Positive-Unlabeled Learning for User-Centred XAI: a Case Study in Healthcare

Iris Heerlien[1](✉), Selin Çolakhasanoglu[2], and Jeroen Linssen[3]

[1] Saxion University of Applied Sciences, Enschede, The Netherlands
i.r.heerlien@saxion.nl
[2] Saxion University of Applied Sciences, Enschede, The Netherlands
s.colakhasanoglu@saxion.nl
[3] Saxion University of Applied Sciences, Enschede, The Netherlands
j.m.linssen@saxion.nl

Abstract. As the aging population grows, coupled with a shortage of healthcare personnel, the demand for innovative solutions becomes imperative. Digital tools, such as medicine dispensers, offer promising avenues for remote healthcare delivery, alleviating the workload on professionals. Nonetheless, home care organizations encounter challenges in implementing and scaling these tools, ranging from a lack of awareness about available options to difficulties in selecting the most suitable tool for specific situations. This study investigates a recommendation methodology for a medicine dispenser based on Omaha profiles from Electronic Patient Dossier (EPD). Using the CRISP-DM methodology, we designed a Positive-Unlabeled learning-based algorithm. We added Explainable Artificial Intelligence (XAI) techniques, showing a feature importance representation based on Shapley values, to enrich the transparency and reliability of suggested interventions. The solution was evaluated with healthcare professionals from two healthcare organizations. Although the technical performance of the algorithm was decent (recall: 0.9), they stated the data is not detailed enough to conclude whether a medicine dispenser could be used, showing the need for human evaluation during the process. This study addresses challenges like a sparse dataset lacking detailed data and iteratively involving users during development when performing research in a real life situation.

Keywords: User-Centered XAI · Elderly care · Digital Healthcare Tools

1 Introduction

Elderly care is facing challenges due to personnel shortages. According to Gupta Strategists, a shortage of 67,300 health personnel in the elderly care in 2031 is realistic in the Netherlands. Medical technologies could be of help in this by lightening tasks and preventing elderly care [15]. Examples include video calling, medicine dispensers, lifestyle monitoring, and daily structure support. However, these technologies are not used as frequently as possible.

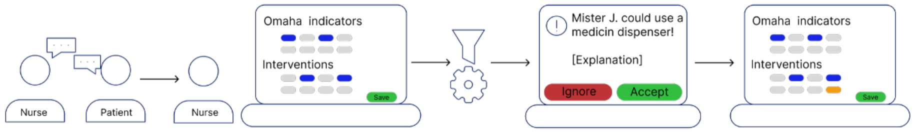

Fig. 1: The imagined workflow.

Research performed by Nedap Healthcare, an IT company that facilitates an Electronic Patient Dossier (EPD) platform, shows that staff involved in patient care are sometimes unaware of the available healthcare technologies within their organization. In addition, they find it challenging to determine which patients would benefit from these technologies. This lack of awareness and difficulty in identifying suitable patients is the core problem preventing the successful scaling of healthcare technology.

We performed a case study to explore the use of machine learning for recommendations on available healthcare technology based on EPD data. The goal is to develop a tool that gives advice whether a medicine dispenser could be used by a elderly patient based on EPD data. Besides the technical implementation and evaluation, we evaluated the results with the healthcare professionals.

The process including this advice will look like the diagram shown in Figure 1. The nurse will perform an intake call with the patient. Secondly, the nurse will create a plan of approach by filling in the signs and symptoms, and the interventions that need to be taken. Based on this, the algorithm decides whether a healthcare technology could be used. As the healthcare professional is in the lead, they can decide to follow this approach or to neglect it. If accepted, the technology is added to the to be performed interventions.

In this paper, we explain our methodology, discuss the challenges we faced, and present our key findings. We aim to use explainable AI to support and substantiate decision making.

Our main contributions are: (1) applying PU learning to handle incomplete labels, (2) evaluating machine learning models and their interpretability, and (3) involving healthcare professionals in the evaluation. The remainder of this paper is structured as follows. In Section 2, we discuss related work, followed by the methodology and results in Section 3. We discuss and conclude this study in Sections 4 to 6.

2 Related Work

This section covers the two main topics related to this study. First, we discuss Explainable AI and its relevance as we will use this in understanding the algorithms. Afterwards, we explain techniques related to PU learning as our dataset has no negative labels.

2.1 Understandable decision-making

It is important to understand why and how machines make certain decisions. Explaining decision-making to humans builds trust in AI-systems, which encourages the use of AI-based algorithms as assistant in various domains [14,7,5]. This study investigates different techniques of making the outcome of the machine learning algorithm interpretable and understandable for healthcare workers.

Explainable AI techniques There are several types of XAI techniques to make a machine learning model interpretable and understandable [8]. These techniques help users understand how models generate their predictions and enable trust in their outputs. Well-known techniques are feature importance (for global interpretation) and SHAP (for local interpretation) [10].

Feature importance techniques states the relative importance when making a prediction showing the most influential features. One common method is Permutation Importance. It evaluates the impact of shuffling a feature's values on model performance, where a significant decrease in performance indicates the feature's importance to the model. Feature importance and permutation importance are easy to compute and interpret, however they may be less effective for models with complex interactions between features.

Another technique is SHAP. This technique is a game-theoretic approach to explain output of an ML model, providing both local and global interpretability. SHAP values explain the contribution of each feature to a particular prediction by calculating the difference between the prediction made with and without the feature. [9]. According to Dwivedi et al. [3], an important advantage of SHAP values is that it is transparent and locally interpretable. This makes SHAP valuable in situations where understanding individual model decisions are crucial. For example, in healthcare a physician might want to understand the factors contributing to a patient's predicted risk to a disease. SHAP can provide insight in individual features such as age and medical history. Feature importance as discussed above provides insight of how important features are for the entire population, ignoring the individual scenarios.

User-centered explainable AI While algorithm-centric methods, such as SHAP and feature importance techniques, provide technical insight into model behavior, they often fail to consider whether these explanations are meaningful or useful to the end users of systems [14,7]. Research has shown that explanation techniques are often designed from the perspective of AI experts and do not necessarily align with the needs of end users, such as healthcare professionals [4]. Explanations that make sense to AI experts may not be as intuitive for domain experts or users, who rely on AI for decision making.

Different stakeholders interact with AI systems in various ways, requiring tailored explanations to meet their specific needs. Burnett et al. [2] emphasize that explanations should be adapted to the expertise of the users, whereas Wang et al. [16] discuss how user expectations influences the perception on explainability.

Human-centered design methods focus on identifying what end users need and how the design can respond to these needs [11].

Liao and Varshney [7] argue that a well-crafted explanation does not necessarily translate into effectiveness or benefit for the person interacting with an AI-system. There are human-centered design techniques that are important to achieve trust and decision-making. Kim et al. [5] propose three key factors for ensuring meaningful explanations. First, the contextualized quality of the explanation, e.g., is it satisfying, useful or trustworthy? Second, its contribution to human-AI interaction, e.g., improvement of user's perception on trustworthiness of AI system. Third, the contribution of the explanation to human-AI performance: does it help users complete tasks more effectively?

These factors highlight the necessity of shifting from a solely algorithm-centric approach to one that develops explanations that are not only technically relevant and logical, but also useful for end users. Ensuring that explanations align with the end user needs strengthens trust in AI systems and support for more effective decision-making for those who rely on them.

2.2 Positive Unlabeled learning

PU learning refers to a learning technique in which the training set contains only positively labeled and unlabeled data points, with no explicitly labeled negative samples [1]. Unlike traditional binary classification, where the training set consist of positive and negative labels, PU learning assumes that the unlabeled data can belong to either the positive class or the negative class but it is unknown. The primary challenge is distinguishing between true negative samples and unobserved positive samples.

There are several examples of applications in which PU data is being used. For example, personalized advertising methods label visited pages and clicks as a positive instance, but this does not mean that all other pages and ads are not necessarily uninteresting, therefore it cannot be labeled as a negative instance. Another common area with PU data is the medical domain. Medical records contain information about diseases that patients have and not diseases that patient do not have. The unlabeled diseases in this case do not mean that the patient does not have the disease. It can be unnoticed or not diagnosed by a health professional [1]. Several studies show different techniques on how to handle data that only have positive or unlabeled labels. In this study we evaluated two types of PU learning techniques: two-step approach and PU bagging approach. Both approaches differ in complexity and performance depending on the data. We compare both approaches to obtain information about which technique suits this application best.

Two-step approach This method aims to find reliable negative samples from unlabeled data in two steps: (1) train a classifier on positive samples to identify reliable negative labels based on low predicted probabilities, and (2) use these reliable negatives with the positive samples to train a (semi-)supervised model.

The two-step approach assumes separability [13,1] and smoothness, which means that it assumes that all positive samples are similar to the labeled samples and the negative samples are very different from the labeled samples, so basically all positive instances have similar behaviour [6]. A challenge within this approach is defining the reliable negative samples. If the defined reliable negatives are not representable as real negatives, the system will not learn sufficiently.

PU bagging approach Another technique can be described as a bagging strategy. Mordelet and Vert [12] studied this approach by using a bagging Support Vector Machine (SVM) to learn from positive and unlabeled data. They state that the bagging methodology performs as good as the non-bagging methods. They compare their performance to the biased SVM, which directly discriminates between positive and unlabeled examples by rebalancing misclassification costs. They also compare against one-class SVM and a baseline ranking approach. The ranking approach ranks the unlabeled data by their similarity to the average positive data samples. Their bagging approach outperforms state-of-the-art methods on simulated data. However, the difference in performance is minimal on real world data. Performance of the bagging approach relies on the quality of the random created subsamples drawn from the unlabeled instances. If the random subsamples do not represent the distribution of the data sufficiently, the classifier will find it difficult to generalize the results [12].

The PU Bagging approach operates as follows. First, a training set is constructed using all known positive data points along with a random sample of the unlabeled data points. A classifier, such as an SVM, is then trained using the positive and unlabeled data points, where the unlabeled samples are initially treated as negative. Once the classification model is trained, it is applied to the remaining unlabeled samples (those excluded from training) to obtain classification scores, commonly referred to as 'Out-of-the-Bag' (OOB) scores. This process is repeated multiple times, averaging the OOB scores for each unlabeled data point. Importantly, the classifier is retrained in each iteration using a newly sampled subset of unlabeled data. This iterative approach ensures that every unlabeled sample receives an OOB score by the end of the cycle, allowing for the computation of the likelihood that an unlabeled data point is either positive or negative [12].

3 Method and Results

The iterative method Cross-Industry Standard Process for Data Mining (CRISP-DM) is used in this research. [4] The first steps in this method are business and data understanding, after which the data is prepared, the models are trained and evaluated. As it is an iterative method, earlier steps are revisited when necessary. When the model is optimized, it can be deployed. The overview in Figure 2 visualizes this process.

[4] `https://www.ibm.com/docs/it/SS3RA7_18.3.0/pdf/ModelerCRISPDM.pdf`

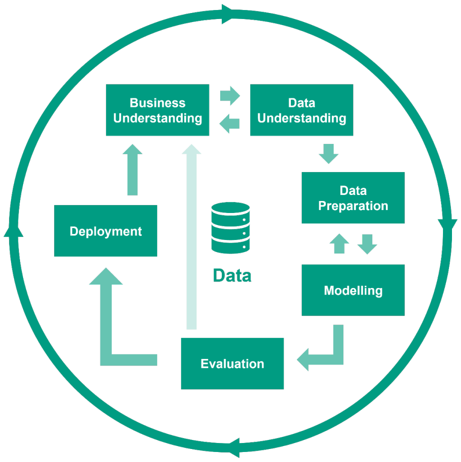

Fig. 2: The CRISP-DM cycle.

To effectively visualize the decision-making process of the ML algorithm, XAI methods were used. It is important for us as researchers to understand the decision-making of the model, but also for the end user to ensure trust in systems. In the next sections, we will explain the performed CRISP-DM steps in more detail, followed by the Explainable AI and generalizability components.

3.1 Business Understanding

It is important to start this research with the business understanding phase to clearly define the goals and objectives. To gain insights into the context from the perspectives of both healthcare organizations and the EPD software company, we interviewed a data scientist and an implementation expert from the company.

EPD software company The EPD is developed by a software company named Nedap Healthcare and they are committed to increase its performance. They observed that, despite positive feedback on the use of digital tools, healthcare devices were not frequently used. They discovered that healthcare professionals are often not aware that a digital tool is suitable for a patient. Therefore, Nedap Healthcare aims to create an algorithm that advises healthcare professionals on when to use a specific tool.

Elderly care organizations Multiple elderly care organizations use the EPD system of Nedap Healthcare, called ONS. They recognize the benefits of using healthcare devices, such as reducing the workload. The elderly care organizations agree with the software company that the care professionals are not aware when to use these tools. They believe that incorporating advice within the EPD would enhance awareness.

Table 1: An overview of the number of medicine dispensers per organization.

Organization	Total number of cases	Positive labels (%)	Negative labels
1	4,028	160 (3.97%)	Unknown
2	29,489	375 (1.27%)	Unknown
3	4,824	166 (3.44%)	40
4	9,439	5 (0.05%)	125

3.2 Data Understanding

The data used to develop the ML algorithm is derived from the Omaha profiles. These profiles are an outcome of the Omaha system which is a standardized framework for reporting and documenting symptoms and corresponding actions in the healthcare sector.[5] These Omaha profiles provide an overview of why an individual receives elderly care. The profiles includes multiple levels, such as symptoms, actions, and focus areas.

Elderly care professionals manually annotated the data, identifying instances where a patient could benefit from using a medicine dispenser. This is annotated as ‘positive’. As a result, the dataset includes both positive and unlabeled instances. The unlabeled data could be either positive or negative, resulting in the need for special data preparation methods. Only a small part of the data contains positive labels, as can be seen in Table 1 shown in percentages. The negative labels identify situations in which a medicine dispenser can not be used.

For the Exploratory Data Analysis, the library YData Profiling is used.[6]. This shows in one overview the basic statistics, such as missing values, minima and maxima, and the distribution. Additionally, the inclusion and exclusion criteria for using a medicine dispenser were explored in collaboration with the domain experts. In general, a medicine dispenser is only useful when a patient is still well enough to take the medication by themselves, possibly with help of a partner. The sweet spot between being not too good and not well enough is the situation in which a medicine dispenser should be advised. Criteria such as ‘terminal’, ‘just included in care’, ‘severe cognitive issues’, and ‘wandering’ were seen as exclusion criteria for not being able to use a medicine dispenser anymore. Criteria to denote the phase before a medicine dispenser is helpful were not known.

The dataset comprises multiple json files, each containing a report for individual patients. These reports include information such as symptoms, scores, and interventions. These data are all single data points, however there is a relationship between the symptom of a patient and the corresponding intervention. We captured this relationship by creating a path including the symptom and the corresponding intervention, which is discussed in more detail in the next subsection.

[5] https://www.omahasystem.nl/over-omaha-system/werken-met-het-omaha-system

[6] https://docs.profiling.ydata.ai/latest/

3.3 Data Preparation

In this phase, the raw data is converted to data that is suitable for the modeling phase. After exploring the data and consulting with the domain experts, we made the conclusion that the areas of interest (symptoms, scores, and interventions) were not properly connected in a cause-and-effect relationship. To resolve this, pathways of the different features were created to establish the cause-and-effect connections. For example, we combined the following features: symptoms 'unable_to_take_medication_without_help' and intervention 'medication _administration' together as a pathway. Additionally, PU learning, data balancing, and feature selection techniques were used to prepare the data.

PU Learning The data only consists of positive labels or unlabeled data. To still be able to train an algorithm on this data, we used PU learning. As explained in Section 2.2, two different methods were evaluated to obtain negative labels: the two-step approach and the bagging approach. During training, the two-step approach was much faster: this method took approximately three seconds, while the bagging approach took three hours. The reason for this is that the bagging approach relies on bootstrapping, which involves repeatedly resampling the data and retraining the model multiple times.

Data balancing To overcome the issue of unbalanced data, where the unlabeled data significantly outweighs the group of positive data points, a random undersampling technique is used in which data points from the unlabeled data were removed from the dataset until the dataset was balanced. [7] As can be seen in Table 1, the data is unbalanced as only approximately 2% of the data points are labeled.

By using PU learning and the undersampling method, four datasets were created, namely **PU learning - Two-step approach, PU learning Bagging approach**, **Undersampling**, and the **Original dataset.** These datasets were used in the modelling phase and evaluated in the evaluation phase.

Feature selection After integrating paths as features, the dataset resulted in having a large number of features. To reduce the feature set, Chi-squared feature selection was applied.[8] The number of features, ranging from one to 500, were evaluated with steps of twenty. For each set of features (k), the machine learning algorithm ran and its performance was evaluated to identify the optimal number of features that most influence algorithm's outcome. The implementation of the chi-squared feature selection is shown in the pseudocode Algorithm 1.

The result of the different sets of features can be seen in Figure 3. From 250 features onward, the performance did not increase. Therefore, these 250 features were used in model training.

[7] https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html

[8] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html

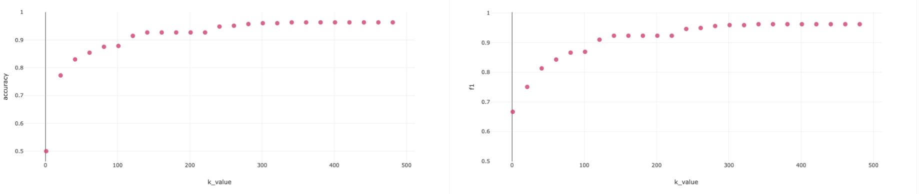

Fig. 3: The accuracy (left) and F1-score (right) stabilizes around 250 features.

Algorithm 1 Feature Selection and Model Evaluation with MLflow Logging

```
for each dataset type in datasets do
    Extract features $X$ and labels $y$
    for each model type and model object in models do
        for each value $k$ in $k\_values$ do
            Apply chi-squared feature selection with $k$ features ($chi2\_selector$)
            Select the $k$ best features from $X$ ($X\_chi2\_selected$)
            Perform cross-validation:
                Compute accuracy, precision, and recall scores using CV
            Log parameters using MLflow:
                Dataset type, model type, and $k$ value
            Log metrics using MLflow:
                Mean accuracy, precision, recall, and F1 scores
            Log the trained model using MLflow
```

3.4 Modelling

Three supervised learning models were trained on the datasets, namely a Decision tree, an SVM, and a Random Forest. We used the scikit-learn package and the latest implementations of these algorithms.[9] These models were chosen since they differ in complexity and transparency. Algorithm 1 shows the pseudocode for modelling. As can be seen, first a dataset is chosen, followed by a model type. The next step was to perform the feature selection and train the model based on these features. The last step is to log the results to MLFlow, as explained in Section 3.5.[10]

The model trained on the 'normal' dataset is our baseline model. This dataset is cleaned, however, no balancing or PU learning techniques have been used.

3.5 Evaluation

The models were evaluated both on a technical level and from the perspective of a domain expert.

[9] `https://scikit-learn.org/stable/supervised_learning.html`

[10] `https://mlflow.org/`

Table 2: An overview of the performance of the models.

Model type	**Accuracy**	**F1**	**Precision**	**Recall**
PU Bagging Decision Tree	0.724	0.753	0.681	0.842
PU Bagging Random Forest	0.740	0.769	0.690	0.867
PU Bagging SVM	0.772	0.790	0.734	0.855
PU two-step Decision Tree	0.942	0.940	0.980	0.903
PU two-step Random Forest	**0.940**	**0.938**	**0.968**	**0.909**
PU two-step SVM	0.882	0.881	0.889	0.873
Undersampling Decision Tree	0.785	0.770	0.826	0.721
Undersampling Random Forest	0.782	0.776	0.796	0.756
Undersampling SVM	0.818	0.817	0.822	0.812
Normal dataset Decision Tree	0.624	0.398	1.000	0.248
Normal dataset Random Forest	0.600	0.338	0.944	0.206
Normal dataset SVM	0.540	0.146	1.000	0.079

Technical evaluation For technical and testing purposes, healthcare professionals have created a separate test set, containing both negative and positive samples, to evaluate the model's performance on unseen data, see Table 1. We used 165 negative samples and 165 positive samples as a test set, which were not used during model training. MLflow was used in the technical evaluation of the model. Performance metrics that were used are precision, recall, F1-score, and accuracy. Also, the feature importance were saved to understand which features were most important in prediction.

An overview of the results are shown in Table 2. We observe that the two-step method resulted in the highest performance. The performance between the different models did not differ much. However, as the focus was on the recall, the random forest performed the best.

Algorithmic explainability The XAI techniques SHAP and feature importance were used to understand the reasoning of the model. The features making the biggest impact denoted information about patients not able to take medication without help, patient who do not follow the dosage schedule, and patients with limited recall of recent events. These insights and decision-making information were also evaluated with the healthcare professionals to evaluate the relation between technical model logic and real-world clinical reasoning.

Evaluation with the domain experts As discussed above, the results and model reasoning were discussed with the healthcare professionals of two healthcare organizations during two separate focus groups. Seven cases were discussed at one organization and three at the other. The difference in the number of cases was due to the depth of the discussions held during the sessions. We chose to delve deeper into these cases to gather more in-depth information about the way of working of the healthcare workers and their decision-making process. The cases were used to understand if they would advise a medicine dispenser for patients based on the features that was used in training. The features represented features with the most influence on the model outcome, according to SHAP and feature importance.

A case contains only the information that was used during training of the algorithm. Specifically the key features that contributed to the final decision which are obtained by using SHAPley values and feature importance techniques. The case was, in the beginning, shown anonymously to prevent the healthcare professional from recognizing the patient and having background information which are not shown in the case. The healthcare professionals were asked to read the case and to decide if a medicine dispenser could be used in this situation. After this, they were asked to open the dossier of this patient and to use information from the EPD to decide if this patient could use a medicine dispenser. Their task was to decide if they would advise something different based on the full data and which information from the EPD they used for this decision. The last step was to compare that to the result from the algorithm and to discuss why this could be different if this was the case.

In all cases the healthcare professionals were not able to decide whether a patient could use a medicine dispenser and needed more information than only the Omaha profiles. Especially the data from the free text fields within the EPD was used to understand the nuance of a selection in the Omaha profile. An example is the Omaha tag that the patient has a limited ability to concentrate. The healthcare professionals needed the additional information from the free text fields to inform them how limited this ability was and if there was, e.g., a partner living with this patient that could help to take the medicines on time. All in all, they needed more detailed context than only standardized features.

User-Centered Explainability To enhance the usability of the AI by the healthcare professionals, adding user-centered XAI is important to create trust in a system [14]. In this case, the healthcare professional needs to understand why a medicine dispenser could be used based on the available data. During our focus group sessions, we explored what types of information healthcare professionals need and how this information should be presented. Through feedback on the cases, which showcased key elements based on SHAP and feature importance values, we found that the presented information is not intuitive enough to support confident decision-making. This means that even though we use explainable AI to clarify a model's reasoning, the current explanation does not meet the requirements nor expectations of the healthcare professionals. They require more contextual information, such as severity of a patient's memory issues, to make well-informed decisions. Unfortunately, these kind of contextual data is not included in the Omaha profiles on which our model is trained. As a result, we are unable to provide the healthcare professionals with explanations that are fully understandable or useful due to limitations in the available data.

Table 3: An overview of models performance where one healthcare organization is left out as test set.

Model type	**Accuracy**	**F1**	**Precision**	**Recall**
PU Bagging Decision Tree	0.700	0.002	0.001	0.600
PU Bagging Random Forest	0.705	0.002	0.001	0.600
PU Bagging SVM	0.788	0.003	0.002	0.600
PU two-step Decision Tree	0.800	0.750	1.000	0.600
PU two-step Random Forest	0.800	0.750	1.000	0.600
PU two-step SVM	0.800	0.750	1.000	0.600
Undersampling Decision Tree	0.600	0.714	0.556	1.000
Undersampling Random Forest	0.500	0.667	0.500	1.000
Undersampling SVM	0.500	0.667	0.500	1.000
Normal dataset Decision Tree	0.500	0.000	0.000	0.000
Normal dataset Random Forest	0.500	0.000	0.000	0.000
Normal dataset SVM	0.500	0.000	0.000	0.000

3.6 Generalizability

Lastly, we explored the generalizability of the trained ML model. The goal is to determine whether a model trained on data from a specific healthcare organization could be effectively applied to a different healthcare organization. To investigate this, the ML model was retrained, this time excluding the data from one healthcare organization (healthcare organization 4 in Table 1) during the training phase. The data from the excluded organization was then used as a test set to evaluate the model's performance. The results are shown in Table 3. As can be seen, the models struggle to perform well in this situation, indicating difficulties in learning and generalizing. This was discussed with the healthcare professionals. They stated that it is very different per organization, and also per team, how the Omaha profiles are filled in. Besides, the user has the freedom in the system to make all different combinations of symptoms and interventions, as this is not restricted in the software. This creates the very diverse and thus sparse dataset.

4 Discussion

During this research project the goal was to advise if a medicine dispenser could be used by patients from elderly care. As this decision is based on the medical status of the patient, reports from an EPD were used as input.

4.1 Methodology and results

The CRISP-DM process model was used to understand the context and the data, and using this information to prepare the data to be able to train and evaluate a model using this data. The following sections reflect on this methodology and the technical results.

We considered the CRISP-DM model as useful. It helped to structure the research and to focus on the context first before diving into the data. However, we would advise to focus more on the business and data understanding phases as this would have revealed earlier in the process that the data would not suffice. The deployment phase was not performed as the evaluation with the domain experts indicated a next iteration is necessary before the model would represent the situation sufficiently. This further highlights the necessity of involvement of all stakeholders in CRISP-DM, especially during the first phases.

The technical results seemed promising: a recall around 0.9 was achieved. However, when evaluating the results with the domain experts, they stated that based on only the Omaha profiles, which were used as input for the algorithm, they could not decide whether a medicine dispenser would indeed be useful. Also, when the trained model was evaluated using unseen data from another organization, the model did not perform well (see Table 3). One explanation could be that it is due to overfitting on the data of an organization, and to finding patterns that are not there. The second explanation is a result of different ways of reporting by the teams, creating a dataset with different paths different than those found in the training data.

As the healthcare professionals indicated that this tool would help them, we advise to improve the tool with free text fields and to evaluate it again. Our hypothesis is that using data from free text fields would improve the model and its explainability, resulting in a higher user satisfaction.

4.2 Challenges

During this research, we encountered challenges due to performing a real world use case.

Positive Unlabeled data The first challenge was handling data with only positive labels. Two PU learning techniques were evaluated: the two-step and bagging approaches. The two-step approach was faster and performed better. The reason it is faster is that the bagging approach relies on bootstrapping, which involves repeatedly resampling the data and retraining the model multiple times. The reason it performs better could be that the percentage of positive labels was very small, which the two-step approach handles better. If sufficient reliable negative samples can be found during the first step of the two-step approach, it is possible to train a decent classifier to recognize patterns between positive and negative samples. In this case, we expect the performance of the two-step approach to be more reliable. The performance of the bagging approach relies on the quality of the random subsamples drawn from the unlabeled data. If the random subsamples do not represent the underlying distribution of the data, the classifier can have a hard time to find patterns and generalize, which can result in lower performance compared to the two-step approach [12].

Sparse dataset Secondly, as multiple paths were possible to state the same medical situation, the dataset was sparse. Although we did not experience this as a problem during modeling, during evaluation we encountered that the algorithm tested on data from other elderly care teams resulted in bad performance as the conditions were reported differently resulting in different paths in the data. This makes it infeasible to make this algorithm generalizable to be used by multiple organizations and teams. A solution is to train an algorithm for every team, however due to the small dataset per team the expectation is that the results will not be trustworthy. Another topic to investigate is evaluating a model that does not take into account the cause-and-action relationship and takes every data point as a single feature. This will solve the issue of a sparse dataset since it does not take into account all possible combinations of features relationships. A second suggestion is to only use the 'symptom' and 'action' data in the path as this seems the core information from the dataset. Using only these two fields in the paths, fewer paths are possible, resulting in a less sparse dataset with still sufficient information. A third solution is to use dimensionality reduction techniques such as principle component analysis. However, this makes it hard to create an explainable model which was one of the goals of this research. Future work should focus on experimenting with dimensionality reduction techniques which keep the explainability high.

Technical evaluation versus real-world evaluation Thirdly, the technical evaluations in comparison with the domain experts evaluation showed different results. Although the (explainability) results from a technical perspective when trained and evaluated on data from the same teams seemed promising, the evaluation with the domain experts showed that it seems not intuitive to create an algorithm based on the data from the Omaha profiles only. The experts needed more contextual data from the open text fields to understand the nuance and to be able to decide if a medicine dispenser could be used by a patient. Although the addition of the open text fields is possible from a technical perspective, there is a privacy perspective that needs to be taken into account as well. Based on this data, there is a chance that the researchers could deduce which medical dossier belongs to which patient, which raises concerns over privacy. One possibility to use this data is to anonymize and pseudonymize it before it is used. Future research should investigate how to integrate contextual data from patient reports while ensuring privacy.

5 Generalizable Insights about Responsible Application of Machine Learning in Healthcare

The main takeaway when looking at a responsible application of machine learning in this domain is that one should always evaluate the results with the domain experts and/or end-users. Although the algorithm may seem technically valuable, it could be the case it is not from a domain experts view. Explainable AI techniques could help in this by showing the reasoning of the underlying model.

This could strengthen the conversation and make clear which mistakes are made by the model. It helps as an interface between the data scientists and the domain experts. Our research has shown this in this use case: the explainable AI component helped by understanding if the algorithm focused on the right information. This also revealed that the information needed by the healthcare professionals to understand the model was not in the data. This is an important evaluation point which would have not been found when not using XAI techniques.

Evaluation with the end-users is also an important aspect since they have to work with the created tools. If the tools do not meet end-users' expectations, adoption is likely to be low, resulting in a less effective product. This would also have been the case when we would have implemented this tool in the workflow without evaluation. The XAI component would give explanations which are not useful for the end-users, resulting in lower trust and thus adoption.

Additionally, it should be recognized by the healthcare professional when the AI provides a wrong advice. If this is missed, the effects could be dangerous. An example from this project is that if a patient gets a medicine dispenser while one is not able to use it, the patient will not take the medicines correctly which creates possibly an unhealthy situation. Evaluating the solution with and keeping the control by the healthcare professionals could decrease the chances of creating this undesirable situation as well.

6 Conclusion

During this research project, a case study was performed using the Omaha profiles from the EPD system to predict if a medicine dispenser could be used by a patient. The CRISP-DM process model was adopted to understand the context, prepare the data and create and evaluate the models. Technically challenges were encountered, examples are a positively unlabeled and sparse dataset, which included different features per organization, creating the necessity to use techniques to create a dataset that could be used for modeling. Although the technical evaluation seemed promising, the evaluation with domain experts showed that the data used did not give enough information to give an advise based on the data. Additionally, the multiple challenges showed the difference between a theoretical case compared to a case from a real life situation. An important takeaway from this study is the need to involve end-users earlier and to evaluate whether the available data provides sufficient information to support decision-making. Future research should focus on exploring the possibility to include more data from the open text fields. The hypothesis is this will improve the algorithm, resulting in a higher user satisfaction.

Acknowledgments. This study was funded by Tech For Future[11]. Additionally, we would like to thank the company Nedap Healthcare and their client organizations for making this research possible.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article

[11] https://techforfuture.nl/

References

1. Bekker, J., Davis, J.: Learning from positive and unlabeled data: a survey. Mach. Learn. **109**(4), 719–760 (Apr 2020)
2. Burnett, M.: Explaining ai: fairly? well? In: Proceedings of the 25th International Conference on Intelligent User Interfaces. p. 1–2. IUI '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3377325.3380623
3. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Rana, O., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., Ranjan, R.: Explainable ai (xai): Core ideas, techniques and solutions. ACM Computing Surveys **55** (09 2022). https://doi.org/10.1145/3561048
4. Ehsan, U., Wintersberger, P., Liao, V., Mara, M., Streit, M., Wachter, S., Riener, A., Riedl, M.: Operationalizing human-centered perspectives in explainable ai. pp. 1–6 (05 2021). https://doi.org/10.1145/3411763.3441342
5. Kim, J., Maathuis, H., Sent, D.: Human-centered evaluation of explainable ai applications: a systematic review. Frontiers in Artificial Intelligence **7** (2024). https://doi.org/10.3389/frai.2024.1456486
6. Lee, W.S., Liu, B.: Learning with positive and unlabeled examples using weighted logistic regression. vol. 20, pp. 448–455 (01 2003)
7. Liao, V., Varshney, K.: Human-centered explainable ai (xai): From algorithms to user experiences (10 2021). https://doi.org/10.48550/arXiv.2110.10790
8. Loh, H.W., Ooi, C.P., Seoni, S., Barua, P.D., Molinari, F., Acharya, U.R.: Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Computer Methods and Programs in Biomedicine **226**, 107161 (2022). https://doi.org/10.1016/j.cmpb.2022.107161
9. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions (2017), https://arxiv.org/abs/1705.07874
10. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
11. Melles, M., Albayrak, A., Goossens, R.: Innovating health care: key characteristics of human-centered design. International Journal for Quality in Health Care **33**, 37–44 (10 2020). https://doi.org/10.1093/intqhc/mzaa127
12. Mordelet, F., Vert, J.P.: A bagging svm to learn from positive and unlabeled examples (2013), https://arxiv.org/abs/1010.0772
13. Ortega Vázquez, C., vanden Broucke, S., De Weerdt, J.: A two-step anomaly detection based method for pu classification in imbalanced data sets. Data Min. Knowl. Discov. **37**(3), 1301–1325 (Mar 2023), 10.1007/s10618-023-00925-9
14. Schoonderwoerd, T.A., Jorritsma, W., Neerincx, M.A., van den Bosch, K.: Human-centered xai: Developing design patterns for explanations of clinical decision support systems. International Journal of Human-Computer Studies **154**, 102684 (2021). https://doi.org/10.1016/j.ijhcs.2021.102684
15. Strategists, G.: Uitweg uit de schaarste, https://gupta-strategists.nl/storage/files/220525-Gupta-Strategists-FME-Uitweg-uit-de-schaarste.pdf, accessed on January 24, 2024.
16. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing theory-driven user-centric explainable ai. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. p. 1–15. CHI '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3290605.3300831