Hybrid Personal Medical Digital Assistant Agents

Sara Montagna[1,*], Christel Sirocchi[1,*]

[1]Department of Pure and Applied Sciences, University of Urbino, Piazza della Repubblica 13, 61029, Urbino, Italy

Abstract

Autonomous intelligent systems are beginning to impact clinical practice as personal medical assistant agents, by leveraging experts' knowledge when needed and exploiting the vast amount of patient data available to clinicians. However, these approaches are seldom integrated. In this paper, we propose an integrated hybrid agent architecture that combines symbolic reasoning with sub-symbolic, data-driven models. Using the PIMA dataset, we demonstrate that this hybrid approach enhances the performance of both approaches when used alone. Specifically, we show that integrating a logical agent, which uses predefined expert knowledge plans, with rules obtained by symbolic knowledge extraction from machine learning models trained on historical data, improves system reliability and clinical decision-making, while reducing misclassified instances.

Keywords

PMDA, Hybrid agent architecture, Symbolic knowledge extraction

1. Introduction

The advent of personal medical digital assistant agents (PMDA) marks a significant milestone in healthcare [1], aiming to provide support and recommendations to both patients and clinicians. However, in healthcare settings, it is essential for systems to be both trustworthy and explainable, as they typically handle safety-critical tasks [2]. Consequently, the design of PMDA agents exhibiting trust and reliability is pivotal for adopting these systems in clinical practice.

Agents that base their recommendations on established medical protocols, usually modelled as the agent's beliefs in its knowledge base, inherently possess a degree of trustworthiness and reliability. When these agents utilise logical, rule-based systems, explaining decisions becomes relatively straightforward, as the explanation is provided by the rule whose conditions were satisfied. However, these protocols may not always deliver the performance required for effective clinical adoption, particularly in grey areas where patient cases do not conform to predefined categories or the clinical evidence is ambiguous [3]. In this context, the literature recognises the advanced capabilities of machine learning (ML) models, which have gained significant attention in recent years [4]. These models can uncover latent patterns and knowledge from data that extend beyond the scope of traditional medical protocols [5]. Thus, trained ML models can be integrated into the agent's internal cycle, delivering robust analysis of its perceptions. However, unlike purely rule-based agents, this integration complicates the explanation of the decision-making process, as black-box ML models inherently lack transparency [2].

WOA 2024: 25th Workshop "From Objects to Agents", July 8-10, 2024, Forte di Bard (AO), Italy

*Corresponding author.

sara.montagna@uniurb.it (S. Montagna); c.sirocchi2@campus.uniurb.it (C. Sirocchi)

0000-0001-5390-4319 (S. Montagna); 0000-0002-5011-3068 (C. Sirocchi)

Given these premises, there is a growing recognition of the need for hybrid models that integrate the robustness of medical protocols with the adaptive learning capabilities of ML [6]. Such integration aims to harness the strengths of both approaches while ensuring the decisions made by PMDA agents are both explainable and reliable [7, 8, 9].

In this paper, we propose a hybrid agent architecture obtained by integrating ML into the reasoning cycle of agents. In particular, the proposed solution is grounded on logical agents whose knowledge can be updated based on new rules extracted by ML models. This hybrid methodology allows the PMDA to navigate the grey areas where medical protocols fall short, and data-driven insights can provide additional support. Additionally, incorporating knowledge extracted from ML models in symbolic form enhances predictive abilities of these agents while maintaining their explanatory capabilities.

The potential of this integrated approach is demonstrated using the PIMA dataset, a widely used benchmark in medical research, particularly in the study of diabetes [10]. The study focuses on reducing the number of false negatives in diabetes diagnosis, where patients at risk are not identified by the clinical protocol due to its limited coverage. By incorporating additional rules extracted from ML models, we aim to refine and improve diagnostic accuracy, particularly for these borderline cases. The integration of ML-derived rules improved the predictive capabilities of the PMDA agents, ensuring both higher diabetes detection and full coverage, while outperforming both the clinical protocol and the standalone ML models. Furthermore, since agent actions are based solely on a logical theory, they inherently retain explainability.

Additionally, agents may have distinct requirements for updating their knowledge, based on factors such as prediction accuracy, explanation readability, and knowledge fragmentation, depending on their application domain. Therefore, we demonstrate how agents can update their knowledge base according to specific internal criteria. Finally, simulations conducted within this framework demonstrate that if a PMDA identifies a patient at high risk of developing diabetes, immediate interventions such as dietary adjustments can be initiated. This proactive approach underscores the significance of combining traditional medical knowledge with cutting-edge ML techniques to enhance patient outcomes and advance personalised medicine.

2. Background

The intersection of artificial intelligence and healthcare is the core of several research efforts and reports significant advancements. ML, in particular, is the most discussed technology in this field [11, 12], as it allows for the exploitation of large datasets by discovering relationships and patterns hidden in data. Its primary objective in healthcare is to develop accurate and robust models capable of making clinical predictions. Moreover, ML models are also widely used to identify risk factors by detecting crucial features in predictions.

ML has achieved remarkable performance in various domains of clinical medicine, outperforming human physicians in some cases and enabling the development of computer-aided diagnosis systems [4]. However, with thousands of studies applying ML algorithms to medical data, only a handful have significantly contributed to clinical care, with few systems receiving FDA approval for healthcare use [13] Resistance to embracing ML in clinical settings can be attributed to the prevalent reliance on evidence-based clinical pathways, guidelines, and protocols

as the foundation for clinical decision-making [14], while ML primarily relies on available data. Novel ML models, even when reporting superior performance compared to current protocols, might be unsuitable for clinical use if they *(a)* fail to correctly predict cases effectively managed by the protocol in place due to potential liabilities, *(b)* make predictions based on confounding variables and erroneous relationships that contradict established clinical knowledge [15] or *(c)* make predictions that cannot be explained to the user [16].

On the other side, medical protocols alone can sometimes fail to detect complex patterns, correlations, causal relationships and little variations in data due to their reliance on predefined rules and thresholds, making them less effective in borderline decision cases [17].

Bridging this gap, the integration of medical knowledge with ML has emerged as a topic of ongoing debate in the literature. Particular attention is given to methods performing knowledge injections into ML models, falling under the paradigm of informed ML [6], which seeks to augment ML models by combining data-driven learning with domain-specific expertise. Furthermore, symbolic knowledge extraction methods are also documented in the literature as a means to derive rules from trained ML models, which can then be used in recommender and decision support systems [18, 19]. For instance in [20] a method to extracting a symbolic model from a Graph Neural Network is presented.

At the same time, there is a growing demand for seamless real-time interaction between human users and digital assistants to ensure continuous and effective support. In response, a dynamic and evolving research field is dedicated to the design and advancement of Personal Medical Assistant Agents (PMDAs) [21]. These are specialised assistant agents [22], designed to support users in their daily activities –from simple tasks such as making calls, reading emails, sending messages, and opening web pages, to more complex tasks like scheduling appointments, interacting with physical objects in the environment, and controlling smart devices–, specifically devised for the medical domain. A distinctive feature of PMDAs is their ability to support continuous bidirectional interaction with the user. This involves acquiring information about the user's state and environment and providing feedback in various forms: these agents can for instance engage in natural language dialogues, offering a more intuitive and effective means of interaction. While traditional recommender systems have paved the way, the future lies in developing more interactive and responsive digital assistants that, by leveraging real-time data, can provide personalised and context-aware support.

In the context of healthcare, these agents aim to provide personalised medical assistance, ranging from health monitoring to medical advice, by delivering precise, reliable, and explainable healthcare recommendations [1]. With the rise in chronic diseases and an ageing population, they have been particularly adopted for assisting patients, offering 24/7 support, thereby reducing the burden on healthcare professionals and improving patient outcomes [23]. These digital assistants can integrate with various health monitoring devices, such as wearable fitness trackers, smartwatches, and home medical equipment, to gather real-time health data. This data is then analysed to provide insights into the user's health status, offer reminders for medication, suggest lifestyle changes, and even predict potential health issues before they become critical. Additionally, PMDAs have been used for supporting the work of caregivers [21].

PMDAs leverage a variety of technologies, integrating wearable devices and the Internet of Medical Things (IoMT) to acquire data from the environment and exploiting different cognitive algorithms to exhibit intelligent functionalities, thereby providing comprehensive support to

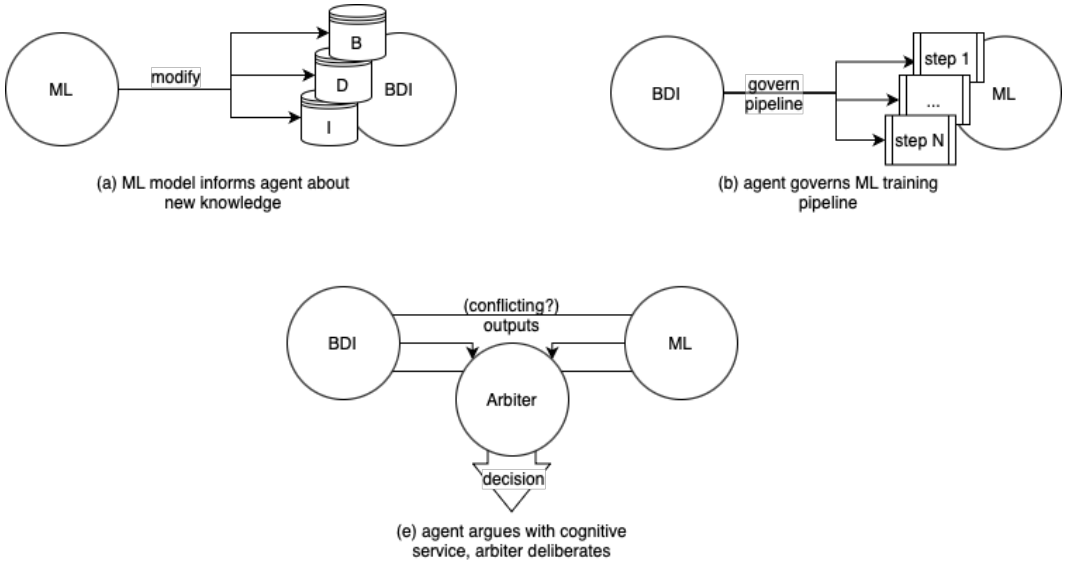

Figure 1: Flows of information and control between the BDI agent and a ML model

users. The literature also reports several examples of embedded data-driven AI. Generally, the concept of learning and adapting plans based on experience is addressed through reinforcement learning techniques, enabling a BDI agent to have some plans explicitly programmed while others are learned during its lifecycle [24]. Another instance of embedded AI is presented by [25], where a decision tree is used to define the optimal set of plans for a BDI agent. Similarly, another study explores using decision trees to enhance BDI agents with learning capabilities for plan selection [26]. Moreover, when discussing these topics, the need for a proper reference ontology is recognised. For instance, the work presented in [27] introduces a multi-agent system architecture that integrates neural network and symbolic methods for constructing and updating ontologies, enhancing autonomous agent cooperation and mutual monitoring through semantic similarity derivation.

In this paper, we focus on the design and discussion of the internal mechanisms of these agents. By delving into the internal architecture of PMDAs, we aim to highlight a model that makes these assistants reliable and interpretable. Given the identified gaps in the literature, we propose an architecture that effectively leverages the advantages of data analytics and the inference capabilities derived from well-grounded medical knowledge. Our proposed architecture integrates real-time health data processing, advanced machine learning algorithms, and established medical guidelines to provide accurate, personalised, and reliable healthcare assistance. This comprehensive approach ensures that the digital assistant not only processes vast amounts of health data efficiently but also applies robust medical reasoning to deliver high-quality care recommendations.

3. An Integrated Architecture

The hybrid integrated agent architecture we are proposing in this paper grounds on previous works of us [28, 21], where we envisioned different possible models for making the outcomes of data-driven models influencing the agent cycle in the definition of the plan, strategies or actions to be performed. These integrations delineate various approaches for how data-driven models may be integrated within the life-cycle of a logical agent, there assuming that the agent is specifically designed as a BDI [29, 30] agent—but the same discussion stands for any architecture of logical agent.

A simplified version, which summarises the main elements of the proposal in [28, 21], is reported in Figure 1. The first architecture (case *(a)*) positions the ML model as an input for the agent's knowledge base, manipulating its constructs like goals or beliefs. Depending on the specific design, the agent's goals define whether it accepts or rejects the KB update proposed by ML. This allows for adaptiveness by expanding or contracting the agent's range of activities based on data collected from the operational domain. Contrarily, case *(b)* inverts the roles of supervision and intervention. The agent can for instance operate on the inner workflow of a ML model, potentially adjusting parameters or modifying the learning process for instance by injecting knowledge at different steps of the learning pipeline. The final architecture (case *(c)*) presents a collaborative model where the agent and data-driven models are peers, providing outputs to an arbiter responsible for making final decisions. This approach facilitates adaptiveness, safety, and other desired properties through argumentation, where both parties engage in dialogue overseen by the arbiter.

Each architecture presents technical challenges, particularly regarding the impact on agent autonomy and integration with other technologies. While not entirely novel, these architectures draw from existing research in belief revision, automated planning, reinforcement learning, and explainable AI paradigms.

In this paper, we present our proposal in alignment with integration type *(a)* and address the various challenges it entails, with a particular emphasis on enhancing the reliability and explainability of agent decisions. Building on our previous work [21], where ML models suggested actions based on their predictive capabilities without explaining the reasoning behind those suggestions, we now focus on methods to enhance predictive capabilities while also providing explanations for those predictions. Moreover, in this improved version, our objective is not to interfere with the agent's actions by prioritising data-driven predictions, but rather to enrich the agent's knowledge, upon which the agent's strategy is defined. We achieve this by approximating the ML black-box model with an interpretable rule-based model, extracting rules from the trained ML, and integrating these rules into the agent's knowledge base. Issues arise when the extracted rules may not align with the predefined plans, and the policy for retaining rules strongly depends on the agent's goals. More details on this are provided in the next section by exemplifying possible solution through the case study.

4. Materials and methods

4.1. Dataset and domain knowledge

The dataset analysed within the study is the Pima Indians Diabetes dataset, compiled by the National Institute of Diabetes and Digestive and Kidney Diseases from a study of the Pima Indian population, known for its high diabetes incidence [10]. The dataset comprises 768 medical profiles of women aged 21 and above, who underwent an Oral Glucose Tolerance Test (OGTT) to measure their glucose and insulin levels at two hours. The target variable is binary, indicating a diabetes diagnosis within five years. Details about the features available in the dataset can be found in Table 1. Missing values are present in the attributes I_{120} (48.70%), ST (29.56%), BP (4.55%), BMI (1.43%), and G_{120} (0.65%), and were imputed in this work with the median value of the respective variable, as reported in the literature [31].

Table 1
Pima Indians Diabetes dataset

Feature name	Code	Description
Pregnancies		Number of times pregnant
Glucose	G_{120}	2-hour plasma glucose concentration in OOGT in mg/dL
Blood Pressure	BP	Diastolic blood pressure in $mmHg$
Skin Thickness	ST	Triceps skin-fold thickness in mm
Insulin	I_{120}	2-hour serum insulin in $\mu U/mL$
Body mass index	BMI	Body mass index as weight/(height)^2 in kg/m^2
Diabetes Pedigree Function	DPF	Likelihood function of diabetes based on family history [10]
Age		Age in years

Public health guidelines on type-2 diabetes risks report that individuals with a high BMI (≥ 30) and high blood glucose level (≥ 126) are at severe risk for diabetes, while those with normal BMI (≤ 25) and low blood glucose level (≤ 100) are less likely to develop diabetes. These guidelines have been utilised to design rules [32] expressed as logic predicates (see Table 2), which represent the Knowledge Base (KB) for the current case study.

Table 2
Knowledge base for predicting risk of type-2 diabetes as formalised by Kunapuli et al. (2010) [32].

Rule 1	$(BMI \geq 30) \wedge (G_{120} \geq 126)$	$\Longrightarrow$ diabetes
Rule 2	$(BMI \leq 25) \wedge (G_{120} \leq 100)$	$\Longrightarrow$ healthy

4.2. Machine learning models and rule extraction

In this study, six ML classifiers were utilised, namely Decision Tree (DT), Gradient Boosting (GB), Multi-Layer Perceptron (MLP), Logistic Regression (LR), Random Forest (RF), and K-Nearest Neighbour (KNN). Performance evaluation encompassed a range of metrics including Accuracy (A), Precision (P), Recall (R), F1 score (F1), Balanced Accuracy (BA), and Matthew's Correlation

Coefficient (MCC), as well as True Positive (TPR), True Negative (TNR), False Positive (FPR), and False Negative Rates (FNR). Cross-validation was employed with an extensive parameter search for hyperparameter optimisation of each model. In particular, a nested cross-validation approach was utilised, comprising 10 outer folds for evaluation and 5 inner folds for hyperparameter tuning. Performance metrics were computed for each outer fold using the model parameters optimised in the inner folds, and the average and standard deviation of these metrics were calculated to provide a comprehensive understanding of the models performance. The models with the highest R and the lowest FNR (GB and DT) were selected for further investigation.

The two selected models were retrained with a two-fold cross-validation procedure, with one half of the dataset used for training, incorporating nested 3-fold cross-validation for hyperparameter tuning, and the other half for testing. The training and test sets were alternated between folds to test over the entire dataset while reducing computational costs. Following training, GB and DT models were converted into rule sets. Rule-based interpretable models approximating the predictions of GB were derived via rule extraction using CART [33] available from the PSyKE library [34], resulting in rule sets denoted as GB-CART. The maximum number of leaves, and hence rules, in the CART rule-extraction process was varied from 5 to 30 and ultimately set to 20 to maximise fidelity, which was evaluated in terms of accuracy and F1-score of the rule set with respect to the black-box model. Conversely, rule extraction was not required for DT, which could be converted into rule sets by translating each root-to-leaf path into an if-then rule. The rule sets representing the trained models (GB-CART and DT) were utilised to predict outcomes in the test set. Additionally, modified rule sets (GB-CART + KB and DT + KB), integrating the two rules from KB, were also used for prediction. In this integration, rules from KB are assigned priority over the ones derived by ML, so that if an instance satisfies the conditions of multiple rules, priority is given to the rules from KB. Performance metrics detailed above, as well as coverage, which measures the proportion of dataset samples accounted for by the rule set, were computed for the clinical protocol, the two ML-derived rule sets (GB-CART, DT) and the two integrated rule sets (GB-CART + KB, DT + KB).

4.3. Personal medical digital assistant with knowledge update

A PMDA was developed in Java, incorporating a knowledge base in tuProlog [35][1]. The system consists of three main components: the *environment*, the *health monitor agent*, and the *knowledge base*. The environment simulates a patient's health data, including glucose levels, blood pressure, BMI, insulin levels, age, diabetes pedigree function, pregnancies, and skin thickness. These parameters are initialised with example values and are dynamically updated to simulate changes over time. The health monitor agent interfaces with a Prolog-based reasoning engine (tuProlog). The agent loads a Prolog knowledge base containing rules for assessing diabetes risk. During each reasoning cycle, the agent performs the following steps: (a) *sensing* - retrieves current health data from the environment and updates the Prolog knowledge base with these values; (b) *reasoning* - evaluates the updated knowledge base against predefined Prolog rules to determine the next action; (c) *acting* - sets the risk status in the environment to either high or low based on the reasoning outcome and communicate an appropriate message to the user to help manage

[1] https://github.com/tuProlog/2p-java

the risk of diabetes. Specifically, the agent communicates "*Alert: Diabetes Risk is High!*" if the risk is evaluated as high, "*Well done! Diabetes Risk is Low.*" if the risk is low, and no message is given if the risk could not be assessed because the patient's parameters did not satisfy the conditions of any rule in the knowledge base (which occurs when rule coverage is incomplete). Additionally, an explanation is provided to the patient, listing the specific conditions that were met, which determined the risk prediction. In the main simulation loop, the health monitor agent runs for 1000 iterations, updating health parameters randomly within specified ranges to simulate real-world variability. Each iteration involves the agent performing its sensing, reasoning, and acting cycle. The agent can update the knowledge base by adding rules extracted and presented by a ML model trained on data.

To this aim, GB was retrained on 50% the dataset, using nested 3-fold cross-validation for hyperparameter tuning, while the remaining 50% was used for testing. Eight rules were extracted from the trained model using CART to balance fidelity and rule set size. Each extracted rule was evaluated on the test set in terms of coverage, accuracy, number of conditions, and the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The rules, along with their corresponding metrics, were proposed to the agent. The agent selected rules for inclusion in the current knowledge base based on the following criteria: (i) include rules that introduce zero false negatives to reduce cases of undiagnosed diabetes; (ii) include rules with at least 50% accuracy to maximise correctness; (iii) include rules with at least 5% coverage to avoid knowledge fragmentation; (iv) include rules with a maximum of three conditions for improved readability of the explanation offered to the patient.

5. Results and discussion

5.1. ML performance

The initial performance comparison of various ML models trained on the Pima Indians Diabetes dataset is summarised in Table 3. All models report a modest accuracy, ranging from 0.706 for the MLP to 0.762 for the LR. Among these models, LR stands out with the highest ability to correctly predict the negative class (healthy individuals), as evidenced by the lowest FPR, highest TNR, and highest P. This strength in evaluating negative instances, more abundant in the dataset, reflects on global performance metrics, making LR also the top scorer in A and MCC. On the other hand, GB reports the best ability to correctly predict the positive class (diabetic individuals) with the highest TPR and the lowest FNR, as well as the best scores for R, F1, and BA. DT is also notable for its diabetes prediction capabilities, being the only other model besides GB to achieve a recall above 0.6. Given the considered clinical scenario, where the correct classification of positive instances is crucial and recall is typically the metric to optimise, GB and DT are identified as the best-performing models and considered for integration into personal medical assistant agents.

A deeper analysis of the predictions made by each model, compared to those made by the clinical protocol and the actual outcomes, is illustrated in Figure 2. The heatmap is divided into regions based on whether the clinical protocol correctly predicts positive and negative instances. It can be observed that the coverage of the clinical protocol is relatively low, at about 34.5%, leaving many cases, primarily healthy individuals, without a diagnosis. Additionally,

Table 3
Evaluation metrics for ML models trained on the Pima Indians Diabetes dataset, averaged over 10 model instances trained during 10-fold cross-validation. The best value for each metric is highlighted in bold, corresponding to the highest value for all metrics, except for FPR and FNR for which it is the lowest.

Metric	A	P	R	F1	BA	MCC	TNR	FPR	FNR	TPR
Gradient Boosting	0.750	0.645	**0.638**	**0.639**	**0.724**	0.450	0.527	0.124	**0.126**	**0.223**
Decision Tree	0.741	0.636	0.623	0.624	0.713	0.432	0.523	0.128	0.132	0.217
K Nearest Neighbor	0.743	0.653	0.589	0.614	0.708	0.429	0.538	0.113	0.143	0.206
Random Forest	0.759	0.686	0.582	0.626	0.718	0.456	0.556	0.095	0.146	0.203
Logistic Regression	**0.762**	**0.714**	0.560	0.619	0.715	**0.464**	**0.566**	**0.085**	0.154	0.195
Multi-Layer Perceptron	0.706	0.608	0.488	0.530	0.655	0.333	0.535	0.116	0.178	0.170

the protocol produces false positives (region 3) but no false negatives, which is highly desirable in a clinical setting where a positive outcome typically leads to further specialised tests for confirmation, whereas a negative outcome usually does not prompt further examination. This characteristic should ideally be preserved in updated models, as models that introduce undiagnosed cases are less likely to be adopted in clinical practice.

Examining the predictions of the ML models in detail reveals several insights. In **region 1**, which includes diabetic cases correctly predicted by the protocol, all models make some mistakes. This suggests that replacing the current protocol with any of these models could pose potential risks, as cases previously predicted correctly might now become cases of undiagnosed diabetes. In **region 2**, which consists of diabetic cases for which the protocol could not make predictions, all models perform poorly, with only a fraction of cases being correctly identified as diabetic. This indicates that the available features are not sufficiently predictive for these cases. Nevertheless, some patients in this region are correctly identified by multiple models, suggesting the possibility of identifying criteria to correctly classify these patients and augmenting the protocol by adding rules to increase coverage in this region. In **region 3**, which includes cases incorrectly classified as diabetic by the protocol, most models also classify these instances as diabetic. This suggests that the features are not sufficiently predictive for this group of patients. However, some patients in this region are correctly identified by multiple models, indicating potential for identifying rules to replace existing ones and reduce false positives, thereby mitigating over-triage. This update, however, takes lower priority as we are mainly concerned with reducing false negatives and will be addressed in future work. In **region 4**, which includes healthy individuals correctly predicted by the protocol, all models also predict these patients as healthy, maintaining consistency with the protocol. In **region 5**, which consists of healthy individuals for whom the protocol cannot give a prediction, all models correctly predict most patients. This again suggests the potential for augmenting the knowledge base with rules to address this region and improve coverage.

5.2. KB integration

Interpretable rule-based models, such as DT, or obtained by rule extraction from black-box models, such as GB with CART, were evaluated against the protocol before (DT, GB-CART)

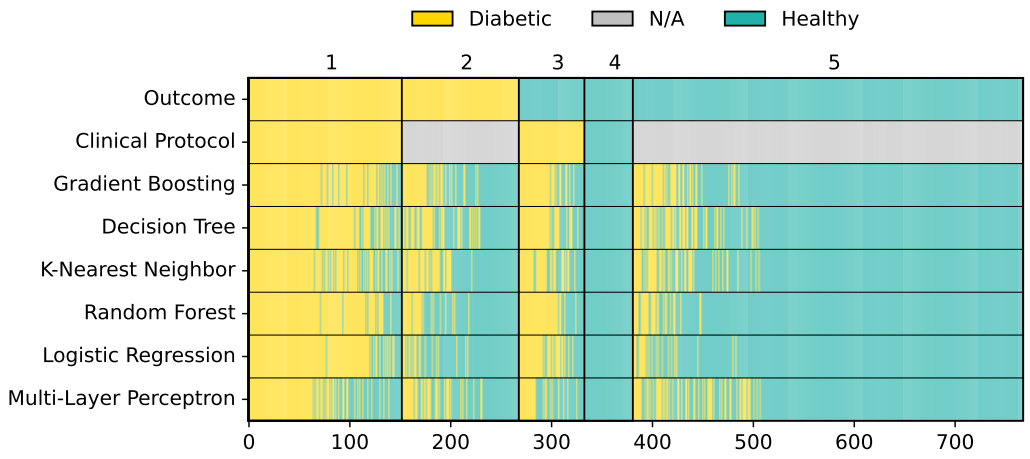

Figure 2: Predictions generated by six ML models trained on the Pima Indians Diabetes dataset using 10-fold nested cross-validation with hyperparameter tuning. Predictions from the clinical protocol and the actual outcomes are also reported. Five regions are highlighted based on protocol predictions.

and after (DT + KB, GB-CART + KB) integrating the rules of the clinical protocol. Performance metrics reported in Table 4 demonstrate the effectiveness of the clinical protocol, achieving very high performance metrics over the covered instances, including a 0% FNR and, consequently, perfect recall. However, this high performance comes with a limited coverage of only 34.5%. In contrast, all ML-derived rule sets offer full coverage. Integrated rule sets, DT + KB and GB-CART + KB, report improved performance across almost all metrics with respect to DT and GB-CART. FNR, which we want to minimise, is reduced by at least 25% for both models and TPR is similarly increased, while R is increased from 0.56 to 0.66 for GB-CART and from 0.58 to 0.71 for DT. The integration increases the overall number of patients predicted diabetic, thus reducing TNR and increasing FPR, although by a lesser amount. Performance metrics evaluating both classes (A, F1, BA, and MCC) all report improvement as a result of integration.

Table 4
Evaluation metrics computed for rule sets on the Pima Indians Diabetes dataset. Included rule sets are the clinical protocol formalising the Knowledge Base (KB), Decision Tree (DT), rules extracted from Gradient Boosting (GB) using CART (GB-CART), and composite rule sets DT + KB and GB-CART + KB.

Metric	A	P	R	F1	BA	MCC	TNR	FPR	FNR	TPR	Coverage
Clinical Protocol	0.755	0.700	1.000	0.824	0.712	0.545	0.181	0.245	0.000	0.574	0.345
GB-CART	0.734	0.636	0.560	0.595	0.694	0.401	0.539	0.112	0.154	0.195	1.000
GB-CART + KB	0.754	0.644	0.660	0.652	0.732	0.462	0.523	0.128	0.118	0.230	1.000
DT	0.711	0.586	0.582	0.584	0.681	0.363	0.508	0.143	0.146	0.203	1.000
DT + KB	0.727	0.590	0.713	0.645	0.723	0.431	0.478	0.173	0.100	0.249	1.000

Figure 3 illustrates how integrating rules from the knowledge base (KB) impacts predictions.

Adding Rule 1 from Table 2 to ML-derived models corrects predictions from healthy to diabetic for patients in region 1, but also introduces false positives in region 3, which are less critical than false negatives in the considered scenario. Rule 2 has minimal impact, as all rule sets agree on this patient subgroup. By incorporating KB rules with priority, the integrated models align perfectly with the predictions of the clinical protocol. Additionally, they provide full coverage, correctly identifying a fraction of diabetic patients in region 2, and most healthy individuals in region 5. This integrated approach leverages the high recall of the original protocol with the full coverage and more complex knowledge base derived from ML.

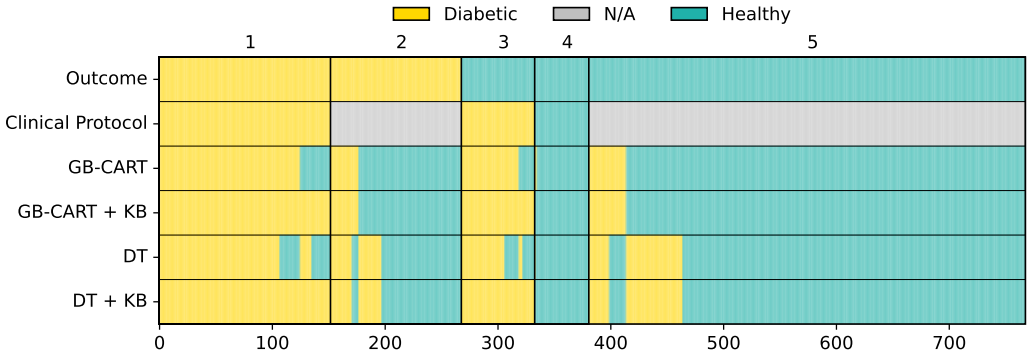

Figure 3: Predictions generated by rule sets over the Pima Indians Diabetes dataset, including the clinical protocol formalising the Knowledge Base (KB), the Decision Tree model trained on data (DT), the rule set extracted from the Gradient Boosting using CART (GB-CART), as well as composite rule sets DT + KB and GB-CART + KB, integrating protocol rules with priority. Additionally, the actual outcomes are included. The highlighted data subsets corresponds to the same regions depicted in Figure 2.

5.3. KB integration with agents requirements

The integration of a clinical protocol with additional rules extracted from ML models offers significant potential for personal medical assistant agents. However, the decision to include new rules may vary depending on the agent's role and specific criteria. For instance, agents tasked with diagnosing critical conditions may prioritise rules with a null FNR to minimise the risk of undiagnosed conditions. Agents aiming to maximise coverage while ensuring rule quality may select rules achieving a minimum level of accuracy. Conversely, agents aiming to prevent rule proliferation and knowledge fragmentation may only consider rules above a certain coverage threshold. Finally, agents providing simple explanations to users regarding prediction rationales may favour rules with a small number of conditions.

To address diverse agent requirements for potential knowledge updates based on ML recommendations, four (non-exhaustive) scenarios were explored. In each scenario, an agent predicted the patient's state based on sensed clinical parameters and an internal knowledge base, with the possibility to update this knowledge base with rules extracted by ML based on internal quality criteria. The eight extracted rules and their performance metrics computed over the dataset are reported in Table 5. Based on the rule metrics and quality criteria, the first scenario includes

rules 1, 3, 5, 6, and 8, while the second scenario incorporates rules 1, 3, 4, 5, and 7. Similarly, in the third scenario, rules 1, 4, and 7 are selected, whereas in the fourth scenario, all rules except 7 and 8 are added. Figure 4 illustrates the predictions made by the four updated knowledge bases over the original dataset. Additional scenarios can be explored, particularly by combining the criteria mentioned above or applying different criteria to rules predicting healthy and diseased outcomes. For example, by considering only rules with a null FNR and an accuracy exceeding 70%, only rule 1 would be added. This rule predicts individuals as diabetic if their $G120$ value is greater than 143.5 and their DPF value exceeds 0.32. Remarkably, this rule correctly predicts 24 individuals, constituting 3% of the dataset and 9% of the diabetic patients in the dataset, who could not be predicted by the original protocol. This illustrates that even minor updates to the knowledge base can greatly enhance the predictive capabilities of the monitoring agent.

Table 5
Performance metrics computed on the test set for 8 rules extracted from a Gradient Boosting model trained on the Pima Indians Diabetes dataset with 50-50 train-test split.

Rule	#Conditions	Outcome	Total	Correct	#TP	#TN	#FN	#FP	Accuracy	Coverage
1	2	Diabetes	53	39	39	0	0	14	0.736	0.138
2	3	Healthy	14	5	0	5	9	0	0.357	0.036
3	3	Diabetes	18	12	12	0	0	6	0.667	0.047
4	3	Healthy	244	191	0	191	53	0	0.783	0.635
5	3	Diabetes	7	4	4	0	0	3	0.571	0.018
6	3	Diabetes	16	7	7	0	0	9	0.438	0.042
7	4	Healthy	30	20	0	20	10	0	0.667	0.078
8	4	Diabetes	2	0	0	0	0	2	0.000	0.005

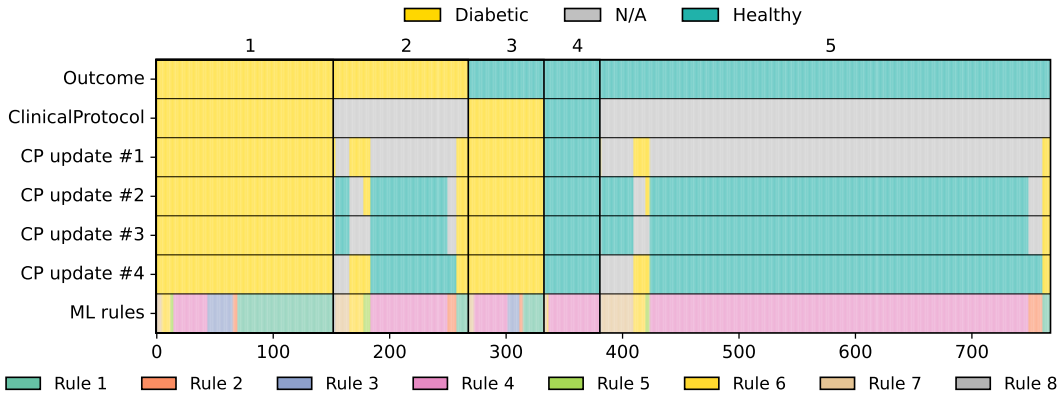

Figure 4: Predictions generated by the clinical protocol and the protocol enhanced with rules extracted from a ML model and selected by the agent according to four different criteria.

6. Conclusions and future work

This study demonstrates the potential of integrating clinical protocols with machine learning ML-derived rules to enhance the performance of PMDA agents. By combining the robustness and trustworthiness of established medical protocols with the adaptive learning capabilities of ML, these hybrid models can offer more comprehensive and accurate diagnostic suggestions. The approach was validated using the PIMA dataset, focusing on reducing the number of false negatives—patients likely to develop diabetes but not identified by medical protocols alone. The integration of additional rules extracted from ML models improved the predictive capabilities of the PMDA agents, ensuring both higher recall and full coverage. Future research will focus on refining integration techniques by investigating more sophisticated methods for combining ML-derived rules with clinical protocols, such as ensemble techniques. We also plan to validate our approach on a broader range of medical datasets to ensure generalisability across different conditions. Additionally, we aim to develop mechanisms for dynamic rule updates in real-time as new data becomes available, maintaining the accuracy of PMDA agents. Finally, enhancing the explainability of PMDA agents by incorporating user-friendly explanations in the form of natural language and visualisations is another avenue for further investigation.

Availability of data and code The dataset analysed is publicly available (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database), and the code to replicate the experiments can be found in the GitHub repository (https://github.com/ChristelSirocchi/hybrid-medical).

References

[1] A. Croatti, S. Montagna, A. Ricci, E. Gamberini, V. Albarello, V. Agnoletti, Bdi personal medical assistant agents: The case of trauma tracking and alerting, Artificial Intelligence in Medicine 96 (2019) 187–197. URL: https://www.sciencedirect.com/science/article/pii/S0933365717306000. doi:https://doi.org/10.1016/j.artmed.2018.12.002.

[2] H. Hagras, Toward human-understandable, explainable ai, Computer 51 (2018) 28–36.

[3] G.-D. Hou, Y. Zheng, W.-X. Zheng, M. Gao, L. Zhang, N.-N. Hou, J.-R. Yuan, D. Wei, D.-E. Ju, X.-L. Dun, et al., A novel nomogram predicting the risk of positive biopsy for patients in the diagnostic gray area of prostate cancer, Scientific Reports 10 (2020) 17675.

[4] F. Piccialli, V. Di Somma, F. Giampaolo, S. Cuomo, G. Fortino, A survey on deep learning in medicine: Why, how and when?, Information Fusion 66 (2021) 111–137.

[5] Z. Obermeyer, T. H. Lee, Lost in thought: the limits of the human mind and the future of medicine, The New England journal of medicine 377 (2017) 1209.

[6] L. Von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy, et al., Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems, IEEE Trans. on Knowledge and Data Engineering 35 (2021) 614–633.

[7] F. Leiser, S. Rank, M. Schmidt-Kraepelin, S. Thiebes, A. Sunyaev, Medical informed machine learning: A scoping review and future research directions, Artificial Intelligence in Medicine 145 (2023) 102676.

[8] S. Kierner, J. Kucharski, Z. Kierner, Taxonomy of hybrid architectures involving rule-based reasoning and machine learning in clinical decision systems: A scoping review, Journal of Biomedical Informatics (2023) 104428.
[9] C. Sirocchi, A. Bogliolo, S. Montagna, Medical-informed machine learning: integrating prior knowledge into medical decision systems, BMC Medical Informatics and Decision Making 24 (Suppl 4) (2024) 186. doi:https://doi.org/10.1186/s12911-024-02582-4.
[10] J. W. Smith, J. E. Everhart, W. Dickson, W. C. Knowler, R. S. Johannes, Using the adap learning algorithm to forecast the onset of diabetes mellitus, in: Proceedings of the annual symposium on computer application in medical care, American Medical Informatics Association, 1988, p. 261.
[11] E. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nature Medicine 25 (2019) 44–56. doi:10.1038/s41591-018-0300-7.
[12] P. Rajpurkar, E. Chen, O. Banerjee, E. J. Topol, AI in health and medicine, Nature Medicine 28 (2022) 31–38. doi:10.1038/s41591-021-01614-0.
[13] S. Benjamens, P. Dhunnoo, B. Meskó, The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database, NPJ digital medicine 3 (2020) 118.
[14] J. J. Clinton, K. McCormick, J. Besteman, Enhancing clinical practice: The role of practice guidelines., American Psychologist 49 (1994) 30.
[15] Z. Qian, W. Zame, L. Fleuren, P. Elbers, M. van der Schaar, Integrating expert odes into neural odes: pharmacology and disease progression, Advances in Neural Information Processing Systems 34 (2021) 11364–11383.
[16] C. C. Yang, Explainable artificial intelligence for predictive modeling in healthcare, Journal of healthcare informatics research 6 (2022) 228–239.
[17] Z. Obermeyer, T. H. Lee, Lost in thought — the limits of the human mind and the future of medicine, New England Journal of Medicine 377 (2017) 1209–1211.
[18] G. Ciatto, F. Sabbatini, A. Agiollo, M. Magnini, A. Omicini, Symbolic knowledge extraction and injection with sub-symbolic predictors: A systematic literature review, ACM Computing Surveys 56 (2024). doi:10.1145/3645103.
[19] M. Magnini, G. Ciatto, F. Cantürk, R. Aydoğan, A. Omicini, Symbolic knowledge extraction for explainable nutritional recommenders, Computer Methods and Programs in Biomedicine 235 (2023) 107536. doi:https://doi.org/10.1016/j.cmpb.2023.107536.
[20] M. Cranmer, A. Sanchez-Gonzalez, P. Battaglia, R. Xu, K. Cranmer, D. Spergel, S. Ho, Discovering symbolic models from deep learning with inductive biases, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.
[21] S. Montagna, S. Mariani, E. Gamberini, Augmenting bdi agency with a cognitive service: Architecture and validation in healthcare domain, Journal of Medical Systems 45 (2021) 103. URL: https://doi.org/10.1007/s10916-021-01780-1. doi:10.1007/s10916-021-01780-1.
[22] P. Maes, Agents that reduce work and information overload, Commun. ACM 37 (1994) 30–40. doi:10.1145/176789.176792.
[23] D. Calvaresi, D. Cesarini, P. Sernani, M. Marinoni, A. F. Dragoni, A. Sturm, Exploring the ambient assisted living domain: a systematic review, Journal of Ambient Intelligence and Humanized Computing 8 (2017) 239–257. doi:10.1007/s12652-016-0374-3.
[24] M. Bosello, A. Ricci, From programming agents to educating agents – a jason-based

framework for integrating learning in the development of cognitive agents, in: L. A. Dennis, R. H. Bordini, Y. Lespérance (Eds.), Engineering Multi-Agent Systems, Springer International Publishing, Cham, 2020, pp. 175–194.

[25] D. Singh, S. Sardiña, L. Padgham, G. James, Integrating learning into a BDI agent for environments with changing dynamics, in: T. Walsh (Ed.), IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011, IJCAI/AAAI, 2011, pp. 2525–2530. doi:10.5591/978-1-57735-516-8/IJCAI11-420.

[26] A. Guerra-Hernández, A. El Fallah-Seghrouchni, H. Soldano, Learning in bdi multi-agent systems, in: J. Dix, J. Leite (Eds.), Computational Logic in Multi-Agent Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 218–233.

[27] D. Rosaci, Cilios: Connectionist inductive learning and inter-ontology similarities for recommending information agents, Information Systems 32 (2007) 793–825. doi:https://doi.org/10.1016/j.is.2006.06.003.

[28] S. Montagna, S. Mariani, E. Gamberini, A. Ricci, F. Zambonelli, Complementing agents with cognitive services: A case study in healthcare, Journal of Medical Systems 44 (2020) 188. URL: https://doi.org/10.1007/s10916-020-01621-7. doi:10.1007/s10916-020-01621-7.

[29] M. E. Bratman, D. J. Israel, M. E. Pollack, Plans and resource-bounded practical reasoning, Computational Intelligence 4 (1988) 349–355.

[30] M. P. Georgeff, A. L. Lansky, Reactive reasoning and planning, in: Proceedings of the Sixth National Conference on Artificial Intelligence - Volume 2, AAAI'87, AAAI Press, 1987, pp. 677–682.

[31] H. B. Kibria, M. Nahiduzzaman, M. O. F. Goni, M. Ahsan, J. Haider, An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable ai, Sensors 22 (2022) 7268.

[32] G. Kunapuli, K. P. Bennett, A. Shabbeer, R. Maclin, J. Shavlik, Online knowledge-based support vector machines, in: Machine Learning and Knowledge Discovery in Databases: European Conference, 2010, Proceedings, Part II 21, Springer, 2010, pp. 145–161.

[33] L. Breiman, Classification and regression trees, Routledge, 2017.

[34] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, Symbolic knowledge extraction from opaque ML predictors in PSyKE: Platform design & experiments, Intelligenza Artificiale 16 (2022) 27–48.

[35] E. Denti, A. Omicini, A. Ricci, Multi-paradigm java–prolog integration in tuprolog, Science of Computer Programming 57 (2005) 217–250. doi:https://doi.org/10.1016/j.scico.2005.02.001.