Discover Psychology

Discover

Review

Applications of machine learning to behavioral sciences: focus on categorical data

Pegah Dehghan[1] · **Hany Alashwal**[2] · **Ahmed A. Moustafa**[3,4]

Received: 22 October 2021 / Accepted: 29 January 2022
Published online: 15 March 2022

Abstract
In the last two decades, advancements in artificial intelligence and data science have attracted researchers' attention to machine learning. Growing interests in applying machine learning algorithms can be observed in different scientific areas, including behavioral sciences. However, most of the research conducted in this area applied machine learning algorithms to imagining and physiological data such as EEG and fMRI and there are relatively limited non-imaging and non-physiological behavioral studies which have used machine learning to analyze their data. Therefore, in this perspective article, we aim to (1) provide a general understanding of models built for inference, models built for prediction (i.e., machine learning), methods used in these models, and their strengths and limitations; (2) investigate the applications of machine learning to categorical data in behavioral sciences; and (3) highlight the usefulness of applying machine learning algorithms to non-imaging and non-physiological data (e.g., clinical and categorical) data and provide evidence to encourage researchers to conduct further machine learning studies in behavioral and clinical sciences.

Keywords Machine learning · Data analytics · Psychology · Statistics · Behavioral studies

1 Introduction

Through the lens of data analysis statistics, the main goal of behavioral sciences researchers was to create models for inferring human behavior for about a century. They applied specific methods such as null hypothesis to draw conclusions and find causalities and underlying mechanisms of behaviors. For about two decades, machine learning methods (which we refer to throughout the article as models of prediction) have gained interest and become widely used in research papers. In this approach, researchers attempt to build models for prediction, meaning their main goal is to design a model that can forecast unseen future behavior with the highest possible accuracy. Both models of inference and prediction provide some advantages and limitations, which we discuss below. We here argue that one of these models or combinations of them should be used in future work. In this study, we aim to review these two type of models (inference and prediction). The article is structured as follows:

✉ Hany Alashwal, halashwal@uaeu.ac.ae | [1]Department of Psychology, School of Literature and Humanities, Shahid Bahonar University of Kerman, Kerman, Iran. [2]College of Information Technology, United Arab Emirates University, Al-Ain, United Arab Emirates. [3]Department of Human Anatomy and Physiology, The Faculty of Health Sciences, University of Johannesburg, Johannesburg, South Africa. [4]School of Psychology, Faculty of Society and Design, Bond University, Queensland, Gold Coast, Australia.

1) In Sect. 2, we describe models investigating mechanistic processes (models built for inference), the way they treat data, methods used in the models, the underlying assumptions, and we explain how their theoretical limitations paved the way for predictive models' popularity.
2) In Sect. 3, we introduce models built for prediction, their similar language and methods with models created for inference, how they treat data, advantages and degrees of freedom they offer (via various examples), and we briefly introduce "deep learning" family as part of machine learning (and models of prediction). Then, we explain the pitfalls of machine learning and the efforts made to address them. Finally, we show how it is possible to combine both types of models in order to complete each other to reach the highest performance. Through these explanations, we also discuss a significant, often ignored, concern called Ground Truth.
3) In Sect. 4, we discuss the fact that although multi-modal data or cognitive measurements can also benefit from using models of prediction, these models are mainly applied to continuous data such as physiological signals and neuro-imaging data.
4) In the final section, we summarize the articles and encourage researchers working with non-imaging and non -physiological signals data to take machine learning into consideration.

2 Models built for inference in behavioral sciences

One of the main goal of behavioral scientists and psychologists is to infer and predict human behavior. Inferring behaviors means explaining the causality of behaviors precisely, explaining relationships among variables, and achieving better mechanistic insights regarding human behavior [1, 2]. To understand mechanistic processes researchers use specific statistical methods to treat input and output variables (also known as dependent and independent variables). These methods are formal mathematics such as confidence intervals and hypothesis testing, and rely on strong assumptions such as normality, linearity, non-collinear variables, variable independence [3, 4]. Scientists apply classical statistics to answer questions such as which input variables, compared to other input variables, exercise contributing effects on output variables in the gathered experimental sample? For this purpose, for instance, models of inference isolate the effects of a single variable to measure the behavioral changes and attribute them as the effects of that variable. This is a common agenda in many well controlled experimental designs [5] like studies comparing the effects of an intervention. Another example is provided by Koing et al. who aimed to understand multitasking performance; they created a regression model and considered fluid intelligence, working memory, attention, polychronicity, and extraversion to predict multitasking performance in 122 participants [6]. Koing et al. compared observed data (here multitasking performance) and expected data generated from their created model and tested their hypotheses such as whether all the variables have important contributions in predicting multitasking performance. Based on how generated data fit in specific statistical data distribution, their hypotheses are either confirmed or rejected, which results in discovering new relationships that presumably are not due to noise. These results are considered as true effects [7]. If available data are sufficient, assumptions such as equal variance become explicitly verified and in case of necessity, the model will be refined [8] as they found polychoronicity and extroversion do not have significant effects on predicting multitasking performance, and therefore, can be eliminated from the model. The advantage of this approach is that it is simple to understand. Scientists use few variables that clinically or theoretically are considered crucial with mainly limited linear interactions among them. Therefore, interpreting results is not complicated. This way of treating variables imposes serious limitations on the models built for inference, paving the way for models built for prediction to address these limitations.

Recently many criticisms have been made against the methods used in models investigating mechanistic processes. These critics argue that having an emphasis on explaining the roots and causal underpinnings of behaviors have ended in mechanistic models that, despite their appealing theoretical underpinnings, have not demonstrated potential in forecasting future behaviors. These methods are mainly concerned with whether the "goodness of fit" between the statistical model and sample data is adequate or whether implications of different theories confirm the magnitudes and directions of regression coefficients. These may restrict models to make better predictions [2].

The generalizability of models designed for inference is also questionable. In social and medical research fields, the inference is mainly drawn based on statistical tests on aggregated data. The underlying assumption is that what is estimated from groups can be generalized to individuals and lead to understanding subjective phenomena such as behaviors. Generalizing group-based level findings to single-subject level is possible as long as the processes in question are ergodic [9]. Ergodicity means that the effects under investigation are homogenous across individuals and stable in the course of time [10]. Since psychological variables and constructs are organized within people over time, if there are

individual exceptions, the generalizations based on groups are not ergodic. This is why the classical statistics applied to groups rather than individuals raises serious questions regarding the extent to which derived results can be generalized [11].

In addition, models of inference were designed for a number of sample sizes and variables that are now considered small to moderate. Furthermore, since most statistical models are based on many assumptions such as limited interactions among variables [12], considering more input variables and associations contributes to weaker explanatory [8].

However, Yarakoni and Westfall argued that one reason for choosing the inferential methods in psychology relied on poor understanding of tools to generate a successful prediction and slow pace of deploying the tools after being developed [2]. Today, data science and artificial intelligence allowed scientists to address classical statistics limitations regarding a large number of assumptions, generalizability, complexity, and a small number of input variables, and poor prediction power.

3 Models designed for prediction in behavioral sciences

Recent advancement in artificial intelligence and data analysis has shed light on models for prediction (i.e., machine learning) as an approach creating a context in which researchers can address previous statistic and methodological limitations, which contributed to considerable improvement in scientific modeling predictions [13–16]. Models of prediction create systems involving advanced statistical and probabilistic techniques that learn from data and detect latent patterns to predict unobserved or out-of-sample data accurately. The questions that machine learning tries to answer are more heuristic, like what variables are helpful to distinguish people with specific traits or characteristics from others [5]. Models built for prediction and models built for inference are both based on relationships in data and it is often the case that both use similar terms with different meaning and purposes. For instance, researchers applying regression algorithms to understand mechanistic processes focus on how well it accounts for the original dataset while in predictive models, the focus is on how accurately these algorithms can predict new data [17].

Models of prediction (i.e., machine learning) offer advantages over models built for inference, contributing to its high growth in science. First, inferential statistics' reliance on strong assumptions such as error distribution, additivity of parameters with linear predictors are not satisfied most of the time in clinical practice and become ignored in the scientific literature. This issue is solved since models of prediction methods rely on minimal a priori assumptions. Second, in contrast to traditional statistics that researchers handpick few variables based on their knowledge to avoid collinearity, machine learning is able to consider all available data on a specific field [3], especially in situations where there are few observations and many predictors. For example, investigating rare mental or physical conditions (e.g., suicide or blind sightedness) requires recording all possible information such as brain structure and function, genetics information, and historical and demographical information about subjects to draw a valid conclusion. Methods used in inferential models like regression models have serious limitations when variables outnumbering subjects. Machine learning and, in general, models of prediction made it possible to apply a number of approaches on small datasets. Thus, machine learning algorithms provide degrees of freedom to make good predictions even when input variables exceed subjects [4].

Furthermore, models of predictions and machine learning can analyze complex systems. Complex systems involve interactions among many factors inside and outside of systems. Mental disorders also have such complexity, and the subsystems such as biology, emotion, cognition, behavior, and environment interact with each other within an individual [10, 18]. Models built for inference analyzing few variables and few interactions at a group level cannot estimate such complex situations accurately [10]. Machine learning and models built for prediction, instead, allows researchers to analyze complex multivariate relationships related to high-dimensional data with known interdependencies [19].

Psychologists and computer scientists have also taken advantage of the models of prediction capability of applying models to individuals rather than groups. For example, Spape et al. [20] developed a model to predict what kinds of faces each individual thinks as attractive. These researchers showed pictures of different faces to participants and asked them to rate their attractiveness. After that, through machine learning and artificial intelligence, Spape et al. detected the patterns based on which each participant called pictures attractive or ugly. Then, they used the patterns and produced novel faces (attractive/ugly) and asked participants again to rate them. The results showed that the accuracy of predicting individuals' answers was high [20].

Machine learning techniques can be divided into supervised learning in which data are categorized and labeled (for example, healthy and diagnostic group). Unsupervised learning can cluster data based on their similarities without prior knowledge or labels. And semisupervised learning which is the intercept of the two previous techniques and includes

labeled and unlabeled data. Psychology can apply a wide range of these techniques based on the problems they are trying to address. For instance, a form of supervised learning is classification involves automatic detection of regularities in data according which the data would be classified into different predefined categories. Psychology researchers apply these algorithms to classify healthy individuals from those suffering from mental disorders or any other types of diseases or abnormalities, called diagnosing. For example, Magnin et al. used the Support Vector Machine algorithm to classify 16 patients with Alzheimer's disease from 22 healthy elderlies based on their MRI data and reached the prediction accuracy of 94% [21]. Today, models of prediction such as machine learning algorithms are developed to predict and diagnose diseases such as Alzheimer's disease, depression, anorexia, anxiety disorders, specific phobia, and substance abuse with the highest accuracy [4, 22–26].

Furthermore, changes in brain structure and function, cognition, and physiology due to many disorders and diseases occur continuously. For instance, Alzheimer's disease gradually progresses in the course of years, and sometimes the disease onset starts years before the last stage. So, in order to manage patients more effectively, it is necessary to estimate the progression of diseases through accurate prediction generated from baseline clinical scores [22, 27]. Machine learning-based methods such as pattern regression have attracted attention for many purposes including prognosis brain disease and mental disorders. Pattern regression recognition involves estimating continuous variables such as cognitive scores rather than categorical. For example, Stonnington et al. [28] studied Alzheimer's disease (AD). They applied relevance vector regression algorithm to participants MRI data to predict several continuous scores evaluating dementia severity, such as Mini-Mental State Examination (MMSE), Alzheimer's disease assessment scale (ADAS), Auditory Verbal Learning Test (AVLT) and dementia rating scale (DRS) in AD, mild cognitive impairment (MCI) and healthy control groups [28]. There are numerous neuroimaging studies that investigated models with high accuracy to predict the transition from mild cognitive impairment to Alzheimer's disease [22, 27]. Neuroimaging and clinical studies also have generated accurate predictions in psychotic transition [29, 30]. Models of prediction are also a helpful tool to aid health care systems and caregivers to make treatment decisions. For example, Whitifield et al. classified patients with social anxiety disorder based on their MRI into those who can or cannot benefit from cognitive behavioral therapy [31]. The significance of these types of studies is that considerable waste of expenditure, time, and energy resulting from testing therapies by trial and error without achieving satisfactory outcomes can be effectively avoided.

As for advantages and advancements in models built for prediction, deep learning models showing promising results in creating models of complicated behaviors. Deep learning is part of broader family of machine learning based on artificial neural network which uses layers to extract progressively higher level features from raw input. Deep neural networks algorithms contain simple units being organized in layers and then stacked to create deep networks. The data are trained on connections and relationships among the units and learn information extraction to solve tasks. A combination of large annotated datasets and complicated network architectures and advances in computer hardware paved the way for addressing many problems in computer vision and behavioral research such as pose estimation [32, 33], which refers to measuring body parts' geometrical configuration. Scientists use videography to approximate poses in the course of time and transform them into dynamic, kinematics and action [34, 35]. The advancement of deep learning contributes to several studies conducted on relatively small datasets and achieve excellent results for different poses such as locomotion reaching, egg lying in flies, trail tracking in mice, and hunting in cheetahs [36–38]. In terms of speed, accuracy and robustness of pose estimation, deep learning was found to be fast accurate and generalizable even when it comes to complicated experiments such as measuring interactions of multiple animals with objects, or social behaviors in bats or bee stingers. In terms of standard game theories, deep learning was also shown to be powerful. Game theory is a framework in which analyzing multi-agent systems and their strategic interactions is under study. For example, security systems investigating how to allocate security guards strike a balance between maximizing the use of scarce resources and rational adversaries. In the most existing studies in this area models are based on the assumption that participant are 100% rational or they try to create models based on cognitive psychology and experimental economics; but deep learning by performing automatically cognitive modeling allows researchers to be self-sufficient and independent of such expert knowledge [39].

Having said that, models for prediction do, however, have a number of limitations. Firstly, since models of prediction use complex nonlinear relationships between variables to find patterns in data, these complicated relationships and algorithms have made it challenging to interpret machine learning results [40]. Interpretability concerns about the extent to which models allow for human understanding and it is a vital factor to understand causes of relationships. Generally, interpretable models include few understandable components, whereas non-interpretable models encompass a large number of complex components [41]. Therefore, models of prediction emphasis on data science rather than mechanistic insight have made it hard to interpret and, in order to avoid misinformed conclusions, expertise in both psychology and

data science is crucial. Secondly, although the goal of models of prediction is to make an accurate prediction, there are studies showing that models built for inference can have similar performance, if not better, to machine learning (models of prediction) in this regard [22, 42]. Thirdly, even when the performance is high, the results can be questionable. One of the trivial examples is studies estimating criminality based on individuals' facial photos. Although the classifiers have succeeded in accurately categorizing images of criminals from non-criminals, it is not clear based on what differences in the categories this discrimination is made. For instance, if in all images, the criminals wear black hats and non-criminal put on no hat, it is possible to reach 100% accuracy in classifying them, but it is not related neither to crime nor to facial structure. Criminal mugshots are generally taken in different conditions from normal non-criminal photos in terms of camera, resolution, illumination, background, angle, and distance. In addition, time in prison can impact aging, facial expression, and a higher risk of facial damage like a broken nose. All of the differences related to situations not facial structure, can lead to higher accuracy despite the fact that they are not related to the relationship between individuals' face and crime. Furthermore, the available datasets come from mugshots of those arrested, while a high percentage of criminals are never caught. Therefore, it is essential not to overlook models of prediction limitations manifested in both poor and high performance [43–45].

Recently, researchers have made efforts to address the lack of interpretability, which has contributed to emerging fields such as Explainable Artificial Intelligence (XAI). The goal of XAI is to provide methods increasing interpretability, transparency, and fairness. Proxy models, introspective models, correlative techniques and saliency, post hoc explanations, and example-based explanations are among these methods which discussing about all of them is beyond the scope of this article. For example, in post hoc methods categories, the Local-Interpretable Model-Agnostic Explanation (LIME) is popular and widely used. LIME is a local surrogate model (usually linear or tree) fitted in adjacent of the instance to explain, and in order to do so, LIME perturbs the instance to explain and generate the required training dataset for fitting the local explainer. The interpretable model is trained by minimizing the loss functions like weighted Root Mean Square Error (RMSE). The weight is based on the distance between the instance to explain and the perturbed data meaning that closer data gain higher weights. In order to have an interpretable model, the loss function should be minimized to the extent to which complexity measures and fidelity allow. Fidelity is defined as the extent to which an interpretable model can approximate the original one, and interpretability refers to complexity of the interpretable model. For example, the number of features that should be applied in the interpretable model is an influential factor. In LIME, the weight and its direction of each feature illustrate the impact of that feature in the explainer. Therefore, features with positive weights push the prediction closer to the selected label, and negative weights push it away. In this way, researchers can also determine and rank features contributions to succeeding the prediction. Therefore, machine learning interpretability can be seen not only as an issue but also as a tool to extract information from predictive machine learning models [46, 47]. For example Posada-Quintero et al. used machine learning interpretability models to interpret SVM and decision tree models to find the differences on risk factors and symptoms of burnout syndrome in two categories of teachers. And they found the most relevant symptoms of burnout were fatigues and headache where the most relevant risk factor was how satisfied they were with their incomes [48]. In addition to previous attempts to address the interpretability of predictive models, a new generation of a discipline called artificial cognition provides insights into better interpretability. In artificial cognition, researchers aim to understand machine behavior similar to cognitive psychology in which scholars try to understand human behavior. Accordingly, they use similar pipelines such as experimentation. They identify behavior and its environmental correlates, infer the causes, and determine its boundary conditions. The experimentations are rooted in Popperian falsificationism tradition in which each theory gains its confirmation by defeating other alternatives. Therefore, it has high explanatory power. For example, Geirhos et al. [49] aimed to understand why a machine learning algorithm can recognize objects easily when the shape is distorted by using its texture. They applied the input's texture to another input's shape (e.g., having a cat shape picture filled by elephant skin and testing their machine learning algorithm to see whether it identifies the picture belongs to cat categories or elephants one [49]. This experimental design was conducted before to understand human cognition too, and it was found that human mind preference in object feature recognition is different with machines. In general, this was an example to show how machine learning behavior can be explained [47]. In general, models for inference and models for prediction both have strengths and limitations, and researchers should choose suitable methods based on their goals and possibilities. For example, they can use models of inference rather than models of prediction when their knowledge regarding their area of interests is substantial or when their priority is to provide mechanistic insights. Furthermore, models of inference can be a good choice when they intend to study limited variables, or the number of observations is much more than variables [3]. On the other hand, ml algorithms can be more helpful than classical statistics when the priority is to accurately make a predictive model of the behaviors. Furthermore, machine learning and models of prediction can be used when there are a large number of different types of variables

involving numerous interactions such as instruments and batteries measuring significant behavioral aspects, reaction time, demographic data, EEG, MRI, genetics, and "omics" data assessing human traits frequency and distributions.

More importantly, scientists can use both approaches to have better results. For example, during feature selection in machine learning, it is vital to select those variables which exert the highest impact on outcomes. Including unrelated features in datasets might decrease accuracy because the tested model does not consider the distinguishing variables. Classical methods such as ANOVA and multidimensional scaling when the target variable is categorical (like classification) help the model consider important features and reduce the time of training datasets to detect the pattern and increase accuracy [3]. Another important example can be evaluating ground truth. Ground truth (GT) is the data reference based on which discriminative models and algorithms are trained. In many machine learning studies, GT was considered to be perfectly accurate (100%); but recently, studies highlighted the point that it is not possible to have 100% accuracy, particularly in medicine or psychology. Since human experts and specialists determine GT labels based on their interpretation and abilities, it is prone to error and disagreement, and they may interpret the same phenomenon differently and consequently label it differently. In order to address these disagreements and increase reliability, inferential methods such as Pearson r and spearman Rho are applied. In addition, when ground truth data is skewed inferential methods such as transformation are used. Therefore these two approaches (machine learning and inferential methods) can promote each other [50–52].

In addition, one can apply machine learning without predicting behavior; thus the distinction between making predictions and inferring behavior should not imply that tools used in one of the two models cannot help the other one. Recently social science researchers used machine learning to discover new concepts, quantify the extent to which these concepts are prevailed and assess the causal effects. For example, there are a large number of studies using machine learning to estimate how the effect of a specific intervention can differ across individuals' characteristics. The derived information can be used to find more effective target treatments. It can also be used as indirect evidence of the underlying mechanism based on which the treatment comes into operation. Machine learning can solve these problems via estimating average effects within strata defined by the covariates [53].

Broadly speaking, models of prediction is a relatively new method that provides significant advantages for behavioral sciences. Based on their goals and possibilities, researchers can decide whether models of inference, models of prediction, a combination of them, or other approaches would be suitable (Table 1). Therefore knowing about the method and its applications in psychology and behavioral sciences can pave the way for future creative studies.

4 Machine learning applications to behavioral data

Although the popularity of models of prediction (e.g., machine learning) is increasing in almost all areas of science, the applications of this approach to behavioral sciences have not been adequately addressed in existing studies. Behavioral sciences are broad fields with various types of assessments, but neuroimaging and physiological signals are the dominant modalities used in machine learning algorithms [22, 54–56]. It may be because models of prediction happen to become important when scientists start to deal with massive unstructured noisy data. Dimension reduction as a method to address this type of data contribute to so called "factors" for which interpretation is neither

Table 1 General comparison between models built for prediction and models built for inference

	Models of prediction (e.g. machine learning)	Models of inference
Goal	Prediction	Inference
Complexity	Considering complex non-linear interactions among variables Large number of input variables	Designed for few interactions among variables Few input variables
Generalizability	Individual participant	At a group level
Data	Up to large scale dataset	Small to moderate sample size
Assumptions	Minimal assumptions	Larger number of assumptions
Mechanism	Detecting patterns in data	Comparing groups with each other and with random sample
Reliance	Relying on data	Relying on theories regarding the essence of the behavior of interest
Limitations	Problems with interpretability	Low degree of freedom to consider complex relationships Necessitating satisfaction of many assumptions

wanted nor necessary. Since the reduced dimensions of such complex datasets do not always require a specific interpretation, it is likely researchers using neuroimage techniques were searching for methods to analyze such complex datasets earlier than other field of psychology and behavioral sciences, which has paved the way for black box models to show exceptional utility [57, 58].

Nonetheless, identifying more comprehensive characterization of behaviors would lead to more accurate predictions. Clinical and cognitive measures, for instance, are inexpensive assessments that play indispensable roles in many scientific purposes such as diagnosing and evaluating mental disorders. Recently, it was shown that using cognitive and clinical data significantly increases the accuracy of machine learning predictions compared to not using these types of data [59]. Following this, some studies incorporated different imaging modalities with clinical, cognitive, demographic, and genomic data [60–63]. For example, Whelan et al. [64] collected a broad domain of data including neural, personality, cognitive, genetic, and demographic data from participants with substance misuse at ages 14 and 16. MRI data were assessed during tasks such as reward processing, motor inhibition control, and emotional reactivity. They developed a model that can predict alcohol misuse based on brain structure and function, individual personality and cognitive differences, environmental factors, life experiences, and candidate genes [64]. In terms of physiological signals there are a number of studies incorporating different modalities such as Heart Rate Variability and Electrodermal Activity to find biomarkers of Autism, identify cognitive tasks, detect risk for emotional eating episode, to name but a few [65–67].

Furthermore, Personality computing (PC) is the intersection of personality and computer science which attempt to extract personality theory driven data such as Big Five trait or HEXACO from machine-sensed information like speech pattern, written text, digital footprint and non-verbal behaviors by machine learning [68]. In order to improve interpretability, Researchers applying machine learning in personality field report terms like variables importance and predictor effects in addition to measures of predictive performance and extend the existing theoretical constructs such as behavioral manifestations of personality traits [69].

Apart from studies incorporating multimodal data, behavioral researchers who work only on non-imaging and non- physiological data can also benefit from machine learning and models of prediction. Koutsouleris et al. [70], for example, built a 189 item questionnaire, based on data from 334 individuals having experienced the first episode of psychosis recorded in 44 mental health centers and managed to predict their subsequent functional outcomes. They used classification (support vector machine) to label outcomes based on good outcomes (Global Assessment of Functioning [GAF] score ≥ 65) vs bad outcomes (GAF < 65). The results indicated questionnaire can predict the outcomes with accuracies above 70% in both a one-month and a one-year periods. This study is an illustrative example of machine learning strength in gathering features parsimoniously to build a new questionnaire [45]. Kessler et al. [71] also gathered baseline information by interviewing 1056 subjects with major depression, re-interviewed them 10–12 years later, and compared the results of the classical method to the results obtained from machine learning algorithms. They asked about the number of years since age-of-onset with episodes lasting more than 2 weeks and lasting most days throughout the year. The researchers also investigated whether respondents were ever hospitalized for depression after their first episode, and if they were disabled at the time of interview because of their depression. In general, they intended to predict persistence severity, chronicity, hospitalization, disability, and suicide attempts in people with major depression. They used regression ensemble trees algorithm and the features included temporally primary comorbid lifetime disorders, parental depression, major depressive disease incident episode symptoms, and other information about the incident episode such as age-of-onset and whether the episode was triggered or endogenous. Eventually, the accuracy in predicting high persistence, hospitalization, disability, suicide attempt were above 70% (that for high chronicity was 63%), which was more precise than predictions made by the classical statistics method (logistic regression) [46].

In general, although the number of non-imaging and non-physiological studies of behavioral sciences that have used models of prediction are significantly smaller than the imaging studies, the results are promising and in some cases non-imagining data generate more accurate outputs than imaging data. For example, Samper-Gonzalez et al. [72] compared the performance of MRI and fluorodeoxyglucose positron emission tomography to cognitive and clinical scores at predicting Alzheimer's disease in patients with mild cognitive impairment (MCI), and they found that the latter assessments made a better prediction than the former [47]. Therefore, a better understanding of what models of prediction have brought about to non-imaging behavioral sciences may pave the way for researchers to conduct more creative experiments using machine learning and generate more accurate models with non-imaging data. In sum, the future work should investigate the applications of machine learning in psychology and behavioral sciences working with non-imaging data.

5 Conclusions

Machine learning is a novel approach that brings considerable benefits to behavioral sciences, such as relatively accurate diagnosis and prognosis of mental disorders and making better treatment decisions. This approach can be applied to either neuroimaging data, non-imaging data, or combinations of different data types. Therefore, behavioral researchers depending on their research interests and the type of data they work with can investigate the applications of machine learning algorithms. In particular, further investigations is suggested in non-imaging assessments to extend the current findings of machine learning applications in this less studied area.

Authors' contributions All authors whose names appear on the submission contributed to the study conception and design. PD wrote the main manuscript text. AAM and HA reviewed the manuscript. All authors read and approved the final manuscript.

Funding This work received financial support from the United Arab Emirates University (Grant No. CIT31T129).

Data availability We do not analyse or generate any datasets, because our work proceeds within a theoretical approach.

Code availability Not applicable.

Declarations

Competing interests The authors declare no competing interests.

References

1. Azzolina D, et al. Machine learning in clinical and epidemiological research: isn't it time for biostatisticians to work on it? 2019. https://doi.org/10.2427/13245
2. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: lessons from machine learning. Perspect Psychol Sci. 2017;12:1100–22. https://doi.org/10.1177/1745691617693393.
3. Rajula HSR, et al. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina. 2020;56(9):455. https://doi.org/10.3390/medicina56090455.
4. Orrù G, et al. Machine learning in psychometrics and psychological research. Front Psychol. 2020;10:2970. https://doi.org/10.3389/fpsyg.2019.02970.
5. Bzdok D, Ioannidis JP. Exploration, inference, and prediction in neuroscience and biomedicine. Trends Neurosci. 2019;42(4):251–62. https://doi.org/10.1016/j.tins.2019.02.001.
6. Konig CJ, Buhner M, Murling G. Working memory, fluid intelligence, and attention are predictors of multitasking performance, but polychronicity and extraversion are not. Hum Perform. 2005;18(3):243–66. https://doi.org/10.1207/s15327043hup1803_3.
7. Ioannidis JP, Tarone R, McLaughlin JK. The false-positive to false-negative ratio in epidemiologic studies. Epidemiology. 2011. https://doi.org/10.1097/EDE.0b013e31821b506e.
8. Ij H. Statistics versus machine learning. Nat Methods. 2018;15(4):233. https://doi.org/10.1038/nmeth.4642.
9. Molenaar PC. A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. Measurement. 2004;2(4):201–18. https://doi.org/10.1207/s15366359mea0204_1.
10. Molenaar PC, Campbell CG. The new person-specific paradigm in psychology. Curr Dir Psychol Sci. 2009;18(2):112–7. https://doi.org/10.1111/j.1467-8721.2009.01619.x.
11. Fisher AJ, Medaglia JD, Jeronimus BF. Lack of group-to-individual generalizability is a threat to human subjects research. Proc Natl Acad Sci. 2018;115(27):E6106–15. https://doi.org/10.1073/pnas.1711978115.
12. Ryo M, Rillig MC. Statistically reinforced machine learning for nonlinear patterns and variable interactions. Ecosphere. 2017;8(11):e01976. https://doi.org/10.1002/ecs2.1976.
13. Alashwal H, et al. Latent class and transition analysis of Alzheimer's disease data. Front Comput Sci. 2020;2:1–13. https://doi.org/10.3389/fcomp.2020.551481.
14. Alashwal H, et al. The application of unsupervised clustering methods to Alzheimer's disease. Front Comput Neurosci. 2019;13:31. https://doi.org/10.3389/fncom.2019.00031.

15. Moustafa AA, et al. Applying big data methods to understanding human behavior and health. Front Comput Neurosci. 2018;12:84. https://doi.org/10.3389/fncom.2018.00084.
16. Moustafa AA, et al. A longitudinal study using latent curve models of groups with mild cognitive impairment and Alzheimer's disease. J Neurosci Methods. 2021;350: 109040. https://doi.org/10.1016/j.jneumeth.2020.109040.
17. Coutanche MN, Hallion LS. Machine learning for clinical psychology and clinical neuroscience. Sciences. 2019;22(3):258–69.
18. Fornito A, Zalesky A, Breakspear M. The connectomics of brain disorders. Nat Rev Neurosci. 2015;16(3):159–72. https://doi.org/10.1038/nrn3901.
19. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol. 2018;14:91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037.
20. Spape M, et al. Brain-computer interface for generating personally attractive images. IEEE Trans Affect Comput. 2021. https://doi.org/10.1109/TAFFC.2021.3059043.
21. Magnin B, et al. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology. 2009;51(2):73–83. https://doi.org/10.1007/s00234-008-0463-x.
22. Arbabshirani MR, et al. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage. 2017;145:137–65. https://doi.org/10.1016/j.neuroimage.2016.02.079.
23. Fu CH, et al. Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. Biol Psychiat. 2008;63(7):656–62. https://doi.org/10.1016/j.biopsych.2007.08.020.
24. Klöppel S, et al. Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method. Brain. 2008;131(11):2969–74. https://doi.org/10.1093/brain/awn239.
25. Lavagnino L, et al. Identifying neuroanatomical signatures of anorexia nervosa: a multivariate machine learning approach. Psychol Med. 2015;45(13):2805–12. https://doi.org/10.1017/s0033291715000768.
26. Visser RM, et al. First steps in using multi-voxel pattern analysis to disentangle neural processes underlying generalization of spider fear. Front Hum Neurosci. 2016;10:222. https://doi.org/10.3389/fnhum.2016.00222.
27. Wang Y, et al. High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables. Neuroimage. 2010;50(4):1519–35. https://doi.org/10.1016/j.neuroimage.2009.12.092.
28. Stonnington CM, et al. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. Neuroimage. 2010;51(4):1405–13. https://doi.org/10.1016/j.neuroimage.2010.03.051.
29. Koutsouleris N, et al. Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. Arch General Psychiatry. 2009;66(7):700–12. https://doi.org/10.1001/archgenpsychiatry.2009.62.
30. Koutsouleris N, et al. Disease prediction in the at-risk mental state for psychosis using neuroanatomical biomarkers: results from the FePsy study. Schizophr Bull. 2012;38(6):1234–46. https://doi.org/10.1093/schbul/sbr145.
31. Whitfield-Gabrieli S, et al. Brain connectomics predict response to treatment in social anxiety disorder. Mol Psychiatry. 2016;21(5):680–5. https://doi.org/10.1038/mp.2015.109.
32. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.
33. Vickers NJ. Animal communication: when i'm calling you, will you answer too*?*. Curr Biol. 2017;27(14):R713–5. https://doi.org/10.1016/j.neunet.2014.09.003.
34. Schaefer AT, Claridge-Chang A. The surveillance state of behavioral automation. Curr Opin Neurobiol. 2012;22(1):170–6. https://doi.org/10.1016/j.conb.2011.11.004.
35. Dell AI, Bender JA, Branson K, Couzin ID, de Polavieja GG, Noldus LP, et al. Automated image-based tracking and its application in ecology. Trends Ecol Evol. 2014;29(7):417–28. https://doi.org/10.1016/j.tree.2014.05.004.
36. Mathis A, Biasi T, Schneider S, Yuksekgonul M, Rogers B, Bethge M, et al., editors. Pretraining boosts out-of-domain robustness for pose estimation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2021. arXiv:1909.11229
37. Mathis A, Warren R. On the inference speed and video-compression robustness of DeepLabCut. BioRxiv. 2018. https://doi.org/10.1101/457242.
38. Nath T, Mathis A, Chen AC, Patel A, Bethge M, Mathis MW. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. Nat Protoc. 2019;14(7):2152–76. https://doi.org/10.1038/s41596-019-0176-0.
39. Mathis MW, Mathis A. Deep learning tools for the measurement of animal behavior in neuroscience. Curr Opin Neurobiol. 2020;60:1–11. https://doi.org/10.1016/j.conb.2019.10.008.
40. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT press; 2016. https://doi.org/10.1007/s10710-017-9314-z
41. Johansson U, Sönströd C, Norinder U, Boström H. Trade-off between accuracy and interpretability for predictive in silico modeling. Future Med Chem. 2011;3(6):647–63. https://doi.org/10.4155/fmc.11.23.
42. Steyerberg EW, Uno H, Ioannidis JP, Van Calster B, Ukaegbu C, Dhingra T, et al. Poor performance of clinical prediction models: the harm of commonly applied methods. J Clin Epidemiol. 2018;98:133–43. https://doi.org/10.1016/j.jclinepi.2017.11.013.
43. Bowyer KW, King MC, Scheirer WJ, Vangara K. The, "criminality from face" illusion. IEEE Trans Technol Soc. 2020;1(4):175–83.
44. Pasquale F. When machine learning is facially invalid. Commun ACM. 2018;61(9):25–7. https://doi.org/10.1145/3241367.
45. Wu X, Zhang X. Responses to critiques on machine learning of criminality perceptions (Addendum of arXiv: 1611.04135). arXiv preprint, arXiv:1611.04135. 2016.
46. Schmidt P, Biessmann F. Quantifying interpretability and trust in machine learning systems. arXiv preprint arXiv:190108558. 2019.
47. Taylor JET, Taylor GW. Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. Psychonomic Bull Rev. 2020. https://doi.org/10.3758/s13423-020-01825-5.
48. Posada-Quintero HF, Molano-Vergara PN, Parra-Hernández RM, Posada-Quintero JI. Analysis of risk factors and symptoms of burnout syndrome in Colombian School teachers under statutes 2277 and 1278 using machine learning interpretation. Social Sciences. 2020;9(3):30. https://doi.org/10.3390/socsci9030030.

49. Geirhos R, Temme CRM, Rauber J, Schütt HH, Bethge M, Wichmann FA. Generalisation in humans and deep neural networks. arXiv preprint, arXiv:1808.08750. 2018.
50. Cabitza F, Campagner A, Albano D, Aliprandi A, Bruno A, Chianca V, et al. The elephant in the machine: proposing a new metric of data reliability and its application to a medical case to assess classification reliability. Appl Sci. 2020;10(11):4014. https://doi.org/10.3390/app10114014.
51. Basile V, Cabitza F, Campagner A, Fell M. Toward a perspectivist turn in ground truthing for predictive computing. arXiv preprint, arXiv:2109.04270. 2021.
52. Campagner A, Ciucci D, Svensson C-M, Figge MT, Cabitza F. Ground truthing from multi-rater labeling with three-way decision and possibility theory. Inf Sci. 2021;545:771–90. https://doi.org/10.1016/j.ins.2020.09.049.
53. Grimmer J, Roberts ME, Stewart BM. Machine learning for social science: an agnostic approach. Annu Rev Polit Sci. 2021;24:395–419. https://doi.org/10.1146/annurev-polisci-053119-015921.
54. Bone D, Lee C-C, Chaspari T, Gibson J, Narayanan S. Signal processing and machine learning for mental health research and clinical applications [perspectives]. IEEE Signal Process Mag. 2017;34(5):196–205. https://doi.org/10.1109/MSP.2017.2718581.
55. Mateos-Pérez JM, Dadar M, Lacalle-Aurioles M, Iturria-Medina Y, Zeighami Y, Evans AC. Structural neuroimaging as clinical predictor: a review of machine learning applications. NeuroImage Clin. 2018;20:506–22. https://doi.org/10.1016/j.nicl.2018.08.019.
56. Sakai K, Yamada K. Machine learning studies on major brain diseases: 5-year trends of 2014–2018. Jpn J Radiol. 2019;37(1):34–72. https://doi.org/10.1007/s11604-018-0794-4.
57. Sejnowski TJ, Churchland PS, Movshon JA. Putting big data to good use in neuroscience. Nat Neurosci. 2014;17(11):1440–1. https://doi.org/10.1038/nn.3839.
58. Parsons T, Duffield T. Paradigm shift toward digital neuropsychology and high-dimensional neuropsychological assessments. J Med Internet Res. 2020;22(12): e23777. https://doi.org/10.2196/23777.
59. Burgos N, Colliot O. Machine learning for classification and prediction of brain diseases: recent advances and upcoming challenges. Curr Opin Neurol. 2020;33(4):439–50. https://doi.org/10.1097/WCO.0000000000000838.
60. Lu P, Colliot O, editors. Multilevel survival analysis with structured penalties for imaging genetics data. Medical Imaging 2020: Image Processing; 2020: International Society for Optics and Photonics. https://doi.org/10.1117/12.2549010
61. Peng J, An L, Zhu X, Jin Y, Shen D, editors. Structured sparse kernel learning for imaging genetics based Alzheimer's disease diagnosis. International Conference on Medical Image Computing and Computer-Assisted Intervention; 2016: Springer. https://doi.org/10.1007/978-3-319-46723-8_9
62. Qiu S, Chang GH, Panagia M, Gopal DM, Au R, Kolachalama VB. Fusion of deep learning models of MRI scans, Mini-Mental State Examination, and logical memory test enhances diagnosis of mild cognitive impairment. Alzheimer's Dement. 2018;10:737–49. https://doi.org/10.1016/j.dadm.2018.08.013.
63. Sørensen L, Nielsen M. Alzheimer's Disease Neuroimaging Initiative. Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination. J Neurosci Methods. 2018;302:66–74. https://doi.org/10.1016/j.jneumeth.2018.01.003.
64. Whelan R, Watts R, Orr CA, Althoff RR, Artiges E, Banaschewski T, et al. Neuropsychosocial profiles of current and future adolescent alcohol misusers. Nature. 2014;512(7513):185–9. https://doi.org/10.1038/nature13402.
65. Alcañiz Raya M, Chicchi Giglioli IA, Marín-Morales J, Higuera-Trujillo JL, Olmos E, Minissi ME, et al. Application of supervised machine learning for behavioral biomarkers of autism spectrum disorder based on electrodermal activity and virtual reality. Front Hum Neurosci. 2020;14:90. https://doi.org/10.3389/fnhum.2020.00090.
66. Posada-Quintero HF, Bolkhovsky JB. Machine learning models for the identification of cognitive tasks using autonomic reactions from heart rate variability and electrodermal activity. Behav Sci. 2019;9(4):45. https://doi.org/10.3390/bs9040045.
67. Juarascio AS, Crochiere RJ, Tapera TM, Palermo M, Zhang F. Momentary changes in heart rate variability can detect risk for emotional eating episodes. Appetite. 2020;152: 104698. https://doi.org/10.1016/j.appet.2020.104698.
68. Phan LV, Rauthmann JF. Personality computing: new frontiers in personality assessment. Soc Personality Psychol Compass. 2021. https://doi.org/10.1111/spc3.12624.
69. Stachl C, Pargent F, Hilbert S, Harari GM, Schoedel R, Vaid S, et al. Personality research and assessment in the era of machine learning. Eur J Personality. 2020;34(5):613–31. https://doi.org/10.1002/per.2257.
70. Koutsouleris N, Kahn RS, Chekroud AM, Leucht S, Falkai P, Wobrock T, Derks EM, Fleischhacker WW, Hasan A. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. Lancet Psychiatry. 2016;3(10):935–46. https://doi.org/10.1016/S2215-0366(16)30171-7.
71. Kessler RC, et al. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. Mol Psychiatry. 2016;21(10):1366–71. https://doi.org/10.1038/mp.2015.198.
72. Samper-Gonzalez J, et al. Predicting progression to Alzheimer's disease from clinical and imaging data: a reproducible study. In: OHBM 2019-organization for human brain mapping annual meeting 2019; 2019. https://hal.inria.fr/hal-02142315.