Applying explanation methods for the iterative refinement of an ANN-based depression screening tool

Cristian E. Sosa-Espadas[1,2], Manuel Cetina-Aguilar[1,2], Jose A. Soladrero[1,2], Jesus M. Darias[3], Esteban E. Brito-Borges[1,2], Nora L. Cuevas-Cuevas[1,2] and Mauricio G. Orozco-del-Castillo[1,2,*]

[1] *Tecnológico Nacional de México/IT de Mérida, Department of Systems and Computing, Merida, Mexico*

[2] *AAAIMX Student Chapter at Yucatan, Mexico (AAAIMX), Association for the Advancement of Artificial Intelligence, Mexico*

[3] *Department of Software Engineering and Artificial Intelligence, Instituto de Tecnologías del Conocimiento, Universidad Complutense de Madrid, Spain*

Abstract

Depression, as one of many mental health disorders, is a serious health and economy problem, affecting over 300 million people of all ages. The diagnosis of depression is a very complex and time-consuming task for mental health professionals, who usually rely on self-report questionnaires as a screening process. However, the items used in these questionnaires are sometimes subjective, particularly to certain demographics, and could require much time and effort from the patient. In recent years, artificial intelligence techniques, such as artificial neural networks, have been commonly used to screen for depression, however, they operate as black-box models, i.e., they lack explainability and interpretability which are mandatory in health-related fields. In this work we propose not only an artificial neural network, but also a set of explainable artificial intelligence techniques to refine a large set of items from a psychological questionnaire into a more concise, explainable one.

Keywords

Explainable Artificial Intelligence, Depression, Mental Health, Artificial Neural Networks

1. Introduction

The rise in mental health conditions in recent years has become a serious health and economy problem, accounting for over a trillion USD each year [1]. Particularly, depression has become one of the most common mental health conditions in the modern era, affecting over 300 million people of all ages [2], and the number of patients and associated medical costs keep increasing [3]. While most of the patients seeking medical treatment for depression are in their 50s or

ICCBR XCBR'22: 4th Workshop on XCBR: Case-based Reasoning for the Explanation of Intelligent Systems at ICCBR-2022, September, 2022, Nancy, France

[*] Corresponding author.

mauricio.orozco@itmerida.edu.mx (M. G. Orozco-del-Castillo)

0000-0002-0933-5152 (C. E. Sosa-Espadas); 0000-0002-2423-0711 (M. Cetina-Aguilar); 0000-0001-5864-275X (J. A. Soladrero); 0000-0002-2173-1422 (J. M. Darias); 0000-0002-4257-944X (E. E. Brito-Borges); 0000-0002-5477-6176 (N. L. Cuevas-Cuevas); 0000-0001-5793-6449 (M. G. Orozco-del-Castillo)

60s [3], these issues develop much earlier. It is estimated that over 20% of the world's children and adolescents suffer from a mental condition [1], including 8% of young adults between the ages of 18 and 22 [2], which can lead to suicide, the second leading cause of death among adolescents and young adults (15-29) [1]. The fact that very few people in this age group seek medical attention is attributed to an unawareness of their illness, and only seek intervention until their symptoms become severe [2]. This makes research focused on developing more effective detection techniques a mandatory task; if found early, depression has a high cure rate [3].

The goal of this research is the development and refinement of a screening tool that allows the early identification of students with high risk of depression, probably due to the COVID-19 pandemic. This tool is based on an Artificial Neural Network (ANN) applied over a dataset collected from 157 users that completed a questionnaire with 102 items related to to the symptoms of Major Depression Disorder (MDD), Generalized Anxiety Disorder (GAD) and Antisocial Personality Disorder (APD). The choice of a ANN model is rooted on the very high performance demonstrated by this technique for classification tasks.

For the current work, however, we want not only to create the ANN model, but to understand the impact of each item, so we can refine the questionnaire and achieve a higher depression screening performance. Here, the ANN models have a major drawback regarding its explainability, as they are considered "black-boxes" that are not interpretable. Therefore, in order to improve our screening tool we propose the application of several explanation methods along an iterative refinement process that mimics the general structure of a case-based reasoning (CBR) system.

Paper runs as follows. Section 2 presents the background of this work. Section 3 describes the dataset and the ANN screening model. Then, Section 4 describes our iterative refinement process using explanation methods. Section 5 presents the evaluation results and section 6 concludes the paper and opens lines of future work.

2. Background

There is a wide range of self-report questionnaires and inventories to assess different mental illnesses and emotional states, such as the Beck Depression Inventory for depression [4], the Attention Deficit and Hyperactivity Disorder (ADHD) Screening Questionnaire (ADHD-SQ) [5], the Aspiration Index [6] for the measurement of intrinsic and extrinsic aspirations, the Adult ADHD Self-Report Scale, the Borderline Personality Questionnaire [7], the Common Beliefs Survey III-Short Form [8], depression, anxiety and antisocial items [9] designed from the Diagnostic Statistical Manual-V (DSM-V), the Generalized Anxiety Disorder 7 (GAD-7) [10], the reduced scale of the Morningness-Eveningness Questionnaire [11], the Rosenberg Self-Esteem Scale [12], the State-Trait Anxiety Inventory [13, 14] the Five Facet Mindfulness Questionnaire [15], etc. Nevertheless, these self-reports usually need to collect a lot of information in the form of a large number of items to be addressed [16], which usually lowers the quality of the answers as the user progresses throughout the questionnaire [17].

On the other hand, ANN models have demonstrated their extraordinary performance for classification tasks, being a very suitable tool for the screening of mental illnesses [9]. However, they are black-box methods that do not allow to understand the internal data patterns that led to classification. This way, eXplainable Artificial Intelligence (XAI) proposes several methods to understand such models [18].

An XAI system must be able to explain what it has done, what is happening, and whats is going to happen [18]. In order to do this, each explanation of this kind of systems depends on the tasks, skill, and expectations that the user has, and the models to be used in this systems must be transparent and accessible for their decisions and recommendations [18].

3. Dataset and Classification Model

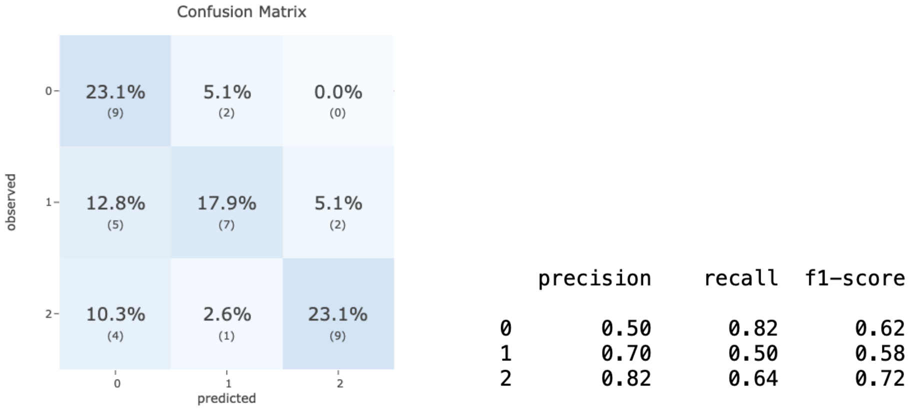
The dataset used on this project corresponds to the one described in [9]. It consists of the answers from 157 users who answered a screening questionnaire of 102 items related to the symptoms of MDD, GAD and APD as reported by the DSM-V. The user demographic consisted of students between 18 and 23 years old, all of them enrolled at the Tecnológico Nacional de México (TecNM)/Instituto Tecnológico de Mérida (ITM), at Mérida, Yucatán, Mexico. Questions have a binary nature, where the users answered either "true" or "false" with respect to their agreement with certain statements. Additionally, there is a "expected answers datasheet" with the values that are supposed to reflect symptoms of mental conditions. This validation artifact was created by the psychologist who designed the questionnaire and contains the expected answers that a person with either MDD, GAD, or APD would most probably choose.

An additional independent variable corresponding to physical symptoms of MDD, GAD, and APD was measured through 15 complementary items. Then, a "potential depression score" (PDS) was calculated by averaging these answers and used to estimate the actual depression risk for each individual. This hypothesis of correlation between the initial 102 items and the PDS value was confirmed by our previous results. This way, an ANN can be trained on this dataset, where PDS is used as the target value for training and prediction.

The PDS variable is codified as a discrete value (ranging from 1 to 5). Initial attempts to estimate this value using a prediction ANN model did not achieve acceptable results. Therefore, the ANN was redesigned as a classification model with five classes. After the initial statistical analysis of the PDS values, a clear imbalance was found, with the highest concentration of users having scores around 3, and the lowest concentrations being located near 1 or 5. Due to this particularity, we decided to combine the five values into three new classes (1,2: Low; 3: Medium; and 4,5: High). This recodification of the target variable also presented a little imbalance, therefore we performed an upsampling process using the SMOTE technique [19] to further diminish this problem.

As reported in [9], a neural classifier was trained on the dataset so relationships between physical symptoms of MDD and the remaining 102 items were found. The ANN follows a multilayer perceptron (MLP) architecture, with paramethers α and *learning-rate* optimized to achieve the highest accuracy. On the test dataset ($1/3$ of the total items), the model presented a global accuracy of 64%. Additional performance results per target class can be appreciated in Figure 1 (right).

However, at this point it is critical to understand the target performance metric. Figure 1 shows the corresponding confusion matrix where we can observe the misclassifications returned by the



```
     precision    recall  f1-score

0         0.50      0.82      0.62
1         0.70      0.50      0.58
2         0.82      0.64      0.72
```

Figure 1: Confusion matrix (left) and performance metrics (right) for the original ANN model. Axis values 0, 1, and 2 represent classes “low”, “medium” and “high” respectively.

ANN. Having in mind that the goal is to achieve the highest depression screening performance, there is a clear error to minimize: false negative ratio or “recall” for class “High PDS”, i.e., those individuals that having an actual high risk of depression were classified by the ANN as low or medium risk, represented in Figure 1 at the lower row of the matrix. This false negative ratio represents individuals with a high PDS that will not be identified by the screening tool, and therefore would not receive further psychological support. Additionally, we can appreciate that precision for class “high” is very good (right column), meaning that when the ANN identifies an individual as high PDS, it is very confident. The other way around –individuals without actual risk of depression being classified as high PDS– does not imply a high risk in terms of their mental health. These individuals will be identified (incorrectly) by the screening tool and receive a deeper evaluation by the psychologists, who will ultimately discard any risk.

4. Iterative Refinement Method

The proposed method to refine the ANN-based depression screening tool is based on the identification of potentially anomalous items through explanation methods that provide insights of the prediction model. This process follows a cycle analogous to CBR, where a set of candidate items are identified (retrieved) and proposed for its revision by an expert user. Finally, once anomalous items are confirmed, they are used to modify the prediction model. This way, we propose an iterative process consisting on the following steps:

1. First, the classification model is introspected with explanation methods to identify those items that may lead to a lower performance (recall for class “high”) according to the “expected answers datasheet”.
2. Next, these answers are proposed to the expert (psychologist) as potential anomalous items. Here we will distinguish between two different anomalous items:

   Inconsistent items that cause an opposite effect to the expected classification (i.e., answers that are expected to predict high PDS but according to the model do decrease the probability to be classified that way).

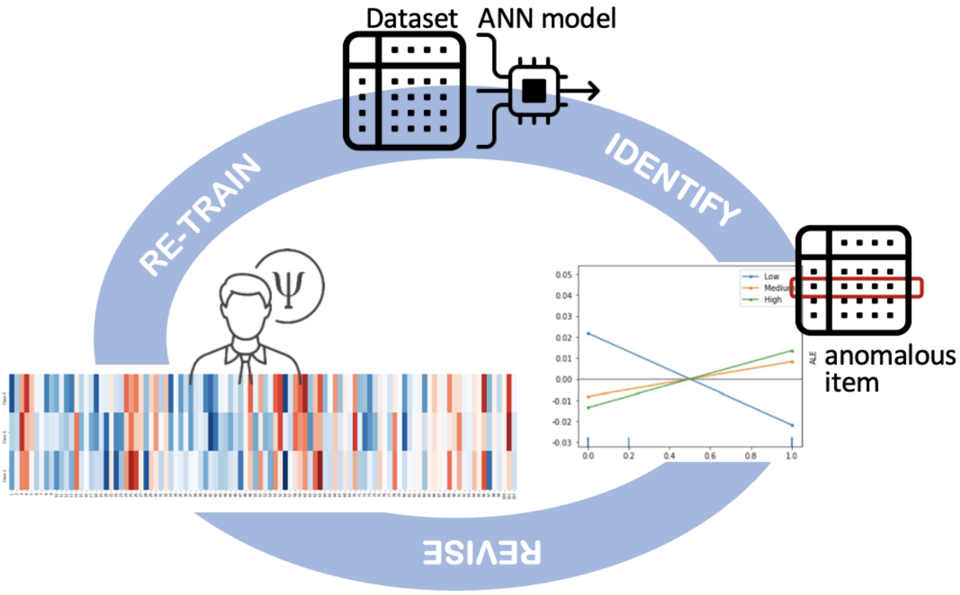


Figure 2: Global schema of the proposed iterative refinement method through XAI methods.

Irrelevant items, on the other hand, are not significant according to the explanation method and do not have any remarkable effect in the classification.

3. The expert will confirm or discard the abnormal behaviour of these items through a deeper analysis supported by additional introspective XAI methods.
4. Those items confirmed by the expert are removed from the dataset, and a new model is trained. Continuing to the next iteration from step 1.

This process is illustrated in Figure 2 and is detailed as follows.

The first step for each iteration is the identification of the anomalous items of the questionnaire. Although there are several explanation methods to perform this task, we have chosen Accumulated Local Effects (ALE). ALE describes how items influence the results of the ANN, based on the differences in the predictions. When applied to our model, ALE returns a plot for each questionnaire item. Each plot contains three lines, representing –on average– the influence of that item for the classification as High, Medium, and Low classes. The x-axis represents the possible answers (0-false, and 1-true) and the y-axis corresponds to the influence of the probability to be classified as each class. Figure 3 shows two of these plots, where we can observe, for example, that answering true (1 in x-axis) to questionnaire item 1 decreases significantly the possibility to be classified as low PDS, whereas it increases (with less influence) the classification as medium or high PDS. If we read the question *"Most of the time I have difficulty concentrating on simple task"*, we can intuitively confirm that the ANN model is working properly, since the "expected answers datasheet" defines that answering "true" is an indicator of higher PDS. If we look at the ALE plot on the right, we can observe that the slopes of the lines is not high enough to consider that item 2 has a clear impact on the classifications.

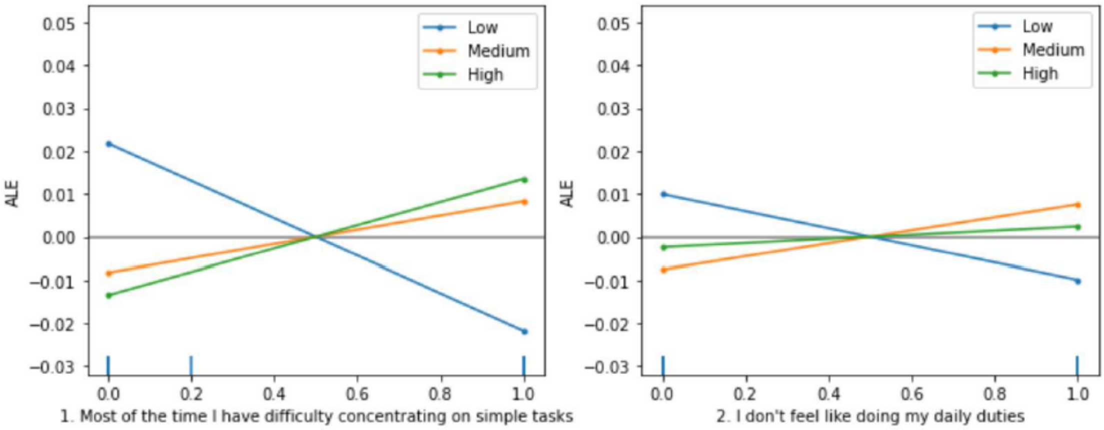


Figure 3: Example of non anomalous questions.

This way, the slope of each line in the ALE plot is significant in order to determine if an item I has an impact on the classification. So, the slope (m) for each class is calculated obtaining the ($\overrightarrow{ALE_I}$) slope vector for each class.

$$\overrightarrow{ALE_I} = \langle m_H, m_M, m_L \rangle. \tag{1}$$

Then, we can define thresholds s^+ and s^- to identify the minimum positive slope and maximum negative slope to consider that an item has impact on the classification of any of the considered classes.

Once defined these thresholds, we proceed to the identification of anomalous questions. To do so, we only need to identify items with influence ($s > s^+$ or $s < s^-$) that are opposite to the values defined int "expected answers datasheet".

To understand this process we can use the examples presented in Figure 4. The expected answer to item 3 *"My friends have told me that I look different"* to diagnose high PDS according to the psychologist is 1. However, the ALE plot (on the left) shows that the behavior of the ANN model is the other way around: answer 1 contributes to be classified as medium PDS but decreases the possibility of belonging to classes medium and high, that is a very anomalous behaviour. The ALE plot on the right side also presents an opposite behaviour where a positive answer increases possibility of classification as low but decreases medium and high.

This way, in order to detect anomalous questions, it is essential to analyze the vectors ($\overrightarrow{ALE_I}$) that contain the slopes of the items. Then we have defined several anomalies in the questionnaire items according to the following conditions:

- *Inconsistent items for class "High"*: Items that, with their theoretical expected answers, promoted false High PDS predictions or false non-High PDS predictions, respectively.
  - Items with $m_H > s^+$ and expected answer False, OR
  - Items with $m_H < s^-$ and with expected True.
- *Inconsistent items for class "Low"*: Items that, with their theoretical expected answers, promoted false Low PDS predictions or false non-Low PDS predictions, respectively.

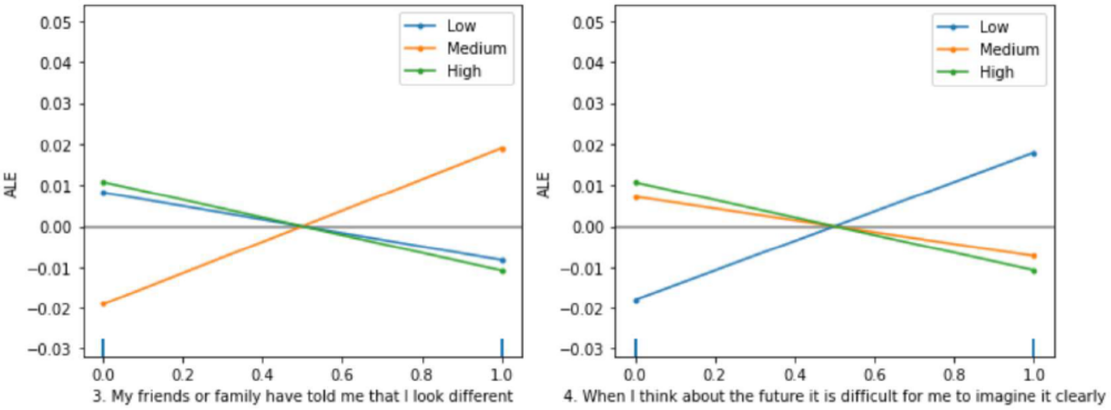


Figure 4: Example of two anomalous questions where the High PDS is negative and the expected answer being positive.

  - Items with $m_L > s^+$ and with expected answer True, OR
  - Items with $m_L > s^-$ and with expected answer False.
- Irrelevant Items: Items that do not make any difference to discern between PDS classes.
  - Items with either high or Low PDS importance values being in the range $[s^-, s^+]$.

We need to clarify that questions can be anomalous for the High PDS class, but that does not automatically makes them anomalous for the Low class. Therefore, items can be anomalous either for one class or both.

At this point, it is very important to note that the identification of anomalous items is based on the comparison to the expected answers according to the expert. However, we should consider that these expectations may be wrong due to many factors: population features, individual disorders, etc. This way, the items identified in the previous step should be considered as "potentially anomalous items" until the expert confirms that the behavior of the ANN model is correct or not. This revision process (similar to the one performed by the CBR cycle) can be really complex for users that may not have any kind of skill on machine learning. Therefore, we propose the use of additional introspective XAI methods to support this task. Concretely, we provide support to this decision making process through several explanation methods such as LIME and SHAP.

The purpose of *SHAP* [20] is to explain the prediction of a particular instance by computing the contribution of each feature to the prediction. These contributions are computed using Shapley values, which are obtained by arranging the features in different coalitions and calculating the average marginal contribution of a feature. In simpler words, SHAP is an indicator of the contribution of a feature to the outputted probability. Local interpretable model-agnostic explanations (*LIME*)[21] is a tool that provides explanations for individual instances relying on local surrogates. A local surrogate is an interpretable model that intends to replicate the behavior of the black-box model for instances similar to the one being analyzed. The explanation gives us the contributions of each feature according to the created surrogate. In our case, both methods are able to measure the impact of each item in the accuracy of the ANN. Therefore, we

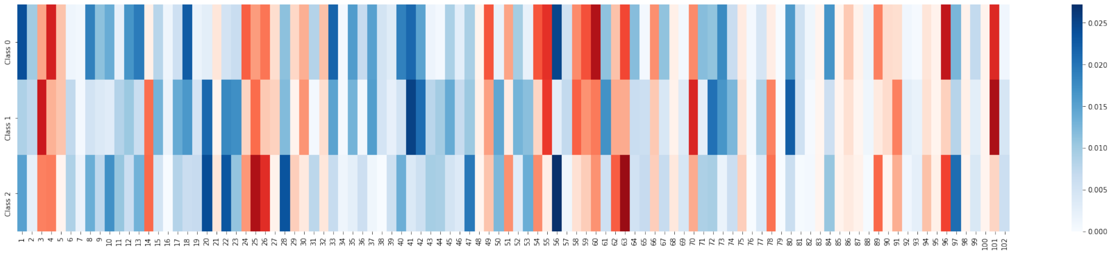


Figure 5: Heat-map reflecting the impact of every item in the model according to SHAP. Red color scale is used for inconsistent items.

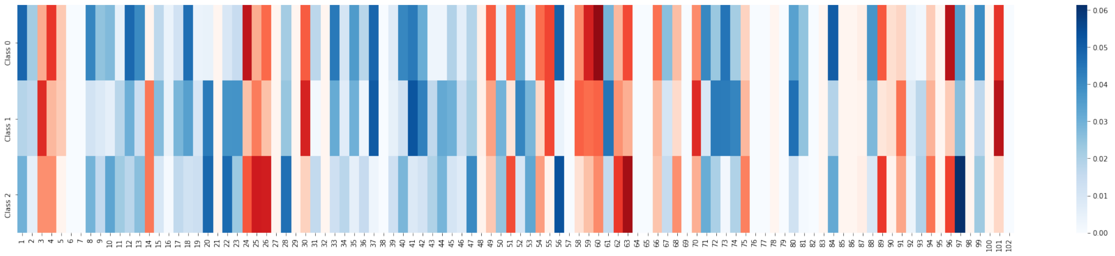


Figure 6: Heat-map reflecting the impact of every item in the model according to LIME. Red color scale is used for inconsistent items.

can plot this effect using a heat-map, as presented in Figures 5 and Figure 6 where we use a blue-color scale to indicate the impact of items behaving correctly according to the "expected answers datasheet" and a red-color scale for inconsistent items. Consequently, irrelevant items appear with a color close to white. As we can see, both XAI methods reflect similar anomalies in the classification model. These heat-maps can be used by the psychologist to decide whether an item is actually an anomaly of the ANN model or it describes a certain pattern that makes sense.

With anomalous items detected and confirmed by the expert, our underlying hypothesis is that it is possible to filter these questions from the dataset in order to achieve a higher prediction performance. Specifically, curating the questionnaire removing High PDS anomalous questions should help the classifier model to more easily identify High PDS instances. This can also degrade the prediction performance of other classes, but in the use case of depression screening, we already mentioned that it is more important to identify those individuals with a high level of depression, even at the cost of having false negative predictions for the Medium PDS and Low PDS classes.

5. Evaluation

Our questionnaire refinement process is evaluated in this section. The removal of items is performed in an iterative fashion over the dataset, creating multiple curated versions of our model. On each iteration, the detected High PDS anomalous questions are removed. Here we are assuming that the psychologist will confirm the inconsistent impact of every potentially

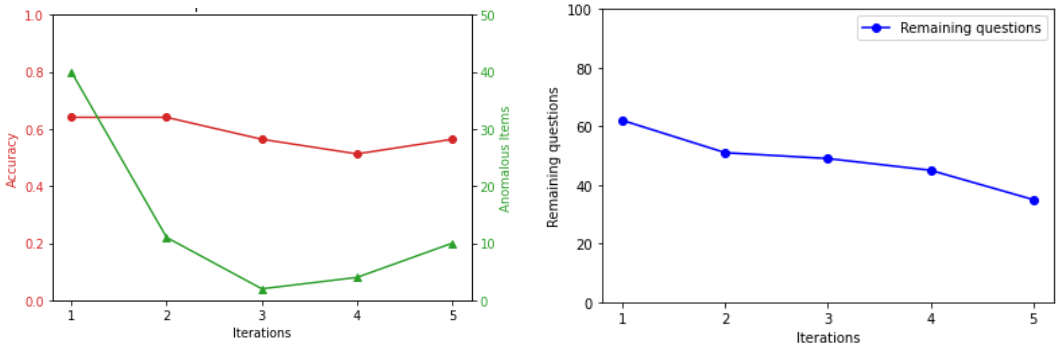


Figure 7: (Left)Global accuracy and number of anomalous questions identified on each iteration. (Right) Number of questions used as inputs on each iteration.

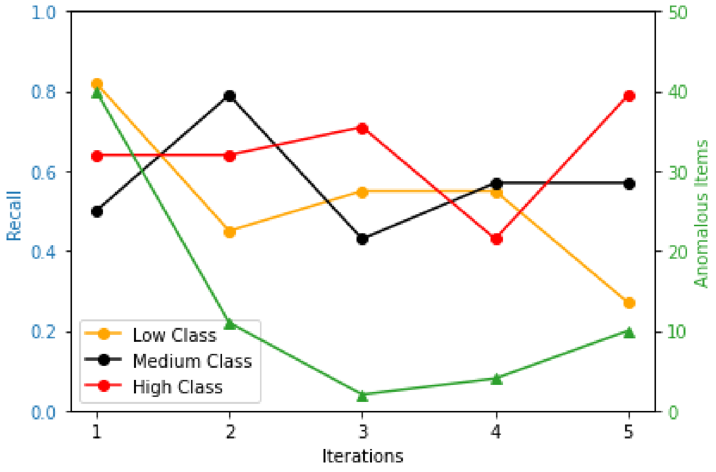


Figure 8: Recall results per each target class and number of anomalous items on every iteration.

anomalous item. Then the ANN is retrained over the remaining questionnaire items to verify the prediction performance of the classifier. Global accuracy performance on each iteration is displayed in Figure 7 (left). As this figure reflects, global accuracy ranges between .5 to .6, reflecting that our process does not have a significant impact on the global performance the model.

Also, the number of inputs used to train the classification model on each iteration is presented in Figure 7 (right). From 102 original items, our refinement process reduced that number down to approximately 40 questions.

Since the main objective of removing questions was to improve the identification of people in the High PDS class, it is also useful to analyse the recall metrics . Corresponding results are presented in Figure 8.

Although recall decreases for the Medium PDS and Low PDS classes, the High PDS presents an important raise, which indicates that the ANN has achieved a better ability to detect individuals with a high presence of depression symptoms, which was the goal of our research.

6. Conclusions

This paper introduces an iterative refinement method that uses XAI techniques, such as ALE, SHAP and LIME, in order to remove irrelevant or anomalous features that could have a negative impact on an ANN-based depression screening tool. Concretely, we have defined a questionnaire refinement process that, in an iterative fashion, uses the ALE technique in order to identify anomalous questions and remove them, producing a curated set of items that enhances the prediction metrics for our target classes, namely, those representing individuals with a high risk of depression.

The main conclusion of this work is that it is possible to use XAI methods not only for the explanation of a black-box model, but also for its improvement and refinement. To do so, we propose an iterative cycle following the global cycle of CBR, where potentially anomalous items are retrieved using the ALE method, and later revised by the clinician using other XAI methods such as LIME or SHAP. The last step consists on retaining this changes into the model to begin a new iteration.

As future work, our paper can be enhanced in multiples ways. First and foremost, a larger dataset of users' answers is always desirable for data-oriented tasks, such as training classification models. This grants more confidence on the results obtained during the evaluation process, as well as stability, due to bias of any type being diluted as the number of instances grow.

In a similar way, another suggestion to increase the reach of the project is to expand the user demographic on which the questionnaire was applied. Students from other universities can manifest different subsets of depression symptoms, and extending the population stratification the data is being collected from, a better understanding on how to perform depression screening can be obtained.

Also, on a more technical level, different feature importance techniques can be applied to measure the relevance each questionnaire item has in the depression screening process. ALE provides one approach, but we could also use LIME or SHAP in this step instead of applying them for the revision.

Finally, removing not only High PS anomalous questions, but anomalous items with Low PS anomalies or even the irrelevant items is our immediate future work to improve the performance of our system.

Acknowledgments

This research is a result of the Horizon 2020 Future and Emerging Technologies (FET) programme of the European Union through the iSee project (CHIST-ERA-19-XAI-008, PCI2020-120720-2) funded by MCIN/AEI/10.13039/501100011033 and European Union "NextGenerationEU"/PRTR. This work has been also partially supported by the PERXAI project PID2020-114596RB-C21 funded by MCIN/AEI/10.13039/501100011033 and the BOSCH-UCM Honorary Chair on Artificial Intelligence applied to Internet of Things of the Universidad Complutense de Madrid. It is also part of projects 10428.21-P and 13933.22-P of the Tecnológico Nacional de México/IT de Mérida.

References

[1] K. Martinez, M. I. Menéndez-Menéndez, A. Bustillo, Awareness, Prevention, Detection, and Therapy Applications for Depression and Anxiety in Serious Games for Children and Adolescents: Systematic Review, JMIR Serious Games 9 (2021) 1–19. doi:10.2196/30482.

[2] M. Y. Wu, C. Y. Shen, E. T. Wang, A. L. Chen, A deep architecture for depression detection using posting, behavior, and living environment data, Journal of Intelligent Information Systems 54 (2020) 225–244. doi:10.1007/s10844-018-0533-4.

[3] J.-W. Baek, K. Chung, Context Deep Neural Network Model for Predicting Depression Risk Using Multiple Regression, IEEE Access 8 (2020) 18171–18181. doi:10.1109/access.2020.2968393.

[4] A. T. Beck, N. Epstein, G. Brown, R. A. Steer, An Inventory for Measuring Clinical Anxiety: Psychometric Properties, Journal of Consulting and Clinical Psychology 56 (1988) 893–897. URL: /record/1989-10559-001. doi:10.1037/0022-006X.56.6.893.

[5] I. Manor, N. Vurembrandt, S. Rozen, D. Gevah, A. Weizman, G. Zalsman, Low self-awareness of ADHD in adults using a self-report screening questionnaire, European Psychiatry 27 (2012) 314–320. URL: http://dx.doi.org/10.1016/j.eurpsy.2010.08.013. doi:10.1016/j.eurpsy.2010.08.013.

[6] T. Kasser, R. M. Ryan, A dark side of the American dream: Correlates of financial success as a central life aspiration., Journal of Personality and Social Psychology 65 (1993) 410–422. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.65.2.410. doi:10.1037/0022-3514.65.2.410.

[7] A. M. Poreh, D. Rawlings, G. Claridge, J. L. Freeman, C. Faulkner, C. Shelton, The BPQ: A scale for the assessment of borderline personality based on DSM-IV criteria, Journal of Personality Disorders 20 (2006) 247–260. doi:10.1521/pedi.2006.20.3.247.

[8] G. L. Thorpe, R. B. Frey, A short form of the common beliefs survey III, Journal of Rational - Emotive and Cognitive - Behavior Therapy 14 (1996) 193–198. doi:10.1007/BF02238270.

[9] M. G. Orozco-del Castillo, E. C. Orozco-del Castillo, E. Brito-Borges, C. Bermejo-Sabbagh, N. Cuevas-Cuevas, An Artificial Neural Network for Depression Screening and Questionnaire Refinement in Undergraduate Students, in: M. F. Mata-Rivera, R. Zagal-Flores (Eds.), Telematics and Computing. WITCOM 2021. Communications in Computer and Information Science, volume 2, Springer Nature Switzerland AG 2021, 2021, pp. 1–13. URL: https://link.springer.com/10.1007/978-3-030-89586-0_1. doi:10.1007/978-3-030-89586-0_1.

[10] R. L. Spitzer, K. Kroenke, J. B. Williams, B. Löwe, A brief measure for assessing generalized anxiety disorder: The GAD-7, Archives of Internal Medicine 166 (2006) 1092–1097. URL: https://jamanetwork.com/. doi:10.1001/archinte.166.10.1092.

[11] A. Adan, H. Almirall, Horne & Östberg morningness-eveningness questionnaire: A reduced scale, Personality and Individual Differences 12 (1991) 241–253. doi:10.1016/0191-8869(91)90110-W.

[12] M. Rosenberg, Society and the adolescent self-image, Wesleyan University Press, Middletown, CT, 1989.

[13] C. D. Spielberger, Theory and Research on Anxiety, in: Anxiety and Behavior, Academic Press, New York, NY, 1966, pp. 3–20. doi:10.1016/B978-1-4832-3131-0.50006-8.

[14] C. D. Spielberger, F. Gonzalez-Reigosa, A. Martinez-Urrutia, Development of the Spanish

Edition of the State-Trait Anxiety Inventory, Interamerican Journal of Psychology 5 (1971) 3–4.

[15] R. A. Baer, G. T. Smith, J. Hopkins, J. Krietemeyer, L. Toney, Using self-report assessment methods to explore facets of mindfulness, Assessment 13 (2006) 27–45. doi:10.1177/1073191105283504.

[16] R. L. Greene, The MMPI-2: An interpretive manual, 2nd ed., Allyn & Bacon, Needham Heights, MA, US, 2000.

[17] K. Brosnan, B. Grün, S. Dolnicar, Identifying superfluous survey items, Journal of Retailing and Consumer Services 43 (2018) 39–45. doi:10.1016/j.jretconser.2018.02.007.

[18] G.-Z. Y. David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, XAI—Explainable artificial intelligence David, Science Robotics (2019) 1.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357. doi:10.1613/jair.953. arXiv:1106.1813.

[20] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774.

[21] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144. doi:10.1145/2939672.2939778.