Agents that argue and explain classifications of retinal conditions

Adrian Groza* · Liana Todorean ·
George Adrian Muntean · Simona Delia
Nicoara

Received: date / Accepted: date

Abstract

Purpose: Expertise for auditing AI systems in medical domain is only now being accumulated. Conformity assessment procedures will require AI systems: 1) to be transparent, 2) not to rely decisions solely on algorithms, or iii) to include safety assurance cases in the documentation to facilitate technical audit. We are interested here in obtaining transparency in the case of machine learning (ML) applied to classification of retina conditions. High performance metrics achieved using ML has become common practice. However, in the medical domain, algorithmic decisions need to be sustained by explanations. We aim at building a support tool for ophthalmologists able to: i) *explain algorithmic decision to the human agent* by automatically extracting rules from the ML learned models; ii) *include the ophthalmologist in the loop* by formalising expert rules and including the expert knowledge in the argumentation machinery; iii) *build safety cases* by creating assurance argument patterns for each diagnosis.

Methods: For the learning task, we used a dataset consisting of 699 OCT images: 126 Normal class, 210 with Diabetic Retinopathy (DR) and 363 with Age Related Macular Degeneration (AMD). The dataset contains patients from the Ophthalmology Department of the County Emergency Hospital of Cluj-Napoca. All ethical norms and procedures, including anonymisation, have been performed. We applied three machine learning algorithms: decision tree

A. Groza (corresponding author) and L. Todorean
Technical University of Cluj-Napoca, Cluj-Napoca, Romania
Tel.: +40-264-401200
Fax: +40-264592
E-mail: Adrian.Groza@cs.utcluj.ro,Todorean.Io.Liana@utcluj.didatec.ro

G. A. Muntean and S. D. Nicoara
Iuliu Hatieganu University of Medicine and Pharmacy, Cluj-Napoca, Romania
Tel.: +40-264-597-256
Fax: +40-264-597-257
E-mail: georgemuntean99@elearn.umfcluj.ro, stalu@umfcluj.ro

(DT), support vector machine (SVM) and artificial neural network (ANN). For each algorithm we automatically extract diagnosis rules. For formalising expert knowledge, we relied on the normative dataset [13]. For arguing between agents, we used the Jason multi-agent platform. We assume different knowledge base and reasoning capabilities for each agent. The agents have their own Optical Coherence Tomography (OCT) images on which they apply a distinct machine learning algorithm. The learned model is used to extract diagnosis rules. With distinct learned rules, the agents engage in an argumentative process. The resolution of the debate outputs a diagnosis that is then explained to the ophthalmologist, by means of assurance cases.

Results: For diagnosing the retina condition, our AI solution deals with the following three issues: First, the learned models are automatically translated into rules. These rules are then used to build an explanation by tracing the reasoning chain supporting the diagnosis. Hence, the proposed AI solution complies with the requirement that "algorithmic decision should be explained to the human agent". Second, the decision is not solely based on ML-algorithms. The proposed architecture includes expert knowledge. The diagnosis is taken based on exchanging arguments between ML-based algorithms and expert knowledge. The conflict resolution among arguments is verbalised, so that the ophthalmologist can supervise the diagnosis. Third, the assurance cases are generated to facilitate technical audit. The assurance cases structure the evidence among various safety goals such as: machine learning methodology, transparency, or data quality. For each dimension, the auditor can check the provided evidence against the current best practices or safety standards.

Conclusion: We developed a multi-agent system for retina conditions in which algorithmic decisions are sustained by explanations. The proposed tool goes behind most software in medical domain that focuses only on performance metrics. Our approach helps the technical auditor to approve software in the medical domain. Interleaving knowledge extracted from ML-models with expert knowledge is a step towards balancing the benefits of ML with explainability, aiming at engineering reliable medical applications.

Keywords Explainable artificial intelligence · argumentative agents · machine learning · AgentSpeak · retina conditions

Mathematics Subject Classification (2010) 68T42 · 68T05 · 68T30

1 Introduction

Patients expect physicians to comprehensibly explain decisions that have an impact on them. The same holds for physicians interacting with decision support systems, which need to explain their recommendations [10]. Explaining algorithmic decisions is not only a trust related issue, but also a legal one. For instance, article 22 of GDPR empowers human agents with the "right to demand an explanation of how AI system made a decision that affects them".

With the current surge of interest in Machine Learning (ML)-based medical software, explaining decisions made based on black-box ML models remains challenging [2]. For many domains [8,9], especially the medical one [17,23], there is need for algorithmic decisions should be sustained by explanations [16,24] and assurance cases. It has been argued that machine learning equals human-level capacity in medical diagnosis [15], while achieving high performance metrics has indeed become common practice [25]. In our view, this high performance obtained through deep learning models – which are black box models – rises at least three practical challenges: 1) How explanations can be provided to the users? 2) How assurance cases can be build for audit and safety approval? 3) How expert knowledge can be included in the loop? These challenges are not distinct, but rather interleave each other: if one extracts knowledge from the learned models, the reasoning can be easily traced to generate explanations and to build safety cases.

Machine learning uses various algorithms to analyse data. The problem is that, if data is analysed with a Decision Tree (DT) algorithm, data reveals a particular pattern. If the same data is analysed with Support Vector Machine (SVM) or Artificial Neural Network (ANN) algorithms, different patterns will be revealed. Which pattern is the correct one, becomes a challenging task. In our approach, each ML algorithm learns from the same data a distinct set of rules. Each rule represents an argument, and the agents exchange part of these arguments. The resolution of the debate forms a diagnosis that is then explained to the stakeholder, by showing the chain of supporting arguments and by building the assurance case [28].

We developed a diagnosis system for retinal conditions. We designed the system to meet the following requirements:

1. *To explain algorithmic decision to the human agent* by automatically extracting rules from the black box ML algorithms.
2. *To include the ophthalmologist in the loop* by formalising expert rules and including the expert knowledge in the argumentation machinery.
3. *To build safety cases* by creating assurance argument patterns for each diagnosis.

This design goes behind most software in medical domain that focuses only on performance metrics. Our approach helps the technical auditor to approve software in the medical domain.

We propose an argumentation-based decision function after receiving results from three ML classifiers and one *expert* agent. This *expert* agent contains rules manually formalised from normative data for retina, such as the data found by Invernizzi et al. [13]. Interleaving knowledge extracted from ML-models with expert knowledge is a step towards balancing the benefits of ML with explainability, aiming at engineering reliable medical applications.

We applied our machinery on distinguishing between three classes: normal retina, Diabetic Retinopathy (DR), and Age Related Macular Degeneration (AMD). We chose DR and AMD because both diseases: i) are highly prevalent; ii) produce changes in the macular area; iii) benefit from an early diagnosis

Fig. 1: Extracting rules from learned models

and treatment: if detected early, we could prevent severe vision loss; iv) modify the volume and the thickness of the macular retina.

2 Diagnosing by machine learning and expert knowledge

Retina conditions can be identified using Optical Coherence Tomography (OCT). From the OCT images, we extract the volume and thickness of the retina, defined as the space between the internal limiting membrane and Bruch's membrane. The segmentation of this space is made by the OCT software installed on our Heidelberg Spectralis device. We use an Early Treatment Diabetic Retinopathy Study (ETDRS) grid, which is built inside the OCT software [7]. The grid is made out of three concentric circles with 1, 3 and 5 mm radius. The grid is further divided into nine zones: one central zone c_0 in the 1 mm circle; four zones in the 3mm circle (n_1, s_1, t_1, i_1); and four zones in the 5mm circle (n_2, s_2, t_2, i_2). These nine zones are depicted in the top left of Figure 1. Here n stands for nasal, s for the superior zone, t for the temporal zone, and i for the inferior zone of the retina. The thickness and volume values inside the grid are automatically calculated, and the grid's position on the macula is manually adjusted by the OCT operator, so that the center of the three concentric circles is overlapped over the center of the fovea.

Fig. 2: Flowchart of the proposed method

Patients are classified in three classes: Age Related Macular Degeneration (AMD), Diabetic Retinopathy (DR), and Normal. For the experiments we used a dataset consisting of 699 images: 126 samples from Normal class, 210 from DR, and 363 from AMD. Each sample has 18 attributes: 9 thickness values and 9 volume values for each zone: $c_0, n_1, s_1, t_1, i_1, n_2, s_2, t_2, i_2$. The dataset contains patients from the Ophthalmology Department of the County Emergency Hospital of Cluj-Napoca, Romania. We included the following patients that visited our clinic and underwent OCT examination: 50 patients with AMD having at least one eye with stage 4 AREDS - exudative type; 5 patients with DR having at least one eye with macular edema; 3 patients with normal retinal architecture for both eyes. All ethical norms and procedures, including anonymisation, have been performed.

Our method for argumentation-based diagnosing has three steps: 1) Augmenting data and training; 2) Extracting rules; 3) Arguing, conflict resolution and explaining decision.

Augmenting data and training. In the first step, we start with an unbalanced dataset of patients with multiple visits: 50 with *AMD*, 5 with *DR* and 3 with *Normal* retina. Since this dataset is small for machine learning, we perform data augmentation. Data augmentation [27,29] is performed by changing thickness and volume with a random value between -0.001 and 0.001 for each zone, so the sample belongs to the same class. For each sample from *DR* and *Normal*, we generate 10 more samples with small variations. After augmentation, the resulted dataset of 699 images is more balanced: 126 *Normal*, 210 *DR*, and 363 *AMD*, with the class distribution 18% *Normal*, 30% *DR*, and 52% *AMD*.

The values for the volume and average thickness for each zone are directly extracted from the OCT images. The dataset was divided into training, validation and test set. The training set contains 60% from the input data, and it is used to learn the model. The validation set contains 20% from the input data, and it is used to tune the parameters of the model. The test set contains the remaining 20% and it is used to check the learned model against data never seen before.

We experimented with three algorithms: Decision Tree (DT), Support Vector Machine (SVM) and Artificial Neural Network (ANN) [22]. These algorithms illustrate three distinct approaches in the world of machine learning: DT is the master algorithm for the symbolists, SVM is the master algorithm for analogizers, while ANN is the favored approach for connectionists [11]. Since these algorithms follow different learning strategies, it fits our aim to show different perspectives on the given retinal case. For each algorithm there is an agent with the role of arguing for its classification decision.

Each internal node in a DT contains a test on a value of one of the input features. The branches of a node correspond to all possible values of the feature. The leaf nodes specify what class value to be returned by the decision tree function. The DT algorithm is a greedy one, since at each step it picks the feature with the highest importance. The importance of a feature can be measured in terms of information gain (i.e., expected reduction in entropy)

SVM has a different perspective on the task. It is a risk averse algorithm, since it finds linear separators that stay as far as possible from both classes. It therefore needs to compute distances between the feature vectors corresponding to the training instances. Since SVM works well when there is a clear margin of separation between classes, it will bring on the table strong arguments for cases in which the diagnosis is clear.

ANN has the ability to detect complex nonlinear relationships. Learning occurs by gradually adjusting the weights between neurons. During training, the error computing for each instance is used to adjust the weights: the error is distributed proportionally by the backpropagation mechanism to the neurons that have contributed to it. Deep learning architectures applied on large data would boost performance. Here, we kept the architecture simple (e.g. only two hidden layers), since we want to be able to break the black box model of ANN and to extract rules that will be presented to the human agent.

For DT, we used Gini impurity to measure the quality of split. For SVM, we used a 3-bin-discretizer, encoding ordinal and uniform strategy for discretisation. For SVM training [6], we used a linear penalty squared hinge loss, a tolerance for stopping criteria $1e-5$, and the regularisation parameter $C = 1.0$. For ANN [20], we used 2 hidden layers of 18 and 3 neurons, with the RELU activation function, ridge regression to avoid overfitting (i.e. L_2 since we have only 18 features to learn from), penalty parameter $\alpha=1e-5$, learning rate 0.0001 and tolerance of $1e-4$. Also, the Limited-memory BFGS optimiser [20] is appropriate for our small dataset of 699 images. The accuracy of the decision tree is 97%, neural network is 75%, while for SVM is 70%. Note that our focus is not on improving performance, but on explainability: how we can explain the algorithm-based diagnosis to the human agent.

Extracting rules. In the second step, we automatically extract rules from the learned models. For each of the nine zones, we use two predicates: $t(zone)$ for thickness, and $v(zone)$ for volume. The rules learned by the DT algorithm are easily extracted by traversing the learned tree. The challenge remains to extract rules from more black-box models like SVM and ANN. To be relevant for the ophthalmologists, the rules should contain only features from the input

Fig. 3: *DT*, *SVM* and *ANN* agents have rules automatically extracted from learned models. *Expert* agent has rules manually formalised from normative datasets on retina. The *master* agent handles conflict resolution among the arguments conveyed by *DT*, *SVM*, *ANN* and *expert*.

space, and not features from the learned model. With SVM and ANN, the learned parameters do not correspond to our predicates of interest $t(zone)$ and $v(zone)$. To transform the algorithm’s parameters space into the input space, we use the Rule Matrix tool [18].

Arguing, conflict resolution and explaining decision. Third, the multi-agent system consists of six agents (see Fig. 3): three agents have rules learned from machine learning: the *DT* agent, the *ANN* agent, and the *SVM* agent. The *expert* agent has rules manually formalised from normative datasets on retina. These four agents are called *argumentative agents*, since they exchange arguments based on their own perspective on the current patient. The *master* agent handles the conflict resolution among arguments and generates explanations. The *ophthalmologist* agent provides new cases for diagnosis. The agents were developed with the Jason platform and its AgentSpeak programming language [3].

We use three classes for retinal disease $\langle AMD, DR, Normal \rangle$ and three learning algorithms (LAs): a decision tree (DT), a support vector machine (SVM), and an artificial neural network (ANN). We also use expert knowledge E , hence $LA = \{DT, SVM, ANN, E\}$. From the learned models, we extract rules R_i with the structure:

$$R_i^{LA(m_j)} : Cond \mid Cond \otimes Cond \rightarrow^{support} \langle AMD, DR, Normal \rangle \quad (1)$$

Here m_j represents the performance metric of the learning algorithm LA, e.g. accuracy (a), precision (p), recall (r), F-measure (f). Each condition $Cond$ has the form:

$$Cond = t(zone) \oplus value \mid v(zone) \oplus value \mid feature(patient) \quad (2)$$

Table 1: Sample of rules extracted from the decision tree (DT)

$R_1^{DT(a=0.97)}:$	$t(s_1) \leq 0.35 \wedge v(s_1) \leq 0.51 \rightarrow^{69} \langle 1, 0, 0 \rangle$
$R_2^{DT(a=0.97)}:$	$t(s_1) \leq 0.35 \wedge v(s_1) > 0.51 \wedge v(c_0) \leq 0.22 \wedge v(n_1) \leq 0.52 \rightarrow^9 \langle 1, 0, 0 \rangle$
$R_3^{DT(a=0.97)}:$	$t(s_1) \leq 0.35 \wedge v(s_1) > 0.51 \wedge v(c_0) \leq 0.22 \wedge v(n_1) > 0.52 \wedge t(n_1) \leq 0.34 \wedge v(s_2) \leq 1.48 \rightarrow^2 \langle 1, 0, 0 \rangle$
$R_4^{DT(a=0.97)}:$	$t(s_1) \leq 0.35 \wedge v(s_1) > 0.51 \wedge v(c_0) \leq 0.22 \wedge v(n_1) > 0.52 \wedge t(n_1) \leq 0.34 \wedge v(s_2) > 1.48 \rightarrow^{37} \langle 1, 0, 0 \rangle$
$R_5^{DT(a=0.97)}:$	$t(s_1) \leq 0.35 \wedge v(s_1) > 0.51 \wedge v(c_0) \leq 0.22 \wedge v(n_1) > 0.52 \wedge t(n_1) > 0.34 \rightarrow^6 \langle 1, 0, 0 \rangle$
$R_6^{DT(a=0.97)}:$	$(s_1) \leq 0.35 \wedge v(s_1) > 0.51 \wedge v(c_0) > 0.22 \rightarrow^{25} \langle 1, 0, 0 \rangle$
$R_7^{DT(a=0.97)}:$	$t(s_1) > 0.35 \wedge v(s_1) \leq 0.57 \wedge t(s_2) \leq 0.3 \rightarrow^{10} \langle 1, 0, 0 \rangle$

where the operator $\oplus \in \{<, \leq, >, \geq, =\}$, while t is the thickness and v the volume of one retinal zone $c_0, n_1, s_1, t_1, i_1, n_2, s_2, t_2, i_2$. With $feature(patient)$, the rules can include any additional information on the current case (e.g. $male(patient)$). Several conditions can be linked by the logical operator $\otimes \in \{\vee, \wedge\}$. The output of each rule represents the distribution of probability over the three classes. Hence $AMD + DR + Normal = 1$. The number of cases supporting the rule is given by the *support* value above the implication operator.

3 Obtained rules

We obtained three sets of rules automatically extracted from the machine learning algorithms, and one set of rules manually formalised from normative datasets on retina.

3.1 Rules extracted from machine learning models

Table 1 lists part of the rules learned with the decision tree algorithm. For instance, the rule R_1^{DT} states that if the thickness of the zone s_1 is less than 0.35 (i.e. $t(s_1) \leq 0.35$), and the volume of the same zone s_1 is less than 0.51 (i.e. $v(s_1) < 0.51$) then the given retina has the *AMD* condition with 100% probability. Since the rule is valid for a number of 69 images in our dataset, its support is 69. The confidence of the rule is given here by the accuracy metric. During learning, the obtained accuracy for decision tree algorithm was $a=0.97$. Apart from the rules extracted from the decision tree, the DT algorithm also computes the most informative feature, which in our case was the s_1 zone. This means that the DT algorithm looks firstly at the s_1 zone to diagnose an OCT image.

We obtained six rules from the SVM model (see Table 2). Since the thickness of the n_2 zone appears in all six rules learned by SVM, this zone is considered by SVM as the most informative parameter. Note also that some retinal zones (e.g. i_2) do not appear in these rules, meaning that the SVM algorithm found no information within the zone i_2 . Note also that the

Table 2: Rules extracted from the SVM classifier

$R_1^{SVM(a=0.7)}$:	$t(n_2) > 0.45 \rightarrow \langle 0.0143, 0.5429, 0.4429 \rangle$
$R_2^{SVM(a=0.7)}$:	$t(n_2) \leq 0.45 \wedge t(t_2) > 0.41 \wedge v(n_2) < 2.41 \wedge v(n_2) > 1.94 \rightarrow \langle 0.0149, 0.5373, 0.4478 \rangle$
$R_3^{SVM(a=0.7)}$:	$t(n_2) \leq 0.45 \wedge (t(t_2) \leq 0.41 \vee v(n_2) > 2.41) \wedge t(c_0) < 0.093 \rightarrow \langle 0.2439, 0.0488, 0.7073 \rangle$
$R_4^{SVM(a=0.7)}$:	$t(n_2) \leq 0.45 \wedge (t(t_2) \leq 0.41 \vee v(n_2) > 2.41) \wedge t(c_0) > 0.093 \wedge v(t_2) \leq 1.83 \wedge t(n_2) \leq 0.36 \rightarrow \langle 0.0037, 0.0168, 0.9795 \rangle$
$R_5^{SVM(a=0.7)}$:	$t(n_2) \leq 0.45 \wedge (t(t_2) \leq 0.41 \vee v(n_2) > 2.41) \wedge t(c_0) > 0.093 \wedge (v(t_2) > 1.83 \vee t(n_2) > 0.36) \wedge t(i_1) > 0.42 \rightarrow \langle 0.0075, 0.0226, 0.9699 \rangle$
$R_6^{SVM(a=0.7)}$:	$\rightarrow \langle 0.0071, 0.2786, 0.7143 \rangle$

Table 3: Rules extracted from the ANN classifier

$R_1^{ANN(a=0.75)}$:	$v(t_2) \leq 1.28 \rightarrow \langle 0.0045, 0.0856, 0.9099 \rangle$
$R_2^{ANN(a=0.75)}$:	$v(t_2) > 1.28 \wedge t(t_1) \leq 0.35 \wedge (v(s_2) > 1.62 \vee v(s_2) \leq 1.86) \rightarrow \langle 0.0127, 0.7089, 0.2785 \rangle$
$R_3^{ANN(a=0.75)}$:	$v(t_2) > 1.28 \wedge (t(t_1) > 0.35 \vee (v(s_2) \leq 1.62 \vee v(s_2) > 1.86) \wedge (v(t_2) > 1.28 \wedge v(t_2) \leq 1.66) \rightarrow \langle 0.0041, 0.2992, 0.6967 \rangle$
$R_4^{ANN(a=0.75)}$:	$v(t_2) > 1.28 \wedge (t(t_1) > 0.35 \vee (v(s_2) \leq 1.62 \vee v(s_2) > 1.86) \wedge (v(t_2) \leq 1.28 \vee v(t_2) > 1.66) \wedge v(t_1) > 0.6 \wedge (t(t_2) > 0.31 \wedge t(t_2) \leq 0.46) \rightarrow \langle 0.0035, 0.6489, 0.3475 \rangle$
$R_5^{ANN(a=0.75)}$:	$v(t_2) > 1.28 \wedge (t(t_1) > 0.35 \vee (v(s_2) \leq 1.62 \vee v(s_2) > 1.86) \wedge (v(t_2) \leq 1.28 \vee v(t_2) > 1.66) \wedge v(t_1) > 0.6 \wedge (t(t_2) \leq 0.31 \vee t(t_2) > 0.46) \rightarrow \langle 0.0047, 0.9065, 0.0888 \rangle$

rules R_4^{SVM} and R_5^{SVM} are the most certain with respect to the *Normal* class (i.e. the probability is some 97% in both cases). The availability of such information is in line with the goal of Explainable AI (XAI), since the system is able to answer questions such as: when can I trust you? For instance, the SVM agent, will be reliable for patients whose OCT images activate the rules R_4^{SVM} and R_5^{SVM} . Note also that the rule R_6^{SVM} is a default rule, which is activated when the conditions of the other rules do not hold for the current OCT image.

The rules extracted from the ANN are listed in Table 3. For instance, the rule R_1^{ANN} states that if the volume of the t_2 zone is less than 1.28 then most probably (i.e. 0.9099), the retina is normal. The confidence of all ANN rules is given by the learning accuracy $a = 0.75$ for the ANN classifier.

3.2 Rules formalised from expert knowledge

To formalise the expert knowledge, we start from normative data for retinal thickness. Invernizzi et al. found that the thickness of retinal zones is influenced by gender, sex, and axial length [13]. In Table 4, we formalised their results as rules to infer normal retina. Retinal thickness was significantly higher in males. The $t(c_0)$ is positively correlated with the age. The mean retinal volume

```

2  true & S1t <= 0.35 & S1v <= 0.51                                %context
3  <- .send(agentMaster, tell, newDiagnosis(2, 100.00,             %actions
4  [[s1t, -1000.00, 0.35], [s1v, -1000.00, 0.51]])).
5
6  +!diagnose([S1t, S1v], 0)                                       %achieve goal
7  <- .send(agentMaster, tell, needs([s1t, s1v, c0v, n1v], 1)).%context

```

Listing 1: Two alternative plans for the argumentative agents

Algorithm 1: Updating diagnosis based on arguments from other agents

```

Result: keep (0) or change (1) diagnosis
 $\mathcal{R} \leftarrow \text{MatrixTool}(ML(\text{dataset}))$  - set of rules extracted from the learned model;
for  $r : a_i \rightarrow c \in \mathcal{R}$  do
   $\mathcal{A}_c^+$  - the set of supporting arguments for conclusion  $c$ ;
   $\mathcal{A}_c^-$  - the set of counter arguments for conclusion  $c$ ;
  if  $c = \text{new\_diagnosis}$  then
    if  $\forall a_i \text{ checked}(a_i)$  then
       $\mathcal{A}_{\text{new\_diagnosis}}^+ = \mathcal{A}_{\text{new\_diagnosis}}^+ \cup \{r\}$ ;
    else
       $\mathcal{A}_{\text{new\_diagnosis}}^- = \mathcal{A}_{\text{new\_diagnosis}}^- \cup \{r\}$ ;
  if  $c = \text{current\_diagnosis}$  then
    if  $\forall a_i \text{ checked}(a_i)$  then
       $\mathcal{A}_{\text{current\_diagnosis}}^+ = \mathcal{A}_{\text{current\_diagnosis}}^+ \cup \{r\}$ ;
    else
       $\mathcal{A}_{\text{current\_diagnosis}}^- = \mathcal{A}_{\text{current\_diagnosis}}^- \cup \{r\}$ ;
   $\text{Keep} \leftarrow |\mathcal{A}_{\text{new\_diagnosis}}^-| + |\mathcal{A}_{\text{old\_diagnosis}}^+|$ ;
   $\text{Change} \leftarrow |\mathcal{A}_{\text{old\_diagnosis}}^-| + |\mathcal{A}_{\text{new\_diagnosis}}^+|$ ;
  if  $\text{Keep} > \text{Change}$  then
    return 0
  else
    return 1

```

Second, the *assurance argumentative pattern* [21] is adapted here for a multi-agent system (Fig. 4). We represent the pattern with the Goal Structuring Notation (GSN) [14]. GSN is a description language consisting of six nodes: *goals* (what to assure), *contexts* (state, environment or conditions of the system), *strategies* (describe how to break down a goal into subgoals), *evidence* (assuring the goal can be reached), *monitoring* (representing evidence available at runtime), or *undeveloped* nodes (status of no evidence or monitoring supporting the goal). In Fig. 4, the *goals* are depicted with yellow rectangles, the *contexts* with rounded rectangles, the *solutions* with the blue circle, while the *strategies* with the parallelogram. The assurance argument points out pieces of evidence (the green shapes) for goals related to the machine learning methodology, the extracted rules, the conflict resolution or the data type used. One can see that the subgoal "Dataset is balanced" is an undeveloped node, signaled by the empty diamond. This empty diamond is an indicator for the ophthalmologist that the result is biased towards the most

Fig. 4: Assurance argument pattern for multi-agent system

represented class in the initial dataset (this class is *AMD*). Or, the auditor can signal that the 699 instances used here do not meet the current threshold for data quantity for the given task. Such safety cases are therefore helpful to structure evidence and therefore to facilitate technical audit of AI systems in medical domain.

4.2 Conflict resolution

The *Master* agent employs different conflict resolution strategies: *voting*, *weighed voting*, *strongest-argument*. These strategies take into consideration three components: number of supporting arguments (i.e. majority vote), confidence in the argument, or the performance metric of the learning process.

With *majority vote*, if three agents reached the same diagnosis and only the fourth one has a different vote, the diagnosis is given by the most common view on the case. The votes can be *weighted* by the confidence in the conveyed arguments. The weight of an argument is given by: 1) its probability distribution over the diagnosis variable $\langle AMD, DR, Normal \rangle$ and 2) by the performance metric achieved during learning. This probability distribution for each rule is extracted automatically from the learned model. For instance, in case of the decision tree algorithm, the nodes in the tree contain the number of samples in each class that satisfy all the conditions from the root to that node. The values from the leaves of the tree are used to compute the probability that the diagnosis is correct. If all instances belong to one class, then the

probability is 100%. If there are instances that belong to another class, then the probability decreases with each sample. In the case of the SVM and ANN algorithms, each classifier provides the probability for a sample to belong to each class.

As an example, assume that the following four arguments have all their premises satisfied:

$$\begin{array}{llll}
R_1^{DT(a=0.97)} & : & t(s_1) \leq 0.35 \wedge v(s_1) \leq 0.51 & \rightarrow^{69} \langle 1, 0, 0 \rangle \\
R_1^{SVM(a=0.7)} & : & t(n_2) > 0.45 & \rightarrow \langle 0.0143, 0.5429, 0.4429 \rangle \\
R_1^{ANN(a=0.75)} & : & v(t_2) \leq 1.28 & \rightarrow \langle 0.0045, 0.0856, 0.9099 \rangle \\
R_{11}^E & : & t(c_0) > 320 \wedge male(patient) & \rightarrow^{55} \langle 0, 1, 0 \rangle
\end{array}$$

Here, the *DT* agent argues in favor of *AMD* (recall the order of the diagnosis in the distribution probability vector $\langle AMD, DR, Normal \rangle$), with a certainty of 100%, accuracy 0.97 and support 60. The *SVM* agent argues in favor of *DR* with a certainty of 0.5429 and an accuracy of 0.7. The *ANN* agent argues in favor of *Normal* with a certainty of 0.9099 and an accuracy of 0.75, while the *expert* agent supports the *DR* diagnosis with 100% certainty based on 55 previous cases. Hence, there are two arguments for *DR* (*expert* and *SVM*), one argument for *Normal* (*ANN*) and one for *AMD* (*DT*). By weighting the certainty of the votes with the performance metric of the algorithms, we obtain $1 \times 0.97 = 0.97$ for *AMD*, $0.5429 \times 0.7 + 1 \times 1 = 1.38$ for *DR* and $0.9099 \times 0.75 = 0.68$ for *Normal*. Based on the weighting votes strategy the diagnosis is $\max(0.97, 1.3, 0.68) = 1.3$ for *DR*.

The *strongest-argument* strategy selects the diagnosis that is supported by the argument with the highest certainty weighted by the agent's performance metric. For instance, let one argument conveyed by the SVM agent and one argument by the ANN agent:

$$\begin{array}{ll}
R_1^{SVM(a=0.7)} : t(n_2) > 0.45 \rightarrow \langle 0.0143, 0.5429, 0.4429 \rangle \\
R_1^{ANN(a=0.75)} : v(t_2) \leq 1.28 \rightarrow \langle 0.0045, 0.0856, 0.9099 \rangle
\end{array}$$

Here, the SVM agent supports the diagnosis *DR* with $0.7 \times 0.5429 = 0.38$. Differently, the ANN agent supports the diagnosis *Normal* with $0.75 \times 0.9099 = 0.68$. Assuming these are the only two arguments, the ANN's argument wins under the strongest-argument strategy. The ophthalmologist is the one responsible to select the conflict resolution strategy.

4.3 Explaining diagnosis

The master agent explains the diagnosis to the ophthalmologist by means of arguments obtained from the arguing agents. These rule-based arguments are verbalised, as Table 5 bears out. Verbalising rules is based in our approach on linguistic patterns, such as: *The patient has the diagnosis D because the volume/avg thickness value for retina zone Zn is larger/smaller than value X.*

[Master]	Final diagnosis is <i>Diabetic Retinopathy</i> Agent <i>DT</i> is 100% sure, and <i>DT</i> 's accuracy is 0.96 Agent <i>SVM</i> is 97% sure, and <i>SVM</i> 's accuracy is 0.75 Agent <i>ANN</i> is conveyed a 95.79% sure, and <i>ANN</i> 's accuracy is 0.95 Agent <i>E</i> had no arguments
[Master]	Diagnosis <i>Diabetic Retinopathy</i> was chosen because: the thickness value in t_1 zone is smaller than 0.34 and the thickness value in t_2 zone is smaller than 0.3 and the thickness value in s_2 zone is greater than 0.3 and the volume value in s_1 zone is smaller than 0.58 and the thickness value in s_1 zone is greater than 0.35.

Table 5: Generating explanations in natural language for the ophthalmologist

Fig. 5: The ophthalmologist signals increasing thickness and other bio-markers (left). Thickness values for each retina zones (right)

The master agent displays also the confidence of the supporting arguments conveyed by the agent. For instance, "DT agent is 100% sure" for the diagnosis *Diabetic Retinopathy*, means that the *DT* agent conveyed a supporting rule with the diagnosis vector $\langle 0, 1, 0 \rangle$. Similarly, "SVM agent is 97% sure", means that its vector can be $\langle 0.02, 0.97, 0.01 \rangle$, where the probability 97% is attached to the *DR* diagnosis. Additionally, the master agent displays the global performance metrics for the learned models, e.g. accuracy is 0.6 in case of decision tree learning. In case of conflicting arguments, the master agent presents the supporting arguments for a diagnosis and generates the assurance argument template.

4.4 Introducing the ophthalmologist in the loop

The opinion of the ophthalmologist is presented here along with the results obtained from agents for a patient diagnosed with diabetic retinopathy (Fig. 5, left). Taking into account the last two rules from the Table 4, we consider that thickness of c_0 above 320 mm for men and 305 mm for women indicates the presence of increased macular thickness related to DR. The average retinal thickness in the central zone for this patient is 420mm, therefore this patient might be affected by diabetic macular edema (Fig. 5 right). Analysing the cross-sectional central image, the normal retinal morphology seems to be

modified, with increased thickness, cystoid fluid-filled spaces in the outer retinal zones (red arrowheads), hard exudates (green arrowheads) and disruption of the external limiting membrane and the photo-receptor layers (yellow arrowhead), all these being consistent with diabetic macular edema.

5 Discussions and Related Work

Expertise for auditing AI systems in medical domain is only now being accumulated. Through the New Legislative Framework (NLU), UE recommends the technical auditors of AI systems to use expertise from established organisation such as agencies for medical devices. According to the new proposal of the EU [1], in each member state, a regulatory agency for artificial intelligence is planned. Such AI agency is envisaged to cooperate and benefit from the experience of existing bodies, for instance when deploying AI software on medical devices. Conformity assessment procedures will require AI systems: 1) to be transparent, 2) not to rely decisions solely on algorithms, or iii) to include safety assurance cases in the documentation to facilitate technical audit.

For diagnosis of the retina condition, our AI solution deals with the above three aspects as follows: First, the learned models are automatically translated into rules. These rules are then used to build an explanation by tracing the reasoning chain supporting the diagnosis. Hence, the proposed AI solution complies with the requirement that "algorithmic decision should be explained to the human agent". Second, the decision is not solely based on ML-algorithms. The proposed architecture includes expert knowledge. The diagnosis is taken based on exchanging arguments between ML-based algorithms and expert knowledge. The conflict resolution among arguments is verbalised, so that the ophthalmologist can supervise the diagnosis. Third, the assurance cases are generated to facilitate technical audit. The assurance cases structure the evidence among various safety goals such as: machine learning methodology, transparency, or data quality. For each dimension, the auditor can check the provided evidence against the current best practices or safety standards.

We proposed here an architecture for explaining classification of retina conditions. Each module of this architecture can be further improved: First, the set of training algorithms that we used (DT, SVM, or ANN) can be extended. For each new learning algorithm, an arguing agent will be enacted. Its arguments are processed by the argumentation machinery which takes a decision and explain it. This flexibility of the multi-agent system to accommodate new agents without architectural alteration of the system is important given that new ML-based are developing fast. Note that, for each new learned model, one needs to include also the mechanism to extract rules. Second, the quantity of data used can be improved. For this aim, one can benefit from i) European Health Data Space initiative that aims to overcome the current underuse of medical data available [1], or ii) from more advanced techniques for augmenting medical data [29]. In this line, it has been argued that data augmentation is more important than learning algorithms for retinal vessel segmentation [27].

Third, the safety cases should be better generated based on some certification procedure. The challenge is that these certification processes, including the corresponding certification centers and regulatory sandboxes, are only now developing [1].

The advantage of argumentative agents is that they can easily interact with each other and with an external user. They react to changes in the environment or to new knowledge and they are more suitable for an interactive system.

Hernandez et al. have used Jason Induction of Logical Decision Trees (JILDT) [12] to create logical decision trees using AgentSpeak plans. The focus is on developing agents that learn while they are executing plans. We are interested here in the reverse process. Instead of building decision trees based on plans, we train first various classifiers with available data and then we generate plans from the learned models.

Vadet et al. have developed a multi-agent military health System in which the diagnosis is distributed using a Belief Desire Intention agent-based model [26]. The system has an autonomous sensor network controlled by agents. The agents have common knowledge that includes diagnosis thresholds and costs for each rescue mission.

A diagnosis system used in a clinical environment should have assurance for its performance. Picardi et al. have designing an assurance argument pattern [21] that can be used for ML decision making. After the models are created using training and validation data sets, their performance is measured on the test data set. The performance gives the level of confidence in the ML model.

Our conflict resolution strategy reassembles with voting methods from ensemble learning. The advantage of our approach over ensemble learning is that the agents use strategies to accept arguments. Plans for accepting an argument are changed depending on the context and based on the knowledge of the agent. The behavior of agents is guided by changes in their plans and belief base. An agent can use different plans according to its own knowledge. Also, we include agents with expert rules that are not learned from models.

6 Conclusion

We developed here a multi-agent system for diagnosis of retinal conditions. The system diagnoses and explains three retina conditions (*AMD*, *DR*, and *Normal*) using 18 parameters extracted from OCT images. One advantage of our approach is the interleaving of arguments from expert knowledge with arguments extracted from machine learning algorithms. The learned arguments were automatically formalised as rules, in order to avoid black box based decisions and to ensure transparency. Moreover, the multi-agent system proposed here automatically generates assurance cases. These assurance cases combined with transparency of our multi-agent system are a step towards balancing the benefits of machine learning with explainability for engineering reliable medical application.

Declarations

Conflict of interest. The authors declare they have no conflict of interest.

Availability of data and material (data transparency). Please contact Simona Delia Nicoara for the OCT images dataset.

Code availability. <https://github.com/lianatoderean/diagnosissys>

Acknowledgments. Authors would like to thank Anca Marginean and the anonymous reviewers for their valuable comments and suggestions.

Authors contribution. A. Groza: Conceptualization, methodology, Writing - original draft, Writing review & editing; L. Toderean: Software, Investigation, Writing - original draft; G. A. Muntean: Investigation, Writing original draft; S. D. Nicoara: Supervision, Resources, Writing original draft.

References

1. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Tech. rep., European Commission (2021)
2. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **6**, 52138–52160 (2018)
3. Bordini, R.H., Hübner, J.F., Wooldridge, M.: Programming multi-agent systems in AgentSpeak using Jason, vol. 8. John Wiley & Sons (2007)
4. Chalam, K.V., Bressler, S.B., Edwards, A.R., Berger, B.B., Bressler, N.M., Glassman, A.R., Grover, S., Gupta, S.K., Nielsen, J.S.: Retinal thickness in people with diabetes and minimal or no diabetic retinopathy: Heidelberg spectralis optical coherence tomography. *Investigative ophthalmology & visual science* **53**(13), 8154–8161 (2012)
5. Chan, A., Duker, J.S., Ko, T.H., Fujimoto, J.G., Schuman, J.S.: Normal macular thickness measurements in healthy eyes using stratus optical coherence tomography. *Archives of ophthalmology* **124**(2), 193–198 (2006)
6. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* **2**(3), 1–27 (2011)
7. Coscas, G., Lupidi, M., Coscas, F.: Heidelberg spectralis optical coherence tomography angiography: technical aspects. *OCT Angiography in Retinal and Macular Diseases* **56**, 1–5 (2016)
8. Čyras, K., Birch, D., Guo, Y., Toni, F., Dulay, R., Turvey, S., Greenberg, D., Hapuarachchi, T.: Explanations by arbitrated argumentative dispute. *Expert Systems with Applications* **127**, 141–156 (2019)
9. Čyras, K., Letsios, D., Misener, R., Toni, F.: Argumentation for explainable scheduling. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 2752–2759 (2019)
10. De Graaf, M.M., Malle, B.F.: How people explain action (and autonomous intelligent systems should too). In: *2017 AAAI Fall Symposium Series* (2017)
11. Domingos, P.: The master algorithm: How the quest for the ultimate learning machine will remake our world. Basic Books (2015)
12. Guerra-Hernández, A., González-Alarcón, C.A., El FallahSeghrouchni, A.: Jason induction of logical decision trees: a learning library and its application to commitment. In: *Mexican International Conference on Artificial Intelligence*, pp. 374–385. Springer (2010)
13. Invernizzi, A., Pellegrini, M., Acquistapace, A., Benatti, E., Erba, S., Cozzi, M., Cigada, M., Viola, F., Gillies, M., Staurengi, G.: Normative data for retinal-layer thickness maps generated by spectral-domain oct in a white population. *Ophthalmology Retina* **2**(8), 808–815 (2018)
14. Kelly, T., Weaver, R.: The goal structuring notation—a safety argument notation. In: *Proceedings of the dependable systems and networks 2004 workshop on assurance cases*, p. 6. Citeseer (2004)

15. Liu, X., Faes, L., Kale, A.U., Wagner, S.K., Fu, D.J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., et al.: A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health* **1**(6), e271–e297 (2019)
16. Marginean, A., Groza, A., Nicoara, S.D., Muntean, G., Slavescu, R.R., Letia, I.A.: Towards balancing the complexity of convolutional neural network with the role of optical coherence tomography in retinal conditions. In: 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 475–482. IEEE (2019)
17. Memari, N., Saripan, M.I.B., Mashohor, S., Moghbel, M., et al.: Retinal blood vessel segmentation by using matched filtering and fuzzy c-means clustering with integrated level set method for diabetic retinopathy assessment. *Journal of Medical and Biological Engineering* **39**(5), 713–731 (2019)
18. Ming, Y., Qu, H., Bertini, E.: Rulematrix: Visualizing and understanding classifiers with rules. *IEEE transactions on visualization and computer graphics* **25**(1), 342–352 (2018)
19. Nieves-Moreno, M., Martínez-de-la Casa, J.M., Cifuentes-Canorea, P., Sastre-Ibáñez, M., Santos-Bueso, E., Sáenz-Francés, F., Morales-Fernández, L., García-Feijoó, J.: Normative database for separate inner retinal layers thickness using spectral domain optical coherence tomography in caucasian population. *PLoS One* **12**(7), e0180450 (2017)
20. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)
21. Picardi, C., Hawkins, R., Paterson, C., Habli, I.: A pattern for arguing the assurance of machine learning in medical diagnosis systems. In: International Conference on Computer Safety, Reliability, and Security, pp. 165–179. Springer (2019)
22. Russell, S., Norvig, P.: Artificial intelligence-a modern approach 4th ed. Pearson Education UK (2020)
23. Sassoon, I., Kökciyan, N., Sklar, E., Parsons, S.: Explainable argumentation for wellness consultation. In: Int. Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, pp. 186–202. Springer (2019)
24. Sendi, N., Abchiche-Mimouni, N., Zehraoui, F.: A new transparent ensemble method based on deep learning. *Procedia Computer Science* **159**, 271–280 (2019)
25. Thulkar, D., Daruwala, R., Sardar, N.: An integrated system for detection exudates and severity quantification for diabetic macular edema. *Journal of Medical and Biological Engineering* **40**(6), 798–820 (2020)
26. van Voorst Vader, P.J.: Multi-Agent Based Military Health System for the Future Battlefield. Master Thesis (2018)
27. Wang, Z., Lin, J., Wang, R., Zheng, W.: Data augmentation is more important than model architectures for retinal vessel segmentation. In: Proceedings of the 2019 International Conference on Intelligent Medicine and Health, pp. 48–52 (2019)
28. Zeng, Z., Fan, X., Miao, C., Leung, C., Jih, C.J., Soon, O.Y.: Context-based and explainable decision making with argumentation. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, pp. 1114–1122. International Foundation for Autonomous Agents and Multiagent Systems (2018)
29. Zhang, C., Tavanapong, W., Wong, J., de Groen, P.C., Oh, J.: Real data augmentation for medical image classification. In: Intravascular imaging and computer assisted stenting, and large-scale annotation of biomedical data and expert label synthesis, pp. 67–76. Springer (2017)