Perspective

Machine learning and serious games: opportunities and requirements for detection of mild cognitive impairment

Kyle Leduc-McNiven, Ryan T. Dion, Shamir N. Mukhi, Robert D. McLeod, Marcia R. Friesen

Department of Electrical & Computer Engineering, University of Manitoba, Winnipeg, MB, Canada

Correspondence to: Robert D. McLeod. Department of Electrical & Computer Engineering, University of Manitoba, Winnipeg, MB, Canada. Email: robert.mcleod@umanitoba.ca.

Abstract: This perspective paper presents a simple serious game on a mobile platform (Smartphone game). The game has the integrated capability to track a person's play by storing player metadata on start time, end time, and moves within the game. These data can be analyzed to infer cognitive processes of strategy learning, retention, and recall over a brief period of time for potential future applications in pre-symptomatic assessment of mild cognitive impairment (MCI). Through machine learning (ML), the data are demonstrated to be of utility in providing a "cognitive fingerprint" of play. The ML methods used to classify play use synthetic data generated by robots (bots), ranging from bots playing perfectly to bots playing with various degrees of impairment. The findings include guidance on the volume of data required, as well as the features deemed effective for ML classification of various degrees of bot impairment. The work illustrates several significant considerations when applying ML to simple serious games and the data they can generate.

Keywords: Artificial neural networks; cognitive fingerprint; machine learning (ML); mobile games; serious games

Received: 19 July 2018; Accepted: 30 July 2018; Published: 09 August 2018.

doi: 10.21037/jmai.2018.07.02

View this article at: http://dx.doi.org/10.21037/jmai.2018.07.02

Introduction

Mild cognitive impairment (MCI) constitutes a clinical entity differentiated from healthy control subjects and those with very mild Alzheimer's disease (1). It is an intermediate stage between anticipated and normal age-related cognitive decline and the significant decline associated with dementia. MCI compromises memory, language, thinking and judgment in ways that are more significant than in normal aging (2). As with most diseases, early recognition of MCI and detecting subtle changes in the brain has the potential to improve the efficacy of therapeutics (3,4). The role of serious games combined with machine learning (ML) has the potential to become a technology-mediated assessment tool that is complementary to clinical early diagnosis and therapeutic assessment tools.

Currently, over 35 M people worldwide live with dementia and that is expected to reach 115 M by 2050, fueled by an aging population (5). The negative impacts on the individual, family, and caregivers are significant. As disease-modifying treatments are discovered, early diagnosis will be essential to assist in introducing therapies that can slow the progression and maintain a longer quality of life (6). Aging, although inevitable, is the biggest risk factor for cognitive decline and dementia.

Mental health apps are part of a much larger mobile health (mHealth) space, and the proposed research addresses mHealth apps as technology-mediated empirical measurement tools for cognitive assessment including memory, learning, problem-solving, and other executive processes. The importance and relevance of the serious games combined with ML is the early identification of cognitive decline, crucial for optimal pharmacological treatment and timely provision of (psycho)social care (4,7-11). The first symptoms of cognitive decline may be present several years before a clinical diagnosis of dementia can be made and thus computer aided tools that detect underlying patterns of brain dysfunction are of great importance.

This work represents a perspective on the strong potential afforded by the intersection of serious games as

an example of mHealth, applied to MCI assessment and enhanced with ML.

Presently, there are few new methods for assessing cognitive difficulties related to dementias such as Alzheimer's, while the use of serious games is suggested as a promising approach (9,12-14). The most closely related method to that described here would be that of a computerized Wisconsin Card Sorting Task (15), albeit based on a computerized version of a traditional assessment method. This work goes beyond developing electronic/mobile presentations or reproductions of existing MCI assessment instruments; rather, the work presents mobile games that inherently provide a cognitive assessment function via the analysis of an individual's game-playing data. The work here is also differentiated from the gamification of engaging but unproven cognitive stimulation ("brain games" like Lumosity—www.lumosity.com).

Gamification of MCI assessment opens up opportunities to collect player metadata on large scales that allow for baseline establishment of cognitive abilities across demographic (age) profiles, longitudinal performance of individuals and of groups, and from there, the potential to detect subtle changes in an individual's cognitive processes over time. It is the self-perception of losses of specific cognitive processes such as recognition and recall that can cause anxiety to individuals being assessed. The proposed tools have been designed to include the ability to objectively assess recognition of a game strategy, recall of the strategy, failure to maintain set (reverting to a different strategy or no strategy at all), and perseveration (reverting to an earlier strategy). By the stochastic nature of the prototype game on this work, there is opportunity for distraction, facilitating temporary lapses of concentration or memory (16-18), which the metadata can track and be applied for analysis via ML.

Methods

Our intersection of serious games as an example of mHealth, applied to MCI assessment and enhanced with ML is grounded in a working prototype of a serious game on an Android mobile platform named WarCAT (War Cognitive Assessment Tool). WarCAT represents a prototypical game that is mobile, concurrent, and competitive; further, short duration social games represent a relatively new genre of online mobile cognitive assessment games.

In its present state, each game consists of five rounds of the card game WAR, each game played is measured in seconds, and feedback is near instantaneous. Presently, three levels of a minimum of 100 games each have been implemented in the prototype, and the human player plays against a bot. The bot maintains a consistent strategy for 100 games, and only after a player has demonstrated that they have beaten the bot by a non-chance margin can the player 'level up'. WarCAT tracks real-time player behaviour (metadata) during play including start time of play and end time of play (from which frequency and duration of play can be inferred), as well as each move made by the player within the game, timestamped to also analyze how quickly or slowly a player played.

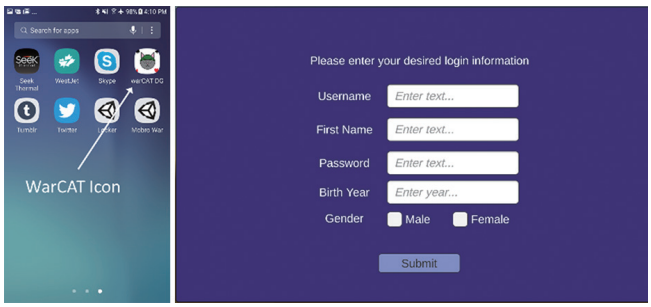

Figure 1 The WarCAT Icon and registration page on Android.

The basic gaming framework has also been further developed during the summer of 2017 to optionally include more traditional cognitive assessments such as a paired associates learning test as an in-app feature which the player completes to proceed. This feature has not proven to be too useful in initial attempts to engage volunteer participation and as such, the following will be more specifically oriented to the game itself.

Figure 1 shows an icon for the game and registration including basic demographic information requested which is essential for the types of ML considered here (i.e., supervised learning). *Figure 2* shows a screenshot of the proof of concept. In *Figure 2*, the players cards are displayed face-up and the player can play them in any order they choose. The objective is for the player to develop, retain, and apply strategy consistently beat a bot that is playing a time invariant strategy. For example, the bot may be repeatedly playing a strategy of low to high cards. A winning and learned strategy would be then playing your second lowest card first, consecutively playing high value cards and playing one's lowest card last.

During play, the game provides the user with immediate feedback as to having won or lost the game (*Figure 3*). Initial feedback from users is that even when losing a hand,

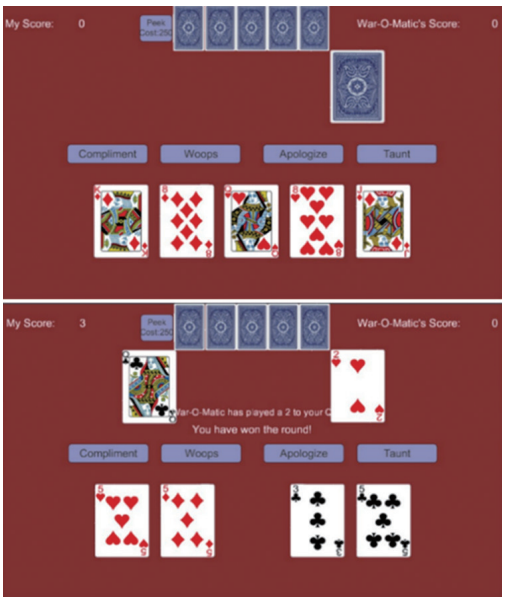

Figure 2 An example play of WarCAT (rudimentary but functional).

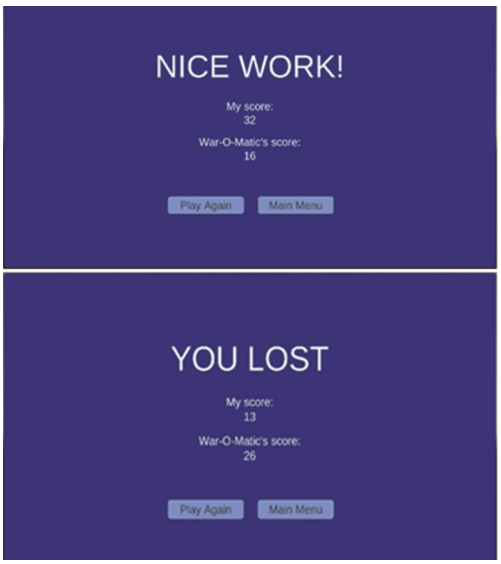

Figure 3 An example of feedback per round (reinforcement learning).

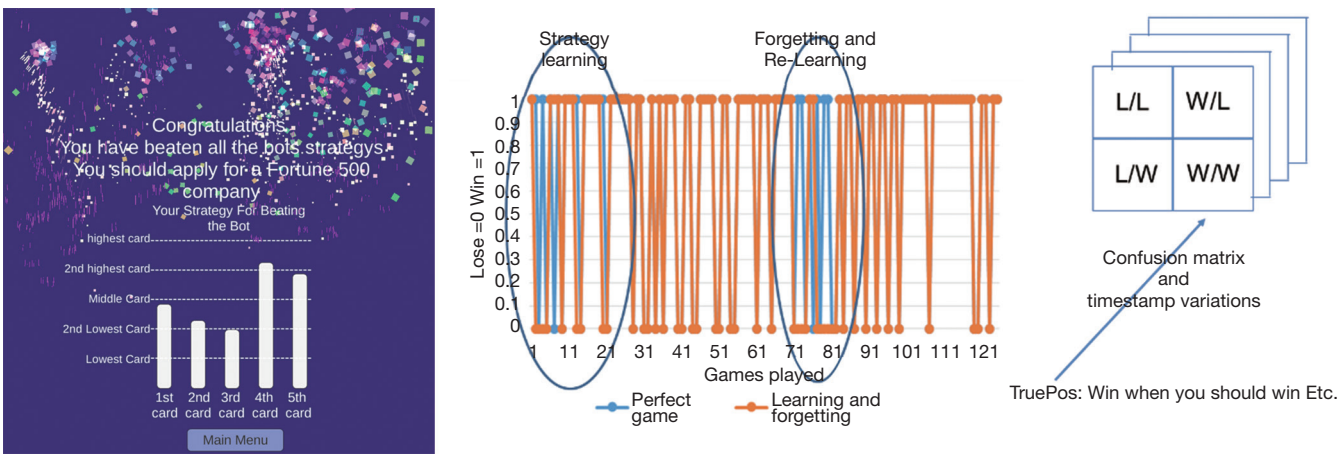

Figure 4 An example of completing a level and illustration of the type of available data.

courteous feedback like "nice try" or "please try again" would encourage further engagement. This is particularly so because by virtue of the fact the cards are dealt in a random or stochastic fashion, there is opportunity to lose even when one is playing correctly. This provides an opportunity for a player to forget their strategy in a moment of distraction.

A critical research element is to determine optimal analytical approaches to interpreting time-series confusion matrices, receiver operating characteristic (ROC) curves, and classification of play on a sufficiently large set of longitudinal player data through ML. Once sufficient data are collected, a neural network can be trained with the data to classify the degree to which a person experiences cognitive difficulties during play. *Figure 4* illustrates a fingerprint of a person's play against a bot, which will be used with ML algorithms. With this groundwork laid, the following ML classification approach has been identified for the initial ML classification investigation.

The inputs for ML techniques are formed by the "fingerprint" of one's cognitive processes as well as meta-data associated with timing variations of play. The ML methods for initial exploration are those associated with the

high-level API accessible through TensorFlow, as these have been successfully used to classify a wide variety of images as well as acoustic signals. *Figure 5* illustrates a more complete classification model.

Preliminary results using a relatively simple dense neural network will be the focus as it clearly illustrates the point without undue complexity. In partial summary, sufficient aspects of the research have been undertaken to lend credibility to the conjecture that serous mobile games combined with ML may be capable of assessing various cognitive states of health. The scale of participants that a mobile gaming framework facilitates (i.e., measured in tens-of-thousands of players, as opposed to more traditional cognitive health assessment methods which are administered to a single individual at a time) can generate baseline of normal age related cognitive decline, against which an assessment of a person suffering early signs of mild cognitive impairment can be made. An example of the game being played can be found at https://www.youtube.com/watch?v=CN-_zz-oIpU&t=30s.

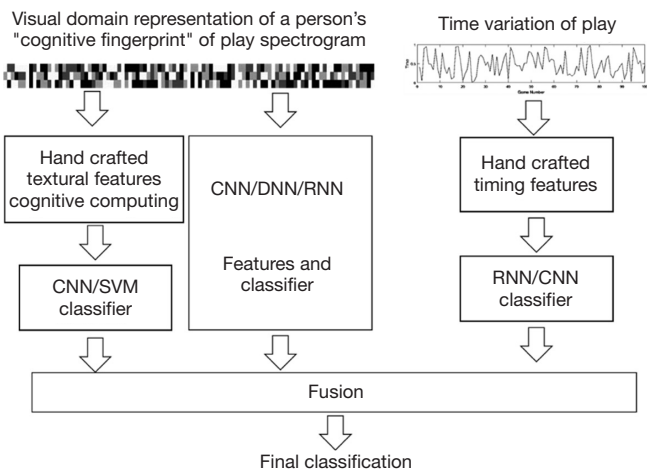

Figure 5 An overview of a proposed and more complete classification scheme.

Preliminary ML results

As with all ML and artificial neural network methods in general, a premium is placed on data and having lots of it. As an illustration of the type of data collected from WarCAT, two rounds of play are illustrated in *Figure 6*. For simplicity, only four entries extracted from each hand's confusion matrix are displayed. These values correspond to a win when a person should win, lose when person should lose, win when a person should lose and lose when a person should win. These are plotted here simply for visualization purposes. As input to a neural network, only a numeric representation of the confusion matrix entries is required.

Figure 6 is included to demonstrate the difference between two players. Further, *Figure 6* demonstrates that it would be extremely difficult or impossible to infer anything regarding normal age-related cognitive decline or MCI from only two people's trajectories of play. However, with massive amounts of data, there is the potential to detect outliers that may point to potential issues with memory or higher level executive function that would otherwise go undetected.

Our compelling argument for this comes from recent results using synthetic data of bots playing against bots, generated in large volumes. One bot will play the game using a predetermined strategy as they would against a person, while other bots emulate the play of a human player with varying degrees of impairment. Impaired bots play probabilistic versions of a perfect strategy. *Figure 7* illustrates confusion matrix entries per hand for a sequence

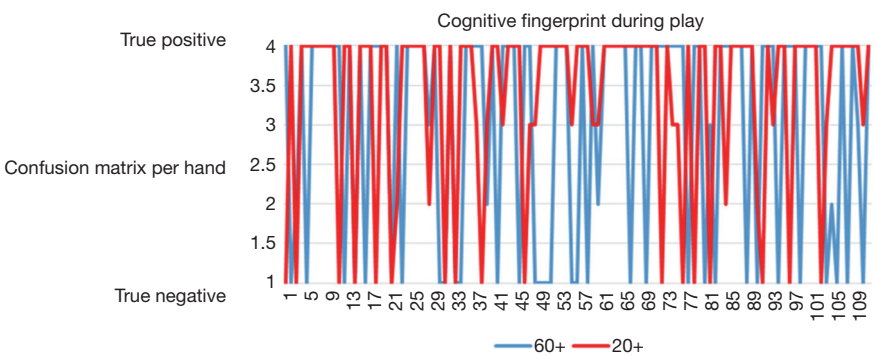

Figure 6 "Cognitive Fingerprints" of play: a 60+ and a 20+ year old player.

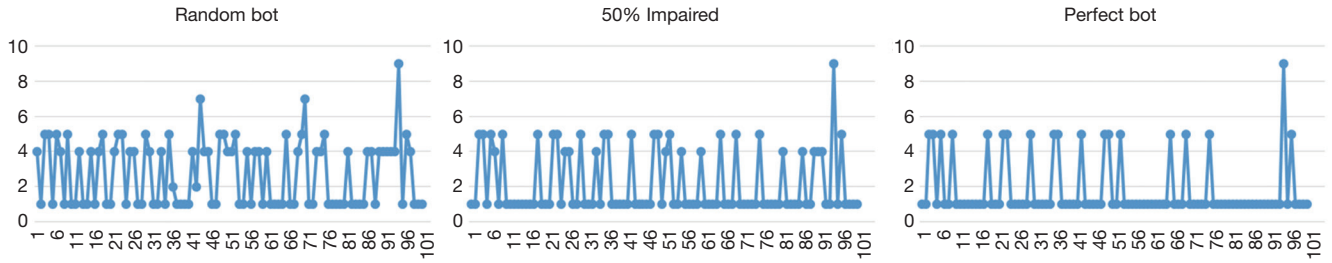

Figure 7 Three visualizations of a bot playing randomly, 50% perfect, and a perfect strategy.

Table 1 Confusion matrix entries inclusive of ties

Win/win
Lose/win
Tie/win
Win/lose
Lose/lose
Tie/lose
Win/tie
Lose/tie
Tie/tie

Table 2 Training and test accuracy after learning using the "cognitive fingerprint" data

Epoch	Training accuracy %	Testing accuracy %
500	85.6	82.8
1000	92.0	87.8
1500	94.6	93.1
2000	97.5	94.2
2500	98.4	96.2

of 100 hands of play for a bot that plays a totally random strategy, 50% impaired, and a perfect strategy.

For preliminary ML training and testing purposes, we used a densely connected neural network with 103 inputs which included the features generated from confusion matrix (e.g., *Figure* 7), as well as aggregate wins, losses and ties. These were normalized to all be in similar ranges. The confusion matrix entries are slightly more complicated to those discussed above to include the possibility of ties. As such, the confusion matrix for each hand is best illustrated by a 3×3 as shown in *Table 1*.

Within this configuration, 110,000 games were played by bots, with five classification labels. The labels were random play, 75% impairment, 50% impairment, 25% impairment, and perfect play. As such, the neural network had five output neurons, and internally two hidden layers of 150 neurons each. The training used stochastic gradient descent, trained on 100,000 training patterns, and tested using 10,000 patterns. As mentioned, the TensorFlow framework and high-level API were used.

Table 2 summarizes the training and testing accuracy of the neural network. These are very respectable values for accuracy and do not provide any evidence for overfitting or biases. It should also be noted that that all the classification errors were of the closest match. For example, if the classification was random play, the error if made was a prediction of 75% impaired. These are errors of the best kind in that they are close to the correct classification.

While the previous model was run using hand-by-hand data, a similar model was run just using aggregate win/loss/tie data (number of wins, losses and ties) as a sanity check. Aggregate data is typically descriptive and easily interpreted by people, however the subtleties of the actual data are lost. *Figure 8* illustrates histograms of play over 2,000 games of play (100 hands each) for each level of impairment. These are aggregate wins per game or cumulative scores which mask the subtleties of hand-by-hand play.

As demonstrated in *Figure 8*, there is considerable overlap among the classifications, and an expectation is that any ML technique would have considerable difficulty in discerning anything other than random play and perfect play. To illustrate this more convincingly, a similar DNN was trained and tested with the aggregates of the data used previously. The results of that test are convincingly illustrated in *Table 3*. It should be noted that there is also considerable dependence between wins and losses and as such they tend to be somewhat redundant features. *Figure 9* illustrates the clustering of data in feature space of Wins and Losses. It is apparent that there is considerable overlap

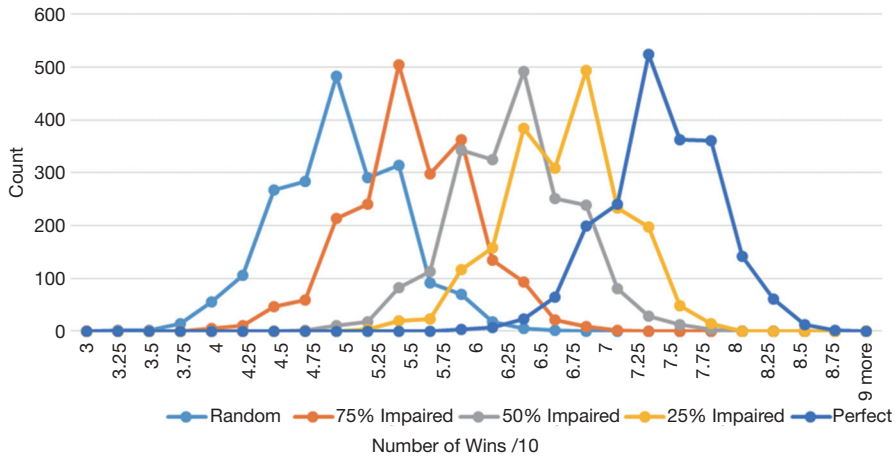

Figure 8 Histograms of wins for various degrees of bot impairment.

Table 3 Training and test accuracy after learning using aggregate win/loss/tie data

Epoch	Training accuracy %	Testing accuracy %
500	59.9	60.3
1000	59.9	60.4
1500	60.0	60.3
2000	59.9	60.4
2500	59.9	60.2

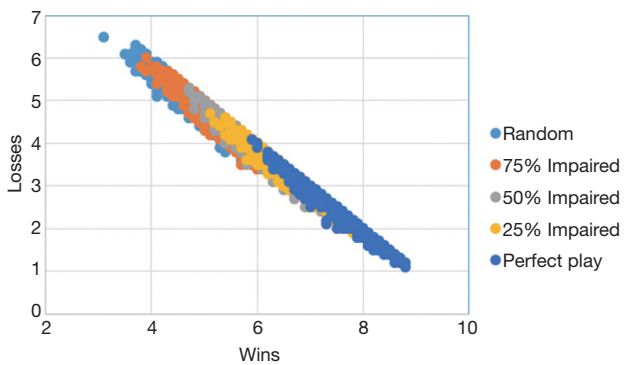

Figure 9 Clustering of labels in win-loss feature space.

which makes it difficult for people or machines to make an accurate classification.

Figure 10 illustrates the histograms associated with classification errors. When the neural network was trained and tested with confusion entries, the number of errors was 384/10,000, and the misclassification was overwhelmingly by an adjacent category. However, when using aggregate Win/Loss/Tie data, the total number of errors was 3971/10,000, with approximately 8% misclassifications greater than an adjacent category.

Discussion

In this perspective paper, we argue that serious games on mobile platforms combined with ML may be a viable technique to help detect pre-symptomatic MCI. To recap, our game WarCAT collects detailed player data and player moves during their game-play, which we have denoted a "cognitive fingerprint" associated with combined executive function, strategy detection, learning, retention, and recall processes that took place during play. The basic data is a 3×3 confusion matrix for each hand played. These data combined with event timing is amenable to ML analytics. To demonstrate this conjecture, we generated synthetic data from bots emulating human players with various degrees of impairment ranging from random play (100% impaired), through 75% impaired, 50% impaired, 25% impaired, through to perfect play.

These data were then presented to a relatively simple Dense Neural Network within TensorFlow. After training, the network was able to correctly classify 96.2% of the test data, with any error being those of the closest classification. To demonstrate that the hand-by-hand entries from the confusion matrix were necessary for classification, the player data were similarly subjected to classification using only aggregate descriptive statistics. Although the network trained reasonably well, the classification of the aggregate

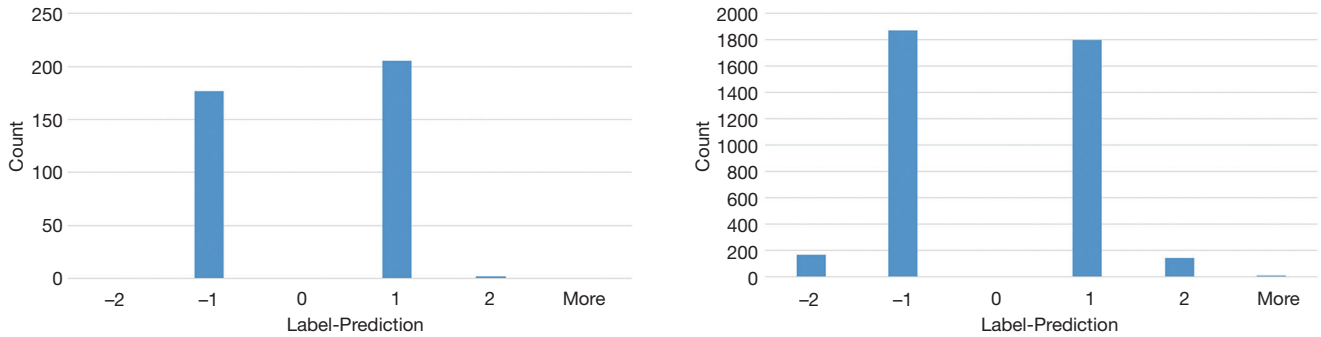

Figure 10 Classification error histograms (left using confusion matrix entries, right aggregate win/loss/tie).

test dataset was only 35% (a random guess would be 20%).

As these data were collected from bots, we did not have the opportunity to record hesitation or any timing-related data which would surely strengthen the inferencing. Further, the bots were stateless, whereas when one experiences a "loss of set," the effect would have some time correlation to subsequent play as one attempts to get back on their winning strategy.

A critical consideration in using ML methods is the volume of synthetic data needed. We generated 100,000 records for training and 10,000 for test purposes. These numbers are often typical for ANNs but more problematic if real data are to come from the general population. There is an opportunity for serious mobile games to overcome data collection difficulties due to their wide penetration into the user population, but they are more easily addressed by commercial gaming companies. Another possibility is to generate additional features for training that may reduce the volume of data required, such as derivatives of the "cognitive fingerprints".

Although this discussion centered on fairly conventional neural networks, we are also investigating the use of more powerful ML techniques such as convolutional neural networks and deep reinforcement learning. Reinforcement learning more closely resembles how people actually learn, and as such various degrees of impairment could more easily and accurately be modeled as compared to the simple approach demonstrated here.

Conclusions

The perspective has presented a simple mobile serious game with the capability to track a person's play, their strategy and ability to recall their strategy, over a brief period of time. The data collected were then demonstrated through ML to be of utility in providing a useful "cognitive fingerprint" of play for classification. Results included guidance for the volume of data required as well as the features deemed effective for classifying various degrees of bot or artificial impairment. The work illustrates several of the more important considerations when combining simple serious games and the data that they provide with ML. ML was demonstrated to be effective in accurately classifying various degrees of bot impairment when using data generated during the course of play as opposed to attempting to classify using aggregate scoring. Data collected during play is amenable to ML and necessary for accurate classification.

Acknowledgments

The authors would like to acknowledge the support of the University of Manitoba, and the Manitoba Liquor and Lotteries Small Grant Program that provided the initial funding to investigate the use of serious games for social gaming/gambling potential issues. We would also like to acknowledge the support of the Canadian Network for Public Health Intelligence (CNPHI) for supporting early forays into ML. We would like to thank Andrew Sadik for his suggestion of reducing the dataset through the use of hand tuned features such as derivatives.
Funding: None.

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/jmai.2018.07.02). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all

aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Petersen RC, Smith GE, Waring SC, et al. Mild cognitive impairment: clinical characterization and outcome. Arch Neurol 1999;56:303-8.
2. Mild cognitive impairment (MCI). Available online: http://www.mayoclinic.org/diseases-conditions/mild-cognitive-impairment/home/ovc-20206082
3. Jack Jr CR, Holtzman DM. Biomarker modeling of Alzheimer's disease. Neuron 2013;80:1347-58.
4. Spaan PE. Episodic and semantic memory impairments in (very) early Alzheimer's disease: The diagnostic accuracy of paired-associate learning formats. Cogent Psychol 2016;3:1125076.
5. World Health Organization (WHO) and Alzheimer's Disease International, Dementia: A Public Health Priority. Geneva: WHO, 2012.
6. Prince M, Bryce R, Ferri C. World Alzheimer Report 2011: The benefits of early diagnosis and intervention. Alzheimer's Disease International; 2011.
7. Luxton DD, McCann RA, Bush NE, et al. mHealth for mental health: Integrating smartphone technology in behavioral healthcare. Prof Psychol Res Pr 2011;42:505.
8. Tong T, Chignell M, Tierney MC, et al. A serious game for clinical assessment of cognitive status: validation study. JMIR Serious Games 2016;4:e7.
9. Hagler S, Jimison HB, Pavel M. Assessing executive function using a computer game: Computational modeling of cognitive processes. IEEE J Biomed Health Inform 2014;18:1442-52.
10. Jimison H, Pavel M. Embedded assessment algorithms within home-based cognitive computer game exercises for elders. Conf Proc IEEE Eng Med Biol Soc 2006;1:6101-4.
11. Basak C, Voss MW, Erickson KI, et al. Regional differences in brain volume predict the acquisition of skill in a complex real-time strategy videogame. Brain Cogn 2011;76:407-14.
12. Vallejo V, Wyss P, Rampa L, et al. Evaluation of a novel Serious Game based assessment tool for patients with Alzheimer's disease. PloS One 2017;12:e0175999.
13. Thompson O, Barrett S, Patterson C, et al. Examining the neurocognitive validity of commercially available, smartphone-based puzzle games. Psychology 2012;3:525.
14. Robert PH, König A, Amieva H, et al. Recommendations for the use of Serious Games in people with Alzheimer's Disease, related disorders and frailty. Front Aging Neurosci 2014;6:54.
15. Nyhus E, Barceló F. The Wisconsin Card Sorting Test and the cognitive assessment of prefrontal executive functions: a critical update. Brain Cogn 2009;71:437-51.
16. Leavitt F, Katz RS. Distraction as a key determinant of impaired memory in patients with fibromyalgia. J Rheumatol 2006;33:127-32.
17. Hay JF, Jacoby LL. Separating habit and recollection: memory slips, process dissociations, and probability matching. J Exp Psychol Learn Mem Cogn 1996;22:1323-35.
18. Miozzo M, Fischer-Baum S, Caccappolo-van Vliet E. Perseverations in Alzheimer's disease: memory slips? Cortex 2013;49:2028-39.

doi: 10.21037/jmai.2018.07.02
Cite this article as: Leduc-McNiven K, Dion RT, Mukhi SN, McLeod RD, Friesen MR. Machine learning and serious games: opportunities and requirements for detection of mild cognitive impairment. J Med Artif Intell 2018;1:7.