Towards Personalized Adaptive Gamification: A Machine Learning Model for Predicting Performance

Christian López and Conrad Tucker *Member, IEEE*

***Abstract*—Personalized adaptive gamification has the potential to improve individuals' motivation and performance. Current methods aim to predict the perceived affective state (i.e., emotion) of an individual in order to improve their motivation and performance by tailoring an application. However, existing methods may struggle to predict the state of an individual that it has not been trained for. Moreover, the affective state that correlates to good performance may vary based on individuals and task characteristics. Given these limitations, this work presents a machine learning method that uses task information and an individual's facial expression data to predict his/her performance on a gamified task. The training data used to generate the *adaptive-individual-task* model is updated every time new data from an individual is acquired. This approach helps to improve the model's prediction accuracy and account for variations in facial expressions across individuals. A case study is presented that demonstrates the feasibility and performance of the model. The results indicate that the model is able to predict the performance of individuals, before completing a task, with an accuracy of 0.768. The findings support the use of adaptive models that dynamically update their training dataset and consider task information and individuals' facial expression data.**

***Index Terms*—Performance; Facial expression; Gamification; Machine learning.**

I. INTRODUCTION

Gamification has emerged as a growing area of interest across a wide range of sectors. In the past seven years, the research community has seen a significant growth of publications related to gamification [1], [2]. Deterding et al. define gamification as "*the use (rather than the extension) of design (rather than game-based technology or other game related practices) elements (rather than full-fledged games) characteristic for games (rather than play or playfulness) in non-game contexts (regardless of specific usage intentions, context, or media of implementation)"* [3, p. 14]. In other words, gamification aims to implement game features (e.g., Points, Leaderboards) in non-game contexts to encourage individuals to perform a task or set of tasks (i.e., promote action or behavior) [4]. The tasks and objectives of a gamified application can vary based on the context of an application, and the designers' intentions. For example, in the health and wellness context, physically-interactive gamified applications such as Active Games, require individuals to use full-body motion to perform a physical task with the objective of increasing their physical fitness or improving their health awareness [5].

Due to the heterogeneity of individuals, researchers have started exploring methods to design personalized and adaptive gamified applications [6]. Current methods are often developed around studies that have explored the relationship between individuals' attributes and their game feature preferences. However, these studies provide guidelines suited for a general demographic of end users and not for unique individuals. Additionally, most of the existing gamification methods are not capable of dynamically capturing data of an individual's interaction with an application (i.e., real-time data capture). Instead, these methods focus on gathering data in discrete time intervals through the use of self-reported questionnaires [7]. This approach ignores the possibility that individuals' attributes and preferences are dynamic in nature and could change over time [8], which could potentially impact the long-term effectiveness of an application [9].

The *Affective Computing* (AC) community has shown how individuals' facial expressions can be systematically captured and used to improve their interaction with an application. Systems capable of capturing individuals' facial expressions have also shown to be suitable for personalization and adaptation [10]–[12]. In light of this, researchers have started to increasingly implement AC methods to improve the user experience in gaming applications [13]. These applications are known as *Affective Games,* and are defined as games in which the "*emotional state and actions of a player can be recognized and used in order to alter the gameplot and offer an increased user experience"* [14, p. 1]. *Affective Games* relate individuals' facial keypoint data to their perceived affective states. This affective state information is used to alter the gameplot or difficulty of the application in order to improve the user experience. However, individual differences in facial expressions can deteriorate the accuracy of existing methods since they employ general models trained with datasets from a limited set of individuals. For these general models, it is challenging to accurately predict the affective state of an individual that it has not been trained for [15].

Moreover, current *Affective Games* aim to recognize individuals' affective states with the goal of improving their

Manuscript received Xxxxxx XX, 20XX; revised Xxxxxx XX, 20XX; accepted Xxxxxx XX, 20XX. Date of publication Xxxxx XX, 20XX; date of current version Xxxxx XX, 20XX. This work was supported in part by the National Science Foundation NSF-NRI #1527148 and NSF CHOT #1624727.

C. López is with the Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA 16802 USA (e-mail: cql5441@psu.edu).

C. Tucker is with the Department of Engineering Design, Pennsylvania State University, State College, PA 16802 USA, the Department of Industrial and Manufacturing Engineering, Pennsylvania State University, State College, PA 16802 USA, and also with the Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802 USA (e-mail:ctucker4@psu.edu).

experience and not necessarily their task performance, which is a key aspect of gamified applications [16]. Studies have shown that the relationship between performance and affective states is mediated by the task and individuals' characteristics [17]. This relationship can limit the effectiveness of current methods in predicting an individual's affective state and adapting an application to improve his/her performance. Therefore, designers should focus on developing models capable of predicting individuals' performance, instead of their affective state. Furthermore, efforts should be taken to develop models capable of systematically updating their training dataset as new data of an individual is acquired and hence, adapting (i.e., learning) to an individual's unique facial expression characteristics.

Given the current limitations, this work presents a method to predict an individual's performance on a gamified task (i.e., tasks of a gamified application). The method enables capturing individuals' facial keypoint data in real-time without affecting their immersion in an application. Furthermore, the training data used to generate the machine learning model is continuously updated each time new data of an individual is acquired. This continuous updating helps improve the model's accuracy and account for variations in facial expressions across individuals. The method has the potential to enable designers to systematically quantify the correlation between an individual's facial keypoint data and his/her performance on a gamified task. This information could potentially be used to adapt the game features and task difficulty of gamified applications [18].

II. Related Work

A. Personalized Adaptive Gamification

Researchers agree that gamified applications should be designed from a highly personalized and adaptive point of view since studies have shown that individuals interact with gamified applications in different ways [19]. As stated by Buckley and Doyle "*individuals do respond differently to gamification, based upon individual attributes*" [20, p. 44]. Even though researchers have begun to explore how different groups with common attributes (e.g., personalities, learning styles) perceive and interact with gamified applications [20]–[23], several limitations still exist. First, these studies have focused on gathering individuals' data through the use of self-reported questionnaires, which can impact the validity of the responses due to individuals' biases [24]. Furthermore, these studies ignore the possibility that individuals' attributes and preferences are dynamic in nature and could change over time [8]. Not considering the dynamic nature of human behavior and preferences can have a negative impact on the effectiveness of gamified applications [9].

Besides individual differences, the characteristics of a task and the effort required to complete it can impact the effects that gamification has on motivating individuals to perform the task successfully. The Fogg's Behavior Model (FBM) [25] suggests that there are some fundamental tasks and individual characteristics that can impact the effectiveness of gamification. For example, in the gamified application presented by Denny [26], in which students generated and answered multiple choice questions, their performance on the number of answers submitted and the number of active days was improved with the gamified application, compared to the control group (i.e., non-gamified). However, there was no significant improvement in the number of questions generated. These results are in line with FBM since the greater effort and time required to generate questions (i.e., greater task complexity) impacted their motivation and performance on that task. Furthermore, Lopez and Tucker's [27] study supports the need to consider task characteristics while designing gamified applications. Their results reveal that there was a negative correlation between the complexity of a task and individuals' performance.

Similarly, the human-computer interaction community has recognized the connection between task properties and individuals' performance, and developed several predictive models of human performance [28], [29]. These models allow designers to evaluate the expected performance of individuals while interacting with an interface, without having to test it. This is done by evaluating task information using models founded on experimental psychology and information theory research [29], [30], or in some cases, even machine learning models [31]. For example, Li et al. [31] used a deep learning algorithm to predict the time individuals spend in a vertical menu selection task. Their model achieved an R^2 ranging from 0.75 to 0.95 when tested with multiple datasets. However, while some of these predictive models do take into consideration individual characteristics (e.g., expert vs. novice) [30], [32], it is still challenging for them to customize their prediction on an individual level.

Recently, a systematic literature review in the field of adaptive gamification was presented [6]. The challenges highlighted in this review illustrate the need for more empirical studies and methods to advance gamified applications. Moreover, the authors stated that machine learning would play a significant role in advancing the field of gamification. For example, Barata et al. [33] presented evidence that suggests that machine learning algorithms can be used to predict *student types.* In a previous study, the authors identified four distinctive *student types* according to their performance, engagement, and behavior on the application [34]. Their results revealed that after nine weeks of interacting with the applications, a participant's performance data could be used to predict his/her *student type* with an accuracy of 0.79. A participant's player type, along with his/her performance data from a five-week period, was only able to predict his/her *student type* with an accuracy of 0.47.

In recent years, researchers have started working on developing methods for personalized adaptive gamified applications with the goal of maintaining individuals' motivation for long periods of time [6]. These methods tend to implement guidelines developed based on a general demographic of end users [35]. Hence, the degree of personalization that they can provide to a unique individual is limited. Furthermore, some of this work only provides conceptual frameworks and little empirical evidence of their implementation or feasibility [9], [36]. Finally, these methods

are not capable of systematically capturing data of individuals' interaction with a gamified application and predicting their task performance. Therefore, due to the limitations of current methods, this work presents a machine learning method to predict an individual's performance on a gamified task. The method captures individuals' facial keypoint data in real-time as they interact with a gamified application without affecting their immersion. Moreover, a benchmark analysis on the performance of the model, generated with multiple machine learning algorithms, is presented. This model has the potential to advance gamified applications by enabling designers to consider task characteristics and individuals' facial expressions.

B. *Affective Computing, Affective Games, and Gamification*

In recent years, researchers have started implementing *Affective Computing* (AC) methods with the objective of improving user experience in gaming applications [13], [14], [37]. AC researchers have been able to infer individuals' affective states by using a wide range of modalities, such as body movements, speech, and facial expressions [38]. Nonetheless, AC applications frequently use facial expressions to infer an individual's affective state [39]. This is because individuals reveal a significant amount of affective state information through their facial expressions [40]. Additionally, facial expressions can be captured with sensors that do not affect an individual's immersion or ability to interact with an application [41]. For example, the *Affective Game* developed by Grappiolo et al. [42], captured individuals' affective state information via facial expressions and the use of self-reported questionnaires. The application used this information to adapt and change its content to improve user experience. Similarly, Shaker et al. [43] presented an *Affective Game* that was capable of adapting its game features and task complexity (i.e., level difficulty) based on individuals' predicted affective states [44], [45]. In a different approach, Athanasiadis et al. [46] incorporated students' scores to predict their "*energy function*" value (i.e., a function of self-reported engagement, boredom, and frustration levels) in an educational application, indicating that students' performance was associated with their affective state. Similarly, others studies have shown a link between individuals' affective state and their task performance, especially in cognitive tasks [47]–[49]. However, research indicates that the affective state that correlates to good performance may vary based on the characteristics of the task and individual [17]. Hence, current applications might adapt based on an individual's affective state, and not observe improvement in his/her performance.

Table I shows a summary of existing methods that researchers have developed to personalize their gamified and non-gamified applications. Most of the methods developed for gamified applications tend to capture individuals' data at discrete times via self-reported surveys. In contrast, *Affective Games* have shown how designers can dynamically capture individuals' data (e.g., facial keypoint data) to predict their affective states. However, most of the current affect-sensitive systems employ general models [14]. The accuracy of these systems might be impacted by the heterogeneity of individuals' facial expressions [50]. As shown by Asteriadis et al. [44], their "*player dependent*" model (i.e., individual model) outperformed their general model in terms of accurately predicting individuals' engagement (i.e., accuracy: 0.71 vs. 0.82). Moreover, existing methods do not update their model's training set dynamically as new data of an individual of interest is acquired. The capability of models to dynamically adapt to individuals has great potential to advance personalized systems [15].

TABLE I
LITERATURE REVIEW SUMMARY

Study	*Dynamic Data Capture* [a]		*Gamified Application* [b]		*Adaptive Individual Model* [c]	
	No	Yes	No	Yes	No	Yes
[43], [46]	X		X		X	
[7], [12], [39], [42], [44], [45], [51]		X	X		X	
[9], [21], [22], [24], [33], [35], [36], [52]	X			X	X	
This work		X		X		X

[a] Data captured dynamically as individuals interact with an application (i.e., facial expression, gestures, voice), not at discrete points in time (i.e., self-reported questionnaires after or before interacting with the application).
[b] Not a full-fledged game intended just for entertainment purposes, but a gamified application intended to promote action or behavior.
[c] Implements a model that systematically updates its training set as new data of an individual of interest is acquired; hence, adapting to a unique individual's characteristics (unlike general models).

Furthermore, current affect-sensitive systems tend to group individuals' affective states into discrete categories or a single function value of their affective states (e.g., engagement, fun, frustration, "*energy function*") [24], [43], [46]. However, individuals' affective state is far more complex and heterogeneous. The assumption of a "*one-to-one correspondence*" between the expression and the experienced affective state of an individual may limit the effectiveness of existing systems [40]. Thus, potentially affecting their adaptability to improve and maintain individuals' motivation and performance over time. Recent studies reveal that individuals' facial keypoint data and machine learning models can be used to bypass the need to group individuals' affective states into discrete categories and predict their performance on a task [12]. For example, a machine learning model that uses students' facial keypoint data captured while reading the instructions of an engineering task, was shown to accurately predict their task completion time [51]. Therefore, in this work, a machine learning method to predict individuals' performance, instead of their affective state, is presented. Specifically, an *adaptive-individual-task* model to predict an individual's performance on a gamified task by using his/her facial keypoint data and task information is presented. The method captures facial keypoint data in real-time as an individual interacts with an application. Furthermore, the method updates the model's training set every time new data of an individual is acquired. The results of this work support the implementation of facial keypoint data and *adaptive-individual-task* models as a potential method to advance gamification.

III. Research Questions

As highlighted in [6] there are many open research questions and challenges in the field of personalized adaptive gamification. Previous studies have shown that machine learning models that implement individuals' facial keypoint data, captured while reading the instructions of a task, can accurately predict individuals' task completion time [51]. However, there is a need for more empirical evidence to support the benefits of implementing machine learning methods to advance the field of gamification. The objective of this work is to bridge the current knowledge gap by exploring fundamental research questions that will provide quantitative evidence in support of implementing facial keypoint data acquisition and machine learning models to predict an individual's performance. In this work the following research questions are addressed:

RQ1. *Can a machine learning model predict the performance of an individual on a gamified task with accuracy greater than random chance by using his/her facial keypoint data and task information?*

Addressing this question will reveal that a machine learning model can predict an individual's performance on a gamified task, with accuracy greater than random chance. Nonetheless, a machine learning model that is trained with data from a limited set of individuals (i.e., general model) will not be able to consider the unique characteristics of a new individual's facial keypoint data. Therefore, the authors propose an *adaptive-individual-task* model capable of updating its training set as new data of an individual is acquired. Consequently, this motivates the following question:

RQ2. *How does an adaptive-individual-task model's performance change as new data of an individual is acquired and the model is re-trained?*

To address ***RQ2***, the *adaptive-individual-task* machine learning model is validated with an iterative cross-validation approach that simulates scenarios in which new data of an individual is acquired. This adaptive process helps account for variation in facial expressions of individuals; hence, enabling the model to adapt (i.e., learn) to an individual's unique facial expression characteristics.

IV. Method

This section introduces a machine learning method to predict an individual's performance on a gamified task (i.e., tasks of a gamified application). Figure 1 presents the outline of the method that includes the *Data Acquisition* (IV.A) of *Task data* (IV.A.1), individuals' *Facial Keypoint data* (IV.A.2), as well as *Performance data* (IV.A.3). Moreover, the method has *Model Generation* (IV.B) and *Model Validation* (IV.C) steps.

A. Data Acquisition

The purpose of this step is to systematically capture an individual's facial keypoint data before performing a gamified task, as well as task and performance data. The data is used to generate the *adaptive-individual-task* model and predict the performance of individuals in a gamified task.

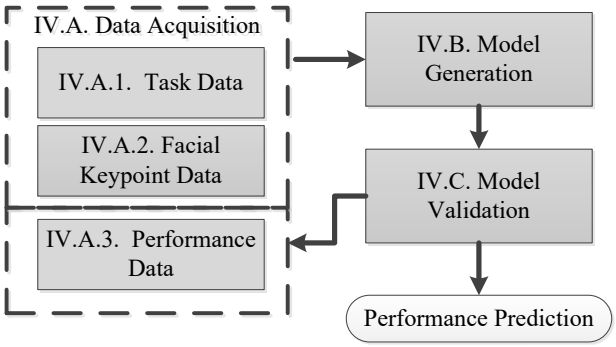


Fig. 1. Method Outline

1) Task data: The efforts required to complete a task can impact the effectiveness of gamification in motivating individuals to perform the task successfully. Hence, the *adaptive-individual-task* model uses as input, data pertaining to the task, as well as data pertaining to individuals. Specifically, the model uses task complexity data as input. Task complexity is frequently modeled with three different approaches (i) *subjective*, which considers an individual's psychological state, (ii) *objective*, which considers task characteristics and properties, and (iii) an integration of the two approaches [53]. However, *subjective* approaches are challenging to implement since their reliability is impacted by individual differences [54]. For example, a math student may perceive complex mathematics problems easy to solve but on the other hand, may perceive aerial work hard. However, individuals with different backgrounds (e.g., construction workers) may perceive the complexity of these tasks differently. Therefore, in this method, a task complexity metric that considers task characteristics and properties is implemented.

Depending on the gamified task (e.g., cognitive task, physical task), different methods that consider task characteristics and properties can be used to measure task complexity (see [27], [54], [55]). For example, Wood [55] proposed a complexity model that described tasks according to three elements: (i) *information cues,* (ii) *products*, and (iii) *acts*. *Information cues* are stimuli that are used to make conscious discriminations. While, p*roducts* are quantifiable outcomes of *acts*, and *acts* are the required steps for creating the *product*. Based on these elements, the model defines task complexity as a function of (i) *dynamic complexity*, (ii) *component complexity*, and (ii) *coordinative complexity*. *Dynamic complexity* relates to the variability between task inputs and *products* over time (e.g., game rules changing over time). *Component complexity* relates to the number of *acts* needed to complete a task (e.g., steps required to complete a task). *Coordinate complexity* relates to the strength between *acts*, *products*, *information cues*, and task inputs (e.g., tasks requiring greater dexterity to perform) [17]. Similarly, in the context of gamification, Lopez and Tucker [27] proposed a task complexity metric to evaluate the physical effort required to perform a task in physically-interactive gamified applications based on the body movements required to perform it (see section V.A.1) .

2) Facial Keypoint data: Facial keypoint data is utilized since it can be captured without affecting an individual's immersion or ability to interact with an application. In this work, a non-

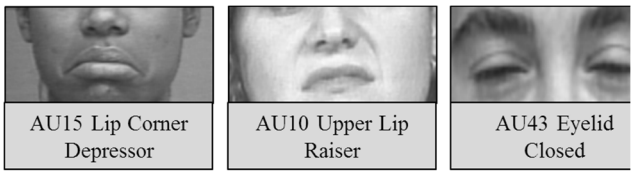


Fig. 2. Actors illustrating a set of Actions Units, from Ref. [56]

wearable sensor is used to collect the facial keypoint data of an individual i before performing a task t ($\mathbf{F}_{it}$). In this work, the facial keypoint data is measured as a relative weight from an Action Unit (AU), ranging from 0-1. This facial keypoint data resembles the Facial Action Coding System [56], in which expert raters code the facial displays of an individual, as illustrated in Fig. 2. The method presented can also be implemented with facial keypoint data measured as two-dimensional coordinates from an image. Nonetheless, in such a case, the facial keypoints need to be regularized and normalized. This normalization can be done via a regularized mean shift algorithm and an ordinary Procrustes analysis, as in [51], [57].

In this work, the facial keypoint data of an individual i consists of j independent facial keypoint time series (for $j \in$ set of facial keypoints). These are collected while individual i interacts with a gamified application *App,* after being introduced to the task t and before completing the task (for $t \in$ set of gamified tasks $\{\mathbf{T}\}$, and *App* $\in$ set of gamified applications). Therefore, the facial keypoint data of an individual i on a task t ($\mathbf{F}_{it}$) is a matrix with n rows and j columns, where n denotes the length of the time series. The length of the time series depends on the duration of the individual's interaction with the gamified application before performing the task and the frequency in which the data is collected. For example, Fig. 3 shows a representation of an individual's facial keypoints q and k (i.e., AU q and k) captured before performing the tasks of an application (i.e., t=$\{1,2,\ldots T\}$). Assuming that the frequency of data captured was 10 frames/sec (i.e., 10Hz) and the tasks were performed every 6 sec, the data captured will generate T matrices (i.e., $\{\mathbf{F}_{i1}$, $\mathbf{F}_{i2},\ldots\ \mathbf{F}_{iT}\}$) with 2 columns (i.e., q and k) and 60 rows (i.e., n=10 frames/sec x 6sec).

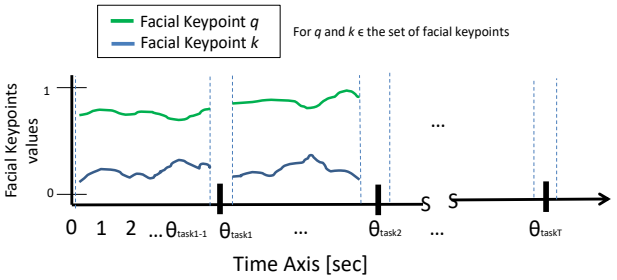


Fig.3. Illustration of facial keypoints data acquisition

3) Performance data: In gamified applications, the tasks are designed such that by successfully performing them, individuals will meet the objective of the application. Due to this relationship, researchers have used individuals' performance on the gamified task as a proxy for measuring their performance in meeting the objective of an application. Therefore, in this work, the same approach is used. For the purpose of this work, the performance of an individual i on a task t is assumed to be a binary variable, where:

Y_{it}= 1, if individual i successfully performed a task t

Y_{it}=0, otherwise.

For,

- $i \in$ set of individuals $\{\mathbf{I}\}$
- $t \in$ set of tasks $\{\mathbf{T}\}$

B. Model Generation

The objective of this step is to build an *adaptive-individual-task* machine learning model to accurately predict the performance of an individual i on a task t (i.e., Y_{it}). The model uses as predictor variables, the mean and standard deviation value of an individual's facial keypoint data captured before performing a gamified task (i.e., $\mathbf{F}\mu_{it}$,$\mathbf{F}\sigma_{it}$), the complexity of the task *(i.e., PC_t)*, as well as individual and application identifier data (i.e., *ID, App*). In order to account for the dynamic nature of facial expressions, and based on previous studies which suggest that reactions are evident in individuals' facial expressions just after one second of stimulus onset [58], the mean and standard deviation of individuals' facial keypoint data is calculated every second (i.e., a 1 second time window). Moreover, the model is first trained with a dataset of a general population of individuals. Then, as new data of an individual of interest is acquired, the training set is updated, and the model is re-trained. This approach allows mitigation of the "*cold start*" problem [59] since before an individual interacts with an application, no prior information of that individual's interaction with the application exists.

In this work, multiple machine learning algorithms are implemented to test their capability to generate a model that can accurately predict an individual's performance on a gamified task. Specifically, in this work, a Logistic Regression, Naïve Bayesian, Support Vector Machines, Random Forest, and a Neural Network classification algorithm are implemented. The performance and computational resources required to train the model using these machine learning algorithms are evaluated. These algorithms were selected since they are frequently used in the *Affective Computing* community, and have different underlying processes for generating classification models (e.g., model-based, decision tree) [40], [60].

C. Model Validation

For the machine learning model to be viable, its accuracy and robustness need to be evaluated. In this work, a cross-validation (CV) approach is implemented. A CV approach requires the partitioning of the dataset into two sets: (i) a training set, and (ii) a testing set. A model is trained using the training set, while the testing set is used to validate the model's accuracy. First, to benchmark the different machine learning algorithms and to address ***RQ1***, a 10-fold CV approach is implemented. In this approach, the dataset is randomly partitioned into 10-folds. In each of the 10 iterations of this CV approach, one fold is used as a testing set while the remaining are used as a training set. To address ***RQ2***, the *adaptive-individual-task* model is evaluated using an iterative leave-one-out CV approach. This

approach is implemented to simulate the scenario in which new data of an individual of interest is acquired, and the model is re-trained. Although the 10-fold CV approach will not evaluate the changes in the model's accuracy as more data of an individual is acquired and the model is retrained, it will help benchmark the performance of the different machine learning algorithms while requiring less computational resources than the iterative leave-one-out CV approach [61]. Moreover, the 10-fold CV approach will produce an accuracy estimator with less variance [61].

For the iterative leave-one-out CV approach, the same testing sets are used in each of the instances to maintain consistency between the iterations of the procedure. Therefore, in each of the leave-one-out instances, the data pertaining to an individual *i* performing the tasks of an application is randomly partitioned into two-thirds for training and one-third for testing. In the first iteration, the training set of the model is composed of a set that does not contain data of the individual of interest (i.e., individual *i*). Hence, in this first iteration, the training and testing sets are *person independent,* which produces a general model. In the subsequent iterations, an extra tuple containing information about individual *i* performing a given task *t* is randomly added to the training set. An extra tuple is added, and the *adaptive-individual-task* model is re-trained. This process is followed until all the tuples from the two-thirds training partition are used. This procedure is performed for all the individuals in the dataset.

Figure 4 illustrates an example of this iterative leave-one-out CV approach. In this example, a dataset of 68 individuals (i.e., *ID*), performing 12 different tasks of different complexity (i.e*., PC),* in two gamified applications (i.e., *App*) is used. Therefore, the dataset is composed of a total of 816 tuples (i.e., 68*12). In the first leave-one-out instance, the 12 tuples of individual *ID=1* are randomly partitioned, 8 are used for training while the remaining 4 are used for testing. In the first iteration, the model

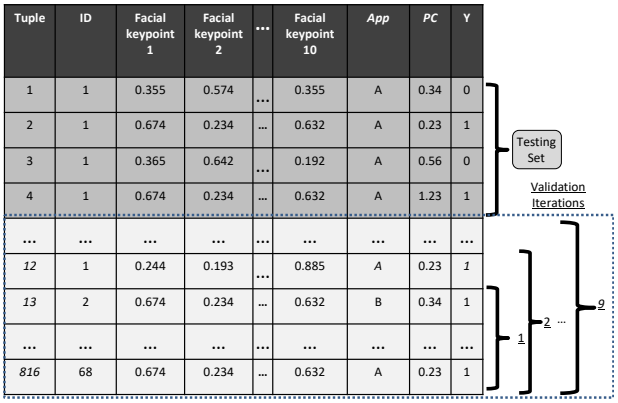


Fig. 4. Example of the iterative leave-one-out cross-validation approach

is trained with a dataset that does not contain any tuple of the individual *ID*=1 (i.e., tuples 13-816). In the remaining iterations, an additional tuple is randomly added to the training set of the previous iteration, one at a time. That is, for iteration 2, the training set consists of the same set as in iteration one, plus the addition of one extra tuple (i.e., 12-816). Therefore, for iteration 9, the training set consists of tuples 5 to 816. This process is carried out for the remaining individuals. This approach helps simulate the scenarios where new data of an individual of interest is acquired, and the model's training set updated.

V. CASE STUDY

The method is implemented in two physically-interactive applications. The goal of the gamified applications is to motivate participants to use full body motions (e.g., bend, extend arm, jump) in order to complete a series of tasks. The objective of each application is to improve individuals' physical performance. Thus, these applications could fall within the umbrella of Active Games. In this work, the gamified tasks consisted of a series of obstacle avoidance tasks. In other words, participants were required to perform certain full body motions to pass through a series of obstacles without touching them, similar to the game show *"Hole in the Wall*[TM]" [62]. In these gamified applications, the authors were able to control for the start and completion time of the tasks. This allowed them to systematically capture the facial keypoint data of participants at equal time points. The applications consisted of 12 different sections, each one with its unique gamified task (i.e., obstacle avoidance). Hence, for this case study, a total of *t*=12 tasks were analyzed.

TABLE II
DESCRIPTION OF GAME FEATURES IMPLEMENTED IN THE APPLICATIONS

Application A
Points- The score measurement of an individual was shown in the top left corner of his/her visual field.
Avatar- The individuals were given the option to change the color of the avatar that will represent them in the virtual environment.
Content Unlocking- Coins were placed throughout the games in different locations. If more than 21 were collected the individual was allowed to change the gaming environment background.
Application B
Win States- At the end of the application, the individuals were told if they had won or lost based on a threshold score level.
Chance- The individuals were given the opportunity to assign a virtual environment background at random.
Achievements- There were three possible achievements individuals could accomplish shown at the beginning of the application. They were: *(i) Lucky Strike:* Get through 3 obstacles in a row without touching*, (ii) Hops:* Jump while going through an obstacle*, (iii) Contortionist:* Pass every obstacle flawlessly.

The two physically-interactive applications used in this case study only differed in the set of game features implemented. The set of game features implemented in each application (i.e., Application A and Application B) were selected based on their presence in successful and unsuccessful applications, respectively (see [63]). The applications consisted of A) 3

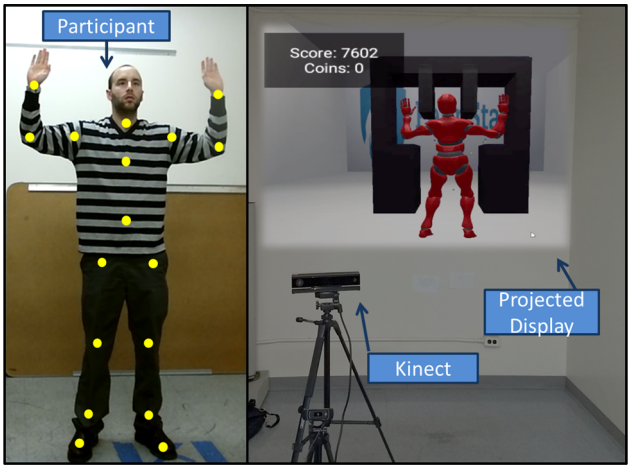


Fig. 5 . Experimental Setup

features common in "*successful*" applications, and B) 3 features common in "*unsuccessful*" applications. Table II shows a brief description of the game features implemented. The applications of this work resemble the ones used in [27].

In this case study, a total of 71 students from the Pennsylvania State University, with ages ranging from 18 to 23 years old (M= 20, and SD= 1.2) participated in the experiment. All of the participants were given an introduction to the applications and experimental setup. After the completion of the informed consent documents, the participants completed a pre-experiment questionnaire, and were then randomly assigned to one of the two applications. Due to technical difficulties in the data acquisition process, only data from 68 participants was analyzed in this work.

A. Data Acquisition

In this case study, the multimodal infrared Kinect sensor was used to capture individuals' facial keypoint data. Moreover, this sensor allowed participants to interact with the physically-interactive applications. The Kinect sensor is used in this work because of its low cost and capability to capture data in real-time without affecting participants' immersion and ability to interact with the applications, as in previous studies [39], [64]. Figure 5 shows the experimental setup used in this case study. Figure 5 shows a Kinect sensor setup in front of a participant and a projected display that allows participants to visualize the applications' virtual environment. As the participant interacted with the application, the Kinect sensor was able to capture a participant's joints location (i.e., yellow dots in Fig. 5) as well as his/her facial keypoint data. Moreover, on the right side of Fig. 5, an illustration of the virtual environment of Application A displayed to the participants is shown.

1) Task data: Due to the physical task of the gamified applications used in this case study, the method for assessing task complexity presented in [27] is implemented. Nonetheless, the *adaptive-individuals-task* model is not constrained to any particular method that measures the complexity of a task based on its characteristics and properties (i.e., "*objective"* approach) (see section IV.A.1). This approach is capable of assessing the physical effort required to perform a task. The approach implements a task complexity metric (*PC*) that is a function of the sum of the Euclidean distances from an individual's joint positions at rest (i.e., X_l^{rest}, Y_l^{rest}, Z_l^{rest}) to the joint positions needed to successfully perform a task t (i.e., X_l^t, Y_l^t, Z_l^t), for $l \epsilon$ the set of joints {**L**} and $t \epsilon$ set of tasks {**T**}, as shown in Eq. (1).

$$PC_t = \sum_{l=1}^{L} \sqrt{(X_l^{rest} - X_l^t)^2 + (Y_l^{rest} - Y_l^t)^2 + (Z_l^{rest} - Z_l^t)^2} \quad (1)$$

In this work, an individual standing up with his/her arms close to the body (e.g., Fig. 6, part A), is considered to be in resting position. Figure 6 part A, shows an illustration of an individual skeletal system with the position coordinates of the right-hand joint. The values of these coordinates are measured as a relative distance from a reference point (e.g., dotted circle in Fig. 6). This joint position data is employed to evaluate the complexity of the task. For example, Fig. 6 part B and C illustrate an individual performing a task (i.e., collecting "*lives*"). It can be seen that the task in part B requires more physical effort to perform than the task in part C. In this example, the coordinates of the right hand joint (i.e., *l*=RH) while at rest, as shown in part A, are X_{RH}^{rest}= Y_{RH}^{rest}= Z_{RH}^{rest} =1m (for this example, the joint coordinates are given as the distance from the reference point: [0,0,0] in meters [m]). Moreover, the coordinates of the right hand joint while performing the task in part B (i.e., t=B), are X_{RH}^{B}=2m, Y_{RH}^{B}=1m, Z_{RH}^{B}=2m, while for part C are X_{RH}^{C}=4m, Y_{RH}^{C}=3m, Z_{RH}^{C}=3m. Using Eq. (1) it can be shown that PC_C=4.1m is greater than PC_B=1.4m. This suggests that less physical effort is needed to perform the task in part B, compared to the task in part C. This is because the location of the item in part B requires individuals to move a shorter distance.

The results presented in [27] suggests that with a limited set of joint position data (i.e., **L**≥13), the physical effort of tasks that require full body motion (e.g., jump, walk) can be accurately capture. However, this metric makes the assumption that the relative difference between the complexities of performing different physical tasks will not change significantly based on an individual's anthropometry. That is, it

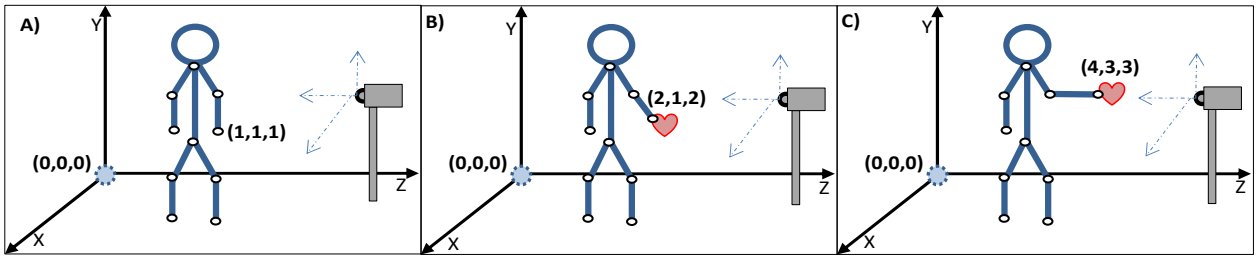


Fig. 6. Example of task complexity assessment

is assumed that a physical task that entails more movement of limbs (e.g., bend) will require more effort to perform (i.e., more complex) than a task that entails less movement of limbs (e.g., extend arm), and that this will be independent of the anthropometry of the individual performing the tasks. Hence, the *PC* values of a task can be acquired from a pilot test of the application, as suggested by [27]. In this pilot test, the joints' position at rest and the position needed to successfully perform a gamified task *t* can be acquired.

TABLE III
PARTICIPANTS' JOINTS TRACKED BY THE MICROSOFT KINECT

1	Right shoulder	7	Right hip	13	Head
2	Right elbow	8	Right knee	14	Neck
3	Right wrist	9	Right ankle	15	Pelvis
4	Left wrist	10	Right toe	16	Left hip
5	Left elbow	11	Left knee	17	Left toe
6	Left shoulder	12	Left ankle		

In this work, the 17 joints tracked by the Kinect sensor (see Table III) were used to calculate the task complexity values of the gamified tasks. The joint locations consisted of X, Y, and Z position data relative to the location of the sensor. Even though both applications implemented the same gamified tasks (i.e., obstacles), the game feature of Content Unlocking impacted the task complexity of Application A. This was due to the locations of the coins used to implement the Content Unlocking feature (see Table II). The coins were positioned before each obstacle of Application A and their position varied by obstacles. To collect the coins individuals had to incur in greater physical effort compared to individuals of Application B, which had no coins. This was due to the location of the coins not being aligned with the obstacles. This additional task of collecting the coins to unlock the new content had a direct impact on the complexity of the tasks of Application A (see Table 6 Ref. [27]). Therefore, a total of 24 measurements of task complexity were calculated (i.e., 12 for Application A, 12 for Application B).

2) Facial Keypoint data: The Kinect sensor captured participants' facial keypoint data as they interacted with the applications. The Kinect SDK is capable of automatically capturing the facial keypoints shown in Table IV by implementing the CANDIDE-3 model [65]. These facial keypoints are able to be captured if a participant had his/her eyebrows or jaw lowered, eyelids closed, and/or lips raised or stretched before completing a gamified task. The AUs presented in Fig. 2 relate to some of the facial keypoints that the Kinect sensor is capable of capturing. For example, if a participant has his/her eyes closed (e.g., similar to the actor shown on the right of Fig. 2), the *Right Eyelid Closed* and *Left Eyelid Closed* facial keypoint values will show as 1; while 0 if the eyes are completely open.

TABLE IV
FACIAL KEYPOINT DATA COLLECTED.

1	Upper Lip Raised	6	Right Eyelid Closed
2	Left Lip Stretched	7	Left Eyelid Closed
3	Right Lip Stretched	8	Jaw Lowered
4	Left Brow Lowered	9	Right Brow Lowered
5	Left Lip Corner Depressor	10	Right Lip Corner Depressor

The equal time intervals of the gamified tasks allows for the systematic capture of facial keypoint data of participants at equal time points; thus, generating equal length time series. The facial keypoint data of each participant *i* on a given task *t* was captured continuously for 6 seconds at a rate of 10 frames/second (i.e., 10Hz). This resulted in a facial keypoint data matrix for each participant *i* on a given task *t* ($\mathbf{F}_{it}$) with 10 columns (j=10) and 60 rows (n=60). The matrices of facial keypoint data values ($\mathbf{F}_{it}$) were collected for each of the 68 participants (i=68) on each of the 12 tasks (t=12). The *adaptive-individuals-task* model uses the average, and standard deviation values of participants' facial keypoint data captured every second while interacting with a gamified application *App,* after been introduced to the task *t* and before completing the task *t.* That is, for each individual *i* on a task *t*, their respective $n \times j$ (i.e., 60x10) facial keypoint data matrix ($\mathbf{F}_{it}$) is transformed to a 6x10 matrix of average values ($\mathbf{F}\mu_{it}$) (i.e., average over 10 data points) and a 6x10 matrix of standard deviation values ($\mathbf{F}\sigma_{it}$) which are used as input for the proposed model.

3) Performance data: In addition to capturing facial keypoint data, the Kinect sensor is capable of capturing individuals' joint location data (see Table III), which enables participants to interact with the applications in the virtual environment. This data also enables the applications to assess in real-time, whether a participant *i* successfully performed a task *t* (Y_{it}= 1) or not (Y_{it}= 0). For example, Fig.7 shows a representation of a participant performing a task in Application B, with the 17 joints tracked by the sensor highlighted. In this figure, the joints highlighted in green indicate the ones within the predefined obstacle avoidance area for that specific task, while the red ones indicate the joints outside this area. Hence, for a participant *i* to successfully perform a task *t*, all of his/her 17 joints have to be within the obstacle avoidance area of that task. Hence, in this example, the participant did not successfully perform the task since not all of his/her joints were within the obstacle avoidance area.

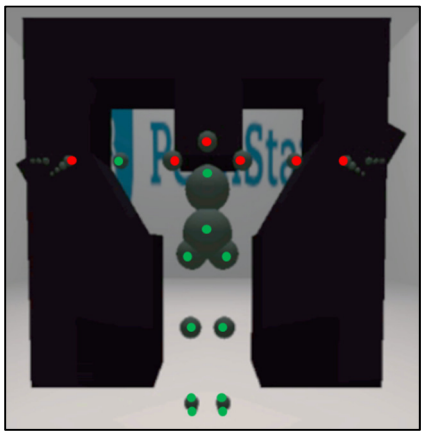


Fig. 7. Illustration of a gamified task with joints highlighted for visualization.

B. Model Generation

The different machine learning algorithms used to generate the model were implemented in R (v.3.5.1) [66]. The Support Vector Machines and Random Forest algorithms were implemented with the R package *e1071* (v. 1.6-7) [67] while the Logistic Regression was implemented with the package *caTools* (v.1.17.1) [68]. The Naïve Bayesian was implemented with *klaR* (v.0.3.3) [69], and the Neural Network algorithm with *nnet* (v.7.3-12) [70]. The hyper-parameters of the algorithms were tuned using a random search approach.

C. Model Validation

To benchmark the performance of the machine learning algorithms, first a 10-fold cross-validation approach is implemented. Subsequently, to address ***RQ2***, an iterative leave-one-out cross-validation procedure, as proposed in section IV.C is used. In this case study, 816 instances are analyzed, one for each participant i on a gamified task t (e.g., *68*12*=816). Therefore, the leave-one-out cross-validation procedure might be understood as training and testing 68 models. Moreover, since data of 12 different tasks was acquired, there are nine validation iterations. In each iteration, the testing set for the models consisted of 4 tuples pertaining to a participant i performing four randomly selected tasks. For the first iteration, the training set consisted of a dataset that did not contain data of the individual of interest (i.e., training and testing sets were *person independent*). Subsequently, the remaining eight tuples were randomly added, one at a time, to the training set during iterations 2 through 9. This approach was implemented to simulate the scenarios where new data of an individual of interest is acquired, and the model is re-trained, similar to the example shown in Fig.4.

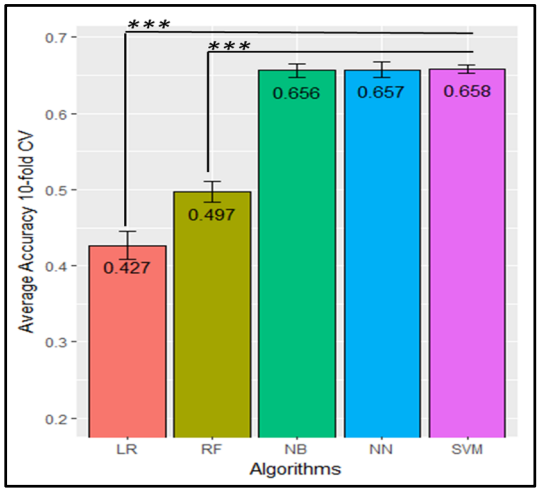
From the 816 tuples of the dataset, 341 corresponded to participants who successfully performed the gamified tasks, while 475 corresponded to participants who did not. Similarly, 444 of these instances corresponded to participants who interacted with Application *A* (e.g., 37 participants on 12 tasks), while 372 corresponded to participants who interacted with Application *B* (e.g., 31 participants on 12 tasks). Each tuple of the dataset was composed of: (i) the performance data of a participant i on a task t (Y_{it}), (ii) the complexity of the gamified task t (PC_t), (iii) participant's average and standard deviation of facial keypoint data ($\mathbf{F}\mu_{it}$, $\mathbf{F}\sigma_{it}$), and identification variable for the (iv) participant *id* and the (v) application he/she interacted with (App_i).

VI. Results and Discussions

The benchmark results of the 10-fold cross-validation revealed that the Support Vector Machine (SVM) algorithm had the greatest average accuracy out of the methods tested (M=0.658, SD=0.018). Figure 8 shows a summary of these accuracy results. The independent *t*-tests indicate that the average accuracy of the SVM was statistically significantly greater than the average accuracy of the Logistic Regression (LR) (M=0.427, SD=0.060, t_{18}=11.66, *p-value*<0.001) and Random Forrest (RF) (M=0.497, SD=0.044, t_{18}=10.71, *p-value*<0.001) algorithms. However, the average accuracy of the SVM was not significantly different than the average accuracy of the Neural Network (NN) (M=0.657, SD=0.031, t_{18}=0.09, *p-value*=0.465), and Naïve Bayesian (NB) algorithms (M=0.656, SD=0.028, t_{18}=0.19, *p-value*=0.426). Nonetheless, it is important to highlight that with the 10-fold cross-validation approach, the SVM algorithm took 107 seconds to train and test, while the NN and NB algorithms took 261 seconds and 145 seconds, respectively. Hence, out of the top performing algorithms, the SVM reached the greatest accuracy and required the least computational resources.

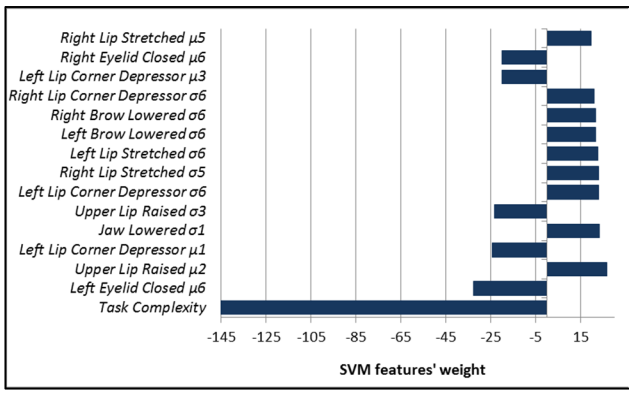
Moreover, the *t*-test results indicate that the accuracy of the models generated with the SVM (t_{19}=27.76, *p-value*<0.0.01), NN (t_{19}=16.02, *p-value*<0.0.01), and NB (t_{19}=17.62, *p-value*<0.0.01) machine learning algorithms were statistically significantly greater than random chance. These findings help address the ***RQ1*,** indicating that a machine learning model that uses individuals' facial keypoint data and task information can accurately predict the performance of individuals prior to completing a gamified task.

For completeness and to assess the value of considering task information and individuals' facial keypoint data, the *adaptive-individual-task* model was benchmarked against a model that only considered task information (i.e., *task* model) and a model that only considered individuals' facial keypoint data (i.e., *individual* model). This benchmark analysis was performed using a 10-fold cross-validation approach. The models were generated using an SVM algorithm. The independent *t*-tests results indicate that the average accuracy of the proposed model (M=0.658, SD=0.018) was statistically significantly greater than the average accuracy of the *individual* model (M=0.578, SD=0.025, t_{18}=8.21, *p-value*<0.001) and *task* model (M=0.594, SD=0.029, t_{18}=5.93, *p-value*<0.001). In addition, the accuracy of both the *individual* model (t_{19}=9.87, *p-value*<0.001) and *task* model (t_{19}=10.25, *p-value*<0.001) were statistically significantly greater than random chance. The results also indicate that there was no significant difference between the average accuracy of the *task* model and the *individual* model (t_{18}=1.32, *p-value*=0.102). However, the *task* model achieved accuracy greater than random change by using only task information (i.e., task complexity). The results of the Kendall correlation test (τ= -0.53, *p-value*<0.001) reveal that task complexity was negatively correlated with participants' performance.



Note: p-value<0.001***

Fig. 8. Benchmark results of the 10-fold cross-validation analysis.



Note: μ_x and σ_x indicate the average and standard deviation of facial keypoint measured between (x-1)*sec* and (x)*sec*, respectively.

Fig. 9. SVM features weight for the *adaptive-individual-task* model.

Moreover, Fig. 9 shows a plot of the weights of the SVM features used in the model. As the previous correlation results indicate, task complexity is an important feature for predicting individuals' performance on a gamified task. The results of this work indicate that on average, participants successfully perform the less complex tasks than, the more complex ones. These findings are in line with previous studies and the Fogg Behavioral Model [25], [27], indicating that task complexity is significantly correlated to participants performance. Nonetheless, while the *task* model results revealed that task complexity is a good indicator of individuals' performance, the benchmark results also indicated that individuals' facial keypoint data provide additional and valuable discriminatory power for predicting the performance of individuals in gamified tasks.

To address the ***RQ2***, the *adaptive-individual-task* model was validated with an iterative cross-validation approach that simulates scenarios in which new data of an individual is acquired, as presented in section IV.C. Since the testing sets consisted of 4 randomly selected gamified tasks per individual, a total of 272 tuples are used for testing (i.e., 68*4=272). Table V shows the confusion matrix for the 1st validation iteration. That is, the one in which the training and testing sets were *person independent,* which generated a general model. While, Table VI shows the confusion matrix for the 9th validation iteration, the one in which the training set contained 8 tuples from the individuals of interest. The results show that the general model (i.e., 1st validation iteration) was able to classify participants' performance with an accuracy of 0.654 (SD=0.24) and with an F_1-score of 0.435. While in the 9th iteration, the accuracy of the *adaptive-individual-task* model increased to 0.768 (SD=0.213) and an F_1-score of 0.909. The independent *t*-test indicated that this accuracy was significantly greater than the general model's accuracy (t_{134}=2.92, *p-value*=0.002). These results reveal that the performance of the proposed model improves as new data of an individual is acquired and the model is re-trained.

TABLE V
CONFUSION MATRIX 1ST ITERATION (GENERAL MODEL)

		Ground truth	
		Y=1 (Pass)	Y=0 (Fail)
Predicted	Y=1 (Pass)	30	9
	Y=0 (Fail)	85	148
	Total	115	157

TABLE VI
CONFUSION MATRIX 9TH ITERATION (ADAPTIVE-INDIVIDUAL-TASK MODELS)

		Ground truth	
		Y=1 (Pass)	Y=0 (Fail)
Predicted	Y=1 (Pass)	79	27
	Y=0 (Fail)	36	130
	Total	115	157

Figure 10 shows a plot of the model's accuracy vs. the validation iterations. The plot shows that in certain validation iterations (i.e., iteration 4, 7, and 8) the model accuracy does not improve or even worsens, in comparison to the previous iteration. These results can be attributed to the randomness of the validation procedure in which the data tuples are randomly partitioned and assigned to the training and testing sets. The plot indicates that, on average, the *adaptive-individual-task* model's accuracy increases as more data of an individual of interest is acquired and used to re-train the model. A linear regression model was fitted to test the significance of this relationship. The model accuracy was used as the response variable and the validation iterations as predictor variables. The participant's identification variable was used as a control variable to account for any possible variation between participants. Table VII shows a summary of the regression model fitted. The results indicate that the regression equation was significant ($F_{68,543}$=28.34, *p-value*<0.001), with an R^2 of 0.78. The results reveal that the coefficient of the *Intercept* and the *Validation Iteration* variable were statistically significantly different than zero (*p-value*<0.001). These results support Fig.10 since they indicate that as the number of validation iterations increase (i.e., model is re-trained with more data from the individual of interest) the accuracy of the *adaptive-individual-task* model increases. These findings address ***RQ2,*** and reveal that the performance of the proposed method improves as new data of an individual is acquired and the model is re-trained.

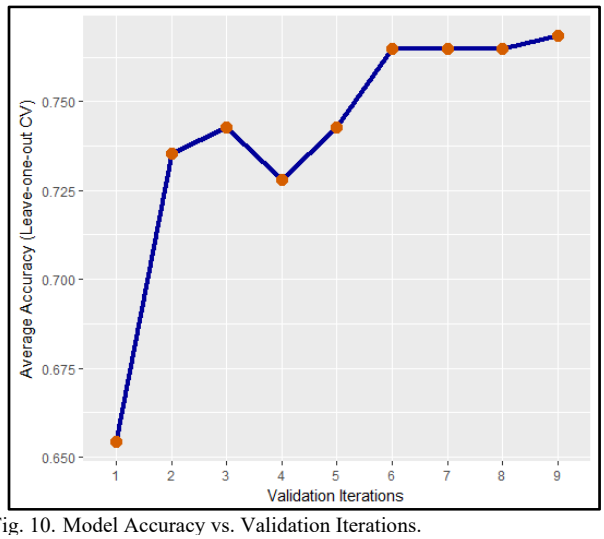


Fig. 10. Model Accuracy vs. Validation Iterations.

TABLE VII
SUMMARY OF LINEAR REGRESSION MODEL FOR ACCURACY

	Estimates	t-value
Intercept	0.83	4.85^{***}
Validation Iteration	0.04	6.15^{***}

Note: p-value<0.001***

In addition, Tables V and VI show that the model tended to correctly classify the instances where the participants did not successfully perform the gamified task more frequently than the instances where they did. In other words, if Y=1 (i.e., successfully performed the gamified task) is considered as the positive condition, the specificity or true negative rate of the models (general model: 0.943, *adaptive-individual-task* model: 0.828) was greater than their sensitivity or the true positive rate (general model: 0.261, *adaptive-individual-task* model: 0.687). In the context of gamification, where designers intend to motivate individuals to successfully perform a task, the results indicate that it is harder to predict if an individual will successfully perform a gamified task than to predict if he/she will not successfully perform it. This difference was more substantial in the general model, which training and testing sets were person independent. These results reveal that the *adaptive-individual-task* model can still be improved. Nonetheless, the model still provides good prediction accuracy with data collected prior the start of the task and not after a participant either fails or succeeds in performing the task, as in previous studies (see section II).

The previous results support the benefits of systematically updating the training set of the *adaptive-individual-task* model as new data of an individual is acquired. This approach allows the model to adapt (i.e., learn) to an individual's unique facial expression characteristics. Nonetheless, if the *adaptive-individual-task* model is re-trained every time new data of an individual is acquired, the computational resources and the time needed to re-train it need to be explored. Hence, the effects that parallelization and the clock speed of CPUs (cores) used has on the time needed to re-train the *adaptive-individual-task* model

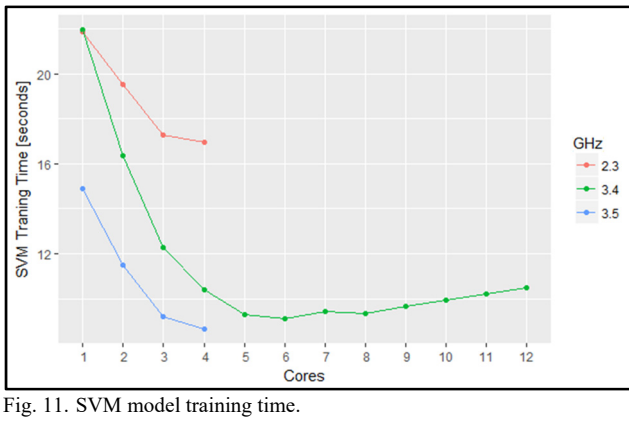


Fig. 11. SVM model training time.

is explored. While other machine learning algorithms have faster training speeds than SVM (e.g., Logistic Regression, Decision trees) [60], the SVM algorithm is used as a benchmark since the results of this work indicate that out of the top performing algorithms, the SVM reached the greatest accuracy and required the least computational resources (see Fig. 8). For this analysis, a (i) 4 Core i5 2.3 GHz Intel™ computer with 6 GB of RAM and Microsoft™ Windows 10, a (ii) 4 Core i5 3.5 GHz Intel™ computer with 8 GB of RAM and Microsoft™ Windows 10, and a (iii) 12 Core i7 3.4 GHz Intel™ computer with 62.8 GB of RAM and Ubuntu 16.04 LTS was used.

Figure 11 shows the time required to train the model using an SVM algorithm, given the number of cores used and the clock speed of the cores in GHz. A second order polynomial model with training time as the dependent variable, and the number of cores and clock speed as the independent variables, was fitted to the data. Table VII shows the summary statistics of the linear regression model. The results indicate that the regression equation was significant ($F_{4,15}$=57.35, *p-value*<0.001), with an R^2 of 0.939. The results reveal that with 4 cores running at a speed of 3.5 GHz, the *adaptive-individual-task* model was trained in 8.67 seconds, which given the conditions of the gamified applications used in this work, is insufficient to provide real-time prediction. However, the regression analysis indicates that with 6 cores running at a speed of 3.6 GHz (e.g., Intel Core i7-7820X, www.intel.com) the *adaptive-individual-task* model can be trained in less than a second, enough to provide real-time predictions if data is collected every 6 seconds as in the case study.

Table VIII
SUMMARY OF LINEAR REGRESSION MODEL FOR TRAINING TIME

	Estimates	t-value
Intercept	-224.59	-3.93^{**}
Number of Cores	-3.39	-9.23^{***}
(Number of Cores)^2	0.20	7.39^{***}
Clock Speed	184.69	4.45^{***}
(Clock Speed)^2	-32.97	-4.57^{***}

Note: p-value<0.001***; p-value<0.01**

VII. CONCLUSIONS AND FUTURE WORKS

Even though researchers are working towards personalized adaptive gamified applications, current methods are not capable of predicting an individual's performance prior to completing a gamified task. This information could be helpful in adapting the game features and task difficulty of gamified applications. Furthermore, current methods are not capable of dynamically capturing an individual's data as he/she interacts with a gamified application. This could limit the degree of personalization and adaptation that current methods can provide. Therefore, due to existing limitations, this work presented an *adaptive-individual-task* machine learning model that uses task information and individuals' facial keypoint data to predict their performance on a gamified task. In this work, individuals' facial keypoint data is captured before completing the task with a sensor that does not affect their immersion or ability to interact with an application. Furthermore, the training data used to generate the machine learning model is updated every time new data of an individual is acquired; hereby, making the model adaptive in nature. A case study involving 68 participants interacting with a set of gamified applications in a virtual environment was presented.

The result of this work provides valuable information about the relationships between individuals' facial keypoint data, as well as their performance and the complexity of gamified tasks.

The results indicate that the *adaptive-individual-task* machine learning model was capable of predicting an individual's performance, with accuracies up to 0.768. While previous studies have focused on developing machine learning models to predict individuals' affective state [44], student type [33], or time spent in performing a vertical menu selection task [31], the performance of the *adaptive-individual-task* model outperformed or closely matches the performance of these existing models. For example, Barata et al. [33] were only able to predict the *students type* with an accuracy of 0.47, even after collecting students' data for a five-week period.

Moreover, the results reveal that the performance of the *adaptive-individual-task* model improves as it is re-trained when more data of an individual is acquired. These results reveal that the model is learning the unique individuals' characteristics and improving its accuracy. Furthermore, the findings presented in this work are in line with previous studies that suggested that the task complexity is of great importance when predicting the performance of individuals.

This work provides quantitative evidence of the feasibility and performance of the *adaptive-individual-task* machine learning models that implement task and facial keypoint data. However, there are several areas for future improvements. For example, the task complexity metric used in this work only consider task characteristics and do not take into account individual's psychological state or their differences. Furthermore, even though a low-cost sensor capable of capturing data in real-time without affecting participants' immersion was implemented in the case study, as in previous works [39], [64], the effect of the applications' tasks on the sensitivity of the sensor was not explored. Nonetheless, the method proposed in this work is not constrained to the sensors implemented in the case study. Current advancements in facial recognition algorithms are allowing researchers to capture facial keypoint data with the use of more widely available sensors (i.e., RGB sensors or webcams) and still maintain high levels of accuracy [71]. Likewise, researchers are currently working on methods to capture body movement and biometrics data with RGB sensors that do not affect an individual's interaction with an application [72], [73]. This type of data should be considered as input for the *adaptive-individual-task* model in future works, as the results of previous studies indicate that this could improve the model's accuracy [39]. Similarly, individuals' in-game behavioral data (e.g., number of coins collected) should be explored since it might provide a better understanding of individuals' attitude towards the gamified application.

Finally, future works should focus on implementing this method with other gamified applications in different environments and tasks (e.g., educational with cognitive tasks), and using this knowledge to tailor the game features and task. This could be an area for future research that will help to test the generalizability of the model and its potential capability to transfer the learning gained from this application to other types of applications. Nevertheless, this work presents initial groundwork towards implementing *adaptive-individuals-task* machine learning models that take advantage of task information and individuals' facial keypoint data to predict their performance in gamified tasks.

ACKNOWLEDGMENTS

This research is funded in part by NSF NRI #1527148 and the NSF I/UCRC Center for Healthcare Organization Transformation (CHOT), NSF I/UCRC award #1624727. Any opinions, findings, or conclusions found in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] D. Dicheva, C. Dichev, G. Agre, and G. Angelova, "Gamification in Education : A Systematic Mapping Study," *Educ. Technol. Soc.*, vol. 18, no. 3, pp. 75–88, 2015.

[2] A. T. Ferreira, A. M. Araújo, S. Fernandes, and I. C. Miguel, "Gamification in the workplace: A systematic literature review," in *World Conference on Information Systems and Technologies*, 2017, pp. 283–292.

[3] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From Game Design Elements to Gamefulness : Defining ' Gamification ,'" *ACM MindTreck'11*, 2011.

[4] T. Aldemir, B. Celik, and G. Kaplan, "A qualitative investigation of student perceptions of game elements in a gamified course," *Comput. Human Behav.*, vol. 78, pp. 235–254, 2018.

[5] E. Biddiss and J. Irwin, "Active video games to promote physical activity in children and youth," *Arch. Pediatr. Adolesc. Med.*, vol. 164, no. 7, pp. 664–672, 2010.

[6] M. Böckle and M. Bick, "Towards adaptive gamification: A synthesis of current developments," in *Proceedings of the 25th European Conference on Information Systems (ECIS).*, 2017, vol. 2017.

[7] C. Pedersen, J. Togelius, and G. N. Yannakakis, "Modeling player experience for content creation," *IEEE Trans. Comput. Intell. AI Games*, vol. 2, no. 1, pp. 54–67, 2010.

[8] D. Hooshyar, C. Lee, and H. Lim, "A Survey on Data-Driven Approaches in Educational Games," *IEEE 2nd Int. Conf. Sci. Inf. Technol.*, pp. 291–295, 2016.

[9] C. S. González, P. Toledo, and V. Muñoz, "Enhancing the engagement of intelligent tutorial systems through personalization of gamification," *Int. J. Eng. Educ.*, vol. 32, no. 1, pp. 532–541, 2016.

[10] G. Chanel, C. Rebetez, M. Betrancourt, and T. Pun, "Emotion assessment from physiological signals for adaptation of games difficulty," *IEEE Trans. Syst. Man, Cybern. A Syst. Humans*, vol. 41, no. 6, pp. 1052–1063, 2011.

[11] S. K. D'Mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, p. A43, 2015.

[12] C. E. Lopez and C. S. Tucker, "From mining affective states to mining facial keypoint data: The quest towards personalized feedback," in *ASME Int. Design Eng. Technical Conf. and Computers and Infor. in Eng. Conf.*, 2017, p. V001T02A039-V001T02A039.

[13] E. Hudlicka, "Affective computing for game design," in *Proceedings of the 4th Intl. North American Conference on Intelligent Games and Simulation (GAMEON-NA)*, 2008, pp. 5–12.

[14] I. Kotsia, S. Zafeiriou, and S. Fotopoulos, "Affective gaming: A comprehensive survey," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 663–670, 2013.

[15] G. N. Yannakakis and J. Hallam, "Real-time adaptation of augmented-reality games for optimizing player satisfaction," in *2008 IEEE Symposium on Computational Intelligence*

and Games, CIG 2008, 2008, pp. 103–110.

[16] M. D. Hanus and J. Fox, "Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance," *Comput. Educ.*, vol. 80, pp. 152–161, 2015.

[17] T. Hu, Q., Bezawada, S., Gray, A., Tucker, C., & Brick, "Exploring the link between task complexity and students' affective states during engineering laboratory activities," in *ASME 2016 Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf.*, 2016, p. V003T04A019-V003T04A019.

[18] A. Knutas, M. Granato, R. Van Roy, J. Kasurinen, T. Hynninen, and J. Ikonen, "Profile-based algorithm for personalized gamification in computer-supported collaborative learning environments," in *In GHITALY17: 1st Workshop on Games-Human Interaction*, 2017, vol. 1956.

[19] J. Hamari and J. Tuunanen, "Player types: A meta-synthesis," *Trans. Digit. Games Res. Assoc.*, vol. 1, no. 2, pp. 29–53, 2014.

[20] P. Buckley and E. Doyle, "Individualising gamification: An investigation of the impact of learning styles and personality traits on the efficacy of gamification using a prediction market," *Comput. Educ.*, vol. 106, pp. 43–55, 2017.

[21] Y. Jia, B. Xu, Y. Karanam, and S. Voida, "Personality-targeted gamification: A survey study on personality traits and motivational affordances," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 2016, pp. 2001–2013.

[22] D. Codish and G. Ravid, "Personality based gamification: How different personalities perceive gamification," in *Proceedings of the 22nd European Conference on Information Systems (ECIS)*, 2014.

[23] J. Koivisto and J. Hamari, "Demographic differences in perceived benefits from gamification," *Comput. Human Behav.*, vol. 35, pp. 179–188, 2014.

[24] J. L. Sabourin and J. C. Lester, "Affect and engagement in game-based learning environments," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 45–56, 2014.

[25] B. J. Fogg, "A Behavior Model for Persuasive Design," in *Proceedings of the 4th International Conference on Persuasive Technology*, 2009, p. 40:1–40:7.

[26] P. Denny, "The Effect of Virtual Achievements on Student Engagement," *ACM CHI'13*, pp. 763–772, 2013.

[27] C. E. Lopez and C. S. Tucker, "A quantitative method for evaluating the complexity of implementing and performing game features in physically-interactive gamified applications," *Comput. Human Behav.*, vol. 71, pp. 42–58, 2017.

[28] G. Bailly, A. Oulasvirta, D. P. Brumby, and A. Howes, "Model of visual search and selection time in linear menus," in *Proceedings of the 32nd annual ACM CHI*, 2014, pp. 3865–3874.

[29] P. M. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement," *J. Exp. Psychol.*, vol. 47, no. 6, pp. 381–391, 1954.

[30] T. S. Jastrzembski and N. Charness, "The model human processor and the older adult: Parameter estimation and validation within a mobile phone task," *J. Exp. Psychol. Appl.*, vol. 13, no. 4, pp. 224–248, 2007.

[31] Y. Li, S. Bengio, and G. Bailly, "Predicting human performance in vertical menu selection using deep learning," *CHI 2018 Conf. Hum. Factors Comput. Syst.*, 2018.

[32] S. K. Card, T. P. Moran, and A. Newell, "Keystroke-level model for user performance time with interactive systems," *Commun. ACM*, vol. 23, no. 7, pp. 396–410, 1980.

[33] G. Barata, S. Gama, J. Jorge, and D. Gonçalves, "Early prediction of student profiles based on performance and gaming preferences," *IEEE Trans. Learn. Technol.*, vol. 9, no. 3, pp. 272–284, 2016.

[34] G. Barata, S. Gama, J. Jorge, and D. Gonçalves, "Engaging engineering students with gamification," in *5th international IEEE conference on Games and virtual worlds for serious applications (VS-GAMES)*, 2013, pp. 1–8.

[35] B. Monterrat, M. Desmarais, E. Lavou, and S. George, "A player model for adaptive gamification in learning environments," in *International Conference on Artificial Intelligence in Education*, 2015, pp. 297–306.

[36] B. Monterrat, É. Lavoué, and S. George, "Motivation for learning: Adaptive gamification for web-based learning environments," in *Proceedings of the 6th International Conference on Computer Supported Education*, 2014, pp. 117–125.

[37] R. Lopes and R. Bidarra, "Adaptivity challenges in games and simulations: A survey," *IEEE Trans. Comput. Intell. AI Games*, vol. 3, no. 2, pp. 85–99, 2011.

[38] B. Schuller, E. Marchi, S. Baron-Cohen, H. O'Reilly, P. Robinson, I. Davies, O. Golan, S. Friedenson, S. Tal, S. Newman, N. Meir, R. Shillo, A. Camurri, S. Piana, S. Bölte, D. Lundqvist, S. Berggren, A. Baranger, and N. Sullings, "ASC-Inclusion: Interactive emotion games for social inclusion of children with autism spectrum conditions," *Proc. 1st Int. Work. Intell. Digit. Games Empower. Incl.*, no. May, 2013.

[39] A. Psaltis, K. C. Apostolakis, K. Dimitropoulos, and P. Daras, "Multimodal student engagement recognition in prosocial games," *IEEE Trans. Comput. Intell. AI Games*, no. c, pp. 1–1, 2017.

[40] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, 2010.

[41] T. Christy and L. I. Kuncheva, "Technological advancements in affective gaming: A historical survey," *GSTF Int. J. Comput.*, vol. 3, no. 4, pp. 7–15, 2014.

[42] C. Grappiolo, Y. G. Cheong, J. Togelius, R. Khaled, and G. N. Yannakakis, "Towards player adaptivity in a serious game for conflict resolution," in *Proceedings 3rd Int. Conf. on Games and Virtual Worlds for Serious Applications,* 2011, pp. 192–198.

[43] N. Shaker, G. N. Yannakakis, and J. Togelius, "Towards Automatic Personalized Content Generation for Platform Games.," in *Proceedings of the 6th Conference on Artificial Intelligence and Interactive Digital Entertainment.*, 2010, pp. 63–68.

[44] S. Asteriadis, N. Shaker, and K. Karpouzis, "Towards player's affective and behavioral visual cues as drives to game adaptation," in *LREC workshop on multimodal corpora for machine learning, Istanbul.*, 2012, pp. 1–4.

[45] N. Shaker, S. Asteriadis, G. N. Yannakakis, and K. Karpouzis, "Fusing visual and behavioral cues for modeling user experience in games," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1519–1531, 2013.

[46] C. Athanasiadis, E. Hortal, D. Koutsoukos, and C. Z. Lens, "Personalized , affect and performance-driven Computer-based Learning," in *CSEDU 2017, 9th International Conference on Computer Supported Education*, 2017.

[47] M. Ben Ammar, M. Neji, A. M. Alimi, and G. Gouardères, "The affective tutoring system," *Expert Syst. Appl.*, vol. 37, no. 4, pp. 3013–3023, 2010.

[48] G. C. Marchand and A. P. Gutierrez, "The role of emotion in the learning process: comparisons between online and face-to-face learning settings," *Internet High. Educ.*, vol. 15, no.

3, p. 150–160., 2012.
[49] F. Tian, P. Gao, L. Li, W. Zhang, H. Liang, Y. Qian, and R. Zhao, "Recognizing and regulating e-learners' emotions based on interactive Chinese texts in e-learning systems," *Knowledge-Based Syst.*, vol. 55, pp. 148–164, 2014.
[50] A. L. Mondragon, R. Nkambou, and P. Poirier, "Evaluating the effectiveness of an affective tutoring agent in specialized education," *Eur. Conf. Technol. Enhanc. Learn.*, pp. 446–452, 2016.
[51] C. Lopez and C. Tucker, "Towards personalized performance feedback: Mining the dynamics of facial keypoint data in engineering lab environments," in *ASEE Mid-Atlantic Spring Conf.*, 2018.
[52] R. Orji, J. Vassileva, and R. L. Mandryk, "Modeling the efficacy of persuasive strategies for different gamer types in serious games for health," *User Model. User-adapt. Interact.*, vol. 24, no. 5, pp. 453–498, 2014.
[53] D. J. Campbell, "Task complexity: A review and analysis," *Acad. Manag. Rev.*, vol. 13, no. 1, pp. 40–52, 1988.
[54] P. Liu and Z. Li, "Task complexity: A review and conceptualization framework," *Int. J. Ind. Ergon.*, vol. 42, no. 6, pp. 553–568, 2012.
[55] R. Wood, "Task complexity: Definition of the construct," *Organ. Behav. Hum. Decis. Process.*, vol. 37, no. 1, pp. 60–82, 1986.
[56] P. Ekman and W. V. Friesen, "Manual for the facial action coding system," *Consult. Psychol. Press*, 1978.
[57] S. Bezawada, Q. Hu, A. Gray, T. Brick, and C. Tucker, "Automatic facial feature extraction for predicting designers' comfort with engineering equipment during prototype creation," *J. Mech. Des.*, vol. 139, no. 2, p. 021102, 2017.
[58] U. Dimberg, M. Thunberg, and S. Grunedal, "Facial reactions to emotional stimuli: Automatically controlled emotional responses," *Cogn. Emot.*, vol. 16, no. 4, pp. 449–471, 2002.
[59] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *Proceedings of the 25th An. Int. ACM SIGIR Conf. on Research and Development in Inf. Retrieval*, 2002, p. 253.
[60] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Informatica*, vol. 31, pp. 249–268, 2007.
[61] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Proc. 14th Int. Jt. Conf. Artif. Intell. - Vol. 2*, 1995.
[62] Ludia, "Ludia & Fremantle media enterprises: Ink deal to bring Hole in the Wall games exclusively for KinectTM on Xbox 360(R)," *Electronics Business Journal*, vol. 64, 2011.
[63] A. Bharathi, A. Singh, C. S. Tucker, and H. B. Nembhard, "Knowledge discovery of game design features by mining user-generated feedback," *Comput. Human Behav.*, vol. 60, pp. 361–371, 2016.
[64] Sujono and A. A. S. Gunawan, "Face expression detection on Kinect using active appearance model and fuzzy logic," in *Procedia Computer Science*, 2015, vol. 59, pp. 268–274.
[65] J. Ahlberg, "Candide-3. An updated parameterised face," Sweden, 2001.
[66] R. R Development Core Team, *R: A Language and Environment for Statistical Computing*, vol. 1, no. 2.11.1. 2011.
[67] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel, "Misc functions of the Department of Statistics (e1071), TU Wien," *R Packag.*, vol. 1, pp. 5–24, 2008.
[68] J. Tuszynski and M. H. Khachatryan, "caTools: Tools for moving window statistics, GIF, Base64, ROC AUC. R Package, Version 1.17.1," 2013.
[69] C. Roever, N. Raabe, K. Luebke, U. Ligges, G. Szepannek, and M. Zentgraf, "klaR–Classification and Visualization. R package, Version 0.3-3.," 2004.
[70] B. Ripley, "nnet: Feed-forward neural networks and multinomial Log-linear models. R Package , Version 7.3-12.," 2013.
[71] T. Baltrušaitis, P. Robinson, and P. Morency, "OpenFace : an open source facial behavior analysis toolkit," in *IEEE Winter Applications of Computer Vision (WACV)*, 2016, pp. 1–10.
[72] G. Papandreou, T. Zh;, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," *ArXiv, arXiv Prepr.*, vol. 1701.01779, 2017.
[73] A. Sikdar, S. K. Behera, and D. P. Dogra, "Computer-vision-guided human pulse rate estimation: A review," *IEEE Reviews in Biomedical Engineering*, vol. 9. pp. 91–105, 2016.

Christian López is currently a Ph.D. candidate at the Pennsylvania State University. He holds an M.S. in Industrial and Systems Engineering from the Rochester Institute of Technology, NY. He has worked as an Industrial Engineer in both the Service and Manufacturing sectors before pursuing his Ph.D. His research interests are the design and optimization of intelligent decision support systems and persuasive technologies to augment human proficiencies.

Conrad Tucker is an Associate Professor of Engineering Design and Industrial and Manufacturing Engineering at Penn State University. He is also Affiliate Faculty of Computer Science and Engineering and directs the Design Analysis Technology Advancement (D.A.T.A) Laboratory. His research focuses on the design and optimization of complex systems and intelligent assistive technologies through the acquisition, integration and mining of large scale, disparate data. He is the Principal Investigator and Site Director of Penn State's NSF Center for Health Organization Transformation (CHOT), an NSF I/UCRC: Industry/University Cooperative Research Center at Penn State. In February 2016, he was invited by National Academy of Engineering (NAE) President Dr. Dan Mote, to serve as a member of the Advisory Committee for the NAE's Frontiers of Engineering Education. He received his Ph.D., M.S. (Industrial Engineering), and MBA degrees from the University of Illinois at Urbana-Champaign, and his B.S. in Mechanical Engineering from Rose-Hulman Institute of Technology.