Review

Semi-Automatic Adaptation of Diagnostic Rules in the Case-Based Reasoning Process

Ľudmila Pusztová *[ID], **František Babič** [ID] and **Ján Paralič** [ID]

Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 9, 042 00 Košice, Slovakia; frantisek.babic@tuke.sk (F.B.); jan.paralic@tuke.sk (J.P.)
* Correspondence: ludmila.pusztova.2@tuke.sk

Abstract: The paper presents a new approach to effectively support the adaptation phases in the case-based reasoning (CBR) process. The use of the CBR approach in DSS (Decision Support Systems) can help the doctors better understand existing knowledge and make personalized decisions. CBR simulates human thinking by reusing previous solutions applied to past similar cases to solve new ones. The proposed method improves the most challenging part of the CBR process, the adaptation phase. It provides diagnostic suggestions together with explanations in the form of decision rules so that the physician can more easily decide on a new patient's diagnosis. We experimentally tested and verified our semi-automatic adaptation method through medical data representing patients with cardiovascular disease. At first, the most appropriate diagnostics is presented to the doctor as the most relevant diagnostic paths, i.e., rules—extracted from a decision tree model. The generated rules are based on existing patient records available for the analysis. Next, the doctor can consider these results in two ways. If the selected diagnostic path entirely covers the actual new case, she can apply the proposed diagnostic path to diagnose the new case. Otherwise, our system automatically suggests the minimal rules' modification alternatives to cover the new case. The doctor decides if the suggested modifications can be accepted or not.

Keywords: medical diagnostics; Case-Based reasoning; rules adaptation; cardiovascular disease

1. Introduction

Machine learning promises significant advancement in the healthcare domain up to personalized medicine [1]. Currently, the most often used are predictive models, trained on various sorts of medical data. Predictions include the use of advanced analytical technology, modeling, machine learning, and data mining algorithms, running simulations, and pattern searching. Some machine learning models provide excellent predictive accuracy but do not offer understandable insights about their predictions. This is a severe problem in the medical domain. Explaining data patterns is useful in every way to understand and trust the system and its predictions. Data scientists interpret large amounts of data, use analytical algorithms, and convey knowledge through visual and textual descriptions to help people understand details.

Today, healthcare providers generate and collect large amounts of data and gradually deploy analytical solutions of various types. This situation requires effective data processing and analytics with results in a simple, understandable form for target users represented, e.g., by general practitioners, doctors, specialists, or researchers. It is not easy for a doctor to consider a large amount of data when diagnosing or determining the right treatment. Due to its complexity, this process is prone to errors and widespread misdiagnosis. There are many cases in which the symptoms and diseases overlap and may be characteristic of several diseases. According to the World Health Organization (WHO), up to 5% of outpatients were misdiagnosed in 2015, which is a very worrying result. Thus, the correct diagnosis of a disease or condition relies on the availability of all the necessary

information. The volume, complexity, and dynamics of this situation challenge doctors and other health professionals. The number of symptoms is non-specific and variable depending on the case; many diagnostic tests are expensive, irregular, and sometimes do not give the correct answer. Also, physicians usually rely on their cognitive abilities and experience in diagnosing similar or identical symptoms. Given the growing diversity of treatments and medicines, the speed of data collection, the availability of diagnostic tests, and the differences of opinion between doctors, the use of a diagnostic decision support system can help.

Traditional decision support models use rules extracted from the historical data to evaluate new cases. However, what is to be done if the new case does not meet any of the existing rules? The typical approach represents the re-generation of all rules based on the extended data sample and requires the time and effort of a data or knowledge engineer who may not be available at the time. It is not ideal in clinical practice, where new patients or the same patients with different medical conditions come every day. This problem is partially solved by case-based reasoning method, which uses the knowledge and solutions used in past cases based on their similarity to the new case.

The case-based reasoning (CBR) method predicts future situations based on similar historical conditions. This approach itself is not entirely new, but after reading a number of studies dealing with case-based reasoning [2–8] and its applications in medical data [9–21], we found that this area still offers several research problems that have not yet received sufficient attention, or a suitable solution has not previously been proposed and tested [9]. One of them is a semi-automatic adaptation of decision models using expert knowledge. In medicine, the adaptation phase of the CBR is a significant challenge, which is avoided by several systems using the CBR method. This fact motivated us to propose an approach to solving this problem without a need to regenerate the decision models for each new case. We published the initial idea in our work "How to improve the adaptation phase of the CBR in the medical domain" [22], presented at the International Cross-Domain Conference for Machine Learning and Knowledge Extraction in 2019.

In this paper, we aimed to create a new mechanism for the semi-automatic adaptation of the case base in the CBR cycle. The logic in the background is still the same; the character of the input data is essential. We describe a pilot case of the data representing patients with cardiovascular diseases. We created a decision support system to test our mechanism using this data. Since we used the rule based CBR system, we chose the decision trees method to generate the decision models as basis for the case base. We believe that this approach will increase the accuracy of cardiovascular disease's diagnostic process, choose the right treatment method, or identify potentially useful and, at first glance, hidden knowledge.

The paper is organized as follows. The first section describes our motivation and identified challenges in this field. The second section presents the methodology, specifically, a description of the case-based reasoning method, the decision tree algorithm, and a brief introduction to decision support systems. The last two sections are devoted to the developed decision support system, integrating our innovated CBR method into this system, and its experimental evaluation on a cardiovascular disease data sample.

2. Materials and Methods

This section briefly presents basic principles of clinical decision support systems, case-based reasoning, and decision tree machine learning model, which are the tree main approaches and methods used in the proposed solution.

2.1. Decision Support System (DSS)

DSS is an interactive and flexible computer system that helps subjects make decisions and solve various types of problems. There are many types of DSS, but for the medical field, the most well-known are knowledge based DSS, also called expert systems (ES). These systems are equipped with the knowledge of an expert a specialist in a specific area (the knowledge base), and an inference mechanism able to derive valid conclusions based

on the input information and the knowledge base. As a result, ES can propose solutions similar as the human expert would do.

In recent years, the use of DSS in the healthcare domain has increased. The systems assist medical staff in day-to-day decision-making processes such as diagnosing one or more diseases, proposing appropriate treatment, diagnosing the absolute absence of some illness, or looking for their causes. Older expert systems were designed to handle only one problem [23–25], such as heart disorders [26]. Current expert systems already cover a wide range of medical issues, from diagnosis and response to test results to prescribing drugs for various diseases [27].

The first medical expert systems have existed since 1960 [23,28,29] and are still being developed with gradual improvements for today. Typical representatives of these systems are DENDRAL [24]—its primary goal was to study the formation of hypotheses and discoveries in science, which resulted in a specific task: to assist chemists identifying unknown organic molecules by analyzing their spectra using knowledge from chemistry. MYCIN [23]—its main task was to help doctors identify bacteria causing serious infectious diseases and to recommend treatment with antibiotics at a dose adjusted according to the patient's body weight (the name MYCIN derived from antibiotics with the suffix "-mycin"). MOLGEN [25]—the best-known ES in genetics, provides an intelligence consultant to a molecular geneticist in designing experiments, including DNA treatment. INTERNIS-1 [30]—is an educational project developed at the University of Pittsburgh in the early 1970s. This system aimed to assist healthcare professionals in the diagnosis of diseases in general internal medicine. The Expert Disease Program [26] focused on treating heart disease. It allowed the physician to write-in information about the patient's history, physical examination, and laboratory tests, and then generate a detailed explanation of the differential diagnoses indicating the clinical data that support each diagnosis of heart disease. PUFF [31]—expert system PUFF (Pulmonary Function System) diagnoses the results of a pulmonary function (PF) test, such as lung volume, the patient's ability to move air to and from the lungs, as well as the lung's ability to get oxygen to the blood and carbon dioxide. The result is a set of interpreted patient statements and diagnoses. Other expert systems include, for example, ONCOCIN (aimed at treating patients with chemotherapy), or TRAUMA Advisor, whose primary role is to help clinicians in rural areas in the diagnosis and treatment of trauma patients with head injuries. The work of Singla et al. [32] provides a brief overview of medical diagnostic expert systems based on already existing studies.
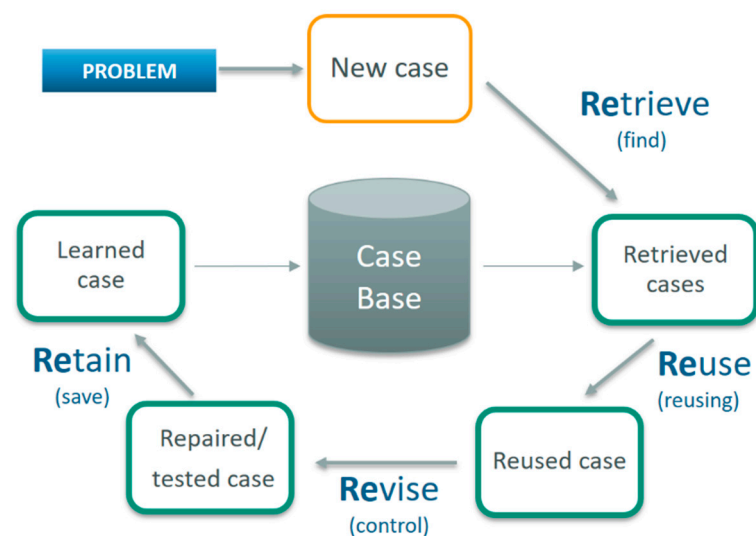
2.2. Case-Based Reasoning (CBR)

Case-based reasoning has become a more well-known approach based on solving new problems using solutions of previous similar problems. Memory, learning, planning, and problem resolution are essential features of the CBR approach, providing the basis for new intelligent computing technology. This process involves adapting used solutions to meet the new requirements and using historical cases to explain the new solutions. The CBR process gained a lot of attention because the basic idea of reusing experience in solving previous problems imitates human thinking. Relying on previous experience is an intuitive and commonly used way of solving problems. In the late 20th century, researchers examined people's ability to solve problems. They found that most people gather solutions to issues based on historical experience with similar situations. Unlike other approaches, which rely only on general knowledge of the problem area, CBR can use specific knowledge about previous problem situations [21], as also confirmed by the authors of [3].

According to Kolonder [4], the CBR cycle consists of two main tasks. At first, the most similar case is retrieved and selected from the knowledge base. The individual cases must be compared to find the most similar ones to the new. If the knowledge base is relatively small, it is possible to perform sequential retrieval or use classification algorithms. The second task is to adapt the existing solution to match the current case and store it in the case database. If the new and retrieved cases are similar, a simple transfer of the

solution is most appropriate. But if it is not the case, the adaptation is needed, which is a very complicated process. It has not yet been published a general adaptation method, and this step is entirely domain-dependent for the best of our knowledge. In this paper, we proposed an original semi-automatic adaptation model.

The typical CBR life cycle consists of the RETRIEVE phase (searching for and retrieving the most similar cases), REUSE (reusing the proposed solution), REVISE (checking the proposed solution), and RETAIN (preserving) the learned case for future use. The REUSE and REVISE steps many researchers combine into a single phase called ADAPT (adaptation).

The existing literature contains several models of operation of this cycle, such as the Hunt model, the Allen model, the model by Kolodner and Leake [5], and the R4 model designed by Aamodt and Plaza [5]. Model R4 is one of the most widely used and is at the highest level of generality. It can be represented by a schematic cycle containing "four REs," as shown in Figure 1. The case database includes experiences that have solved problems in the past.

Figure 1. Case-based reasoning (CBR) cycle designed by Aamodt and Plaza.

Each step in this figure has its place and importance for the successful functioning of the CBR system.

RETRIEVE phase—finding and retrieving the most similar cases. According to several authors [7,8], this phase is one of the most important because it forms the basis for the system's overall work. The retrieving includes finding cases within those cases that are most similar to the currently examined new case. Techniques such as nearest neighbor search, inductive approaches, knowledge approaches, and validated search are the most used for retrieving in this phase.

REUSE and REVISE phases—reuse and control of the proposed solution. In many practical applications, it is often difficult to distinguish between these two steps. Many researchers combine them into a single-phase called ADAPT [1]. In the early 1990s, CBR systems focused only on the retrieval step. The study of various aspects of adaptation only started after this period [10]. Most progress has also been made in the CBR cycle, retrieve, and reuse step [11]. In the reuse phase, progress has been made depending on the system's purpose, i.e., diagnosis, classification, learning, and planning (such as therapy support). In terms of diagnostics and classification, most systems rely on adaptation methods, which consist of copying the solution of the most similar case or a combination thereof, i.e., reuse of the solution [12].

RETAIN phase—preservation of the learned case for future use. After reusing and checking the proposed solution, the next step in the CBR cycle is to save the new (just resolved) case. There are many approaches to achieving this goal. Many systems only store a

previous problem solution, while some systems also store a new case solution. In many cases, this process leads to uncontrolled growth of the case base, which causes a deterioration in system performance in terms of speed, as mentioned in the work of Lawanna and Daengdej [33].

The most critical limitation in the use and implementation of CBR systems is the adaptation phase problem. D'Aquin et al. [13] argue that the adaptation process is relatively demanding because it must address the lack of relevant information about the case, the applicability, and the consequences of the decision. Other limitations include, for example, the complex generalization of similar cases and the processing of noisy data.

CBR in the medical domain is currently an active area of research. The base of medical knowledge is constantly changing; sometimes there is more than one solution, and doctors have different approaches and treatments. This area is expected to grow because the health sector is continually expanding and is increasingly accepting decision support systems. The very fact that the methodology of CBR systems is very similar to the physician's thought process suggests the successful use of CBR in the medical domain [14]. According to work by Choudhury [1], CBR systems in the medical field have a wide range of applications. Most of these systems are developed specifically to address a disease; most systems function as prototypes and not as a final product (as mentioned by Blanco in work [15]). Such systems require a human expert to interpret the result. Another visible trend is the successful hybridization of CBR with various computational methods.

We also used the CBR method in our work [34], which worked only as a retrieval system. The system allowed the user to find the most similar historical patients to a new patient and suggested the most likely potential outcome of an examination called coronary angiography. It also provided useful data visualizations through tables and graphs for the cardiologist who is responsible for the final decision on recommending a new patient for coronary angiography or not.

Some systems avoid entirely automatic adaptation, so they only work as search engines. However, systems that support the adaptation phase often have a reasoning mechanism (CBR) associated with rule based reasoning (RBR) or various computational methods (classification tree, neural networks, etc.). The problem of adaptation, therefore, also occurs in the medical domain. In the adaptation process, it is necessary to process the lack of information about patients, which is traditionally done manually by experts in the field. As a result of these challenge, most of the progress made in CBR medical systems focus mainly on the search phase, and the adaptation phase is therefore limited to planning tasks [12]. No general models for adaptation have been developed yet because they depend primarily on the domain's characteristics and application. Systems in the medical field that apply the full CBR method (retrieve, reuse, revision, retain) are still very seldom.

Some authors tried to solve the problem of adaptation in the medical domain. The CASEY medical expert system [16] used rule-based domain theory. Some of the newer systems successfully perform adaptation using computational techniques, e.g., eXiT * CBR.v2 [17] revises and reuses cases using genetic algorithms; EquiVox developed by Henriet et al. [18] performs adaptation using artificial neural networks. Sharaf-El-Deen et al. [21] and Saraiva et al. [19] decided to solve adaptation by creating a CBR system that integrated CBR and RBR. They made a hybrid CBR model for medical diagnostic procedures that automatically applied the adaptation process using adaptation rules. Salem and Bagoury [20] proposed a hybrid adaptation model combining transformation and hierarchical adaptation techniques with artificial neural networks and specific factors to diagnose thyroid cancer. Zubi and Saad [35] used combined data mining techniques with neural networks to diagnose lung cancer early. Keles et al. [36] used neuro-fuzzy rules to diagnose breast cancer, while Shafar-El-Deen and Moawad [21] introduced a hybrid approach that also combined CBR and RBR reasoning. The major drawback of existing approaches is the lack of their ability to provide explanations (particularly in the case of hybrid approaches making use of black box type of artificial intelligence methods). Moreover, it is usually only the new case which is retained for future learning and improving of the system. In this paper we

propose a new adaptation method, which has strong explanation capabilities because of using decision rules. Also, it also learns the new knowledge acquired when solving the new case in form of rule modifications, which is a more general form of knowledge than just on the case level.

2.3. Decision Tree

Decision trees are among the most widely used data analysis techniques in medicine because of their straightforward interpretation and understanding. The result is presented in a simple, understandable graphic form like a tree. Each structure starts at the root of the tree and goes through the individual nodes to the leaf node, determining the instance's classification. Internal nodes specify a test performed on the selected instance attribute, with each possible test result represented by one branch. In such a way a decision path from the root to a specific leaf node represents one particular decision rule. Decision trees are usually presented also in a graphical representation.

Authors Sharma and Om [37], Wang et al. [38], and Zolbanin et al. [39] applied the decision tree method to examine medical data, the rules of which they used to predict the disease and evaluate them for accuracy. We use a similar approach in our research group; we use the decision tree method to support the diagnosis of Mild cognitive impairment, Metabolic syndrome [40–42], or Parkinson's disease [43]. However, in the work presented here we use it differently. Decision paths are used as a form of summarized knowledge extracted from previous cases. Moreover, decision rules corresponding to those decision paths are used as a basis for the adaptation method. We use in our implementation the rpart package in R.

Typical representatives of the decision rules generation are C4.5 (can work with missing values, uses trimming, uses numeric attributes, branching criterion is proportional information gain [44]); C5.0 (improved algorithm C4.5, better work with memory, ability to work with continuous and discrete values, the criterion of branching is information gain and entropy [45]). Rpart (Recursive PARTitioning) is an R implementation of the main ideas from well-known CART (Classification and Regression Tree) algorithm. It is based on a recursive division of the data set; the subsets created during the division are further subdivided until a predetermined termination criterion is reached [45]. Exporting rules [46] using this algorithm is very popular, as it provides the ability to display the generated rules (as a plot), to create a list of all rules and a list of sub-rules in the form of tables.

3. Results and Discussion

As mentioned before, we proposed a semi-automatic adaptation method of the existing case base according to a new case and expert's decision based on automatically proposed decision rule modifications. Our motivation was to prevent continuously re-generating decision models using machine learning algorithms. Finally, in this study we integrated the innovative CBR approach into our DSS and experimentally tested on available medical samples about cardiovascular disease patients.

This section will briefly introduce our approach (Figure 2), published in one of our previous papers [22].
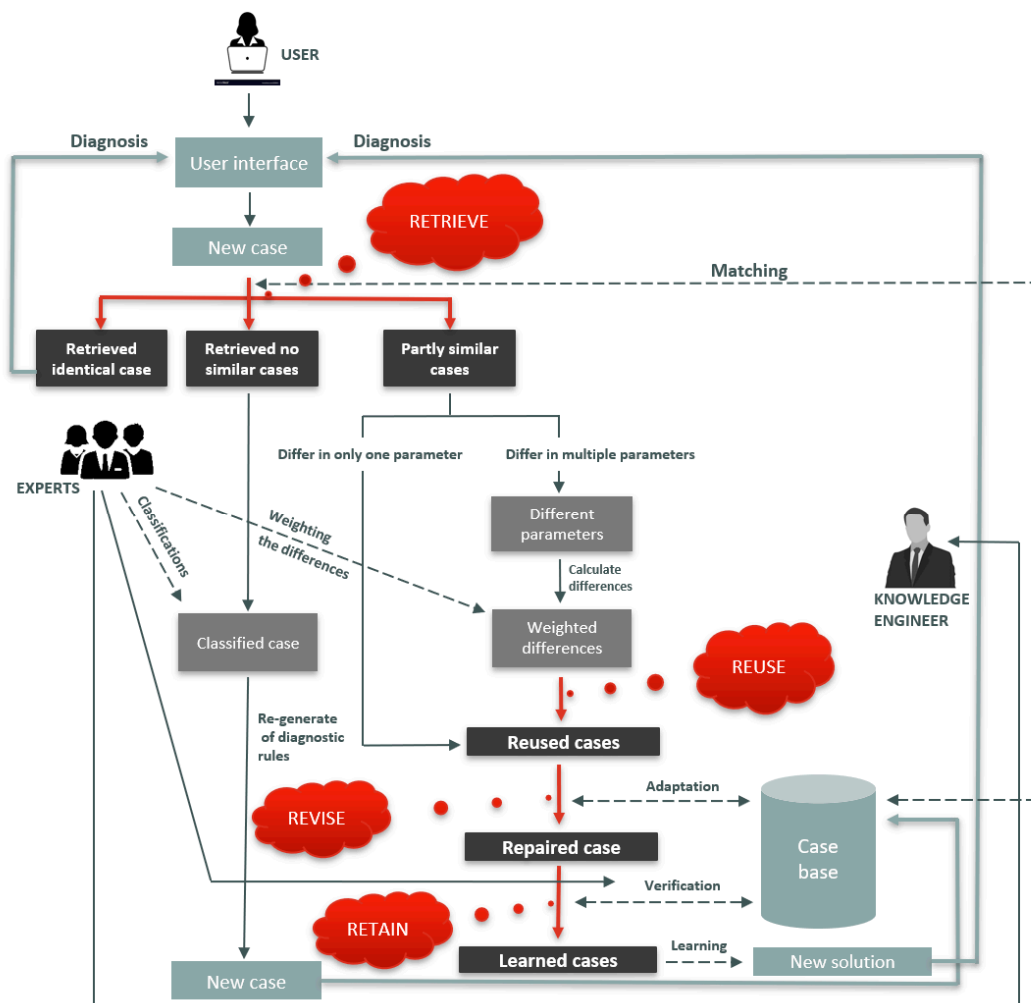
Figure 2. CBR cycle with semi-automatic adaptation of the stored cases.

The cycle starts with information provided about a new patient (case). The user will add information about this patient by means of the user interface in the predefined format containing all necessary characteristics (input variables). In the RETRIEVE step, the new case is compared with existing historical cases stored in the case base using inference mechanism (IM). The cases are stored as IF-THEN rules extracted from the decision model generated by the decision trees algorithm described in previous section. The comparison is determined by calculating the distances between the cases using original similarity metric based on decision (sub)rules transcripts. The comparison can result in one of the following three alternatives:

1. The IM finds a perfectly fitting decision path for the new case. The value of the new case's target attribute will be the same as corresponding decision rule suggests. The system will return the diagnosis to the user.
2. The IM does not find any at least partially relevant decision rule, i.e., all existing cases stored in the case base differ significantly from the new one. This situation requires the intervention of an expert who classifies the new case (determines the value of the target attribute). Then, the system will generate the decision rules again from the extended data set.
3. The IM returns a list of partially similar decision rules with different target attribute values. The CBR cycle will proceed with the steps REUSE, REVISE, and RETAIN.

Before the cycle continues to these steps, it is necessary to examine and determine the differences between the new and similar cases. Two situations can occur:

- The cases differ in only one condition (parameter) on the left side of the rule. The expert may consider modifying this rule (ADAPT phase), check the proposed classification of the new case according to this rule (REVISE phase), and then the new case will be classified according to the value of the target attribute of this rule (REUSE phase). The new case can be saved/not saved (RETAIN phase) in the case base according to the user's decision. If the system finds several cases that differ from the new case in only one condition, the expert decides on the rule containing the difference in the less severe attributes. It is also possible to assign weights to the difference according to the expert's decision.

- The cases differ in several conditions (parameters) on the left side of the rule. Initially, a list of different conditions (parameters) is identified. The difference between a similar case and a new case is calculated using the similarity and score metrics. In this case, an expert can also assign the scales according to their importance to specific differences. The expert determines the value of the new case's target attribute, while our system automatically proposes the rules with the highest similarity and the lowest score to the new case (ADAPT, REUSE, REVISE phase), which can help the expert to decide. As needed, it can adapt the rules (adjust boundaries or change the target value), which can also be saved and will be immediately available for the next iteration of adding a new patient. The new case can be saved/not saved in the case database (RETAIN phase).

The DSS software implementation consists of two modules (Figure 3) and is realized within the R programming language as an interactive web application.

First, we created module 1 (upper part in Figure 3) providing typical analytical features like data visualization in a table or in suitable graphical form like boxplot for numerical variables and histogram for categorical ones. Also, it runs the decision tree algorithm to generate the models.

The user can extract a set of rules and sub-rules from the model. One rule can be formed by one or more sub-rules connected by the logical operator "AND". A set of sub-rules and rules can be downloaded in the form of a table. In this way, a case base contains a list of sub-rules. The number of sub-rules depends on the size of the input data file entering the 1st module. Inputs to this module are entered only by the knowledge engineer.

The second module (lower part in Figure 3) represents an interface to the decision support system using our case-based reasoning cycle. This module uses the most successful model and case base from module 1. Most of the results in this model are displayed in tabular form for easy understanding. For larger data, clustering can be applied before generating the decision models. This will reduce the search space and make more tailored results available to users.
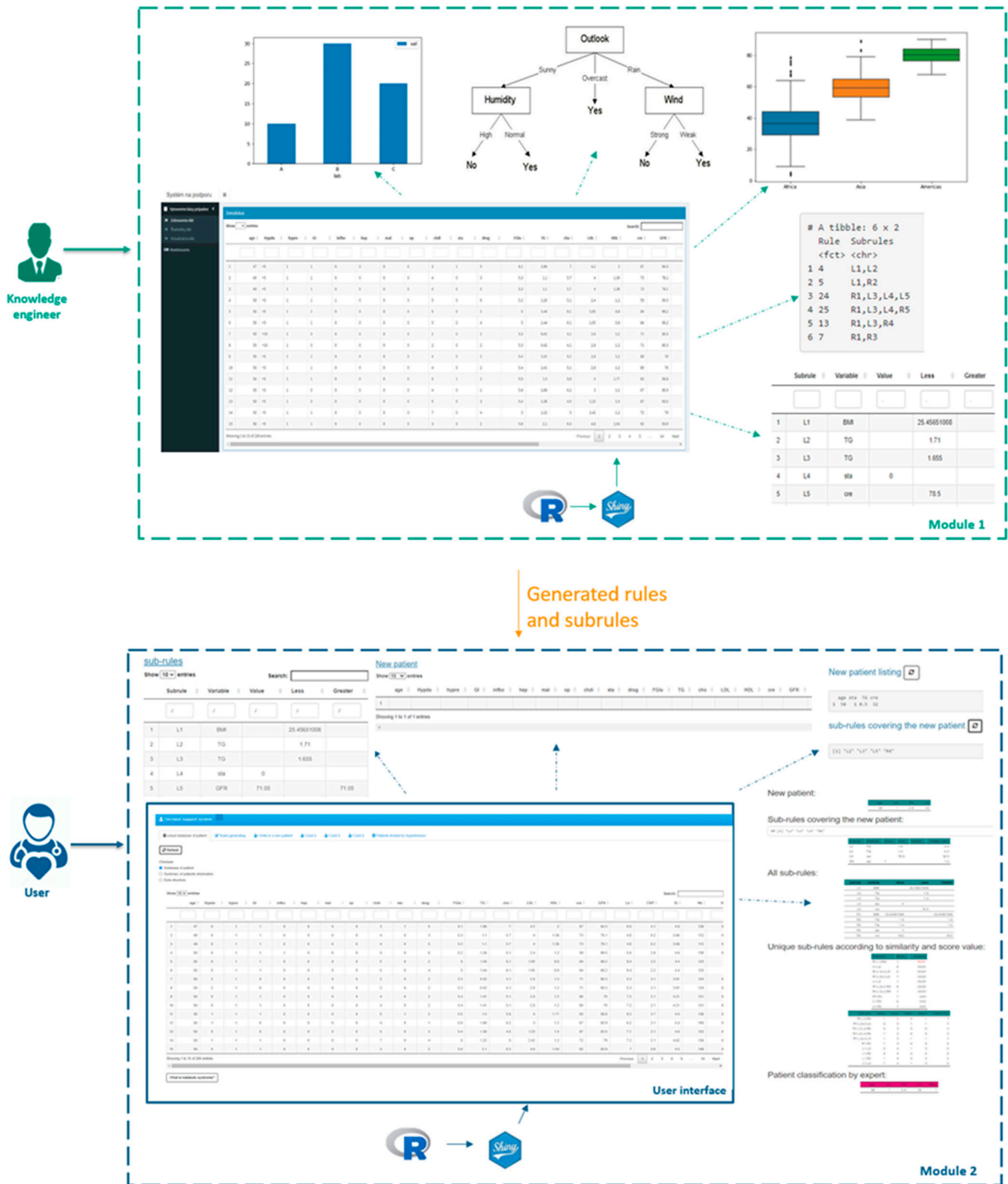
Figure 3. The implementation of updated CBR cycle into our decision support system.

3.1. Decision Support Systems Function

We generated several decision models in module 1 with a holdout method (60/40, 70/30, 80/20, 90/10) and the following algorithms: C5.0, rpart, and C4.5. If the expert has available a large volume of input records, he can apply clustering at first to divide the patients into some groups based on their similarity. This step can help the expert

to provide more personalized diagnostics. Next, the system offers the best decision tree model, and the corresponding decision rules are generated. The graphical form of the decision tree is often quite problematic for practical use. There is a high probability that the tree will be broadly branched and not readable. We prevent this situation with a table form visualization. The sub-rules generation is one of the essential functions of the DSS, without which it would not be possible to continue working with the system. In the Figure 4, we can see 10 sub-rules exported from the best decision model.

List of all sub-rules

	Subrule	Variable	Value	Less	Greater
1	L1	BMI		25.45651008	
2	L2	TG		1.71	
3	L3	TG		1.655	
4	L4	sta	0		
5	L5	cre		78.5	
6	R1	BMI	25.45651008		25.45651008
7	R2	TG	1.71		1.71
8	R3	TG	1.655		1.655
9	R4	sta	1		
10	R5	cre	78.5		78.5

Showing 1 to 10 of 10 entries

Figure 4. The Decision Support System (DSS) visualizes the subrules generated from the decision tree as a table (L—left branch, R—right branch).

The first column in Figure 4 contains the given sub-rule designation in Li or Ri's format, depending on whether it is the left branch in the tree (L) or the right one (R). The second column contains a specific attribute representing the corresponding intermediate node from the decision tree. The next three columns describe the interval condition for a sub-rule of the corresponding decision variable. Based on the corresponding value condition, it can have a value "equal to", "less than", or "greater than" the value in the respective row and column. If there is value in more than one column in one row, it means a conjunction of these corresponding intervals.

The generated sub-rules are not only used to determine the "coverage" of a new patient, but the cardiologist can also see the current database of patients extended by a column, expressing which rule and conjunctions of the sub-rules cover each historical patient (past cases) in this database. One can see it in Figure 5. The first columns represent the data samples attributes, and the last column Subrules represents the sub-rule/conjunctions of the sub-rules covering a particular record (case).

The system also provides the results of matching between the new case and the most relevant decision (sub)rules in form of a table (see Figure 6). The first column in Figure 6 entitled "similarity" represents the similarity value between the new case and historical cases, which can take values between 0 and 100. In means the overlap between the set of subrules valid for the new case and the set of subrules of the compared historical case in percent. The second column represents the score value calculated using the Levenshtein distance [47]. It is a real value representing the minimum number of operations that we need to do in order to make the historical case (in form of covering subrules set encoded as string) match the current new case (string constructed in the same way as for historical case). The number and type of necessary operations contain the next three columns—insert, delete, and substitution. For example, if the score is three, it can represent a combination of one insert, one delete, and one substitution operations, but also two insert, one delete and zero substitution operations.

	age	sex	trestbps
611	67	0	120
612	58	0	100
613	47	0	110
614	52	0	125
615	62	0	128
616	57	0	110
617	64	0	128
618	43	0	115
619	42	1	120
620	67	1	106

thal	num	Rule	Subrules
0	1	45	L1,R2,R3,L4,R5
2	1	3	R1
0	1	4	L1,L2
2	1	3	R1
0	0	10	L1,R2,L3
1	0	3	R1
2	0	3	R1
0	0	4	L1,L2
0	0	4	L1,L2
0	0	10	L1,R2,L3

Figure 5. The DSS visualizes a patients table with coverage by sub-rules.

similarity	score	insert	delete	substitution
50	3	1	0	2
50	3	1	2	0

Figure 6. The DSS calculates the similarity between new and historical patient.

Identified matches are automatically sorted based on their similarity and score. The most interesting case from the adaptation point of view is partial similarity, when expert is presented by the DSS with the most promising subrules Expert will consider respective subrules and can adapt them according to the new case. In the following two tables in Figure 7, we can see such an adaptation of 4 sub-rules with the same attribute.

	Subrule	Variable	Value	Less	Greater
1	L1	BMI		25.45651008	
2	L2	TG		1.71	
3	L3	TG		1.655	
4	L4	sta	0		
5	L5	cre		78.5	
6	R1	BMI	25.45651008		25.45651008
7	R2	TG	1.71		1.71
8	R3	TG	1.655		1.655
9	R4	sta	1		
10	R5	cre	78.5		78.5

Showing 1 to 10 of 10 entries — Previous 1 Next

	Subrule	Variable	Value	Less	Greater
1	L1	BMI		25.45651008	
2	L2	TG		1.9	
3	L3	TG		1.8	
4	L4	sta	0		
5	L5	cre		78.5	
6	R1	BMI	25.45651008		25.45651008
7	R2	TG	1.9		1.9
8	R3	TG	1.8		1.8
9	R4	sta	1		
10	R5	cre	78.5		78.5

Showing 1 to 10 of 10 entries — Previous 1 Next

Figure 7. The DSS visualizes the sub-rules adaptation process (left—previous version, right—adapted one).

The left table represents the sub-rules before adaptation. The expert decided to adapt the TG attribute in which he changed the boundaries from TG < 1.71 to TG < 1.9; from TG < 1.655 to TG < 1.8; from TG ≥ 1.71 to TG ≥ 1.9 and from TG ≥ 1.655 to TG ≥ 1.8. Boundary values can be changed by double-clicking on a given value in the table. Subsequently, it is important to store this table with adapted sub-rules.

The system stores adapted sub-rues in the knowledge base, ensuring up-do-date diagnostics for new patients, without the need to regenerate decision models.

We used the publicly available Heart Disease Dataset (HDD) for experimental evaluation and verification, representing anonymized medical data on suffering/not suffering from heart disease. The system's primary goal is to help the cardiologist (the target user) determine a new patient's final diagnosis based on the most similar known patterns from the case base. Also, he can perform adaptation steps of stored rules, according to his knowledge and actual conditions.

The proposed system is not intended for working with raw medical data. Its use includes a team consisting not only of medical but also analytical experts. They will prepare the input data in the required form.

3.2. Data Description and Preparation

HDD (https://www.kaggle.com/redwankarimsony/heart-disease-data) was created by integrating 4 data samples collected from four different clinics, with the individual data samples containing the same attributes but additional records. The personal records represent medical information about people with a positive/negative diagnosis of heart disease. We started with data preparation. First, it was necessary to join the individual data samples into one common database, which contained 924 records and 14 attributes (including the target). In the preprocessing phase we did not fil out the missing values because of the medical domain specificity avoiding incorrect values imputation. We aggregated the initial target values into binary target attribute, i.e., 1 to 4 were replaced by 1. After that, the database contained 415 patients without heart disease and 509 patients with heart disease. Table 1 describes all the input variables.

Table 1. Description of individual attributes and their current values after the transformation.

Attribute Name	Attribute Description	Attribute Type	Range of Values
age	Patient's age in years	Numerical	<28, 77>
sex	Patient's sex	Binary	(0) men, (1) women
Cp	Chest pain type	Categorical	0 (asymptomatic patient), 1 (atypical angina), 2 (non-anginal pain), 3 (typical angina)
Trestbps	Patient's Blood pressure value measured after admission to the hospital (this is the systolic (upper) blood pressure given in units of mm Hg)	Numerical	<80, 200>
Chol	Blood cholesterol level	Numerical	<85, 603>
Fbs	Fasting blood sugar (the value indicates whether the patient has an increased blood glucose level after 8 h of fasting, when the patient ate nothing and drank only clean water)	Binary	0 (patient has a blood sugar level of less than 120 mg/dl and therefore does not suspect diabetes), 1 (patient has a blood sugar level greater than 120 mg/dl and therefore has suspected diabetes)
Restecg	Resting EKG results	Categorical	0 (no abnormality was found in the results), 1 (T wave inversion was found in the results and/or ST segment elevation or depression was greater than 0.05 mV), 2 (the patient was diagnosed with left ventricular hypertrophy, ie enlarged and densified walls of the left ventricle of the heart)

Table 1. *Cont.*

Attribute Name	Attribute Description	Attribute Type	Range of Values
Thalach	Maximum heart rate achieved (maximum number of beats the heart can perform in one minute)	Numerical	<60 do>
Exang	Occurrence of chest pain during exercise	Binary	0 (no), 1 (yes)
Oldpeak	ST depression induced by exercise relative to rest	Numerical	<−2.9, 6>
Slope	The slope of the peak exercise ST segment	Categorical	0 (downsloping), 1 (flat), 2 (upsloping)
Ca	Number of major vessels (0-3) colored by flourosopy	Categorical	0,1,2,3
Thal	Thallium heart scan	Categorical	0 (the patient did not have a heart problem), 1 (patient overcame a heart problem), 2 (blood flow is blocked)
Num	Diagnosis of heart disease (angiographic disease status)	Binary	0 (the patient does not have heart disease), 1 (the patient has heart disease)

We tested both modules and the whole DSS using the prepared data sample.

3.3. Experiments

We generated several decision models using selected algorithms like ctree, C5.0, rpart, and C4.5 with the results presented in Tables 2 and 3.

**Table 2.** The models' accuracy without the clustering method.

Ratio	DT Type	%	Ratio	DT Type	%	Ratio	DT Type	%	Ratio	DT Type	%
60/40			70/30			80/20			90/10		
	ctree	80.6%		ctree	80.4%		ctree	82.2%		ctree	83.3%
	C5.0	80.6%		C5.0	78.5%		C5.0	76.2%		C5.0	84.5%
	rpart	86.2%		rpart	**86.6%**		rpart	85.4%		rpart	84.9%
	C4.5			C4.5			C4.5			C4.5	

Table 3. The models' accuracy with the clustering method.

Number of Clusters	60/40	70/30	80/20	90/10
2	85.1%	86.5%	87%	87.3%
3	87.7%	86.3%	88.2%	**89.4%**
4	85.1%	88.2%	85.8%	86.1%

The clustering functionality is primarily designed for large datasets. The HDD is a relatively small sample; therefore, we used the best model from Table 2. It is visualized in Figure 8, where we can see that number of rules is 6. Among the most important decision attributes used for division in not leaf nodes were used attributes thal, age, slope, exang, and sex.

From this model we extracted the rules and sub-rules. The total number of sub-rules is 12, and their combinations create a specific rule. For example, the first rule consists of sub-rules L1, R2, and L3, see Figure 9. The sub-rules mainly contained categorical variables like thal, slope, sex, exang, and numerical one age.
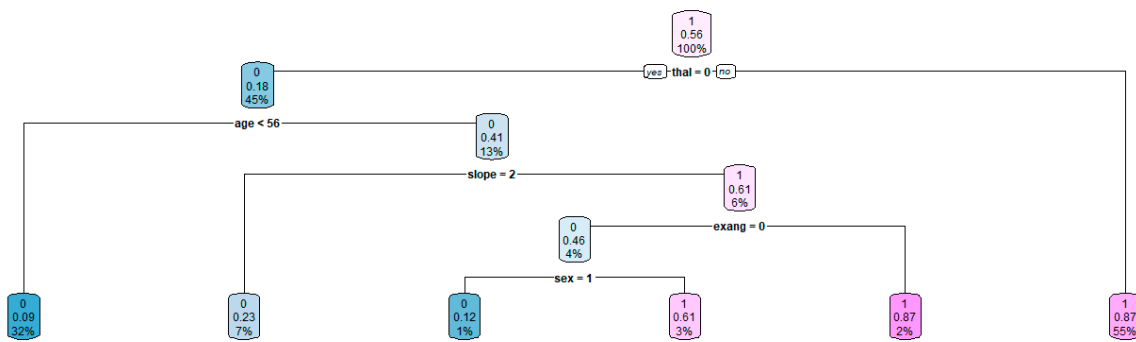
Figure 8. The DSS visualizes best decision model based on the accuracy metrics. The color and shade of nodes signalize the prevalent class in that node (blue for patients without and pink for patients with heart disease). The box's first value represents the target attribute value—1 (patient suffers from heart disease), 0 (patient doesn't suffer from heart disease). The second value is the probability of occurrence of heart disease. The third value is what percentage of patients in each branch who suffer/not suffer from heart disease.

	Subrule	Variable	Value	Less	Greater
	All	All	All	All	All
1	L1	thal	0		
2	L2	age		55.5	
3	L3	slope	2		
4	L4	exang	0		
5	L5	sex	1		
6	R1	thal	1		
7	R1	thal	2		
8	R2	age	55.5		55.5
9	R3	slope	0		
10	R3	slope	1		
11	R4	exang	1		
12	R5	sex	0		

Showing 1 to 12 of 12 entries   Previous 1 Next

Figure 9. The DSS visualizes a list of sub-rules generated from the best model.

The other rules were formed by following conjunctions: 2(L1,R2,R3,R4), 3(R1), 4(L1,L2), 5(L1,R2,R3,L4,L5), and 6(L1,R2,R3,L4,R5). Most patients were covered by rule 4, i.e., sub-rule L1 and L2.

We tested the similarity calculation for three new patients.

Patient no.1

```
age exang thal
1  56    1    0
```

DSS found these sub-rules covering the patient 1:

```
[1] "L1" "R2" "R4"
```

The system calculated the highest similarity for L1, R2, R3, R4, namely 75% and score value 1 (operation of insertion in the sub-rule R3). This condition covered 15 historical

patients. The database also contained patients with a 66.67% similarity and 0% covered by sub-rule R1 with a score value of 3.

Tables in Figure 10 have been shown to the cardiologist to support his decision, suggesting a new patient's target attribute classification as 1 (the patient has heart disease). This situation would occur if the cardiologist did not decide to modify the rules. If the cardiologist decides to adapt sub-rules L2 and R2 with the attribute "age", the result is visualized on the right side. The boundaries were changed from age < 55.5 to age < 57 and from age $\geq$ 55.5 to age $\geq$ 57. Subsequently, the adapted rules were stored.

	Subrule	Variable	Value	Less	Greater		Subrule	Variable	Value	Less	Greater
	/	/	/	/	/		/	/	/	/	/
1	L1	thal	0			1	L1	thal	0		
2	L2	age		55.5		2	L2	age		57	
3	L3	slope	2			3	L3	slope	2		
4	L4	exang	0			4	L4	exang	0		
5	L5	sex	1			5	L5	sex	1		
6	R1	thal	1			6	R1	thal	1		
7	R1	thal	2			7	R1	thal	2		
8	R2	age	55.5		55.5	8	R2	age	57		57
9	R3	slope	0			9	R3	slope	0		
10	R3	slope	1			10	R3	slope	1		
11	R4	exang	1			11	R4	exang	1		
12	R5	sex	0			12	R5	sex	0		
Showing 1 to 12 of 12 entries			Previous	1	Next	Showing 1 to 12 of 12 entries			Previous	1	Next

Figure 10. The DSS visualizes the sub-rules of the best model before (**left**) and after adaptation (**right**).

The system stored the adapted sub-rules in the case base, and sub-rules L1, L2, and R4 covered the new patient. The highest similarity 100% was calculated for the previous patients covered by the L1 and L2 rules with target diagnosis 0. The second option represents positive diagnosis (1) with similarity 50%, score 2 (substitution and insertion), and the combination L1, R2, R3, and R4. Also, the target diagnosis 0 was covered by the similarity of 33,3%, and sub-rules L1, R2, L3.

Patient no.2

```
age thalach oldpeak thal
1 45    173      5    0
```

DSS found these sub-rules covering the patient 2:

```
[1] "L1" "L2"
```

The system calculated the similarity as 100%, and the score value as 0. This condition covered 207 historical patients. The remaining patients already had a similarity of less than 35% and score value from 2 to 4. However, the cardiologist could classify the new patient according to identical patients whose target value was 0 (the patient has no heart disease) in 88%.

Patient no.3

```
    sex chol thalach oldpeak
1   1  170    173      5
```

DSS found these sub-rules covering the patient 3:

```
[1] "L5"
```

The system calculated only minimal similarity, i.e., the cases were very different from the new patient. The highest similarity was 20%, with the score value 4 in patients covered by sub-rules L1, R2, R3, L4, and L5. However, in 95%, the similarity was equal to 0%, and the score value was 1 in patients covered by sub-rule R1.

In this case, the DSS provided additional tables to the cardiologist with unique patients, but the target attribute value's prevalence was approximately the same. The cardiologist had to classify the new patient based on his knowledge. The classified patient is stored in the database and available for the re-generation of the decision models.

We designed the system based on our previous or existing collaborations with various medical experts from Faculty of Medicine Osijek, University JJ Strossmayer, Osijek, Croatia, and East Slovak Institute of Cardiovascular Diseases, Košice Slovakia. We have experience with different medical data analysis tasks like a prediction of Metabolic Syndrome or Mild Cognitive Impairment through suitable machine learning methods, or an analysis of the associations among physical frailty or cognitive impairment through clustering. The performed experiments and obtained experiences helped us to understand how to effectively provide analytical results to the end-users and save the time for the patients. Due to the complicated COVID-19 pandemic situation, we focused primarily on the Minimum Viable Product (prototype) to test its potential. Subsequently, we will work on its completion and adaptation to specific deployment conditions.

4. Conclusions

This paper aims to improve the existing opportunities for medical diagnostics towards a better understanding of analytical results. This step forward covers explainable decision models generated by machine learning methods in processing, analyzing, and using effectively available medical data. It is crucial for future personalized medicine [1]. The case-based reasoning process imitates human problem solving, which uses the experience of previous cases based on similarity. One of the current research tasks in this domain is to design suitable adaptation method of the stored cases based for the new case. We proposed a new semi-automatic approach to the CBR adaptation phase and integrated it into our Decision Support System. The method is based on the automatic extraction of the most relevant decision rules and their fragments suitable for the new case. Moreover, it offers comprehensive options for a user who is not a passive recipient of the information and can actively influence the diagnostic procedure based on his knowledge and experience. We successfully evaluated this approach using the free available Heart Disease Dataset. Our future work includes a more complex evaluation in cooperation with doctors or cardiologists, as part of implementing one of our national projects.

Author Contributions: F.B. and J.P. created a concept of study. Ľ.P. and F.B. designed a new approach for rules adaptation. Ľ.P. created a CDSS. Ľ.P. analyzed the data and performed experiments in CDSS. F.B. and J.P. evaluated experiments and stated the conclusion of the work. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fröhlich, H.; Balling, R.; Beerenwinkel, N.; Kohlbacher, O.; Kumar, S.; Lengauer, T.; Maathuis, M.H.; Moreau, Y.; Murphy, S.A.; Przytycka, T.M.; et al. From hype to reality: Data science enabling personalized medicine. *BMC Med.* **2018**, *16*, 150. [CrossRef] [PubMed]
2. Sharaf-El-Deen, D.A.; Moawad, A.F.; Khalifa, M.E. A New Hybrid Case-Based Reasoning Approach for Medical Diagnosis Systems. *J. Med. Syst.* **2014**, *38*, 1–11. [CrossRef] [PubMed]
3. Althoff, K.D.; Barletta, B.; Manago, M.; Auriol, E. *A Review of Industrial Case-Based Reasoning Tools*; AI Intelligence: Oxford, UK, 2001.
4. Kolodner, J. *Case-Based Reasoning (Morgan Kaufmann Series in Representation and Reasoning)*; Morgan Kaufmann: San Francisco, CA, USA, 1993.
5. Leake, D.B. *Case-Based Reasoning: Experiences, Lessons and Future Directions*; MIT Press: Cambridge, MA, USA, 1996.
6. Aamodt, A.; Plaza, E. Case-based reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Commun.* **1994**, *7*, 39–59. [CrossRef]
7. De Mantaras, R.L.; Mcsherry, D.; Bridge, D.; Leake, D.; Smyth, B.; Craw, S.; Faltings, B.; Maher, M.L.; Forbus, K.; Keane, M.; et al. Retrieval, reuse, revision and retention in case-based reasoning. *Knowl. Eng. Rev.* **2005**, *20*, 215–240. [CrossRef]
8. Pal, S.K.; Shiu, S.C. *Foundations of Soft Case-Based Reasoning*; John Wiley Sons Inc.: Hoboken, NJ, USA, 2004.
9. Choudhury, N. A survey on case-based reasoning in medicine. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 136–144. [CrossRef]
10. Schmidt, R.; Vorobieva, O. The Adaptation Problem in Medical Case–Based Reasoning Systems. *Success. Case-Based Reason. Appl.* **2010**, *1*, 117–141. [CrossRef]
11. Nilsson, M.; Sollenborn, M. Advancements and Trends in Medical Case-Based Reasoning: An Overview of Systems and System Development. In Proceedings of the FLAIRS Conference, Miami, FL, USA, 12–14 May 2004.
12. Pous, C.; Gay, P.; Pla, A.; Brunet, J.; Sanz, J.; Cajal, T.R.; Lopez, B. Modeling reuse on case-based reasoning with application to breast cancer diagnosis. *Int. Conf. Artif. Intell. Methodol. Syst. Appl.* **2008**, *5253*, 322–332. [CrossRef]
13. D'Aquin, M.; Lieber, J.; Napoli, A. Adaptation knowledge acquisition: A case study for case-based decision support in oncology. *Comput. Intell.* **2006**, *22*, 161–176. [CrossRef]
14. Macura, R.T.; Macura, K. Case-based reasoning: Opportunities and applications in health care. *Artif. Intell. Med.* **1997**, *9*, 1–4. [CrossRef]
15. Blanco, X.; Rodríguez, S.; Corchado, J.M.; Zato, C. Case-based reasoning applied to medical diagnosis and treatment. *Distrib. Comput. Artif. Intell. Adv. Intell. Syst. Comput.* **2013**, *217*, 137–146. [CrossRef]
16. Koton, P. A medical reasoning program that improves with experience. *Comput. Methods Programs Biomed.* **1989**, *30*, 177–184. [CrossRef]
17. Pla, A.; LóPez, B.; Gay, P.; Pous, C. Distributed case-based reasoning tool for medical prognosis. *Decis. Support Syst.* **2013**, *54*, 1499–1510. [CrossRef]
18. Henriet, J.; Leni, P.E.; Laurent, R.; Salomon, M. Case-Based Reasoning adaptation of numerical representations of human organs by interpolation. *Expert Syst. Appl.* **2014**, *41*, 260–266. [CrossRef]
19. Saraiva, R.M.; Bezerra, J.; Perkusich, M.; Almeida, H.; Siebra, C. A Hybrid Approach Using Case-Based Reasoning and Rule-Based Reasoning to Support Cancer Diagnosis: A Pilot Study. *Stud. Health Inform.* **2015**, *216*, 862–866.
20. Salem, A.B.M.; Bagoury, B.E. A Case-Based Adaptation Model for Thyroid Cancer Diagnosis Using Neural Networks. In Proceedings of the FLAIRS Conference, St. Augustine, FL, USA, 12–14 May 2003; pp. 155–159.
21. Sharaf-elDeen, D.A.; Moawad, I.F. A Breast Cancer Diagnosis System using Hybrid Case-based Approach. *Int. J. Comput. Appl.* **2013**, *72*, 14–19.
22. Pusztová, Ľ.; Babič, F.; Paralič, J.; Paraličová, Z. How to Improve the Adaptation Phase of the CBR in the Medical Domain. *Int. Cross-Domain Conf. Mach. Learn. Knowl. Extr.* **2019**, *3*, 168–177. [CrossRef]
23. Melle, W. The Structure of the MYCIN System. *Int. J. Man. Mach. Stud.* **1979**, *10*, 313–322.
24. Rashid, F.U. A Review on Dendral Expert System. *Int. J. Res. Appl. Sci. Eng. Technol.* **2018**, *6*, 1089–1091.
25. Stefik, M. Planning with Constraints (MOLGEN 1). *Artif. Intell.* **1981**, *16*, 111–139. [CrossRef]
26. Al-Ajlan, A. *Medical Expert Systems HDP and PUFF*; King Saud University: Riyadh, Saudi Arabia, 2007.
27. Ajani, A.; Ahmed, A.; Ajani, A.A.; Jimada-Ojuolape, B.; Ahmed, M.O. Expert system in rural medical care. *Int. J. Eng. Sci. Res. Technol.* **2018**, *6*, 440–450. [CrossRef]
28. Meera, A.K.; Kumar, S. Review: Use of Expert System in Medical Science. *Int. J. Adv. Res. Comput. Sci. Softw.* **2015**, *5*, 371–373.
29. Tan, C.F.; Wahidin, L.S.; Khalil, S.N.; Tamaldin, N.; Hu, J.; Rauterberg, G.W.M. The Application of Expert System: A Review of Research and Applications. *ARPN J. Eng. Appl. Sci.* **2016**, *4*, 2448–2453.
30. Miller, R.A.; McNeil, M.A.; Challinor, S.M.; Masarie, E.R.; Myers, J.D. The internist-1/quick medical reference project-status report. *West. J. Med.* **1986**, *145*, 816–822. [PubMed]
31. Aikins, J.S.; Kunz, J.C.; Shortliffe, E.H.; Fallat, R.J. PUFF: An expert system for interpretation of pulmonary function data. *Comput. Biomed. Res.* **1983**, *16*, 199–208. [CrossRef]
32. Singla, J.; Grover, D.; Bhandari, A. Medical expert systems for diagnosis of various disease. *Int. J. Comput. Appl.* **2014**, *93*, 36–43. [CrossRef]
33. Lawanna, A.; Daengdej, J. Methods for case maintenance in casebased reasoning. *Int. J. Comput. Inf. Eng.* **2010**, *4*, 10–18.
34. Tocimáková, Z.; Pusztová, Ľ.; Paralič, J.; Pella, D. Case-Based Reasoning for Support of the Diagnostics of Cardiovascular Diseases. *Stud. Health Technol. Inform.* **2020**, *270*, 537–541. [CrossRef]

35. Zubi, Z.S.; Saad, R.A. Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer. In Proceedings of the 10th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Base, Stevens Point, WI, USA, 20–22 February 2011; pp. 32–37.

36. Keles, A.; Keles, A.; Yavuz, U. Expert system based on neuro-fuzzy rules for diagnosis breast cancer. *Expert Syst. Appl.* **2011**, *38*, 5719–5726. [CrossRef]

37. Sharma, N.; Om, H. Data mining models for predicting oral cancer survivability. *Netw. Model. Anal. Health Inform. Bioinform.* **2013**, *2*, 285–295. [CrossRef]

38. Wang, K.J.; Makond, B.; Wang, K.M. An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data. *BMC Med. Inform. Decis. Mak.* **2013**, *13*, 124–129. [CrossRef]

39. Zolbanin, H.M.; Delen, D.; Zadeh, A.H. Predicting overall survivability in comorbidity of cancers: A data mining approach. *Decis. Support Syst.* **2015**, *74*, 150–161. [CrossRef]

40. Babič, F.; Majnarić, L.; Lukáčová, A.; Paralič, J.; Holzinger, A. On Patient's Characteristics Extraction for Metabolic Syndrome Diagnosis: Predictive Modelling. *Int. Conf. Inf. Technol. Bio Med Inform.* **2014**, *5*, 118–132. [CrossRef]

41. Pusztová, Ľ.; Babič, F.; Paralič, J. Data analytics for metabolic syndrome diagnostics. *World Congr. Med Phys. Biomed. Eng.* **2018**, *1*, 311–314. [CrossRef]

42. Sabanovic, S.; Majnaric, L.; Babič, F.; Vadovský, M.; Paralič, J.; Vcev, A.; Holzinger, A. Metabolic syndrome in hypertensive women in the age of menopause: A case study on data from general practice electronic health records. *BMC Med. Inform. Decis. Mak.* **2018**, *18*, 1–24. [CrossRef]

43. Vadovský, M.; Paralič, J. Parkinson's Disease patients classification based on the speech signals. *Int. Symp. Appl. Mach. Intell. Inform.* **2017**, *15*, 321–326.

44. Navada, A.; Ansari, A.N.; Patil, S.; Sonkamble, B.A. Overview of Use of Decision Tree algorithms in Machine Learning. *IEEE Control Syst. Grad. Res. Colloq.* **2011**, 37–42. [CrossRef]

45. Hamburg, M.A.; Collins, F.S. The path to personalized medicine. *N. Engl. J. Med.* **2010**, *363*, 301–304. [CrossRef]

46. Wiliams, G. *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*; Springer: Berlin/Heidelberg, Germany, 2011. [CrossRef]

47. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys.* **1966**, *10*, 845–848.