collected dataset, this approach is possible because we collected motion data from each infant two or three times; however, it may not always be possible to collect multi-session data. Thus, we conducted an additional analysis to compare the accuracy of labels 1) based on all individual infant movement recordings separately, 2) based on the majority vote label across all recordings for each infant, and 3) based on the third recording for each infant that completed all three sessions. Table III lists the overall accuracy of each of these approaches.

For all strategies, the SVM approach performs better than all other models, reinforcing our selection of this classifier type for the task of distinguishing between TD and AR infants. Longitudinal observations yield better classification accuracy than considering all infant observations separately. This consistent trend demonstrates that in future efforts, considering longitudinal observations together in a majority vote will likely

TABLE II
CLASSIFIER RESULTS FOR DISTINGUISHING BETWEEN TD AND AR INFANTS USING INFORMATION GATHERED DURING THE DAY-LONG RECORDING SESSIONS (LOOCV APPROACH). THE BOLDED ROW REPRESENTS THE TOP-PERFORMING CLASSIFIER IN OUR ANALYSIS.

Model Type	Accuracy	Precision	Recall	F1 Score
DT	0.806	0.769	0.833	0.800
KNN	0.778	0.767	0.958	0.852
SVM	**0.889**	**0.954**	**0.875**	**0.913**
Ensemble	0.833	0.846	0.917	0.880

improve classifier accuracy compared to determining infant status based on a single measurement taken at an arbitrary time during the developmental period we are considering (early life before 24 months of age).

In most cases, using the third movement recording as the sole observation for each infant yields a lower classification accuracy, but this strategy performs better in the SVM classifier case. It is important to also consider the precision, recall, and F1 score of the third-observation-only labels before adopting this approach. Table IV reveals that solely considering the final infant observation yields better results in every category.

This final result should be interpreted with caution; although observing motion later into infant development yields more accurate classification results, waiting longer may have implications on the impact interventions can have on infants. Additionally, the recall value still falls short of the KNN and ensemble approach performances discussed previously, so the updated SVM approach may not be preferable in circumstances where identifying all AR infants is the top priority. Furthermore, the decrease in classifier performance for all other models could indicate that a change from considering all infant observations to the final observation alone may lead to overfitting. Thus, more data should be collected before researchers and clinicians turn to this approach alone for distinguishing between TD and AR infants.

D. Effects of Other Cross-Validation Techniques

Another important consideration in assessing our TD vs. AR classifier is the cross-validation approach, which can have major implications for the accuracy of the labels and generality of the approach. For each classifier type in the initial LOOCV approach proposed above, 106 distinct models were *trained* using 105 observations each (and omitting one observation, which rotated through the 106 total movement observations). The remaining observation was then used as the *test* set for the classifier trained without those data. For a relatively small dataset such as ours, this approach is reasonable, but it is easy to imagine that overfitting could result from a strategy that includes two observations from a given infant to classify that infant's third motion recording as TD or AR.

Accordingly, we also evaluated the performance of a leave-one-subject-out cross-validation (LOSOCV) approach that is likely to be less vulnerable to overfitting. In this approach, all movement recordings from one infant are omitted during the training of each model, and the models are then tested on the omitted data. As mentioned before, note that two AR infants did not complete a third session, so some infants had only two observations. Thus, 36 distinct models were *trained* using 103 or 104 observations each (the three - or occasionally two - observations from 35 of the infants). Each trained model omitted one set of infant observations, which rotated through the 36 total infants. The omitted three (or occasionally two) observations were then used as the *test* set for the classifier trained without that data. This approach is commonly used to strengthen model performance on new observations [41].

TABLE III
CLASSIFIER RESULTS FOR DISTINGUISHING BETWEEN TD AND AR INFANTS WHEN EVALUATING THE ACCURACY OF EACH INDIVIDUAL OBSERVATION VS. THE MAJORITY VOTE ACROSS ALL OBSERVATIONS VS. THE FINAL OBSERVATION COLLECTED FOR A PARTICULAR INFANT. THIS APPROACH USED LOOCV.

Model	Ind. Acc.	Maj. Vote Acc.	Third Obs. Acc.
DT	0.783	0.806	0.765
KNN	0.707	0.778	0.765
SVM	0.877	0.889	0.941
Ensemble	0.811	0.833	0.792

TABLE IV
COMPARISON OF ACCURACY AND OTHER MEASURES FOR LABELS OBTAINED USING THE MAJORITY VOTE FROM ALL OBSERVATIONS FOR A PARTICULAR INFANT VS. LABELS BASED ONLY ON THE THIRD OBSERVATION FOR A PARTICULAR INFANT. BOTH OF THESE APPROACHES USED LOOCV.

Model Type	Accuracy	Precision	Recall	F1 Score
SVM Maj. Vote	0.889	0.954	0.875	0.913
SVM Third Obs.	0.941	1.000	0.909	0.952

TABLE V
CLASSIFIER RESULTS FOR DISTINGUISHING BETWEEN TD AND AR INFANTS, AFTER USING THE LOSOCV APPROACH. THE BOLDED ROW REPRESENTS THE TOP-PERFORMING CLASSIFIER.

Model Type	Accuracy	Precision	Recall	F1 Score
DT	0.722	0.690	0.833	0.755
KNN	0.722	0.724	0.875	0.792
SVM	**0.889**	**0.954**	**0.875**	**0.913**
Ensemble	0.750	0.778	0.875	0.823

The classification results for the LOSOCV approach appear in Table V. As foreshadowed above, the overall accuracy suffers somewhat for most of the classifiers after the change in cross-validation approach; however, for the top-performing SVM classifier, the label accuracy did not change. This consistency validates the original cross-validation approach for this classification problem.

E. Model Interpretability

In addition to the above considerations, model interpretability is also important; interpretable results can increase the impact of this work by helping medical professionals to gain insights about patients while maintaining patient (or parent) trust [42]. The machine learning models discussed previously increase in complexity while decreasing in interpretability. Simpler approaches like the DT and KNN models are easier to interpret compared to the more powerful and complex SVM and ensemble approaches.

As an exercise in generating highly interpretable models, we identified the most informative single feature using the random stumps approach discussed in Section III. Then, using only splits in the data based on this one feature, we identified the highest-performing DT classifier. A DT model is easy to visualize and very similar to dichotomous keys or other scientific tools based on a series of binary splits. Thus, such a model is a helpful tool for explaining how certain infant motion features might help medical specialists to distinguish between TD and AR infants.

We found that in the TD vs. AR classification problem, the most meaningful feature for splitting the data was the mean left

leg average peak acceleration. Using just a *single, threshold-based split (decision stump)* on this feature yielded a 75.0% overall classification accuracy. Furthermore, other performance metrics in Table VI show that a simple, highly-interpretable decision stump can achieve as high of a recall value as more complex models.

Another approach that is accessible and interpretable is using only features that clinicians would typically have access to in the standard of care (age and AIMS score) to train models. Using this approach, we discovered that a simple 2-feature machine learning approach performs *even better than* the previously proposed top models, as shown in Table VI. Thus, for the task of identifying a child's current status as TD or AR, data presently available to healthcare providers is just as powerful as added insights offered by inertial measurements.

V. Predicting Future Outcomes for AR Infants

The second goal of this work was to predict *future developmental outcomes* for AR infants based on data collected during their first two years of life. This section details our approach to this classification problem.

A. Selected Features

As in the previous classification problem, we used Lasso to eliminate redundant and uninformative features. This time, we found the following features to be most informative:

- session number
- infant age
- AIMS developmental score
- mean movement duration of the left leg
- overall right leg movement rate
- standard deviation of average right leg acceleration divided by mean right leg movement duration
- standard deviation of average right leg peak acceleration divided by mean right leg movement duration
- right leg movement rate divided by mean right leg movement duration

B. Initial Classification Results

As in the previous classification task, we used LOOCV to create binary classifiers, although this time we aimed to label the omitted infant motion observation as having *higher outcomes* at 24 months (HO) or *developmental delays/lower outcomes* at 24 months (LO). To undertake this classification task, we could only determine model accuracy for the nineteen AR infants for whom we had information about later developmental outcomes (55 total observations). Accordingly, 55 distinct models were *trained* using 54 observations each (and omitting one observation, which rotated through the 55 total movement observations). The remaining observation was then used as the *test* set for the classifier trained without those data.

TABLE VI
Comparison of accuracy and other measures for TD vs. AR labels obtained using the top-performing SVM classifier vs. labels based on a simple decision stump or model trained on only two commonly available features (age and AIMS score). This approach used LOOCV.

Model Type	Accuracy	Precision	Recall	F1 Score
SVM	0.889	0.954	0.875	0.913
Decision Stump	0.750	0.778	0.875	0.823
2-Feature Models	0.944	1.000	0.917	0.956

TABLE VII
Classifier results for distinguishing between HO and LO infants (LOOCV approach). The bolded row represents the top-performing classifier in our analysis.

Model Type	Accuracy	Precision	Recall	F1 Score
DT	0.842	0.750	1.000	0.857
KNN	0.684	0.667	0.667	0.667
SVM	**0.895**	**0.818**	**1.000**	**0.900**
Ensemble	0.895	0.818	1.000	0.900

We used the same classification methods in this round of model training and evaluation, and we again used grid search to select hyperparameters. We determined the classification of HO or LO for a given infant by taking the majority vote label across all observations for any given classification approach. Furthermore, the same accuracies and scores helped us to evaluate each model, the the implications of correct anticipation of future diagnoses and various types of labeling errors were similar to those of the previous classification problem.

The classification accuracy of the tuned models varied from 68.4% for the KNN approach to 89.5% for the SVM approach. The ensemble approach achieved an accuracy of 89.5% with an F1 score of 0.900. Since the ensemble approach achieved the same results as the SVM model with significantly more training complexity (training three sets of models, rather than one), the SVM approach is the top classification strategy among the tested approaches. Table VII displays the classification accuracy of each approach, along with the corresponding precision, recall, and F1 scores. For this classification problem, the SVM approach tied for most accurate and also produced perfect recall values.

C. Effects of Considering Longitudinal Data

As in the previous classification problem, it is interesting to consider not only the majority vote label across all infant observations, but also the accuracy of each individual label and classifications produced at a particular point in infant development. Rich data gathered from multiple sessions with infants may lead to improved classifications, but this benefit is not guaranteed. Furthermore, longitudinal data may not always be available. Thus, we conducted another analysis to compare the accuracy of labels produced using different subsets of the infant motion recordings. Table VIII lists the overall accuracy of each approach.

Similarly to the previous classification problem, the majority vote approach surpassed the individual observation approach, although the individual approach outperformed the majority vote for the KNN model. This time, the use of the third infant observation alone seemed to limit the abilities of the classifiers, leading to the same ceiling performance cap across all approaches. For the KNN case, this cap was still an

improvement over the majority vote approach, but the third-observation-only strategy performed worse than the leading majority vote classifiers.

D. Effects of Other Cross-Validation Techniques

Another consideration in assessing the HO vs. LO classifier results is how the cross-validation approach influences model accuracy and generality. For each classifier type in the initial LOOCV approach proposed above, 55 distinct models were *trained* using 54 observations each (and omitting one observation, which rotated through the 55 total movement observations). The remaining observation was then used as the *test* set for the classifier trained without that data.

Since the above approach may be prone to overfitting, we also evaluated the performance of a LOSOCV approach. In this approach, nineteen distinct models were *trained* using 52-53 observations each (the three - or occasionally two - observations from eighteen of the infants). Each trained model omitted one infant's observations, which rotated through the nineteen total infants. The omitted three (or occasionally two) observations were then used as the *test* set for the classifier trained without that data.

The classification results for the LOSOCV approach appear in Table IX. It is not surprising to find that the overall accuracy suffers somewhat for most of the approaches after the change in cross-validation approach, including the formerly top-performing SVM classifier. This indicates that the accuracy of our classifiers with the original cross-validation approach may suffer as the existing models are used to label new observations; however, even if we needed to shift to using a LOSOCV approach while training models, the updated KNN classifier achieves an 84.2% overall classification accuracy with a recall level that is far above random. This type of model may still augment clinicians' ability to identify the infants who would benefit most from early intervention.

E. Model Interpretability

Model interpretability is an important consideration in this classification task as well. To again generate highly interpretable models, we identified the most informative single feature using the random stumps approach discussed in Section III. Then, using only splits in the data based on this one feature, we identified the highest-performing DT classifier.

For the HO vs. LO infant determination, the most meaningful feature for splitting the data was the AIMS developmental score. Using just a *single, threshold-based split (decision stump)* on this feature yielded a 78.9% overall classification accuracy. Furthermore, other performance metrics in Table X reveal that a simple, highly-interpretable decision stump achieved a recall value well above random and close to that of more complex models. This demonstrates that, of all the features we tested, the AIMS score is the best single predictor of future outcomes for AR infants. At the same time, if using the AIMS score in this extended capacity, it is important to understand more background information about it (including, but not limited to, the caveats in Section II).

As in the previous classification task, we considered simple classifiers with only two features (age and AIMS score) as another viable option. In the HO/LO classification task, this approach does significantly worse than the top performing models and even the decision stump using only AIMS score, as shown in Table X. This finding is consistent with the past work showing AIMS score alone to be an insufficient predictor of later outcomes. It also indicates that in the present task, inertial data can play an important role.

Although the AIMS score is a powerful tool already used by the medical community, conflicts between our results and related work's findings indicate that it may be best used in combination with other features. The combination of AIMS score with other infant demographic and motion features leads to a uniformly greater classification performance. Thus, using AIMS scores with inertial data has potential to better inform clinicians in situations where a significant improvement in classification accuracy is meaningful.

TABLE VIII
CLASSIFIER RESULTS FOR DISTINGUISHING BETWEEN HO AND LO INFANTS WHEN EVALUATING THE ACCURACY OF EACH INDIVIDUAL OBSERVATION VS. THE MAJORITY VOTE ACROSS ALL OBSERVATIONS VS. THE FINAL OBSERVATION COLLECTED FOR A PARTICULAR INFANT. THESE APPROACHES USED LOOCV.

Model	Ind. Acc.	Maj. Vote Acc.	Third Obs. Acc.
DT	0.745	0.842	0.765
KNN	0.691	0.684	0.765
SVM	0.764	0.895	0.765
Ensemble	0.727	0.895	0.765

TABLE IX
CLASSIFIER RESULTS FOR DISTINGUISHING BETWEEN HO AND LO INFANTS, AFTER THE USE OF A LOSOCV APPROACH. THE BOLDED ROW REPRESENTS THE TOP-PERFORMING CLASSIFIER IN OUR ANALYSIS.

Model Type	Accuracy	Precision	Recall	F1 Score
DT	0.737	0.667	0.667	0.667
KNN	**0.842**	**0.875**	**0.778**	**0.823**
SVM	0.684	0.636	0.778	0.700
Ensemble	0.789	0.778	0.778	0.778

TABLE X
COMPARISON OF ACCURACY AND OTHER MEASURES FOR HO AND LO LABELS OBTAINED USING THE TOP-PERFORMING SVM CLASSIFIER VS. LABELS BASED ON A SIMPLE DECISION STUMP OR MODEL TRAINED USING TWO COMMONLY AVAILABLE FEATURES. THESE TESTS USED LOOCV.

Model Type	Accuracy	Precision	Recall	F1 Score
SVM	0.895	0.818	1.000	0.900
Decision Stump	0.789	0.667	0.889	0.762
2-Feature Models	0.684	0.636	0.778	0.700

VI. DISCUSSION

In this work, conducted an initial evaluation of the possibility of using simple machine learning techniques to 1) *distinguish between TD and AR infants* and 2) *predict developmental outcomes* using infant demographic and motion features. Toward the first goal, we found that by using an eight-feature set and a relatively simple SVM classifier, we could achieve an 88.9% classification accuracy with a recall value of 0.875. In our considerations, recall measures are important because a lower recall represents more AR infants being misclassified as

TD and potentially overlooked needed care. By considering an approach using only the final observation for each infant, we are able to create an even more accurate TD vs. AR classifier with higher recall (94.1% accuracy, recall of 0.909), although some ability to intervene early is lost with this approach, and future investigations are recommended to support or refute this strategy. For the TD vs. AR classification problem, LOOCV and LOSOCV approaches yielded the same performance for our chosen classifier, suggesting that a LOOCV approach may be reasonable for this labeling task. At the same time, this outcome does not always hold, and researchers in this and related areas should carefully consider the appropriateness of the selected cross-validation approach when generalizing to new observations.

For the prospective task of predicting future developmental delays, we found that by using an eight-feature set and a relatively simple SVM classifier, we could achieve an 89.5% classification accuracy with a perfect recall value of 1.000. When considering different longitudinal approaches, we found that the majority vote approach performs best for all classifiers except for the KNN approach. Nevertheless, the overwhelming trend of improvement when using the majority vote leads us to believe that this approach is best for the HO vs. LO classification problem. When considering different cross-validation approaches, on the other hand, we found that the originally proposed SVM classifier may not be optimal for reliably predicting future developmental delays for new infants who are added to the dataset. Although the recall value was 0.778 for both the KNN and SVM approaches when using a LOSOCV strategy, the overall label accuracy dropped from 84.2% in the KNN approach to 68.4% in the SVM approach. This change resulted from an increase in false positives for the SVM strategy, which could be acceptable or detrimental depending on the clinical circumstances. Nevertheless, all evaluated classifiers performed much better than average, even in the LOSOCV case.

We also presented examples of very simple and highly interpretable decision stumps that could be a resource to healthcare providers working with infants. For the TD vs. AR classification task, we found that the mean left leg average peak acceleration was a single feature that can be thresholded to achieve a 75.0% overall labeling accuracy with the same recall level as the top-performing SVM classifier. This result is consistent with the past related findings in [7] and trends seen in another infant motion dataset [43]. The differing acceleration of movement across the two groups of infants indicates that underlying movement patterns may be different (e.g., slower acceleration in the AR group), but further investigation is needed about the position and movements of the children to conclude more. Upon considering an accessible 2-feature model option (using just age and AIMS score), we found that this classifier alternative produces the highest accuracy for TD vs. AR classifications (94.4% accuracy, 1.000 recall). For the prediction of future developmental delays, the AIMS score was the single most impactful feature. Thresholding on the AIMS score led to a 78.9% classification accuracy with a recall value of 0.889. This emphasizes that the AIMS score already being used as a tool by healthcare providers is very powerful, but that the prediction of future developmental challenges can be improved by more than 10% after considering additional infant demographic and motion features. This improvement is promising motivation for using inertial data in combination with AIMS score to improve infant outcomes.

While this work can help inform future steps that will provide new tools and insights to healthcare providers, it is also important to consider possible drawbacks and shortcomings of our efforts. Overall, the size of the dataset is relatively small, and as discussed throughout the previous sections, there is a possibility of overfitting. Thus, the outputs of these machine learning models should be interpreted with the help of medical experts and more data should be collected to strengthen the models. In particular, more data is needed before the authors can confidently recommend using the presented HO vs. LO prediction models to anticipate future challenges for infants who were not part of the original dataset, and a universal measure of outcomes at 24 months would have provided a stronger ground truth label for this assessment. The AIMS score used in this work was administered to infants older than 18 months of age who were not yet walking, which is outside of the established age range norms for the assessment. Our methods are also reliant on human selection of features. With larger datasets, techniques like deep learning could eliminate the need to select important features. In this work, our emphasis was on collecting compact and noninvasive data, but additional data collection methods (including video recordings, RGB-D recordings, and thermal imaging) can yield additional types of features [44]–[49]. Further, no diagnoses can be made about specific future impairments using the present dataset. Although the initial results presented here are promising, additional work is needed before these types of algorithms should be used as the central tool for identifying infants with high risk of developmental delays or need for intervention.

VII. Conclusions and Future Work

Overall, we presented preliminary classifiers that show high accuracy and promise for 1) distinguishing between TD and AR infants and 2) anticipating future developmental challenges for AR infants. These results are presented with certain cross-validation caveats, and we also emphasize the creation of simple models that can augment the usefulness and interpretability of this work for healthcare providers. To further strengthen this work, we hope to recruit a large sample of AR infants and collect longitudinal data from them, including traditional developmental assessment test scores, movement recordings, and a check-up at two years of age to assess if the infant has been diagnosed with developmental delays.

The highest-impact goal of this work is to use machine learning tools to reliably anticipate if an AR infant will be diagnosed with developmental delays. If this information is known, infants can be targeted for early interventions that could make an enormous difference in their later life and health outcomes. Currently, developmental delays are often not diagnosed until an infant is two years old. Current tools, such as the AIMS score, detect early signs of atypical development, but these approaches perform best in extreme cases. Our classifiers can strengthen predictions of developmental delays based

on the general movement of young infants. As a result, those infants could receive earlier and more directed interventions. Accordingly, this work can benefit researchers and healthcare providers who seek improved outcomes for AR infants.

ACKNOWLEDGMENT

The authors acknowledge David Goodfellow, Ruoyu Zhi, and Ivan Trujillo-Priego for their contributions to early stages of this work and Ajitesh Srivastava for his feedback on the manuscript. This work was funded in part by the American Physical Therapy Association Academy of Pediatric Physical Therapy Research Grant 1 and 2 Awards (PI: Smith) and in part by NSF award 1706964 (PI: Smith, Co-PI: Matarić). Study procedures were approved by the Institutional Review Boards of the Oregon Health & Science University and the University of Southern California.

REFERENCES

[1] G. R. Alexander, M. Kogan, D. Bader, W. Carlo, M. Allen, and J. Mor, "Us birth weight/gestational age-specific neonatal mortality: 1995–1997 rates for whites, hispanics, and blacks," *Pediatrics*, vol. 111, no. 1, pp. e61–e66, 2003.

[2] A. Ghassabian, R. Sundaram, E. Bell, S. C. Bello, C. Kus, and E. Yeung, "Gross motor milestones and subsequent development," *Pediatrics*, vol. 138, no. 1, p. e20154372, 2016.

[3] J. Piek, "The influence of preterm birth on early motor development," *Motor Behavior and Human Skill: A Multidisciplinary Approach. United States of America: Human Kinetics*, pp. 233–51, 1998.

[4] S. E. Groen, A. C. De Blécourt, K. Postema, and M. Hadders-Algra, "General movements in early infancy predict neuromotor development at 9 to 12 years of age," *Developmental Medicine and Child Neurology*, vol. 47, no. 11, pp. 731–738, 2005.

[5] M. Hadders-Algra and A. M. Groothuis, "Quality of general movements in infancy is related to neurological dysfunction, ADHD, and aggressive behaviour," *Developmental Medicine and Child Neurology*, vol. 41, no. 6, pp. 381–391, 1999.

[6] H. F. Prechtl, "State of the art of a new functional assessment of the young nervous system. An early predictor of cerebral palsy," pp. 1–11, 1997.

[7] M. S. Abrishami, L. Nocera, M. Mert, I. A. Trujillo-Priego, S. Purushotham, C. Shahabi, and B. A. Smith, "Identification of developmental delay in infants using wearable sensors: Full-day leg movement statistical feature analysis," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 7, pp. 1–7, 2019.

[8] E. J. Gibson and A. D. Pick, *An ecological approach to perceptual learning and development*. Oxford University Press, USA, 2000.

[9] E. Thelen and L. Smith, *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: The MIT Press, 1994.

[10] S. A. Rosenberg, C. C. Robinson, E. F. Shaw, and M. C. Ellison, "Part C early intervention for infants and toddlers: percentage eligible versus served," *Pediatrics*, vol. 131, no. 1, pp. 38–46, 2013.

[11] G. Roberts, K. Howard, A. J. Spittle, N. C. Brown, P. J. Anderson, and L. W. Doyle, "Rates of early intervention services in very preterm children with developmental disabilities at age 2 years," *Journal of Paediatrics and Child Health*, vol. 44, no. 5, pp. 276–280, 2008.

[12] B. G. Tang, H. M. Feldman, L. C. Huffman, K. J. Kagawa, and J. B. Gould, "Missed opportunities in the referral of high-risk infants to early intervention," *Pediatrics*, vol. 129, no. 6, pp. 1027–1034, 2012.

[13] M. A. Lobo and J. C. Galloway, "Assessment and stability of early learning abilities in preterm and full-term infants across the first two years of life," *Research in Developmental Disabilities*, vol. 34, no. 5, pp. 1721–1730, 2013.

[14] N. Rademacher, D. P. Black, and B. D. Ulrich, "Early spontaneous leg movements in infants born with and without myelomeningocele," *Pediatric Physical Therapy*, vol. 20, no. 2, pp. 137–145, 2008.

[15] B. A. Smith, C. Teulier, J. Sansom, N. Stergiou, and B. D. Ulrich, "Approximate entropy values demonstrate impaired neuromotor control of spontaneous leg activity in infants with myelomeningocele," *Pediatric Physical Therapy: The Official Publication of the Section on Pediatrics of the American Physical Therapy Association*, vol. 23, no. 3, pp. 241–247, 2011.

[16] M. Kouwaki, M. Yokochi, T. Kamiya, and K. Yokochi, "Spontaneous movements in the supine position of preterm infants with intellectual disability," *Brain and Development*, vol. 36, no. 7, pp. 572–577, 2014.

[17] S. M. McKay and R. M. Angulo-Barroso, "Longitudinal assessment of leg motor activity and sleep patterns in infants with and without Down syndrome," *Infant Behavior and Development*, vol. 29, no. 2, pp. 153–168, 2006.

[18] J. J. Geerdink, B. Hopkins, W. J. Beek, and C. B. Heriza, "The organization of leg movements in preterm and full-term infants after term age," *Developmental Psychobiology*, vol. 29, no. 4, pp. 335–351, 1996.

[19] R. L. Holt and M. A. Mikati, "Care for child development: basic science rationale and effects of interventions," *Pediatric Neurology*, vol. 44, no. 4, pp. 239–253, 2011.

[20] S. Mazilu, U. Blanke, M. Hardegger, G. Troster, E. Gazit, M. Dorfman, and J. M. Hausdorff, "GaitAssist: a wearable assistant for gait training and rehabilitation in Parkinson's disease," in *IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*. IEEE, 2014, pp. 135–137.

[21] C. Azevedo Coste, B. Sijobert, R. Pissard-Gibollet, M. Pasquier, B. Espiau, and C. Geny, "Detection of freezing of gait in Parkinson disease: preliminary results," *Sensors*, vol. 14, no. 4, pp. 6819–6827, 2014.

[22] A. Mannini, D. Trojaniello, A. Cereatti, and A. M. Sabatini, "A machine learning framework for gait classification using inertial sensors: application to elderly, post-stroke and Huntington's disease patients," *Sensors*, vol. 16, no. 1, p. 134, 2016.

[23] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[24] W. R. Klecka, *Discriminant analysis*. Sage, 1980, vol. 19.

[25] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. ACM, 1992, pp. 144–152.

[26] P. Sollich and A. Krogh, "Learning with ensembles: how overfitting can be useful," in *Advances in Neural Information Processing Systems*, 1996, pp. 190–196.

[27] M. C. Piper, J. Darrah, T. O. Maguire, and L. Redfern, *Motor assessment of the developing infant*. Saunders Philadelphia, 1994, vol. 1.

[28] J. Darrah, M. Piper, and M.-J. Watt, "Assessment of gross motor skills of at-risk infants: predictive validity of the alberta infant motor scale," *Developmental Medicine & Child Neurology*, vol. 40, no. 7, pp. 485–491, 1998.

[29] J. Nuysink, I. C. van Haastert, M. J. Eijsermans, C. Koopman-Esseboom, P. J. Helders, L. S. de Vries, and J. van der Net, "Prediction of gross motor development and independent walking in infants born very preterm using the test of infant motor performance and the alberta infant motor scale," *Early Human Development*, vol. 89, no. 9, pp. 693–697, 2013.

[30] P. L. d. Albuquerque, A. Lemos, M. Q. d. F. Guerra, and S. H. Eickmann, "Accuracy of the alberta infant motor scale (aims) to detect developmental delay of gross motor skills in preterm infants: a systematic review," *Developmental Neurorehabilitation*, vol. 18, no. 1, pp. 15–21, 2015.

[31] A. J. Spittle, K. J. Lee, M. Spencer-Smith, L. E. Lorefice, P. J. Anderson, and L. W. Doyle, "Accuracy of two motor assessments during the first year of life in preterm infants for predicting motor outcome at preschool age," *PLoS One*, vol. 10, no. 5, 2015.

[32] C. Einspieler and H. F. Prechtl, "Prechtl's assessment of general movements: A diagnostic tool for the functional assessment of the young nervous system," *Mental Retardation and Developmental Disabilities Research Reviews*, vol. 11, no. 1, pp. 61–67, 2005.

[33] M. Bosanquet, L. Copeland, R. Ware, and R. Boyd, "A systematic review of tests to predict cerebral palsy in young children," *Developmental Medicine & Child Neurology*, vol. 55, no. 5, pp. 418–426, 2013.

[34] APDM Wearable Technologies, Portland, OR, USA, "Opals," https://www.apdm.com/wearable-sensors/.

[35] California Legislative Information, "California Early Intervention Services Act," https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=GOV&division=&title=14.&part=&chapter=4.&article=.

[36] C. Jiang, C. J. Lane, E. Perkins, D. Schiesel, and B. A. Smith, "Determining if wearable sensors affect infant leg movement frequency," *Developmental Neurorehabilitation*, vol. 21, no. 2, pp. 133–136, 2018.

[37] B. A. Smith, I. A. Trujillo-Priego, C. J. Lane, J. M. Finley, and F. B. Horak, "Daily quantity of infant leg movement: wearable sensor algorithm and relationship to walking onset," *Sensors*, vol. 15, no. 8, pp. 19 006–19 020, 2015.

[38] I. A. Trujillo-Priego and B. A. Smith, "Kinematic characteristics of infant leg movements produced across a full day," *Journal of Rehabilitation and Assistive Technologies Engineering*, vol. 4, no. 1, pp. 1–10, 2017.
[39] R. Tibshirani, "Regression selection and shrinkage via the lasso," *Journal of the Royal Statistical Society Series B*, vol. 58, no. 1, pp. 267–288, 1996.
[40] G. C. Cawley, "Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2006, pp. 1661–1668.
[41] M. Esterman, B. J. Tamber-Rosenau, Y.-C. Chiu, and S. Yantis, "Avoiding non-independence in fMRI data analysis: leave one subject out," *Neuroimage*, vol. 50, no. 2, pp. 572–576, 2010.
[42] A. Vellido, J. D. Martín-Guerrero, and P. J. Lisboa, "Making machine learning models interpretable." in *ESANN*, vol. 12. Citeseer, 2012, pp. 163–172.
[43] C. Jiang, J. T. de Armendi, and B. A. Smith, "The immediate effect of positioning devices on infant leg movement characteristics," *Pediatric Physical Therapy: The Official Publication of the Section on Pediatrics of the American Physical Therapy Association*, vol. 28, no. 3, p. 304, 2016.
[44] N. B. Karayiannis, Y. Xiong, J. D. Frost, M. S. Wise, and E. M. Mizrahi, "Improving the accuracy and reliability of motion tracking methods used for extracting temporal motor activity signals from video recordings of neonatal seizures," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 4, pp. 747–749, 2005.
[45] D. Corbetta, Y. Guan, and J. L. Williams, "Infant eye-tracking in the context of goal-directed actions," *Infancy*, vol. 17, no. 1, pp. 102–125, 2012.
[46] D. P. Dogra, A. K. Majumdar, S. Sural, J. Mukherjee, S. Mukherjee, and A. Singh, "Toward automating hammersmith pulled-to-sit examination of infants using feature point based video object tracking," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 20, no. 1, pp. 38–47, 2011.
[47] M. D. Olsen, A. Herskind, J. B. Nielsen, and R. R. Paulsen, "Model-based motion tracking of infants," in *European Conference on Computer Vision*. Springer, 2014, pp. 673–685.
[48] A. Procházka, H. Charvátová, O. Vyšata, J. Kopal, and J. Chambers, "Breathing analysis using thermal and depth imaging camera video records," *Sensors*, vol. 17, no. 6, p. 1408, 2017.
[49] A. Procházka, M. Schätz, F. Centonze, J. Kuchyňka, O. Vyšata, and M. Vališ, "Extraction of breathing features using ms kinect for sleep stage detection," *Signal, Image and Video Processing*, vol. 10, no. 7, pp. 1279–1286, 2016.