Behavior Research Methods manuscript No.
(will be inserted by the editor)

Malingering Remover: High Accuracy Reconstruction of Honest Responses from Malingered Responses with Machine Learning

Matteo Cardaioli · Stefano Cecconello · Merylin Monaro · Giuseppe Sartori · Mauro Conti · Graziella Orrù

Received: date / Accepted: date

Abstract Malingered responses to psychological testing are frequent when monetary incentives or other forms of rewards are at stake. Psychological symptoms are usually identified through clinical questionnaires which, however, may be easily inflated by malingered responses (fake-bad). A fake-bad response style is usually identified through specialised scales embedded in the personality questionnaires, but no procedure is currently available that reconstructs honest responses from malingered responses.

In this paper, we present a technique for the Millon (MCM III) questionnaire a widely used test for investigating psychopathology. This technique detects malingered MCM III profiles (malingering detector) and removes the intentionally inflated test results (malingering remover). Using the decision tree, an interpretable machine learning model, we demonstrate that validity scales of MCM III can discriminate between malingerer and honest profiles with 90% accuracy. Moreover, our results show that by applying machine learning models (e.g., Autoencoder and multi-output Regressor) to malingerer tests, we are able to well reconstruct the original honest profile. Our models decrease the RMSE (Root Mean Square Error) of the reconstruction up to 19% compared to base correction procedures. Finally, applying the malingering detector to the reconstructed scales, we show that only 5% were classified as malingerers, demonstrating the validity of the proposed approach.

Keywords Malingering · Malingering Remover · Millon · Machine Learning

Matteo Cardaioli · Stefano Cecconello · Mauro Conti
Department of Mathematics, University of Padua, Padua, Italy

Matteo Cardaioli
GFT Italy, Milan, Italy

Merylin Monaro · Giuseppe Sartori
Department of General Psychology, University of Padua, Padua, Italy

Graziella Orrù
Department of Surgical, Medical, Molecular & Critical Area Pathology, University of Pisa, Pisa, Italy

1 Introduction

Deception to direct questions may take two different forms: faking-bad and faking-good. Faking-bad characterizes some forensic settings (e.g., criminal, insurance claims) in which the examinee is likely to exaggerate or make up his psychological disorder (Sartori et al., 2017). Clinical interviews generally yield low detection rates of malingerers, meaning that many cases will be misclassified if clinicians rely solely on their subjective judgement (Rosen and Phillips, 2004). Indeed, intuitive clinical judgment yields detection rates of faking-bad that are comparable to the disappointingly low hit rates (i.e., 60%) found for intuitive judgment in the broader deception–detection literature (Vrij, 2000).

Malingering is the dishonest and intentional production or exaggeration of physical or psychological symptoms to obtain external gain (Tracy and Rix, 2017). Despite it being categorically coded by both ICD10 (Organization et al., 1992) and DSM5 (Association et al., 2013), malingering is not a binary "present" or "absent" phenomenon: it must be considered within specific domains (e.g., psychological, cognitive, and medical), often coexists with genuine disorders and can be classified into different types. Due to these considerable variations, appraising the prevalence of malingering in clinical and forensic populations is difficult. Furthermore, according to estimates by forensic practitioners, malingering likely occurs in 15–17% of forensic cases (Rogers and Bender, 2018; Young, 2014).

Usually, psychological symptoms are identified, in psychopathological inventories, through responses to direct questions (e.g., MMPI-2 or MCM III) where the examinee is required to respond YES/NO to sentences targeting relevant symptoms. However, the evaluation based only on responses to direct questions is failing miserably in some contexts. Specifically, when the responder has an incentive to aggravate his symptoms to gain economic advantage or any other form of gains. To counter this problem, a wide array of tests has been developed that provide scores on the credibility level of the endorsed symptoms. When employing these instruments, empirically-based cut-offs aid in determining whether symptoms are likely to be genuine or not (Merten and Merckelbach, 2013).

As regards the detection of malingering, several detection techniques for psychological testing are based on validity scales embedded in general psychopathological questionnaires (e.g., MMPI II and MCM III - Millon Clinical Multiaxial Inventory that are among the most used tests to evaluate psychiatric disorders) or specific tests (e.g., SIMS) as reported by Orrù et al. (2020b) and Mazza et al. (2019). Such detection strategies usually evaluate the endorsement of very atypical symptoms. For example, the SIMS may distinguish malingerers from honest responding with good accuracy (van Impelen et al., 2014), collecting responses to questions that cover a broad spectrum of pseudo-psychopathology (e.g., items indexing atypical depression, improbable memory problems, pseudo-neurological symptoms, hyperbolic signs of mental retardation).

Malingering is a continuous variable and the level of malingering is modulated by the stake, and by the strategy under implicit or explicit control of the malingerer. For this reason, efforts have been made to develop specific tests that flag the responder as a faker. Such tests may be specific (e.g., SIMS) or may take the form of a validity scale embedded in a psychopathological questionnaire (e.g., MMPI, MCM III). While such procedures may spot the faker with decent accuracy, to the best of our knowledge no procedure has been proposed to reconstruct the honest response profile once a faker has been identified and only the faked profile is available. In short, a non-depressed subject who wants to appear as depressed may be spotted as a faker. However, there is no valid procedure that may be used to uncover his true level of depression resulting from honest responses. Suppose, for example, that the true depression level of the responder is 0.2 standard deviations above the mean of the depression scale. However, because of feigning, he appears to be 2.5 standard deviations above the mean. Available techniques today may flag that he is a faker but no technique is available that derives from the observed standard score of 2.5, the real score (unobserved 0.2).

2 Method

The MCM III (Millon and Davis, 1997) is a widely used questionnaire-type test that assesses a variety of psychopathological dimensions. The format of the test requires the examinee to respond to sentences that index psychopathological symptoms. Statements addressing homogeneous psychopathological symptoms (e.g., depression) are added together, leading to a scale score which is high in the case of psychopathology and in a lower range for non-pathological subjects. Malingerers can easily alter their true response from non-pathological to pathological, thus inflating the pathological significance of the resulting score. To highlight the effect of such intentional (and unintentional) distortions, the MCM III is equipped with a number of validity scales. As with other modern psychological tests, the MCM III has three validity scales that are devised to capture exaggeration (X scale) and symptoms denial also called social desirability (Y scale). It has been shown that from scores at these two scales the faker can be identified with an accuracy that depends on several factors (Daubert and Metzler, 2000). Apart from the three validity scales reported above, the MCM III has 11 scales indexing personality patterns, three scales indexing severe personality disorders, three severe clinical syndromes and finally seven clinical syndromes for a total of 24 clinical scales.

Our work aims to reconstruct honest MCM III profiles starting from dishonest malingered profile. The procedure we propose consists of two steps: the Malingerer Detection and the Malingerer Removing, as reported in Figure 1. The malingerer detector takes in input the 24 clinical and the three validity scales of the MCM III questionnaire. This first step consists of a binary classifier that labels the input in honest or malingerer. If the profile is classified as honest, no further elaborations are needed, and the final output corres-

ponds to the original output of clinical and validity scales. On the contrary, if a malingerer profile is identified, the original scales are processed by the malingering remover. The malingering remover consists of a regression algorithm that filters the input, removing the malingering distortion, and providing a reconstructed honest profile as output.

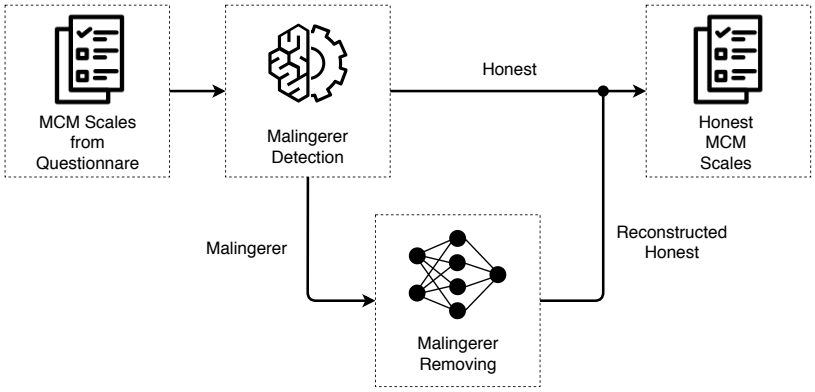


Figure 1: Workflow of the proposed approach. The Malingering Removing step is applied only to those profiles classified as malingered from the Malingerer Detection step.

2.1 Data collection

One hundred healthy participants were required to respond to the MCM III honestly and also to respond faking depression in order to sustain an insurance compensation seeking claim. 40 participants were males and 60 females. Age ranged between 20 and 61 (mean = 27.45, sd = 7.87) and schooling between 8 and 22 (mean = 16.56 and sd = 2.67). All the participants were Italian native speakers and the MCM III was given in Italian (Zennaro et al., 2008). All the subjects did not report previous psychological or psychiatric assessments. Half of them responded in the honest condition first and half in the honest condition second. The participant responses were collected using a computer presentation of the MCM III with one of the experimenter supervising in the room. Instructions for the condition requiring malingering were the following: *"You are now asked to fake a severe depression due to a family mourning. Please, respond to the questionnaire pretending to be depressed. The final goal is to obtain insurance compensation for the psychological damage you had after the mourning. Be careful to respond in a way that the depression is credible"*.

At the end, for each participant, two MCM III raw results were available. The first with standard instructions was regarded as honest responding and was used as ground truth in the development of the malingering remover. The

second collected with fake-bad instructions was regarded as the malingered MCM III to be corrected by the model.

3 Data analysis and results

In this Section, we first give insight into the dataset analysing the statistical distribution of honest and malingerer profiles. We then evaluate the results of different malingerer detection classifiers. Finally, we report and compare the performance of different approaches for the malingering remover.

3.1 Descriptive statistics

A first analysis was carried out by examining the statistical differences between malingerer and honest tests. We applied Kolmogorov-Smirnov test (Massey Jr, 1951) (with α fixed to 0.05), which rejects the null hypothesis, suggesting that the scales were not normally distributed. For this reason, honest and malingered test results were compared using (i) the Wilcoxon signed-rank test (Woolson, 2007), and (ii) the Cliff's d effect-size measure (Cliff, 1996). As recommended by the guidelines given by Grissom and Kim (2005), we interpret the effect size as *small* for $|d| < 0.33$, *medium* for $0.33 \leq |d| < 0.474$, and *large* for $|d| \geq 0.474$.

The Wilcoxon signed-rank test resulted in significant differences on all the scales ($p < 0.05$) except for Scale N (Bipolar Disorder), suggesting that, when faking a depression, also scales not related to depression change significantly. As regards to the effect size, 17 out of 24 scales showed large values of d. In particular, scales CC (Major Depression) and D (Persistent Depression) reported the highest effect size (0.96 and 0.95 respectively), confirming that participants have successfully faked a depression profile in the malingered test. Small values of d ware reported for four scales: 6A (Antisocial), 6B (Aggressive), T (Drug Use), and PP (Delusional Disorder). Finally, only Scale 7 (Compulsive) reported having a medium effect size. As reported in Figure 2, most scales present higher scores in malingering tests. Further, two scales had a reduction in score after malingering: Scale 5 (Narcissistic) and Scale SS (Thought Disorder).

A correlation analysis was carried out to assess if the dependencies between scales change in malingerer and honest condition. Figure 3a depicts the scales correlation in honest condition. In particular, the 5 couples of scales that present highest r values are: Z (Debasement) - D (Persistent Depression) $r = 0.93$, Z (Debasement) - 2B (Melancholic) $r = 0.92$, Z (Debasement) - H (Somatic Symptom) $r = 0.89$, CC (Major Depression) - D (Persistent Depression) $r = 0.89$, and CC (Major Depression) - H (Somatic Symptom) $r = 0.88$.

Similarly, we analysed the correlation matrix in the malingering condition (see Figure 3b). The results are the following: X (Disclosure) - 8A (Negativistic) $r = 0.87$, X (Disclosure) - P (Paranoid) $r = 0.86$, 8B (Masochistic)

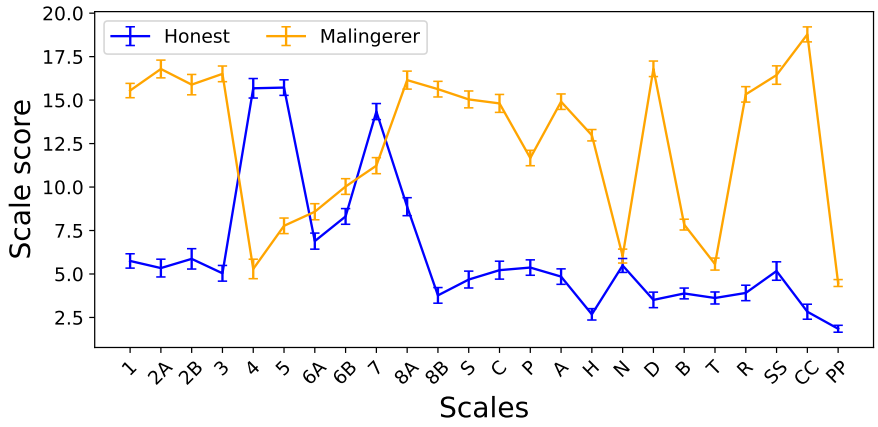
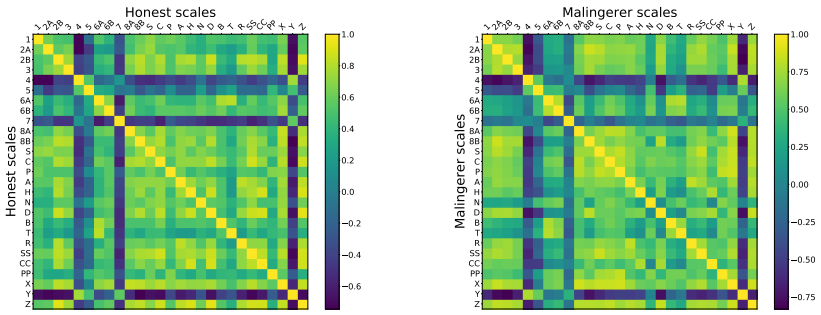


Figure 2: Average profile of the 100 participants (with the corresponding standard error) responding to the Millon questionnaire for the 24 clinical scales.

- 2A (Avoidant) $r = 0.86$, X (Disclosure) - Z (Debasement) $r = 0.86$, and S (Schyzotypal) - SS (Thought Disorder) $r = 0.85$. If we compare the two correlation matrices, there are no pairs of scales in common between the two top-5 correlations, suggesting a different response strategy depending on the task (honest vs malingering) that alters the correlation structure among the scales.



(a) *Honest condition.* (b) *Malingering condition.*

Figure 3: Autocorrelation matrices between pair of scales in both honest and malingering condition.

Analysing the cross-correlation between honest and malingering, it is possible to notice how the values of r drop significantly. The five pairs of scales

with the highest r values are respectively: N_honest (Bipolar Disorder) - N_malingerer (Bipolar Disorder) $r = 0.52$, PP_honest (Delusional Disorder) - N_malingerer (Bipolar Disorder) $r = 0.49$, PP_honest (Delusional Disorder) - 6B_malingerer (Sadistic) $r = 0.43$, S_honest (Schizotypal) - S_malingerer (Schizotypal) $r = 0.43$, and SS_honest (Thought Disorder) - SS_malingerer (Thought Disorder) $r = 0.42$. There are no strong correlations between corresponding scales in the two conditions, indicating that the reconstruction of the honest profile from the malingering profile cannot be based on a clear relation between corresponding scales.

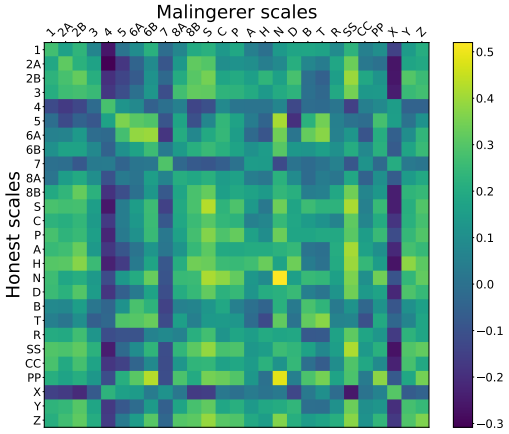


Figure 4: Correlation matrix between pair of scales in the honest and malingering conditions.

3.2 Malingering detection

The discrimination of honest and malingerer profiles represent the first step of our method (see Figure 1). To perform this task, five machine learning (ML) algorithms were tested using leave-one-out cross-validation (LOOCV) on the collected dataset: decision tree, logistic regression, Support Vector Machine (SVM), random forest, and KNN. LOOCV provides a reliable and unbiased estimation of model performance, and it is commonly used with small datasets. In LOOCV, the number of folds is equal to the number of instances in the dataset (Sammut and Webb, 2017). Thus, for each fold, the test set is composed of only one instance, and the training set is composed of all the other instances. Further, an inner 5-fold cross-validation was used on the training set to tune the hyper-parameters using grid search. In particular, for the decision tree *max_depth* was set in $[2, 3, 4]$, for the logistic regression *penalty*

was set in [l1, l2], for the SVM C was set in $\{10^{-3}, 10^{-2}, \ldots, 10^{2}\}$ and γ in $\{10^{-4}, 10^{-3}, \ldots, 10^{1}\}$, for the random forest *max_depth* was set in $[3, 4, 5]$ and *n_estimators* in $[5, 10, 20, 50, 100]$, and finally, for the KNN *n_neighbors* was set in $\{3, \ldots, 12\}$. Malingering detector was trained using only the three validity scales X, Y, Z.

As reported in Table 1, the decision tree achieved an accuracy of 90%, resulting in the best classifier to discriminate between honest and malingerer profiles. Moreover, the decision tree is an easily interpretable classifier, where the importance of a feature is identified from its depth (with the most important feature being at the root) (Hastie et al., 2009). Figure 5 reports the decision tree structure of a single LOO split in the discrimination of honest and malinger profiles on our dataset.

Table 1: Average classifiers performance in discriminate between honest and malingerer profiles on MCM III validity scales.

Model	Test Accuracy (%)
Decision Tree	**90**
Logistic Regression	89
SVM (Kernel RBF)	90
Random Forest	88
KNN	87

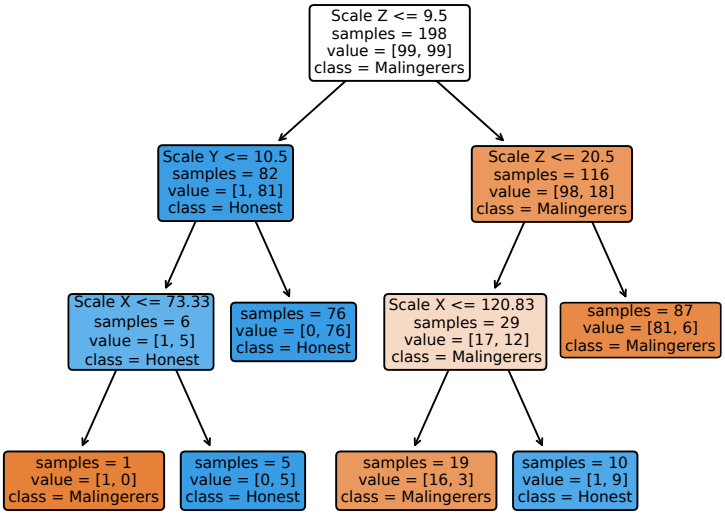


Figure 5: Decision tree structure for a single split of the leave-one-out. Each node of the tree is coloured in blue if it classify a sample as honest or orange if the sample is classified as malingerer.

3.3 Malingering remover

In clinical and forensic evaluations a good accuracy at single subject level is required. This objective is particularly important given that it has been shown, in many datasets, that the number of single cases that behave differently from the trend observed in the group is high (Fisher et al., 2018). As already mentioned, an important but unaddressed issue in malingering research is the reconstruction of the honest test profile on the sole basis of the malingered test profile. To address this issue, we developed a technique that efficiently reconstructs the honest responses of a participant taking, as input, his malingered profile. To deal with this problem, we introduced three malingering removing algorithms: Average removing, Denoising Autoencoder, and Multi-output regressor. We applied the LOOCV procedure in all the reported analysis. One honest and one malingerer test of the same participant were excluded iteratively from the training set. The malingerer trial was used as test while honest as the ground truth. The training set consisted of the remaining 99 malingering trials and the corresponding 99 honest trials. In the following, we describe the three proposed malingering removing techniques.

3.3.1 Average removing

A simple technique removing malingering consists in correcting each malingered profile by subtracting the average score of the honest responses for each scale. To avoid inconsistency, values that were out of their specific scale range, were set to the closer scale bound (i.e., values lower than 0 were set to 0). Consider, for example, how the score of Scale 1 for subject one is corrected with the average removing. Subject one had a score of 11 on Scale 1, and the average of malingered responses of this same scale is 15.6. The average for honest responses is 5.7. The estimated corrected score for subject one is 1.1 (11-(15.6-5.7)). In short, this method assumes that, on each specific scale, malingering has the same effect for all the participants. Moreover, possible correlations between scales are not considered using this method. The average Root Mean Square Error (RMSE) achieved by this trivial technique is 4.05 ± 1.78.

3.3.2 Denoising Autoencoder

Autoencoders (LeCun et al., 1989) are self-supervised learning models based on neural networks. In their most common configuration, autoencoders perform two tasks: compressing input data into a lower dimension (encoder) and use this lower-dimensional representation of the data to recreate the original input (decoder) (see Figure 6. Since our problem is to reconstruct honest profiles from malingered test responses a variant of the classic autoencoders, called denoising autoencoder (DAE) was used. Differently to classic autoencoders, the goal of DAEs is to minimise the difference between the reconstructed output (i.e., reconstructed honest profile) and a specific target (i.e., real honest

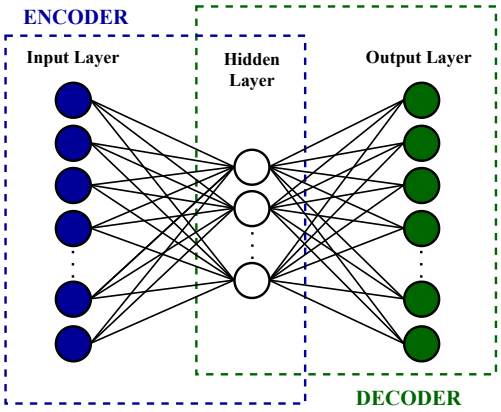


Figure 6: Structure of a Denoising Autoencoder (DAE). For our task, the input layer corresponds to the malingerer profile, while the output layer corresponds to the reconstructed honest profile.

profile). We trained a DAE that takes as input both clinical and validity scales of a malingered profile, and produce as output the corresponding honest scales. All the input data were standardised before training. Moreover, a grid-search with an inner 10-fold cross validation on the training set was used to tune the hyper-parameters of the model. From that, the optimal number of hidden units resulted to be 10. The average RMSE achieved by the proposed DAE is 3.49 ± 1.50.

3.3.3 Multi-output regressor

Similarly to the previously described autoencoder, a multi-output regressor was developed to predict all the honest scales based on the malingered test results of the same participant. While the autoencoder simultaneously produces the estimation of the $24 + 3$ honest results, the multi-output regressor estimates the honest scale scores one-by-one. As reported in Table 2, we tested different regression models using a grid-search with an inner 5-fold cross validation on the training set to tune the hyper-parameters. The best performing model resulted to be a Support Vector Regressor (SVR) with Radial Basis Function (RBF) kernel. A SVR is a generalisation of a SVM for predicting continuous numeric values rather than classes (Awad and Khanna, 2015). The best regression model (SVR with RBF kernel) achieved an average RMSE of 3.27 ± 1.51.

3.4 Reconstruction performance analysis

This work aims to reconstruct a honest profile from a malingered profile. To evaluate the accuracy this reconstruction, we considered several aspects.

Table 2: Performances in leave-one-out and hyper-parameter domain for tested regressor models.

Model	**Parameters**	**Values**	**AVG RMSE on test**
Random Forest	n_estimators max_depth min_samples_leaf	[10, 20, 30] [4, ..., 7] [5, ..., 9]	3.41
Kernel Ridge	α	$[10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}]$	3.33
Ridge	α	$[200, 230, 250, 265, 270, 275, 290, 300, 500]$	3.38
Lasso	α	$[0.02, 0.024, 0.025, 0.026, 0.03]$	3.41
SVR RBF	c γ	$[10^{-3}, 10^{-1}, 1, 10^{2}]$ $[10^{-4}, 10^{-3}, 10^{-1}, 10^{0}, 10^{1}]$	**3.27**
SVR Linear	c	$[10^{-3}, 10^{-1}, 1, 10^{2}]$	3.31
SVR Poly	c γ	$[10^{-3}, 10^{-1}, 10^{0}, 10^{2}]$ $[10^{-4}, 10^{-3}, 10^{-1}, 10^{0}, 10^{1}]$	3.33
KNN Regressor	n_neighbors	$[2, \ldots, 12]$	3.34

Firstly, we assess the performance of our methods comparing the average RMSEs between the reconstructed profiles and the honest profiles. In particular the average RMSEs resulted: 4.05 ± 1.78 for the Average remover, 3.49 ± 1.50 for the DAE, and 3.27 ± 1.51 for the SVR RBF. Machine learning algorithms such the DAE and the SVR, showed an improvement of 14% and 19% respectively, compared to the Average remover. These results confirm what suggested by the statistical analysis (Section 3.1), that highlighted the presence of moderate correlations between honest scales and malingered scales. Table 3 report the results for the Average RMSE for our three malingering removing techniques.

Another method to evaluate the quality of the reconstruction is to perform the malingering detection to the reconstructed honest profile. If a malinger removing technique succeed, the malingering detector should classify the reconstructed profile as honest. We applied this procedure to all the reconstructed profile with average removing, DAE, and SVR. As reported in Table 3, machine learning algorithms outperformed the average remover. In particular, DAE resulted to be the best model, with a 17% improvement compared to the average remover.

In evaluating the MCM III questionnaires, one factor that is commonly considered is the order of the scales when rearranged by increasing value. Based on this consideration, we developed a metric that evaluates the capacity of our malingerer removing algorithms to reproduce the order of the honest profile scales. This metric uses the Top N accuracy defined in (Boyd et al., 2012). Firstly, we normalised the scales on their upper-bound value in MCM III questionnaire. Than, we calculated the Top N accuracy as the percentage of common Top N scales between honest and reconstructed profiles. In Figure 7, we show the Top N accuracy results for the three proposed malingerer removing algorithms. The metric has been calculated for values of

N ranging from 1 to 5. This choice is motivated by the consideration that, among the ordered scales. We also report the Top N accuracy values obtained by directly comparing the honest profiles with the malingerer profiles (without using any sort of malingering remover). The results obtained show that the Top N accuracy calculated on the malingerer profiles are significantly lower than those calculated for the three malingering removing methods. This result confirms that the order of the scales changes significantly between the honest and malingerer tests. Regarding the methods of malingering removing, we can see how SVR and DAE always obtain better performance when compared with the average remover. For values of N up to 3 (which are the most interesting), DAE and SVR perform significantly better compared to average removing, obtaining 14% and 18% improvement respectively. Table 3 report the results for the Top 5 Accuracy.

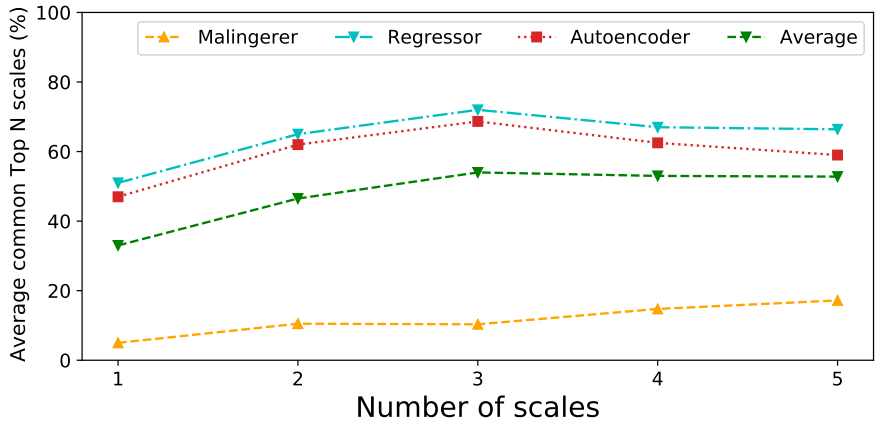


Figure 7: Percentage of average common Top N scales between honest/reconstructed profiles and honest/malingerer profiles. Note that the Top 1 accuracy of the malingered test is near to zero. This indicates that the number of times that the scale with the highest score in the malingered test corresponds to the scale with the highest score in the honest test is almost zero.

In Figure 8, we compared the average performances of the three proposed malingering removing techniques for the 24 clinical scales. The Figure also reports the average values, for the same scales, also for honest and malingerer profiles. From Figure 8, it is clear that both the autoencoder and the regressor techniques reconstruct the scores of the honest profiles, with higher accuracy compare to the trivial average removing method. This result indicates that both the autoencoder and the Regressor seem to capture the specificity of the single subject above and beyond the similarity of the single subject to the overall group.

To better describe the behaviour of our malingering removing algorithms at the single subject level, we report in Figure 9 three examples. The examples correspond to three subjects chosen based on their RMSE: low RMSE (Figure 9a), RMSE close to the RMSE 50 percentile (Figure 9b), and high RMSE (Figure 9c).

Table 3: Performance comparison of malingerer removing techniques using: RMSE between honest and reconstructed profiles, malingering detection accuracy on reconstructed profiles, and average common Top 3 scales accuracy.

Model	**Average RMSE**	**Accuracy (%)**	**Top 3 Scales Accuracy (%)**
Average	4.05	78	54
DAE	3.49	**95**	68
SVR RBF	**3.27**	91	**72**

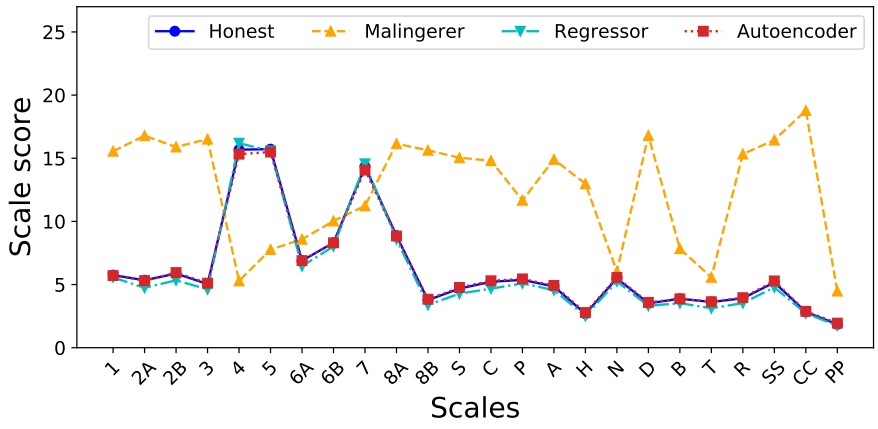
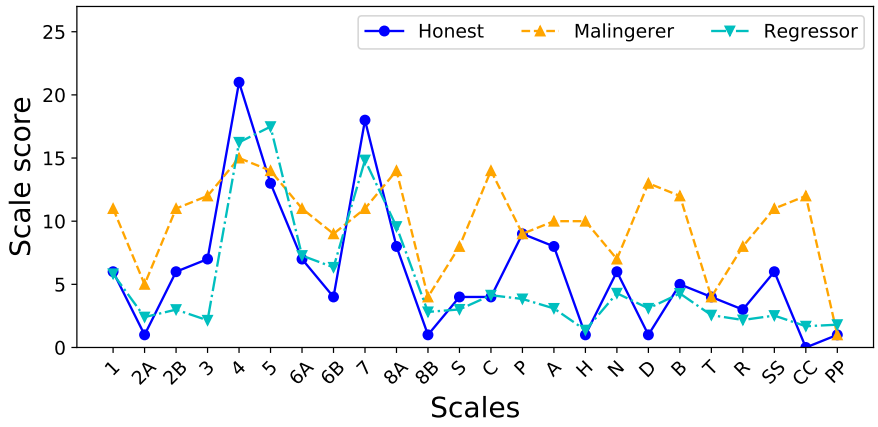


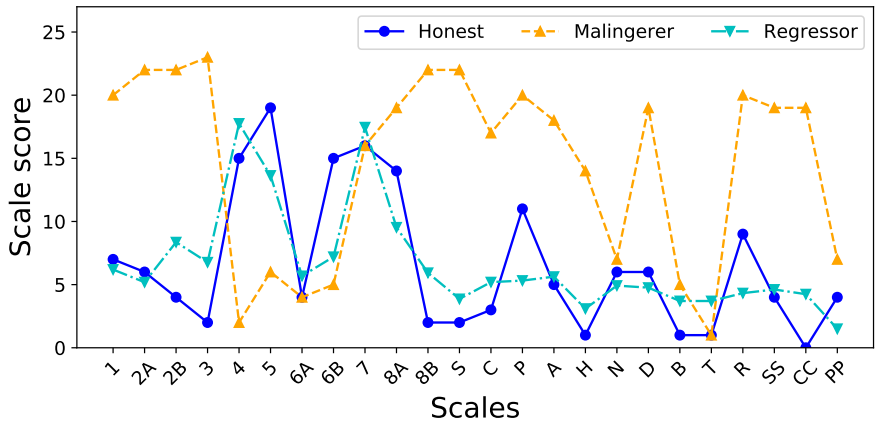
Figure 8: Average values of the 24 clinical scales for honest, malingerer and reconstructed profiles with average removing, DAE and SVR regressor.

4 Discussion

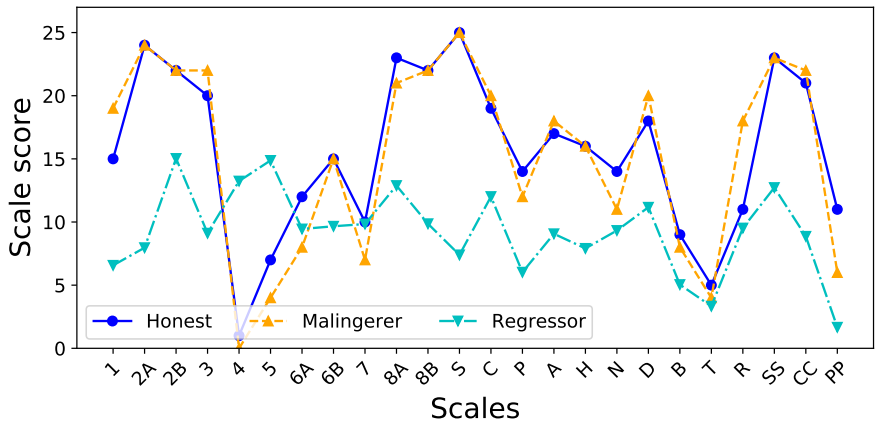
Doctoring responses to direct questions to achieve advantages is ubiquitous in forensic settings. For this reason, to detect malingering, most psychological tests based on the endorsement of symptoms-describing sentences, are complemented with validity scales which may help in identifying a malingering response style.



(a) *Low RMSE (2.17). The reconstructed profile well approximate the honest profile.*



(b) *50 percentile RMSE (2.93). The approximation error increase, but the performance are still good.*



(c) *High RMSE (7.99). The quality of the reconstruction decrease, but the similarity between honest and malingerer profiles suggest that the subject did not follow the instructions.*

Figure 9: Comparisons between honest, malingerer, and reconstructed profiles for three meaningful cases.

While validity scales may succeed in identifying the malingerer, what is currently unaddressed is the reconstruction of honest responses when only the malingered test is available (this is the typical condition in a forensic setting). In other words, malingering detection, as currently achieved using the validity scales (for example, in the MMPI and MCM III), can solve only in part the problem of diagnosis in forensic settings. We are primarily interested in the honest psychological profile depurated from malingering and not only in establishing the existence of malingering. To the best of our knowledge, no proposals have been previously attempted to solve this problem, and this lack of research motivated our effort to develop a technique that removes malingering from psychological tests. We have presented here a proof-of-concept using the Millon, a widely used test for investigating psychopathology that is complemented with validity scales that are used for detecting malingering.

Hundred healthy participants were required to complete the Millon both with standard honest instructions and faking-bad instructions to appear depressed in a hypothetical insurance claim for personal damages.

The models presented here were based on ML and were taking as input the raw scores of all the 24 MCM III clinical scales complemented with the validity scales X, Y, Z. ML models were trained to output the corresponding scores collected during honest responding. We developed two different ML models, one based on an DAE and another using multi-output regressors.

As regards to the descriptive statistics, the main findings were the following:

- The malingered profile was systematically higher with respect to the honest results almost on all scales;
- Malingering is not only confined to depression-related scales (e.g., CC, D), but also extends to other scales (e.g., P, C);
- After malingering the order of the original scales is altered, and very few if any subject has the scale with the highest score when malingering, which corresponds with that observed when responding honestly. In other words, the scale with the highest honest score is never also the scale with the highest score after malingering.

As regards to the ML models used for malingering removal, to avoid overfitting and achieve generalisable results (also with small dataset), all the models were developed using the leave-one-out cross-validation procedure.

The main results regarding the malingering removal procedures were:

- The malingered profile, after malingering removal, is identified by a classifier as an honest profile with high accuracy (95% for DAE and 91% for SVM regressor);
- ML models were very good at group level in removing malingering and approximating the honest test results;
- In predicting individual responses, ML-based models were better than a simple correction strategy (average removing) consisting in subtracting to the subjects scale score the average difference between the group score in

the honest and malingerer condition. In short, ML models succeeded in personalising the process of malingering removal;
- The multi-output regressor (SVM regressor) yielded slightly better predictions with respect to the DAE;
- Both ML models successfully maintained, after malingering removal, most of the original order of the honest responding. In other words, the ranking of the scale in the honest condition was mostly maintained after malingering removal. The highest three scales were identified correctly 72% of the times.

It is relevant to stress that the malingering removers proposed here permit individualised modulation of the prediction. This is relevant to the current debate about the lack of group-to-individual generalisability that has been shown to undermine the validity of scientific research in many fields (Fisher et al., 2018). It has been shown that the credence that an effect at group level generalises at single-subject level is greatly unfounded given that *"Only 68% of all individual correlational values fall within a range that would be predicted by group data to cover 99.7% of all possible correlations–a discrepancy of nearly 32%."* In short, the ML models used for malingering removal have shown not only extremely good reconstruction accuracy at group level but also good reconstruction accuracy at individual level. Given the correction of the ML models, SMV regressor reduces the error by 19% with respect to the correction using the Average remover (that is correcting all the subjects with the same procedure), we can say that this strategy gives the desired individualised predictions. It is worth noting that the personalisation of results is not a trivial task given that different subjects may fake with different levels of intensity and on different symptoms for a variety of reasons.

A qualitative analysis conducted on single case profiles indicated that the few subjects that have poor reconstruction results are those that failed to follow the instructions and had a faked Millon profile which overlapped the honest profile. The proof-of-concept reported here shows that removing malingering from psychological tests may be achieved using ML models. ML models are entering the psychometric field and may now be considered as part of the psychometric toolbox (Orrù et al., 2020a,c). Frontiers in Medicine, 6, 319. Once this avenue of research has been established, to develop a fully functional model of malingering removal, further steps are required and specifically at least the following:

- Showing that malingering removal is possible also for pathological cases that aggravate rather than blatantly fake their profile. In short, psychopathological cases that over-report their symptomatology;
- Increasing the number of cases on which the models are developed;
- Evaluating the malingering remover in other conditions of malingering. Here participants faked a non existent depression, they could also fake other psychopathological conditions such as anxiety, post-traumatic stress disorder. Malingered responses change given different malingering objectives.

Open practices statement The study was not preregistered; the full dataset and code are available at `https://spritz.math.unipd.it/datasets/malingerer_removal/MalingererRemoval.zip`.

References

Association AP, et al. (2013) Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub

Awad M, Khanna R (2015) Support vector regression. efficient learning machines. Apress, Berkeley, CA

Boyd S, Cortes C, Mohri M, Radovanovic A (2012) Accuracy at the top. In: Advances in neural information processing systems, pp 953–961

Cliff N (1996) Answering ordinal questions with ordinal data using ordinal statistics. Multivariate Behavioral Research 31(3):331–350

Daubert SD, Metzler AE (2000) The detection of fake-bad and fake-good responding on the millon clinical multiaxial inventory iii. Psychological Assessment 12(4):418

Fisher AJ, Medaglia JD, Jeronimus BF (2018) Lack of group-to-individual generalizability is a threat to human subjects research. Proceedings of the National Academy of Sciences 115(27):E6106–E6115

Grissom RJ, Kim JJ (2005) Effect sizes for research: A broad practical approach, 2nd edn. Lawrence Earlbaum Associates

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media

van Impelen A, Merckelbach H, Jelicic M, Merten T (2014) The structured inventory of malingered symptomatology (sims): A systematic review and meta-analysis. The Clinical Neuropsychologist 28(8):1336–1365

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. Neural computation 1(4):541–551

Massey Jr FJ (1951) The kolmogorov-smirnov test for goodness of fit. Journal of the American statistical Association 46(253):68–78

Mazza C, Orrù G, Burla F, Monaro M, Ferracuti S, Colasanti M, Roma P (2019) Indicators to distinguish symptom accentuators from symptom producers in individuals with a diagnosed adjustment disorder: A pilot study on inconsistency subtypes using sims and mmpi-2-rf. PloS one 14(12):e0227113

Merten T, Merckelbach H (2013) Symptom validity testing in somatoform and dissociative disorders: A critical review. Psychological Injury and Law 6(2):122–137

Millon T, Davis RD (1997) The mcmi–iii: present and future directions. Journal of personality assessment 68(1):69–85

Organization WH, et al. (1992) The icd-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire 67(30):227–227

Orrù G, Gemignani A, Ciacchini R, Bazzichi L, Conversano C (2020a) Machine learning increases diagnosticity in psychometric evaluation of alexithymia in fibromyalgia. Frontiers in Medicine 6:319

Orrù G, Mazza C, Monaro M, Ferracuti S, Sartori G, Roma P (2020b) The development of a short version of the sims using machine learning to detect feigning in forensic assessment. Psychological Injury and Law pp 1–12

Orrù G, Monaro M, Conversano C, Gemignani A, Sartori G (2020c) Machine learning in psychometrics and psychological research. Frontiers in Psychology 10:2970

Rogers R, Bender S (2018) Clinical assessment of malingering and deception . new york, ny

Rosen GM, Phillips WR (2004) A cautionary lesson from simulated patients. Journal of the American Academy of Psychiatry and the Law Online 32(2):132–133

Sammut C, Webb GI (2017) Encyclopedia of machine learning and data mining. Springer

Sartori G, Zangrossi A, Orrù G, Monaro M (2017) Detection of malingering in psychic damage ascertainment. In: P5 medicine and justice, Springer, pp 330–341

Tracy DK, Rix KJ (2017) Malingering mental disorders: clinical assessment. BJPsych Advances 23(1):27–35

Vrij A (2000) Detecting lies and deceit: The psychology of lying and implications for professional practice. Wiley

Woolson R (2007) Wilcoxon signed-rank test. Wiley encyclopedia of clinical trials pp 1–3

Young G (2014) Malingering, feigning, and response bias in psychiatric/psychological injury. International Library of Ethics, Law, and the New Medicine 56:817–856

Zennaro A, Ferracuti S, Lang M, Sanavio E (2008) Millon clinical multiaxial inventory-iii [mcmi-iii italian adaptation]. Firenze: Giunti OS