





RESEARCH

Open Access



# How accurately can supervised machine learning model predict a targeted psychiatric disorder?

Haitham Jahrami<sup>1,2\*</sup>, Amir H. Pakpour<sup>3</sup>, Waqar Husain<sup>4</sup>, Achraf Ammar<sup>5,6</sup>, Zahra Saif<sup>1</sup>, Ali Husain Als Salman<sup>1,2</sup>, Adel Aloffi<sup>1,2</sup>, Khaled Trabelsi<sup>7,8</sup>, Seithikurippu R. Pandi-Perumal<sup>9,10</sup> and Michael V. Vitiello<sup>11</sup>

## Abstract

**Background** Hoarding disorder (HD) is characterized by a compulsion to collect belongings, and to experience significant distress when parting from them. HD is often misdiagnosed for several reasons. These include patient and family lack of recognition that it is a psychiatric disorder and professionals' lack of relevant expertise with it. This study evaluates the ability of a supervised machine learning (ML) model to match the diagnostic skills of psychiatrists when presented with equivalent information pertinent to symptoms of HD.

**Methods** Five hundred online participants were randomly recruited and completed the Hoarding Rating Scale-Self Report (HRS-SR) and the Generalized Anxiety Disorder 7-item (GAD-7) scale. Responses to the questionnaires were read by an ML model. Responses to the HRS-SR were then converted into anonymized, random-equivalent texts. Each of these individual texts was presented in random order to two experienced psychiatrists who were independently asked for a provisional diagnosis - e.g.; the presence or absence of HD. In case of disagreement between the two assessors, a third psychiatrist broke the tie. A decision tree classification model was employed to predict clinical HD using self-report data from two psychological tests, the HRS-SR and GAD-7. The target variable was whether a participant had clinical HD, while the predictive variables were the continuous scores from the HRS-SR and GAD-7 tests. The model's performance was evaluated using a confusion matrix, which compared the observed diagnoses with the predicted diagnoses to assess accuracy.

**Results** According to the psychiatrists, approximately 10% of the participants fulfilled DSM-5 diagnostic criteria for HD. 93% of the clinician-identified cases were identified by the ML model based on HRS-SR and GAD-7 scores. A decision tree plot model demonstrated that about 60% of the cases could be detected by the HRS-SR alone while the rest required a combination of HRS-SR and GAD-7 scores. ML evaluation metrics showed satisfactory performance, with a Matthews Correlation Coefficient of 55%; Area Under Curve (AUC), 79%; a Negative Predictive Value of 76%; and a False Negative Rate of 24%.

**Conclusions** Study findings strongly suggest that ML can, in the future, play a significant role in the risk assessment of psychiatric disorders prior to face-to-face consultation. By using AI to scan big data questionnaire responses, wait time for seriously ill patients can be substantially cut, and prognoses substantially improved.

**Keywords** Artificial Intelligence, Disease Assessment, Healthcare Analytics, Hoarding disorders, Machine learning

\*Correspondence:

Haitham Jahrami  
haitham.jahrami@outlook.com

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

## Introduction

Hoarding disorder (HD) is defined as a compulsion to collect and store belongings and is associated with considerable distress at even the thought of throwing anything away [1]. The eventual accumulation of stored possessions makes living difficult [1]. The extent of hoarding and the degree of insight into the anomalous nature of the symptoms varies across HD sufferers [2]. A recent meta-analysis of eleven studies of 53,378 individuals estimated a pooled prevalence rate for HD in adults at 2.5% (95% CI (Confidence Interval) 1.7 – 3.6%) [3]. With each additional five years of age, the prevalence of HD rises by 20% [4], peaking at 6–8% in people aged 70 and older [4, 5]. HD usually first emerges in late adolescence, takes a chronic, progressive course, and is linked to negative outcomes, such as poor quality of life and a higher-than-normal mortality rate [1, 6, 7]. Older persons with HD run the risks of food contamination, malnutrition, medication error, falls, and potential homelessness [5]. Moreover, homes of people with HD can present safety issues for their neighborhood owing to pest problems, health code violations, fire hazards, and costs associated with clean-ups and local support services [8, 9]. A comorbid mood or anxiety disorder is present in almost two-thirds of those with HD, which adds a significant additional layer of complexity to their situation [10].

HD is underdiagnosed [11]. A major reason for underdiagnosis is the relative lack of awareness of HD among the general public and even among healthcare providers as a formal psychiatric disorder [1, 11]. Often, HD is attributed to a personality quirk or a housecleaning deficit [11]. Diagnostic criteria for HD have changed over time, and this disorder has only been classified as a separate and formal condition since 2013 in the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [12].

Psychiatric diagnoses are made in a variety of ways. The ‘gold standard’ is by way of a full face-to-face evaluation, which includes a standard mental status examination performed by a trained psychiatrist [13]. In addition, psychometric testing can aid in diagnosing specific conditions and in quantifying the severity of symptoms [13]. Recently, Artificial Intelligence (AI) and Machine Learning (ML) have been introduced to accelerate the detection and diagnosis of psychiatric disorders [14]. Recent research has shown that a set of 28 responses to specific questions scanned by a machine can detect and categorize numerous psychiatric disorders with an accuracy level of approximately 90% [15]. Such AI/ML-assisted ‘provisional’ diagnoses increase efficiency, accuracy, and speed of diagnosis. The algorithms used to arrive at these diagnoses are based on wide experience and knowledge that can be updated in keeping with results of new

research. Thus, they are more up-to-date than the ordinary clinician, provide diagnostic support, and offer consistency and standardization, vital to the quality conduct of evidence-based clinical trials [14, 15].

The existing literature highlights the underdiagnosis of HD due to factors such as lack of public awareness and limited expertise among healthcare professionals [11]. Accurate and timely diagnosis is crucial for early intervention and improved treatment outcomes. However, achieving a definitive diagnosis can be challenging, often requiring extensive clinical evaluations and home visits [11]. Delayed diagnosis can lead to further deterioration of the condition and exacerbation of associated impairments [1, 6, 7]. Therefore, there is a pressing need for efficient screening tools that can aid in the initial identification of individuals at risk of HD, facilitating prompt referral for comprehensive clinical assessment and appropriate management.

While previous studies have explored the potential of ML models for psychiatric disorder diagnosis [14, 15], there is a need to evaluate their diagnostic consistency with clinician assessments specifically for HD. Our study aims to address this gap by directly comparing the diagnostic performance of a supervised ML model with that of experienced psychiatrists in identifying HD cases. By leveraging self-report data from validated psychological tests, we investigate the potential of ML as a supplementary screening tool to assist clinicians in the initial assessment of HD. This approach could facilitate prompt identification of at-risk individuals, reduce diagnostic delays, and ultimately improve access to appropriate care for those suffering from HD.

Our explicit aim was to examine the clinical utility of a supervised ML model in detecting HD using the self-report data of two validated psychological tests. The output of the ML model was compared with the results of HD diagnoses of two psychiatrists based on reading anonymized accounts of symptoms drawn from one of the tests (the hoarding test). It is important to note that the purpose of our study was not to make independent predictions of HD diagnoses. Instead, our goal was to assess the diagnostic consistency between the ML model and the psychiatrists. The ML model analyzed coded numerical data, while the psychiatrists based their evaluations on detailed history descriptions of the same cases. Given previous reports of successfully employing similar methods for identifying other psychiatric disorders [14, 15], we hypothesized that the ML model would achieve acceptable levels of diagnostic consistency (over 90%) with the psychiatric assessments. This comparison was intended to explore the potential of ML models as supplementary tools in clinical settings, rather than to replace human judgment.

## Methods

### Research design

The data collection design was cross-sectional, observational, and quantitative.

### Participants

A total of 500 adults recruited online, aged 18 years and above from Bahrain, Egypt, Jordan, and Tunisia were the participants. Three hundred and sixteen (64%) were female and 74 (15%) were  $\geq 35$  years of age, no other demographic information was collected. Recruits completed a brief screening skip-logic questionnaire to determine eligibility. Those who endorsed serious mental illness and/or a chronic medical condition (diagnosed by a trained physician and receiving treatment) were excluded. This included the exclusion of major depression, bipolar disorders, schizophrenia, neurological disorders, endocrine disorders, autoimmune disorders, and cardiovascular disorders. The sample size required for this study was determined to be 400 participants based on the following assumptions: Z-value corresponding to the desired confidence level ( $Z=1.96$ ); known prevalence rate of HD ( $PR=2.5\%$ ); and a margin of error of 0.05 ( $E=0.05$ ), using the following formula  $n = (Z^2 * p * (1 - p)) / (E^2)$  [16].

### Recruitment and data collection

Recruitment was done online. The study was advertised to the general public on many social media sites, such as Instagram, Facebook, Discord, and Twitter, as well as on messaging apps like WhatsApp, Viber, and Signal. Would-be participants were directed to the online survey in Arabic, made available using a Google form. The main survey included the Hoarding Rating Scale-Self Report (HRS-SR) [17] and the Generalized Anxiety Disorder 7-item (GAD-7) [18] scales, both available in Arabic language. The inclusion of GAD-7 in this study was based on the availability of the data; it had originally been used to assess the convergent validity of the Arabic HRS-SR [19].

The HRS-SR is a self-report measure based on DSM-5 HD criteria, designed to assess HD symptoms [17]. The five-item scale provides a comprehensive assessment of hoarding-related behaviors and the distress that is associated with them [17]. The frequency and severity of each symptom are rated by respondents on a scale from 0 (not at all) to 8 (severe). The HRS-SR total score ranges from 0 to 40, with higher values indicating more severe hoarding symptoms [17]. The Arabic language version has a Cronbach alpha of approximately 0.80 and a test-retest reliability coefficient of approximately 0.90 [19].

Similarly, the GAD-7 is a self-report measure based on the DSM (Diagnostic and Statistical Manual) criteria designed to assess generalized anxiety symptoms [18].

The severity of each symptom is rated by respondents on a scale from 0 (not at all) to 3 (nearly every day) [18]. The GAD-7 total score ranges from 0 to 21, with higher values indicating more severe anxiety [18]. The Arabic language version has a Cronbach alpha of approximately 0.80 and a test-retest reliability coefficient of approximately 0.90 [20].

Responses of the participants to the HRS-SR survey were then converted into anonymized written, random-equivalent texts using AI to mimic mini-case reports. Each text represented responses to the HRS-SR by a single person and was presented in random order. We used AI to generate the texts of the mini-case reports for two reasons: First, to allow multiple textual equivalents of the same response. The HRS-SR requires that participants respond to a Likert-like scale from 0 (not at all) to 8 (severe). The response of 8 would indicate that it is "severe"; ten textual equivalents would be: "it is intense for me" OR "it is extreme for me" OR "it is serious for me" OR "it is significant for me" OR "it is profound for me" OR "it is critical for me" OR "it is substantial for me" OR "it is severe in my case" OR "it poses a severe challenge for me". These variations allow some "noise" in the conversations, which is natural in clinical settings. Second, AI was used to randomly present the order of the questions to avoid presentation order bias.

The AI tool used to generate the mini-case report texts was a language model fine-tuned on a large corpus of clinical texts and psychiatric case reports. Specifically, we employed a transformer-based neural network architecture, Generative Pre-training Transformer [GPT-3], which has demonstrated remarkable capabilities in generating human-like text based on prompts or contextual information. To generate the mini-case reports, we first provided the language model with the original questions from the HRS-SR as prompts. Then, for each response option on the Likert scale (ranging from 0 to 8), we fed the corresponding numerical score along with a seed phrase describing the severity level (e.g., "severe," "moderate," "mild"). The language model then generated multiple variations of textual descriptions that coherently expressed the specified severity level while preserving the semantic context of the original HRS-SR question. This approach allowed us to create diverse yet clinically relevant mini-case report texts representing the participants' responses to the HRS-SR. It is important to note that the generated texts were solely based on the participant's responses to the HRS-SR and did not incorporate any additional clinical information. The primary purpose of using the AI language model was to transform structured numerical data into naturalistic textual descriptions, mimicking how a clinician might document a patient's self-reported symptoms during an initial assessment.

**Table 1** An example of the original questions (HRS-SR) and their AI-Assisted generated Textual equivalents

Item	Sample original question (HRS-SR)	Sample generated text
Clutter	Because of the clutter or number of possessions, how difficult is it for you to use the rooms in your home? Questionnaire response: ①.	Question 1: How would you rate the difficulty of using the rooms in your home due to clutter or the number of your possessions? Answer 1: It is extremely difficult.
Difficulty Discarding	To what extent do you have difficulty discarding (or recycling, selling, giving away) ordinary things that other people would get rid of? Questionnaire response: ①.	Question 2: How challenging is it for you to discard ordinary things that other people would typically get rid of? Answer 2: It is incredibly difficult to get rid of the items.
Acquisition	To what extent do you currently have a problem with collecting free things or buying more things than you need or can afford? Questionnaire response: ①.	Question 3: Do you currently face any issues with collecting free items or purchasing more items than you need or can afford? Answer 3: Yes, it is exceedingly hard for me stop collecting items.
Distress	To what extent do you experience emotional distress because of clutter, difficulty discarding or problems with buying or acquiring things? Questionnaire response: ①.	Question 4: How much emotional distress do you experience because of clutter, difficulty discarding, or problems with acquiring things? Answer 4: It is tremendously demanding for me.
Impairment	To what extent do you experience impairment in your life (daily routine, job/school, social activities, family activities, financial difficulties) because of clutter, difficulty discarding, or problems with buying or acquiring things? Questionnaire response: ①.	Question 5: To what extent does clutter, difficulty discarding, or problems with acquiring possessions affect your daily routine, job/school, social activities, family activities, or financial situation? Answer 5: It has an extreme impact on all aspects of my life.

HRS-SR response keys= 0–1 None/Not at all; 2–3 Mild difficulty; 4–5 Moderate difficulty; 6–8 Severe/Extreme difficulty

Table 1 supplies complete examples of the original questions (HRS-SR) and their AI-generated text equivalents.

Two trained psychiatrists (Arab Board Certified in Psychiatry, each with over 20 years of clinical practice); Psychiatrist 1 completed residency training in general adult psychiatry and has extensive experience diagnosing and treating neuropsychiatric conditions including HD. Psychiatrist 2 completed a fellowship in geriatric psychiatry after finishing a psychiatry residency and has expertise in managing HD cases in elderly populations. Psychiatrist 1 and Psychiatrist 2 were independently invited to supply a provisional diagnosis of HD (or not) for each text/mini-case report. If the two psychiatrists disagreed, a third certified psychiatrist made the determining choice. The intraclass correlation coefficient was 0.96 between the first and second psychiatrists.

The supervised ML model was performed first on a randomly selected 400 of the 500 participants, and, later, results were confirmed using the remaining 100 participants. To elaborate, in machine learning models, the training set is used to “train” the model, allowing it to learn the patterns and relationships between the input features (in this case, the HRS-SR and GAD-7 scores) and the target variable (presence or absence of HD). The model adjusts its internal parameters based on the training data to optimize its ability to make accurate estimates. After the model is trained, its performance is evaluated on a separate validation or test set, which consists of data points that were not used during the training process. This step is crucial to assess the model’s ability to generalize to unseen data and to obtain an unbiased estimate of its performance metrics, such as accuracy, precision, and recall. In our study, we randomly selected 400 participants (80% of the total sample) to serve as the training set for the supervised machine learning model. The model was trained on this subset of data, learning the mapping between the HRS-SR and GAD-7 scores and the presence or absence of hoarding disorder as determined by the psychiatrists’ assessments. Subsequently, to confirm the model’s performance and ensure its robustness, we evaluated it on the remaining 100 participants (20% of the total sample), which served as the validation or test set. This independent evaluation of unseen data provided an unbiased assessment of the model’s diagnostic capabilities and allowed us to report reliable performance metrics, such as accuracy, precision, recall, and other evaluation metrics mentioned in the results section.

This train-test split approach is a standard practice in machine learning to ensure that the model’s performance is not overly optimistic due to overfitting on the training data and to obtain a realistic estimate of its generalization ability to new, unseen cases.

### Ethical considerations

The study procedures followed the ethical guidelines specified in the 1964 Helsinki Declaration and its later revisions. The Institutional Review Board at the Psychiatric Hospital, Bahrain approved the research. Participation was voluntary, and withdrawal was possible at any time. No incentives were offered to the participants. Informed consent was electronically obtained prior to the survey response.

All transformations and computing were performed in local environments, and no information was uploaded to external servers.

### Data analysis

Data were visualized using quantile-quantile plots (Q-Q plots) to examine the data structure, to detect outliers, and to check normality assumptions. Descriptive statistics were used to summarize the findings. Means and standard deviations were used for continuous data. Numbers and percentages were used for categorical data. Independent samples t-tests or  $\chi^2$  ( $\chi^2$ ) statistics were used to compare the results of participants clinically diagnosed with HD vs. not so diagnosed. Cohen’s d or Carmer’s V were used to determine effect sizes [21].

A decision tree classification modeling was employed as a predictive model to move from observations about an item (represented in the tree’s roots) to inferences about the item’s target value (expressed in the tree’s endpoints) [22]. In our analyses, the target variable was the categorical variable (clinical HD vs. no clinical HD). The predictive variables consisted of the continuous variables of HRS-SR vs. GAD-7.

Results reporting the confusion matrix showed the observed classes against the predicted classes [22].

The confusion matrix compares the observed classes to the expected classes in a table that is displayed, it is used to evaluate the model’s accuracy.

Class proportions showed the proportions of each class in the data set, training (and validation), and test set [22, 23]. Finally, we reported fifteen Evaluation Metrics [22, 23] including support; accuracy; precision (positive predictive value); recall (true positive rate); false positive rate; false discovery rate; F1 score; Matthew’s correlation coefficient datasets; area under curve (AUC); negative predictive value; true negative rate; false negative rate; false omission rate; threat score; and statistical parity [22, 23].

Statistical analyses were performed using R statistical software version 4.3.1 (Beagle Scouts) released on 2023-06-16 [24]. Statistical significance was defined as a  $p$ -value < 0.05. The statistical packages “rpart” [25] and “ROCR” [26] were used in AI/ML modelling.

**Results**

Table 2 represents a comprehensive report of the main findings. The mean HRS-SR score for all participants was 13.07, SD (Standard Deviation)=7.55. The HRS-SR total score was significantly higher for the HD group vs. the no HD group, *p*-value < 0.001; Effect Size (Cohen’s *d*)=2.0.

The mean GAD-7 score for all participants was 10.87, SD=4.57. The GAD-7 score was significantly higher for the HD group vs. the no HD group, *p*-value < 0.001; Effect Size (Cohen’s *d*)=1.2.

The confusion matrix showed that among the HD group, 5% were incorrectly identified as cases, while 3% were incorrectly classified as non-cases. For the no HD group, 4% were incorrectly classified as cases, while a large majority (88%) were accurately classified as non-cases.

A total of 51 cases met the clinically defined criteria for HD, a prevalence rate of 10.2% (or 0.102) with a 95% confidence interval of 4.23–16.17%. The ML class proportion analysis showed that, in the HD group category, the training set included 10% of HD cases, and the test set included 11% of HD cases, comparable rates of HD. The ML accuracy rate was 93% in detecting clinical cases of HD. The relative importance of variables included in the model was 88.89% for the HRS-SR scores and 11.11% for the GAD-7 scores.

The average false positive rate was 0.21, the average F1 score was 0.93, and the area under the curve (AUC) was 0.79. The entire details of the evaluation metrics

**Table 3** Evaluation metrics of the supervised machine learning algorithm used in validation sample (*n*= 100)

Metric	Case	Not case	Average/Total
Support	8	92	100
Accuracy	0.93	0.93	0.93
Precision (Positive Predictive Value)	0.56	0.97	0.93
Recall (True Positive Rate)	0.63	0.96	0.93
False Positive Rate	0.04	0.38	0.21
False Discovery Rate	0.44	0.03	0.24
F1 Score	0.59	0.96	0.93
Matthews Correlation Coefficient	0.55	0.55	0.55
Area Under Curve (AUC)	0.79	0.79	0.79
Negative Predictive Value	0.97	0.56	0.76
True Negative Rate	0.96	0.63	0.79
False Negative Rate	0.38	0.04	0.21
False Omission Rate	0.03	0.44	0.24
Threat Score	0.45	8.8	4.63
Statistical Parity	0.09	0.91	1.00

All metrics are calculated for every class against all other classes

are presented in Table 3. Figure 1 presents the decision tree plot using the training dataset. Using the training dataset of 400 participants detailed examination of the decision tree plot demonstrates that, based on the HRS-SR only, the projected prevalence rate of HD is estimated to be 60% (20/31). An added approximately 40% (15/31) of the cases was estimated when using a combination of HRS-SR and GAD-7 scores.

**Table 2** Descriptive results of the main findings in all sample (*n* = 500)

Item	All respondents ( <i>n</i> = 500)	Hoarding disorder ( <i>n</i> = 51)	No hoarding disorder ( <i>n</i> = 449)	<i>p</i> -value; ES (Effect Size)
HRS-SR 1 (Clutter)	2.66±2.01	5.31±1.36	2.36±1.84	< 0.001; Cohen’s <i>d</i> =1.64
HRS-SR 2 (Difficulty Discarding)	2.8±2.06	5.18±1.44	2.53±1.95	< 0.001; Cohen’s <i>d</i> =1.39
HRS-SR 3 (Acquisition)	2.61±1.98	4.0±1.72	2.45±1.95	< 0.001; Cohen’s <i>d</i> =0.80
HRS-SR 4 (Distress)	2.81±2.22	5.47±1.39	2.51±2.09	< 0.001; Cohen’s <i>d</i> =1.46
HRS-SR 5 (Impairment)	2.18±2.08	4.73±1.15	1.89±1.96	< 0.001; Cohen’s <i>d</i> =1.5
HRS-SR Total Score	13.07±7.55	24.69±4.51	11.75±6.64	< 0.001; Cohen’s <i>d</i> =2.0
GAD-7 Score	10.87±4.57	15.61±3.63	10.34±4.35	< 0.001; Cohen’s <i>d</i> =1.23
Sex				0.60; Cramer’s <i>V</i> <0.1
Female	316 (64%)	33 (65%)	283 (63%)	
Male	179 (36%)	18 (35%)	166 (37%)	
Age				0.30; Cramer’s <i>V</i> <0.1
<35 years	425 (85%)	40 (78%)	385 (86%)	
≥35 years	74 (15%)	11 (23%)	64 (14%)	

Values are expressed as Mean ± Standard Deviation

HRS-SR The Hoarding Rating Scale – Self Report, GAD-7 The Generalized Anxiety Disorder Assessment, ES Effect size

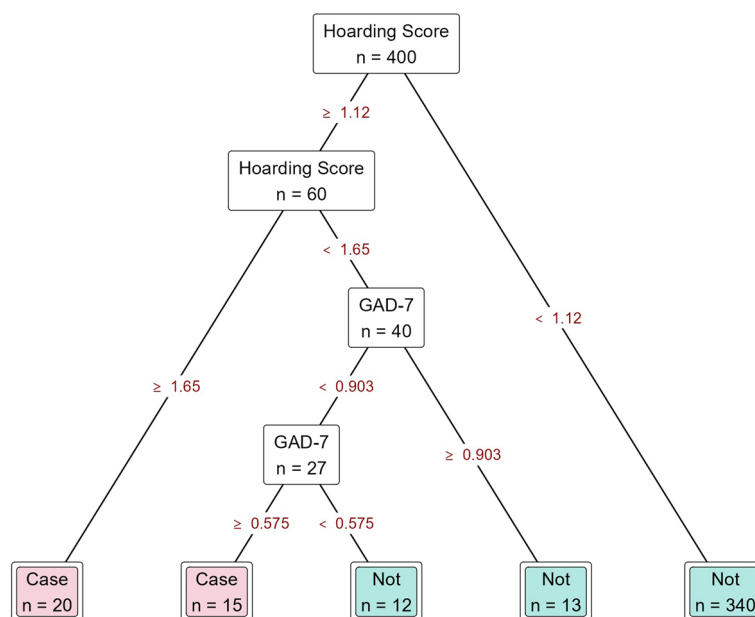


Fig. 1 Decision tree plot

**Discussion**

We found that 51 (10.2%) of study participants were identified by a pair of psychiatrists as meeting HD diagnostic criteria. ML correctly identified 93% of these cases using the HRS-SR and GAD-7 scores.

Our study prevalence rate of 10% appears to be higher than expert consensus reports of population-based prevalence studies, which generally report rates between 2% and 6% [5, 12, 27, 28]. Estimated prevalence, of course, depends on the population being studied. In a clinical population, Ong and colleagues reported that over 30% of a sample of 500 psychiatric outpatients in Singapore reported hoarding symptoms and about 14% met DSM-5 criteria for the disorder [29].

It is plausible, even probable, that the recruitment information posted about the study might have differentially attracted individuals with HD symptoms (i.e., selection bias). If we assume a globally accepted median community-based prevalence rate of approximately 5% for HD [5, 12], the added 5% seen here might also be attributed to the brevity and, therefore, ambiguity of the information made available to the psychiatrists.

Our ML model showed a consistency-with-clinical-diagnosis rate of 93%, suggesting the model was very accurate. The decision tree plot showed that only 4% of the cases needed further evaluation beyond the HRS-SR, suggesting that anxiety symptoms might have contributed to the relatively high prevalence of diagnosed HD found in our sample.

A recent systematic review of AI versus clinicians in disease diagnosis found that AI/ML can sometimes

outperform clinicians, particularly relatively inexperienced clinicians [30]. AI is already showing great promise in passing medical exams [31].

Concerning HD, a full clinical interview is, of course, superior to reading a brief text when determining whether diagnostic criteria have been met [32]. Follow-up questions are very revelatory -e.g. “How difficult is it for you to discard or part with possessions?” or “How much space in the main rooms of your home is filled with clutter?” [32, 33]. A home visit is better yet. The extent of the problem, the impairment it causes, and the potential risks can then be directly evaluated [32]. Obtaining information from a third party (a spouse or close relative) is also informative for a diagnosis of HD [32].

AI/ML cannot, at this time, provide a definitive diagnosis. However, this approach can be used in two-stage large-scale epidemiological surveys to reduce the time and cost of collecting data on specific psychiatric disorders. Our findings further suggest that AI/ML screening preceding a formal evaluation can facilitate a prompt intake and an early start to treatment, cutting wait times in high-risk patients; as well as monitoring therapy outcomes and facilitating care decisions such as discharge and transfer of care.

**Strengths and limitations**

Our study has several strengths. First, it utilized a relatively large study sample. Second, it used a head-to-head comparison between psychiatric and ML diagnoses of HD. Third, it demonstrated how a dedicated ML model can offer a novel approach to diagnosing psychiatric

illness using self-report data in a large cross-sectional sample.

The major limitation of our approach is that psychiatrists' diagnoses were based on written texts derived from HRS-SR responses and not on a standard, face-to-face mental status exam. In our dataset, traceability of cases to determine progress and outcomes was not possible. The ability to track cases should be a key feature of future work in this area. Another limitation was the restricted numbers of the questionnaires used (i.e., HSR-SR and GAD-7). Future research needs to consider more comprehensive tests such as the Symptom Checklist-90-R (SCL-90-R) to better address the question of diagnostic accuracy. A key limitation is the assumption that the psychiatric diagnosis, based on limited texts, is accurate. Furthermore, the sample was ethnically homogeneous (all Arab). Future research needs to confirm the model with diverse clinical populations who have received prior clinically established diagnoses.

## Conclusion

Based on online questionnaires and texts constructed from questionnaire responses, approximately 10% of general population study participants fulfilled diagnostic criteria for HD. Analysis showed that, in 93% of cases, an HD diagnosis consistent with that of well-trained psychiatrists could be reached using HRS-SR and GAD-7 scores interpreted by machine learning. This strongly suggests that ML can, in the future, play a significant role in the risk assessment of psychiatric disorders prior to face-to-face consultation. By using AI to scan big data questionnaire responses, wait time for seriously ill patients can be substantially cut, and prognoses substantially improved.

## Abbreviations

AI	Artificial Intelligence
AUC	Area Under Curve
DSM	Diagnostic and Statistical Manual
GAD-7	Generalized Anxiety Disorder 7-item
HD	Hoarding Disorder
HRS-SR	Hoarding Rating Scale-Self Report
ML	Machine Learning
SD	Standard Deviation

## Acknowledgements

We would like to acknowledge and thank the participants for their time. We would like also to thank Dr. Nour Mohammad Hussain, Dr. Dalal Hasan AlMansour, Dr. Muneera AlGhareeb, Dr. Yaser Mansoor Almutawa, Dr. Omayma Khaled Bucheer, and Dr. Mai Helmy for their role in collecting data and making it useable in this study. We dedicate this paper to the enduring legacy of Dr. Mary Seeman (late) in tribute to her indelible contributions. She contributed in this manuscript before her death. She was a beacon of inspiration for us and many other researchers worldwide. As Professor Emerita in the Department of Psychiatry at the University of Toronto, Dr. Seeman dedicated her career to unraveling the intricate biopsychosocial variances between genders, particularly in the context of psychotic disorders. Her groundbreaking research into gender disparities in schizophrenia not only earned her international acclaim but also revolutionized our understanding of women's mental health. May her soul rest in peace.

## Clinical trial number

Not applicable.

## Author contributions

Author contributions: HJ, AHP, WH, AA, ZS, KT involved in conception and performed experiment. HJ, ZS, KT collected data. AHP and HJ performed all analyses. HJ, AHP, WH, AA, ZS, AHAS, ARA, KT, SRPP, MVV wrote the main manuscript text. HJ, AHP, WH, AA, ZS, AHAS, ARA, KT, SRPP, MVV reviewed the first draft and provided critical revisions. All authors reviewed and approved the manuscript.

## Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## Data availability

Availability of data and materials: The data that support the findings of this study are available from the corresponding author upon request.

## Declarations

### Ethics approval and consent to participate

The research received approval from the Institutional Review Board at the Psychiatric Hospital, Bahrain (Approval number: REC/11/76; Date: November 30, 2023). All procedures were conducted in accordance with relevant guidelines and regulations. The study adhered to the ethical guidelines outlined in the Helsinki Declaration of 1964 and its later amendments (1975, 1983, 1989, and 1996). Informed consent was obtained from all participants.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing interests.

### Author details

<sup>1</sup>Government Hospitals, Psychiatric Hospital, Manama, Bahrain. <sup>2</sup>Department of Psychiatry, College of Medicine and Medical Sciences, Arabian Gulf University, Manama, Bahrain, Manama, Bahrain. <sup>3</sup>Department of Nursing, School of Health and Welfare, Jönköping University, Hälsohögskolan, Jönköping 55318, Sweden. <sup>4</sup>Department of Humanities, COMSATS University Islamabad, Islamabad Campus, Park Road, Islamabad, Pakistan. <sup>5</sup>Department of Training and Movement Science, Institute of Sport Science, Johannes Gutenberg-University Mainz, Mainz, Germany. <sup>6</sup>Research Laboratory, Molecular Bases of Human Pathology, LR19ES13, Faculty of Medicine of Sfax, University of Sfax, Sfax, Tunisia. <sup>7</sup>High Institute of Sport and Physical Education of Sfax, University of Sfax, Sfax 3000, Tunisia. <sup>8</sup>Research Laboratory: Education, Motricity, Sport and Health, EM2S, LR19JS01, University of Sfax, Sfax 3000, Tunisia. <sup>9</sup>Centre for Research and Development, Chandigarh University, Mohali, Punjab 140413, India. <sup>10</sup>Division of Research and Development, Lovely Professional University, Phagwara, Punjab 144411, India. <sup>11</sup>Department of Psychiatry & Behavioral Sciences, University of Washington, Seattle, USA.

Received: 7 April 2024 Accepted: 8 October 2024

Published online: 16 October 2024

## References

1. Bratiotis C, Muroff J, Lin NXY. Hoarding disorder: development in conceptualization, intervention, and evaluation. *Focus (Am Psychiatr Publ)*. 2021;19(4):392–404.
2. Mathews CA, Delucchi K, Cath DC, Willemsen G, Boomsma DI. Partitioning the etiology of hoarding and obsessive–compulsive symptoms. *Psychol Med*. 2014;44(13):2867–76.
3. Zaboloski BA, Merritt OA, Schrack AP, Gayle C, Gonzalez M, Guerrero LA, Duenas JA, Soreni N, Mathews CA. Hoarding: a meta-analysis of age of onset. *Depress Anxiety*. 2019;36(6):552–64.
4. Davidson EJ, Dozier ME, Pittman JOE, Mayes TL, Blanco BH, Gault JD, Schwarz LJ, Ayers CR. Recent advances in Research on Hoarding. *Curr Psychiatry Rep*. 2019;21(9):91.



5. Cath DC, Nizar K, Boomsma D, Mathews CA. Age-specific prevalence of hoarding and obsessive compulsive disorder: a population-based study. *Am J Geriatr Psychiatry*. 2017;25(3):245–55.
6. Ayers CR, Saxena S, Golshan S, Wetherell JL. Age at onset and clinical features of late life compulsive hoarding. *Int J Geriatr Psychiatry*. 2010;25(2):142–9.
7. Ong C, Pang S, Sagayadevan V, Chong SA, Subramaniam M. Functioning and quality of life in hoarding: a systematic review. *J Anxiety Disord*. 2015;32:17–30.
8. Ayers CR. Age-specific prevalence of hoarding and obsessive compulsive disorder: a population-based study. *Am J Geriatr Psychiatry*. 2017;25(3):256–7.
9. Nguyen BK, Zakrzewski JJ, Sordo Vieira L, Mathews CA. Impact of Hoarding and obsessive-compulsive disorder symptomatology on quality of life and their interaction with Depression Symptomatology. *Front Psychol*. 2022;13: 926048.
10. Nakao T, Kanba S. Pathophysiology and treatment of hoarding disorder. *J Neuropsychiatry Clin Neurosci*. 2019;73(7):370–5.
11. Mataix-Cols D, Frost RO, Pertusa A, Clark LA, Saxena S, Leckman JF, Stein DJ, Matsunaga H, Wilhelm S. Hoarding disorder: a new diagnosis for DSM-V? *Depress Anxiety*. 2010;27(6):556–72.
12. American Psychiatric Association D, Association AP. Diagnostic and statistical manual of mental disorders: DSM-5, vol. 5. Washington, DC: American psychiatric association; 2013.
13. Nesse RM. Evolutionary psychiatry: foundations, progress and challenges. *World Psychiatry*. 2023;22(2):177–202.
14. Rashid B, Calhoun V. Towards a brain-based predictive model of mental illness. *Hum Brain Mapp*. 2020;41(12):3468–535.
15. Tutun S, Johnson ME, Ahmed A, Albizri A, Irgil S, Yesilkaya I, Ucar EN, Sengun T, Harfouche A. An AI-based Decision Support System for Predicting Mental Health Disorders. *Inform Syst Front*. 2023;25(3):1261–76.
16. Daniel WW, Cross CL. Biostatistics: a foundation for analysis in the health sciences. Hoboken (NJ): Wiley; 2018
17. Tolin DF, Frost RO, Steketee G. A brief interview for assessing compulsive hoarding: the Hoarding Rating Scale-Interview. *Psychiatry Res*. 2010; 178(1):147–52.
18. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med*. 2006;166(10):1092–7.
19. Hussain NM, AlMansouri DH, AlGhareeb M, Almutawa YM, Bucheeri OK, Helmy M, Trabelsi K, Saif Z, Jahrami H. Translating and validating the hoarding rating scale-self report into Arabic. *BMC Psychol*. 2023;11(1):233.
20. AlHadi AN, AlAteeq DA, Al-Sharif E, Bawazeer HM, Alanazi H, AlShomrani AT, Shuqdar RM, AlOwaybil R. An arabic translation, reliability, and validation of Patient Health Questionnaire in a Saudi sample. *Ann Gen Psychiatry*. 2017;16:32.
21. Poom L, af Wählberg A. Accuracy of conversion formula for effect sizes: a Monte Carlo simulation. *Res Synth Methods*. 2022;13(4):508–19.
22. Kotsiantis SB. Decision trees: a recent overview. *Artif Intell Rev*. 2013;39:261–83.
23. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: springer; 2013.
24. R-Core-Team. 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>. In., 4.3.1 edn.
25. Therneau T, Atkinson B, Ripley B, Ripley MB. Package 'rpart'. <https://doi.org/10.32614/CRAN.package.rpart>.
26. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCr: visualizing classifier performance in R. *Bioinformatics*. 2005;21(20):3940–1.
27. Mueller A, Mitchell JE, Crosby RD, Glaesmer H, de Zwaan M. The prevalence of compulsive hoarding and its association with compulsive buying in a German population-based sample. *Behav Res Ther*. 2009;47(8):705–9.
28. Iervolino AC, Perroud N, Fullana MA, Guipponi M, Cherkas L, Collier DA, Mataix-Cols D. Prevalence and heritability of compulsive hoarding: a twin study. *Am J Psychiatry*. 2009;166(10):1156–61.
29. Ong C, Sagayadevan V, Lee SP, Ong R, Chong SA, Frost RO, Subramaniam M. Hoarding among outpatients seeking treatment at a psychiatric hospital in Singapore. *J Obsessive Compulsiv Relat Disorders*. 2016;8:56–63.
30. Kiliç ME. AI in medical education: a comparative analysis of GPT-4 and GPT-3.5 on Turkish medical specialization exam performance. *medRxiv*. 2023. <https://doi.org/10.1101/2023.07.12.23292564>.
31. Shen J, Zhang CJP, Jiang B, Chen J, Song J, Liu Z, He Z, Wong SY, Fang PH, Ming WK. Artificial Intelligence Versus clinicians in Disease diagnosis: systematic review. *JMIR Med Inf*. 2019;7(3):e10010.
32. Mataix-Cols D, de la Cruz LF. Hoarding disorder has finally arrived, but many challenges lie ahead. *World Psychiatry*. 2018;17(2):224.
33. Postlethwaite A, Kellett S, Mataix-Cols D. Prevalence of Hoarding Disorder: a systematic review and meta-analysis. *J Affect Disord*. 2019;256:309–16.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.