An Illustration of Rheumatoid Arthritis Disease Using Decision Tree Algorithm

Uma Ramasamy and Santhoshkumar Sundar
E-mail: seen.uma25@gmail.com, santhoshkumars@alagappauniversity.ac.in
Alagappa University, Karaikudi, Tamil Nadu, India

Student paper

Keywords: rheumatoid arthritis, decision tree, entropy, information gain, gain ratio

Received: November 13, 2020

The Data Mining domain integrates several partitions of the computer science and analytics field. Data mining focuses on mined data from a repository of the dataset to identify patterns, discover knowledge, additionally to predict probable outcomes. Decision tree belongs to classification techniques is a well-known method appropriate for medical diagnosis. Iterative Dichotomiser 3 (ID3) is the general significant algorithm to construct a decision tree. C4.5 is the successor of ID3 that handles dataset contains different numerical attributes. Although many studies have described and compared different decision tree algorithms, some studies have confined paper with analysis and comparison of the decision tree algorithm without the output of the decision tree. One of the inflammatory diseases is Rheumatoid Arthritis (RA) caused by specific autoantibodies with the destruction of synovial joint autoantibodies. Medical dataset applied to construct a decision tree as output has become seldom study. This study elucidates to explore the medical dataset with the decision tree approach and exhibit the derived decision tree output from the RA dataset. The objective of this paper is to construct a decision tree and display the prominent features that predict RA from the RA dataset using the decision tree algorithm.

Povzetek: Za predstavitev bolezni revmatoidnega artritisa so uporabili metodo za gradnjo odločitvenih dreves.

1 Introduction

Rheumatoid Arthritis (RA) is a rheumatic disease. The word 'Rheumatoid' implies 'rheumatism' relates to a musculoskeletal illness, 'arthr' means 'to joints,' and 'itis' denotes 'inflammation.' It is an inflammatory disorder that mostly impairs the joints, as well as other organs like the skin and lungs. Well-defined and reliable estimation of RA symptoms circumvents durable destruction to the patient’s joints and bones if treated earlier, or else it affects the patient’s quality of life. The research gap has found in the field of Rheumatoid Arthritis using data mining [1, 2].

A dataset is an indispensable component in the discussion of the classification algorithm. The dataset features or attributes are qualitative (nominal) and quantitative (numeric). Many researchers have applied various datasets [3-6] on different classification algorithms and have processed different results based on it. The dataset was utilized as a training set. From the training set decision tree is built.

'Playing tennis’ is often used dataset in the decision tree illustration [7-10]. Preferably the next used dataset in the decision tree example is the student performance [11]. Similarly, dataset like 'a dog represents a risk for citizens [7, 12],' 'reservoir inflow forecasting [13],' 'PEP (Portfolio Evaluation Plan) [14],' 'rainfall forecasting [15],’ and 'college scholarship evaluation [16],' are some illustrations in the classification algorithm that rarely handled by many research authors. A few authors only have examined and published medical datasets for the decision tree illustration.

The medical dataset created for this study is named the 'RA dataset.' The RA dataset was obtained from a new approach of the 2010 ACR/EULAR (American College of Rheumatology / European League Against Rheumatism) classification criteria of RA, which was formed, by two active groups of the ACR and the EULAR [17]. It contains qualitative attributes in a binary category (yes / no). This dataset aims to diagnose whether the patients have Rheumatoid Arthritis or do not have Rheumatoid Arthritis.

Most RA patients experience abhor pain on the joints of the hands, legs, hip, spine, and shoulder. It would be beneficial for medical practitioners to predict the prominent features responsible for RA disease. The feasible attributes to identify RA patients are displayed in Figure 3. Among these feasible attributes, the optimal attributes for the RA patient are predicted in this study using the RA dataset.

Information gain was determined to find the dominant attributes from the dataset to build the decision tree for the iterative dichotomiser 3 (ID3) algorithm. C4.5 is another algorithm to construct the decision tree by calculating the gain ratio. Decision tree algorithms such as ID3 and C4.5 (modified version of ID3) are popular and efficiently used classifiers for RA prediction from a RA dataset**.** Only a few authors practiced the decision tree illustration with medical datasets [18,19]. Although many authors have described and compared the decision tree algorithms, some confined their papers without the relevant decision tree result.

2 Related study

Data mining is the method to classify models from massive databases, that being broad, applied to learn and analyze, and obtain information [20-23]. The decision tree algorithm falls under the type of supervised learning. It is the most familiar data mining technique used frequently to build the classification model. They are used to solve both regression and classification problems. All classification model, function with the classifier, which is a supervised learner that automatically perform the learning process for the training dataset, to predict its target attribute. Data mining techniques are widely used for classification and prediction of the healthcare domain so that it can be an aid for the doctors to identify complex diseases precisely and design a more reliable Decision Support System (DSS) [24].

The Electronic Health Record (EHR) of RA patients were studied for early prediction and diagnosis of the RA disease. Moreover, the comparative study made on several machine learning algorithms identifies which algorithm suites well for the prediction of RA disease [25, 26]. rheumatoid factor (RF), anti-cyclic citrullinated peptide (Anti-CCP), swollen joint count (SJC), and erythrocyte sedimentation rate (ESR) are four essential judging factors for rheumatoid arthritis [27]. Once a patient is diagnosed with RA, the probability of getting heart failure is higher compared to the non-RA patient [28-31]. Medical research and biological research are the ever-growing fields where many biological data are collected, classified, estimated, predicted, associated, clustered, and finally visualized through reports and patterns using data mining techniques [32, 33].

The application of data mining is always in the progress of continuous development. The ID3 algorithm has some issues to handle multi-valued attributes and requires a high amount of computational complexity. A novel approach has been introduced to split attributes in the ID3 algorithm [34]. In the field of bioinformatics, data mining has some challenges like sequencing technologies and data analysis skills. Under analysis estimation instruments, a review of data mining methods performs with the combination of examination tools suitable in research tasks. The literature review finalized the merits and demerits of data mining in bioinformatics [35].

After simulation analysis, ID3 decision tree classification accuracy was higher 6-7 percentage compared to other classifiers. The author proposed an optimized ID3 algorithm that constructs a tree with a minimum node so that it can improve the efficiency and reduce the error rate [36]. Using the Gaussian mixture model, the analysis done using different clinical and laboratory data displayed results with various distributions. The patient global assessment (PGA) and health assessment questionnaire (HAQ) collected after three months of RA diagnosis, SJC, and tender joint count (TJC) considered being the functional attribute for RA diagnosis [37]. Regarding Arthritis disease, women are affected at a higher rate when compared to men [38]. The RA prediction and the RA diagnosis development done by the machine learning approach, it is mandatory to diagnose the essential features for RA prediction [39, 40].

The earlier study practiced the decision tree computation technique to investigate the selection of the second-line drug DMARDs (Disease Modifying Antirheumatic Drug) by rheumatologists which depend on the factor of disease rigor to treat RA patients after the failure of Methotrexate [41]. A few years back the immune suppression effects of DMARDs are systematic and lead to various side effects. Medical experts improved autoimmune response produced from RA by customizing a good care plan and predicting the prognosis of the disease [42]. A recent study was made to support clinical RA treatments using the decision support system to predict a model that can support medical people to give suitable decisions in the early stage of RA disease [43].

The specific proteomic biomarkers have identified for RA diagnosis using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS) combined with weak cationic exchange (WCX) magnetic beads. The classification tree model has been considered an innovative diagnostic tool for RA [44]. The combination of proteomic fingerprint technology and magnetic beads obtained efficient biomarkers and discovered the diagnostic patterns for RA. The biomarker C-C motif chemokine 24 (CCL24) has considered as a significant diagnostic indicator for RA [45].

The author states anti-citrullinated protein antibodies (ACPAs) are specific for RA and, RF was observed in health and elder people with other autoimmune diseases, which indicate immune response for RA development. The shared epitope alleles dwell in the major histocompatibility complex (MHC) class II region involved in a genetic risk factor for RA development. ACPA is the spectrum of autoantibodies that aims for posttranslational modification (PTM) [46].

The authors declare that in the future machine learning (ML) will support rheumatologists to analyze and predict the development of the disease and discover significant disease agents. Furthermore, the authors affirm ML will perform treatment propositions and evaluate their predicted outcome. The shared decision-making combines the patient's viewpoint, rheumatologist's suggestion, and also machine-learned evidence in the future [47].

The general methodologies applied to examine the intensity of RA are the clinical, laboratory, and physical examinations. The authors proposed a hybrid optimization strategy called rheumatoid arthritis disease using weighted decision tree approach (REACT), which combines the features of ID3 and particle swarm optimization (PSO) for feature selection and classification of RA to improve the efficiency and reliability of RA diagnosis [48].

It is necessary to develop therapies for RA patient's treatment at each stage of the disease progress using pathological mechanisms that urge the deterioration of RA progress in individuals. Several modern pharmacologic therapies play a vital role in disease relief without joint deformity. The RA pathogenesis, disease-modifying drugs, and views on next-generation therapeutics for RA have been discussed in this review [49]. Though joint connection, serology, levels of acute-phase reactants, and the duration of the symptoms are marked to be the primary diagnosis classification criteria for RA, yet the diagnosis requires well trained specialists who can discern early symptoms of RA from additional pathology [50].

The paper [51] developed a model for the flare prediction on the RA patients, with reduced intake of biological disease-modifying anti-rheumatic drugs (bDMARDs) in sustained remission. This proposed model used nested cross-validation and optimal hyper-parameters for a suitable model selection approach with machine learning algorithms like Logistic Regression, k-Nearest Neighbors, Naïve Bayes and Random Forest. A dose reduction, feature was selected to be the predominant flare predictor attribute.

A new method [52] focused to promotes the treatment selection in RA patients using GUIDE (Generalized, Unbiased, Interaction Detection and Estimation) decision tree, which matches with predefined rules to predict treatment response to sarilumab and adalimumab. The result classified the presence of Anti-CCP and C-reactive protein (CRP) with a threshold greater than 12.3mg/l exposed as a biomarker pattern to predict response to sarilumab.

Since RA diagnosis is prominently challenging because of reliable biomarkers, the authors [53] identified nine hub genes namely CFL1 (Cofilin 1), COTL1 (Coactosin Like F-Actin Binding Protein 1), ACTG1 (Actin Gamma 1), PFN1 (Profilin-1), LCP1 (Lymphocyte Cytosolic Protein 1), LCK (lymphocyte-specific protein tyrosine kinase), HLA-E(Major Histocompatibility Complex, Class E), FYN (Proto-oncogene tyrosine-protein kinase), and HLA-DRA (Human Leukocyte Antigen – DR isotype) biomarkers that probably distinguished the RA samples out of 52 differentially expressed genes (DEGs) from 112 RA patients. Further, Machine Learning models namely logistic regression and random forest were applied based on the identified genes.

This paper [54] presents a review that summarized the healing treatment for RA. The objective was to highlight, polypeptides, small intermediate or end products of metabolism, and epigenetics regulators as the new targets for healing RA. And prominent molecular targets for medication design were identified, which lessen the early RA and determine nonresponses followed by the partial responses and severe effects for modern DMARDs.

Algorithm Pipeline Development and Validation Study were conducted on this paper [55] using EHR to identify patients with RA. Patients' records who had their first visits were suggested as input from EHRs, and Natural Language Processing (NLP) text processing was applied from randomly selected EHRs. Moreover, Six Machine Learning Methods were utilized in the training and 10- fold cross-validation dataset to identify patients with rheumatoid arthritis from format-free text fields of EHRs.

In this paper [56] dataset taken from The Korean College of Rheumatology Biology (KOBIO) Registry, nearly 1204 RA patients were treated with biologic disease-modifying anti-rheumatic drugs (bDMARDs). To predict remission machine learning techniques included Lasso, Ridge, SVM, Random Forest, and Xgboost and explainable artificial intelligence (XAI) were used to identify the essential clinical features correlated with remission. The accuracy and area under the receiver operating characteristic (AUROC) curve were analysed for prediction.

Treatment guideline for RA patients is given in this paper [57], many references and research work associated with vaccination were collected from precise literature reviews formed by ACR guidelines to deal with RA. These studies recommend services to assist the clinician and patient decision-making and relieve them from RA disease anxiety. In this study, let us analyze the RA dataset using the decision tree model and predict the efficient features that diagnose the disease.

3 About decision tree

A tree structure classifier is the decision tree with a decision node or internal node, a branch, and a leaf node. The test of the attribute has denoted by each internal node.Each leaf node predicts the target classification. Each branch corresponds to the attribute value. To classify training dataset using the decision tree, begin from the root node, follow the suitable decision branches corresponding to the attribute values, and finally reach a leaf-node predicted with the target class. The conjunction of attribute tests corresponds to each path from the root to the leaf. Further, as a whole, the disjunction of these conjunctions represents the tree [*58*]. The dominant attribute is the best attribute classifier from the training set. The internal node represents the dominant attribute that supports to build the decision tree. The dominant attribute is the attribute with the highest information gain and gain ratio, which is discussed in sections 4.2 and 4.3.

3.1 Algorithm

3.1.1 ID3

A set of training examples are processed to learn and construct the decision tree. Furthermore, with the learned classifier, the decision tree classifies the new training examples. The algorithm technique employed is from the basic top-down greedy approach. The fundamental algorithm to build the decision tree is the ID3 algorithm developed by Quinlan in 1973 based on the Concept Learning System (CLS) algorithm. ID3 finds the dominant attribute that classifies the training examples by applying a greedy search and never backtrack [*58*], [*59*] (p.55).

3.1.2 C4.5

ID3 cannot handle practical issues such as attributes with missing values in the training dataset and attributes with continuous values. Additional problems to handle are a small sample of data leads to overfitting, to select an attribute for the decision node, one feature tested at the moment is time-consuming, and it is sensitive with a greater number of attribute values. Practical issues in ID3 overcome by the C4.5 algorithm, stated by Ross Quinlan, create the decision tree. C4.5 is a continuation of Quinlan's earlier ID3 algorithm [59] (p.55).

4 Metrics of ID3 and C4.5

Decision tree metrics are a set of measurement support to draw a decision tree with some parameters quantitative assessment derived from the dataset.

4.1 Entropy

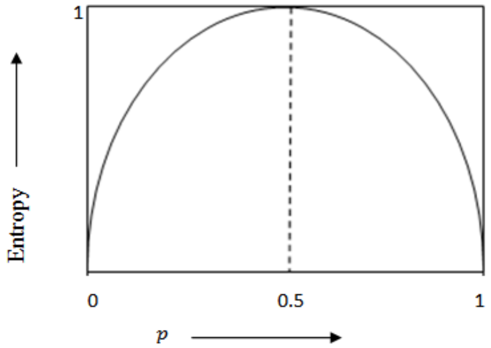

Figure 1: Entropy function relative to binary classification.

S is the sample of training examples (size =10). In the S dataset, positive proportion examples denoted as 'p,' and negative proportion examples denoted as 'n.' Entropy(S) is zero, if the proportion of positive examples (10+, 0-) is the same as the size of the training examples, similarly if the proportion of negative examples (0, 10-) is the same as the size of the training examples. Suppose, positive and negative examples are of equal size (5+, 5-), the impurity in the dataset S is maximum, i.e., entropy is one as shown in Figure 1. Therefore, it is distinct that the impurity of dataset S is measured by entropy.

Entropy(S) is the expected number of bits needed to encode class (true or false, + or -, yes or no, low or medium or high) of randomly drawn members of S. A novel way to assign $-\log_2 p$ bits to messages having probability 'p' introduced in the Information Theory concept of optimal length code [58]. So the expected numbers of bits to encode (yes or no, true or false, + or -) a random member of S is$-p\log_2 p - n\log_2 n$, where positive examples proportion denoted as 'p,' and negative examples proportion denoted as 'n.' Entropy characterizes the impurity of a collection and measures the information content from the sample of training examples. If the number of unique target feature values assigned as m, then the entropy of S w.r.t n-wise classification is equated as

$$Entropy(S) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (1)$$

Where,

$p_i -$ *proportion of S belonging to class* c_i

4.2 Information Gain

Let S be the sample of the training examples with $A_1, A_2, \ldots, A_n$ are the non-target attributes. All the features in the dataset calculated using the information gain formula as shown in Equation 2. Attribute with the highest information gain is the best classifier because the expected reduction is laid out by the information gain in entropy formed by partitioning the records of the dataset using the attribute. How effectively an attribute classifies the training examples according to their target classification has been defined in the information gain measure [59] (p.57-58). WA(A) defines the weighted sum of the information content of each subset of the examples partitioned by the possible values of the attribute. It measures the total disorder or in-homogeneity of the leaf nodes. The minimum WA (A) or maximum information gain(S, A) shows attribute A as the best attribute at a node [58]. The best attribute to select in growing the tree using each step of the ID3 algorithm, a precise measure is the information gain. The calculation of information gain is briefly described in Section 7.

$$\begin{aligned} Gain(S,A) &= Entropy(S) - WA(A) \\ &= Entropy(S) - \textstyle\sum_{v\in Values(A)} (\frac{S_v}{S})\ Entropy(S_v) \quad (2) \end{aligned}$$

Where,

$S_v -$ *subset from S for which attribute A has value v*

$Value(A) -$ *set of all possible values for attribute A*

4.3 Gain ratio

The gain ratio is a ratio between information gain and the split information. Rather than considering the entropy(S) on the target attribute, entropy(S) is concerned about all possible values of the attribute A defined to be the split information [59] (p.73-74). Information Gain Ratio is the fundamental information from the required decrease in entropy. The purpose of Quinlan to introduce this was to overcome bias on multi-valued features by considering the count of branches when choosing an attribute [60-62]. Section 7 discussed to implement gain ratio with an example.

$$GR(S,A) = \frac{IG(S,A)}{IV(S,A)} \qquad (3)$$

$$IV(S,A) = \sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \qquad (4)$$

Where,

$GR(S,A)$
$-$ *Information Gain Ratio after splitting set S on attribute A*
$IG(S,A)$
$-$ *Information Gain after splitting set S on attribute A*
$IV(S,A)$
$-$ *Intrinsic Value or Split Information value after splitting set S on attribute A, where* S_i *through* S_c *are the c subsets of examples resulting from partitioning S by the c − valued attribute A*

5 Work flow model for proposed illustration

The proposed illustration workflow model consists of a tree algorithm for RA [17], which is further converted to

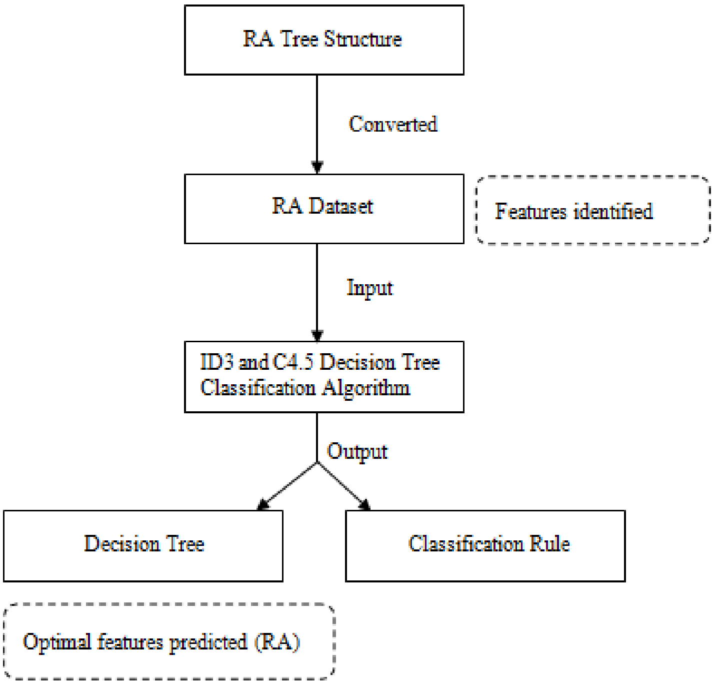

Figure 2: Work flow Model for Proposed Illustration.

a relational database as shown in Table 1. The resultant RA dataset is applied to computational techniques such as ID3 and C4.5 decision tree classifier to obtain decision tree and classification rule. The RA dataset contains all feasible features necessary to identify RA patients, whereas the final result of the decision tree predicts only the optimal features mandatory to predict RA patients.

6 About dataset

As mentioned in the workflow model (Figure 2), the conversion of RA Tree Structure (Figure 3) to RA dataset (Table 1) is done by following each path from the root node to the leaf node. The shape of the root node and the intermediate node is a rectangle, whereas the leaf node is in a circle (Figure 3). Each path represents each row in the RA dataset. There are 60 paths (in Figure 3), so the RA dataset consists of 60 rows. The root node in Figure 3 is '>10 joints (at least one small joint)', and the leaf nodes in the Figure 3 are 'RA' and 'crossed RA' (not RA). The root node and the intermediate node indicate the features/attributes, and the leaf node implies the class label of the RA dataset.

The aim of this dataset is to Classifying patients by diagnosis of Rheumatoid Arthritis or not Rheumatoid Arthritis. The source of our dataset is from the tree flowchart for classifying distinct Rheumatoid Arthritis (RA) given in the 2010 RA classification criteria. Two active groups of the ACR and the EULAR join together to form a new approach for the 2010 ACR/EULAR classification criteria of RA [17]. The number of instances (rows) of the RA dataset – 60. The number of features (columns) of the RA dataset – 9. Number of Classes (unique values of the target feature) – 2. Number of missing values – 0.

The attributes used to diagnosis RA are mixed of both phenotype and genotype. They are '> 10 joints (at least one small joint)', '4-10 small joints', '1-3 small joints', and '2 – 10 large (no small) joints' are four features of phenotype. 'Serology +' (low positive RF or low positive ACPA), 'Serology ++' (high positive RF or high positive ACPA), and 'APR (Acute phase reactants) Abnormal' (abnormal C-reactive protein (CRP) or abnormal ESR) are three features of genotype and the last attribute is ‘Duration of symptoms >=6 weeks’. In Table 1, the features name is followed with a score value to classify RA patients. The cumulative score value of each attribute per record is less than 6 out of 10. Such a score is not classifiable to diagnose RA. Those scores status is yet to be evaluated, and the criteria might be later fulfilled [17].

7 Illustration of RA dataset with ID3 and C4.5 classifiers

RA [Rheumatoid Arthritis] dataset contains the data field of the qualitative binary asymmetric attribute. Binary data has two conditions such as, 'yes or no,' 'affected or unaffected,' ' true or false.' Asymmetric defines binary values are not equally important. Both the predictor (non-target attribute) and response (target attribute) variable in the RA dataset is binary and categorical. Two response variables 'ra' and 'no ra' suggest, diagnosis of rheumatoid arthritis and not rheumatoid arthritis.

7.1 Step-by-step illustration of ID/C4.5 algorithm using RA dataset

Step 1: Find the Entropy for the current RA dataset, S. In RA dataset ‘ra’ and ‘no ra’, two classes are present with the count of 26 and 34, total instances in the dataset are 60. The 'ra' target value informs the patient diagnosed with

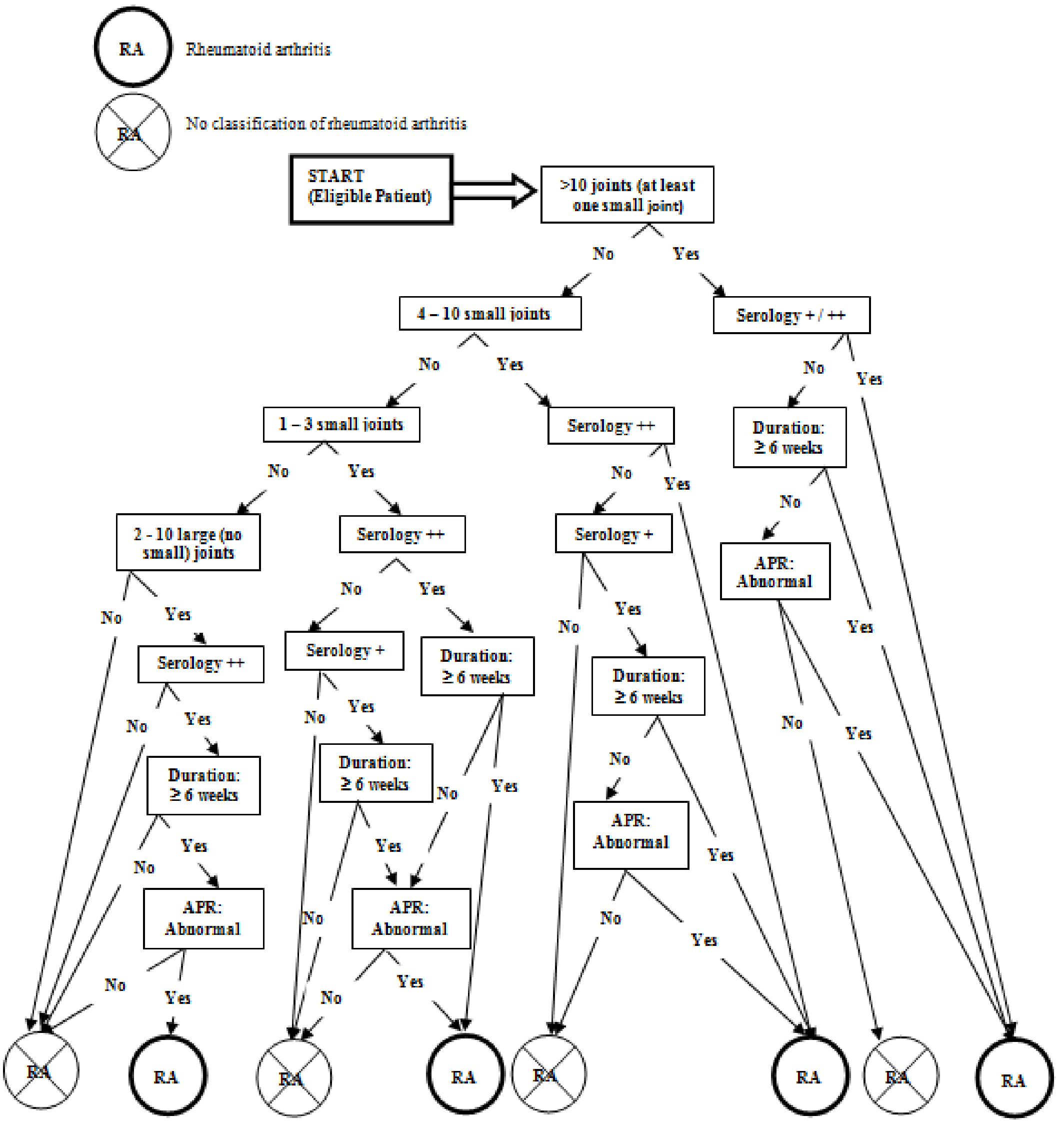

Figure 3 : Tree Algorithm for RA (Rheumatoid Arthritis).

Rheumatoid Arthritis, whereas the 'no ra' target value reveals the patient diagnosed with no Rheumatoid Arthritis. To draw the decision tree initial step is to measure the uncertainty for this dataset, i.e., the Entropy of dataset, S denoted as E(S). Calculate E(S) using Equation 1 discussed in Section 4.

$$E(S) = -\frac{26}{60}\log_2\frac{26}{60} - \frac{34}{60}\log_2\frac{34}{60} = 0.9871$$

Step 2: Find Information Gain by applying Equation 2 for each feature value in the RA dataset.

$$\begin{aligned} &Information\ Gain(S, > 10\ joints) \\ &\quad = E(S) - \sum[p(S, > 10\ joints).E(S, > 10\ joints)] \\ &\quad = E(S) - [p(S| > 10\ joints = yes) \\ &\qquad * E(S, > 10\ joints = yes) \\ &\qquad + p(S| > 10\ joints = no) \\ &\qquad * E(S, > 10\ joints = no)] \end{aligned}$$

$$E(S, > 10\ joints = yes) = -\frac{14}{16}\log_2\frac{14}{16} - \frac{2}{16}\log_2\frac{2}{16} = 0.5436$$

$$E(S, > 10\ joints = no) = -\frac{12}{44}\log_2\frac{12}{44} - \frac{32}{44}\log_2\frac{32}{44} = 0.8454$$

$$\begin{aligned} &Information\ Gain(S, > 10\ joints) \\ &\quad = 0.9871 - (\tfrac{16}{60} * 0.5436 + \tfrac{44}{60} * 0.8454) \\ &\quad = 0.9871 - 0.7649 \\ &\quad = 0.2222 \end{aligned}$$

Furthermore, obtain the information gain for enduring all feature values of the examples.

Step 3: Pick the feature which has the highest information gain. The attribute'> 10 joints' have the highest information gain, as shown in Table 2, '>10 joints' is the best classifier and determined as the root node as shown in Figure 4.

Calculate split information for each attribute using Equation 4.

$$SplitInformation(S, > 10\ joints)$$

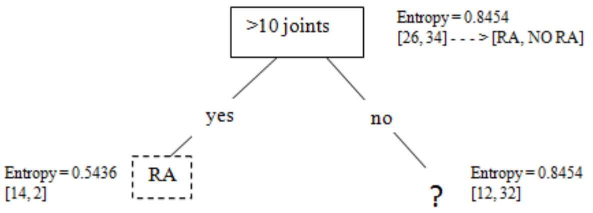

Figure 4 : Root node of the ID3/C4.5 decision tree using RA dataset.

$$= -\frac{16}{60}\log_2\frac{16}{60} - \frac{44}{60}\log_2\frac{44}{60} = 0.8366$$

Calculate Gain Ratio for each attribute using Equation 3.

$$GainRatio(S, > 10\ joints) = \frac{0.2222}{0.8366} = 0.2656$$

Now the decision tree node (root node) is the '>10 joints' attribute with a maximum of information gain (in the Table 2 it is represented as Info Gain). Since the RA dataset is categorical and not in continuous attribute, the decision tree built is the same for the ID3 and C4.5 algorithms. So, here the gain ratio measure is necessary to construct the decision tree using the C4.5 algorithm.

Step 4: Each branch from the attribute '>10 joints' partition the set S into subsets corresponds to the attribute value 'yes' and 'no.' From the root node '>10 joints', the 'yes' branch of the subset has 14 ‘RA’ and 2 ‘NO RA’ examples obtained. Though we can grow a tree further from the 'yes' branch, we have stopped with the target class RA, to avoid overfitting in the decision tree. This approach followed to stop growing the tree earlier before it attains the level to classify the training data perfectly [59] (p.68).

Now recurse (from step 2 to step 3) on the subset (from the root node '>10 joints', the ‘no’ branch of the subset has 12 ‘RA’ and 32 ‘NO RA’ mentioned as ‘?’ in Figure 4) until the ID3 algorithm satisfies the stopping criteria [63] or by following the first-class approach to avoid overfitting [59] (p.68).

Step 5: The classification rule is generated from the decision tree.

7.2 Top-down generalization approach for the decision tree

Figure 5 illustrates the decision tree built from Table 1, which depicts the RA dataset, after applying the ID3 algorithm [58], [59] (p. 56).

The basic steps for the algorithm as follows:

- *Dmat ← dominant attribute for root (initial) / non-leaf node*
- *Set Dmat as dominant attribute for the node*
- *Every unique value of Dmat form new descendant*
- *Classify the dataset records to the leaf node corresponding to the dominant attribute value of the branch*
- *If complete dataset records are ideally classified (target feature has identical values) stop, else iterate over new leaf node*

The dominant attribute (decision attribute) is the best attribute classifier from the training set.

Table 1 : A Sample Dataset of RA [Rheumatoid Arthritis] Derived from Figure 3.

S.No.	>10 Joints (atleast 1 Small Joint) (5)	4 - 10 Small Joints (3)	1 - 3 Small Joints (2)	2 - 10 Large Joints (1)	Serology + (2)	Serology ++ (3)	APR: Abnormal (1)	Duration : >=6 Weeks (1)	Class Label
1	no	no	no	no	no	yes	yes	yes	no ra
2	no	no	no	no	yes	no	yes	yes	no ra
3	no	no	no	no	no	no	yes	yes	no ra
4	no	no	no	no	no	yes	no	yes	no ra
5	no	no	no	no	yes	no	no	yes	no ra
…	…	…	…	…	…	…	…	…	…
56	yes	no	no	no	no	no	yes	no	ra
57	yes	no	no	no	no	no	no	yes	ra
58	yes	no	no	no	no	no	yes	yes	ra
59	yes	no	no	no	no	no	no	yes	ra
60	yes	no	no	no	no	no	yes	yes	ra

Table 2: A sample of Information gain and gain ratio for RA dataset.

Features	**Features Values**	**ra**	**no ra**	**Tot. freq. count**	**E(t)**	**p(t)*E(t)**	**Info Gain**	**Split Info**	**Gain Ratio**
>10 Joints (atleast 1 Small Joint) (5)	yes	14	2	16	0.5436	0.7649	0.2222	0.8366	0.2656
	no	12	32	44	0.8454				
4 - 10 Small Joints (3)	yes	7	5	12	0.9799	0.9708	0.0163	0.7219	0.0226
	no	19	29	48	0.9685				
1 - 3 Small Joints (2)	yes	4	8	12	0.9183	0.9797	0.0074	0.7219	0.0103
	no	22	26	48	0.995				
2 - 10 Large Joints (1)	yes	1	7	8	0.5436	0.9382	0.0489	0.5665	0.0863
	no	25	27	52	0.9989				
Serology + (2)	yes	8	8	16	1	0.9824	0.0047	0.8366	0.0056
	no	18	26	44	0.976				
Serology ++ (3)	yes	12	8	20	0.971	0.9464	0.0407	0.9183	0.0443
	no	14	26	40	0.9341				
APR: Abnormal (1)	yes	16	14	30	0.9968	0.9576	0.0295	1	0.0295
	no	10	20	30	0.9183				
Duration: >=6 Weeks (1)	yes	16	14	30	0.9968	0.9576	0.0295	1	0.0295
	no	10	20	30	0.9183				

ID3 algorithm follows with following input and output.
Input: Datts ←A set of non-target attributes, R← target attribute and D← training examples.
Output: returns a decision tree.

```
ID3(Datts, R, D)
Step 1: If D is null, return a single node with value Failure
Step 2: If D holds the records of the same class, it returns a single leaf node with that value.
Step 3: If Datts is null, then return a node with the value of the most frequent value of R in D.
Step 4: Begin
        4.1: Dmat← the attribute from Atts that best* classifies D
        4.2: tree← a new decision tree with root test Dmat
        4.3: for each value $v_j$ of Dmat do
                4.3.1: $D_j$ ← subset of D with Dmat= $v_j$
                4.3.2: subt ← ID3(Datts-Dmat, R, $D_j$)
                4.3.3: Add a branch to the tree with label $v_j$ and subtree subt
        4.4: return tree.
```

** The highest information gain is the best attribute defined in Equation 2.*

7.3 Extracting classification rule from decision tree algorithm ID3 & C.5

- Classification rules outline the information in the pattern of IF-Then rules
- Single rule is built for each way starting from the root node to a leaf node
- Each attribute-value pair along a path makes an association
- The leaf node contains the predicted class [64]

Classification Rule extracted from Figure 5 decision tree:

Rule 1: If *> 10 joints =”yes”* then class = “RA”

***Rule 2:** If *>10 joints =”no”* AND *4-10 small joints = ”yes”* AND *Serology++ =”yes”* then class=”RA”

Rule 3: If *>10 joints =”no”* AND *4-10 small joints =”yes”* AND *Serology++ =”no”* AND *Serology+ =”yes”* then class=”RA”

***Rule 4:** If *>10 joints =”no”* AND *4-10 small joints =”yes”* AND *Serology++ =”no”* AND *Serology+ =”no”* then class=” NO RA”

Rule 5: If *>10 joints =”no”* AND *4-10 small joints=”no”* AND *1-3 small joints =”no”* then class=”NO RA”

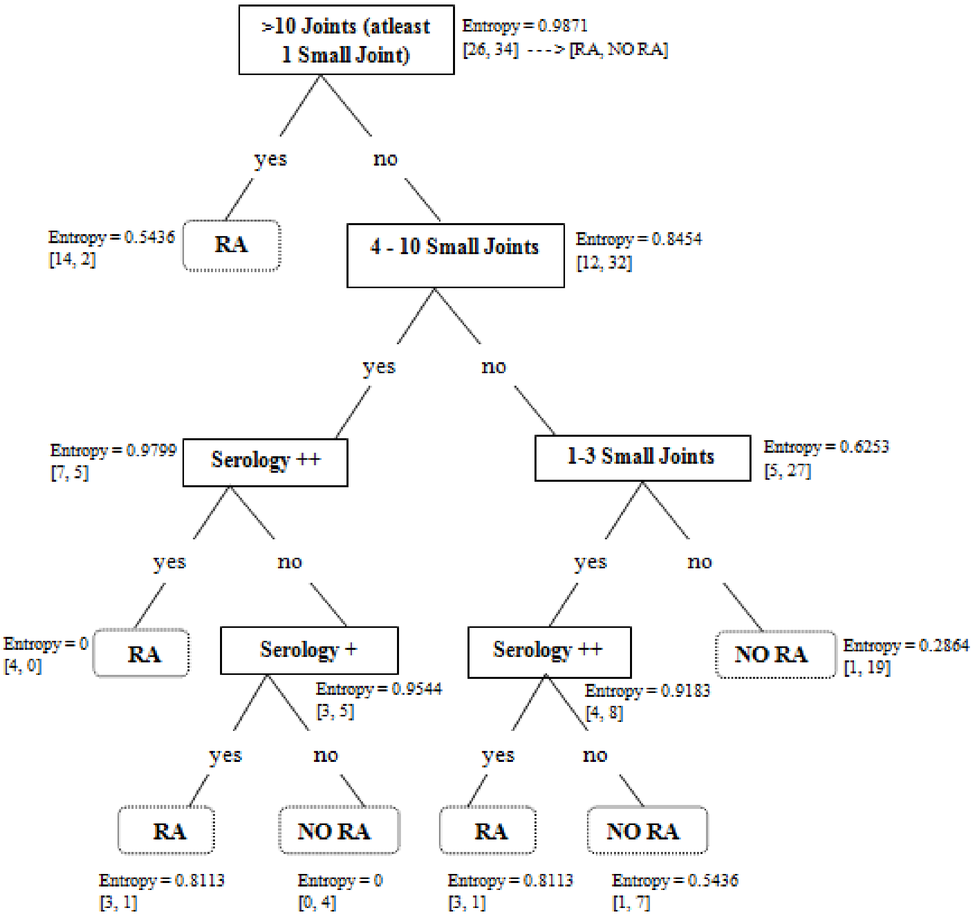

Figure 5: ID3 and C4.5 Final decision tree for RA dataset.

Rule 6: If *>10 joints ="no"* AND *4-10 small joints="no"* AND *1-3 small joints ="yes"* AND *Serology++ ="yes"* then class="RA"

Rule 7: If *>10 joints ="no"* AND *4-10 small joints="no"* AND *1-3 small joints ="yes"* AND *Serology++ ="no"* then class="NO RA"

* in rule denotes **Rule obtained from pure class**

Table 3 : Levelwise Leaf Node Membership for ID3 & C4.5 obtained from Figure 5

Level	**Leaf Node Class Membership**
Level 1	[14,2]
Level 2	-
Level 3	[4, 0]
	[1,19]
Level 4	[3,1]
	[0,4]
	[3,1]
	[1,7]

8 Illustration analysis report

The RA dataset consists of all possible feasible features from a RA patient. The predicted optimal features for RA disease are obtained using the classifier ID3 and C4.5. The Figure 5, describes the first predictor variable, '>10 joints' is achieved from level 1, the second predictor variable, ' 4-10 small joints' is identified from level 2, the third and fourth predictor variables namely *'serology ++' and '1-3 small joints'* exhibited from level 3 and finally, the fifth predictor variable, *'serology +'* is obtained from level 4. Therefore, five optimal features (predictor variables) are *'>10 joints', '4-10 small joints', 'serology ++', '1-3 small joints', and 'serology +' plays a vital role to predict RA patients.* The accuracy is 90% (54/60) for both ID3 and C4.5 decision tree. The performance is identical for both ID3 and C4.5 because the RA dataset contains categorical data. As shown in Table 2 (first level), for all the remaining levels in the decision tree, the information gain and gain ratio are simultaneously highest as displayed in Figure 6.

Table 4 : Performance of Class rulesets for ID3 & C4.5 in RA Dataset.

Class	**Generalized Rule**	**Pure Rule**	**Instances covered**	**False Positive**	**False Negative**
RA	4	1	24	0	4
NO RA	3	1	30	2	0

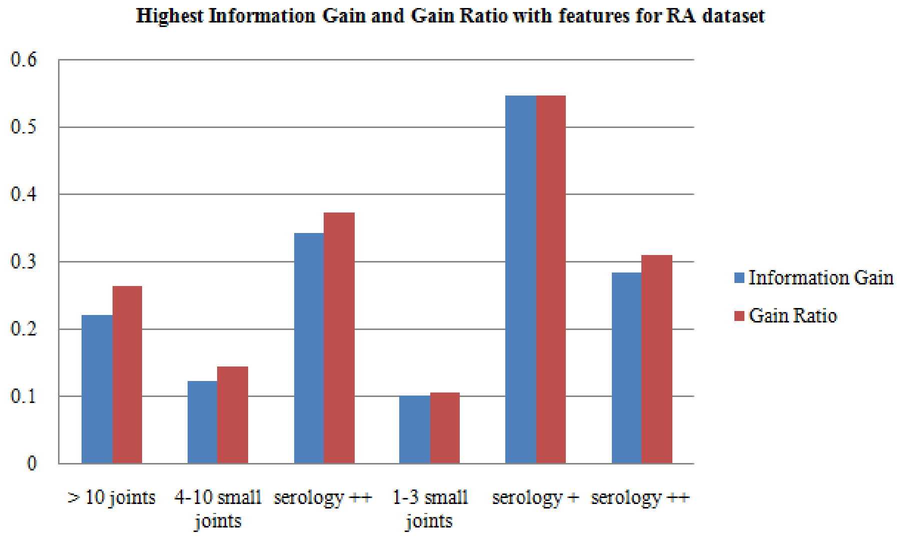

Figure 6 : Highest Information Gain and Gain Ratio for RA optimal features

9 Conclusion

The tree-structured data is converted to a relational database (RA dataset), to identify all feasible features for RA disease. Furthermore, the RA dataset is fed into the decision tree algorithm to obtain optimal features for RA disease. Therefore, we have explored the medical dataset to elucidate with the decision tree approach, and derived decision tree and classification rule as the output from the RA dataset. To summarize the work, ID3 and C4.5 decision tree algorithms construct the same decision tree with a classifier accuracy level of 90% for the RA dataset derived from the tree flowchart for diagnosing precise Rheumatoid Arthritis given in the 2010 RA classification criteria. ID3 and C4.5 classifiers result are equal in performances when considered with RA dataset.

Acknowledgment

This article has been published under AURF Start-up Grant 2018, Alagappa University, Karaikudi, Tamil Nadu, India, Dt. 23.03.2018.

This article has been published under RUSA Phase 2.0 grant sanctioned vide letter No. F. 24-51/2014-U, Policy (TN Multi-Gen), Dept. of Edn. Govt. of India, Dt. 09.10.2018.

References

[1] Zahra Shiezadeh, Hedieh Sajedi, and Elham Aflakie, “Diagnosis of Rheumatoid Arthritis using an Ensemble Learning Approach,” Computer Science and Information Technology. © CS & IT-CSCP 2015, pp. 139–148. https://doi.org/10.5121/csit.2015.51512

[2] Rohini Handa, Rao, U. R. K., Juliana F. M. Lewis, Gautam Rambhad, Susan Shiff, and Canna J. Ghia, “Literature review of rheumatoid arthritis in India International Journal of Rheumatic Diseases,” vol. 19, pp. 440-451, 2016. https://doi.org/10.1111/1756-185x.12621

[3] Jena, Monalisa, and Satchidananda Dehuri. "DecisionTree for Classification and Regression: A State-of-the Art Review." Informatica 44, no. 4 (2020). https://doi.org/10.31449/inf.v44i4.3023.

[4] Yang, Fen. "Decision tree algorithm-based university graduate employment trend prediction." Informatica 43, no. 4 (2019). https://doi.org/10.31449/inf.v43i4.3008.

[5] Kalyani, G., MVP Chandra Sekhara Rao, and B. Janakiramaiah. "Decision tree-based data reconstruction for privacy preserving classification rule mining." Informatica 41, no. 3 (2017). https://doi.org/10.1007/s13369-017-2834-2.

[6] Wang, Xi. "Research on Recognition and Classification of Folk Music Based on Feature Extraction Algorithm." Informatica 44, no. 4 (2020). https://doi.org/10.31449/inf.v45i4.3819.

[7] Halima ELAIDI, Zahra BENABBOU, Hassan ABBAR, A comparative study of algorithms constructing decision trees: ID3 and C4.5, LOPAL’18, May 2-5, 2018, 1-5, Rabat, Morocco © 2018 Association for Computing Machinery, https://doi.org/10.1145/3230905.3230916

[8] Badr HSSINA, Abdelkarim MERBOUHA, Hanane EZZIKOURI, Mohammed ERRITALI, A comparative study of decision tree ID3 and C4.5, International Journal of Advanced Computer Science and Applications, pp: 13-19, 2014. https://doi.org/10.14569/specialissue.2014.040203

[9] Sonia Singh, Manoj Giri, Comparative Study Id3, Cart and C4.5 Decision Tree Algorithm: A Survey, International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682 3(7): 47-52, July 2014, https://doi.org/10.15693/ijaist/2014.v3i7.47-52.

[10] Elaidi, Zahra Benabbou, Hassan Abbar, Using Game Theory to Handle Missing Data at Prediction Time of ID3 and C4.5 Algorithms, (IJACSA) International Journal of Advanced Computer Science and Applications, 9(12): 218-224, 2018. https://doi.org/10.14569/ijacsa.2018.091232

[11] Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao, Predicting Students' Performance Using ID3 and C4.5 Classification Algorithms, International Journal of Data Mining & Knowledge Management Process (IJDKP) 3(5) , 39-52, September 2013. https://doi.org/10.5121/ijdkp.2013.3504

[12] Y.Wang, Y.Li, Y.Song, X.Rong, S.Zhang, Improvement of ID3 Algorithm Based on Simplified Information Entropy and Coordination Degree, Algorithms journal, 10 (4): 124, 2017. https://doi.org/10.3390/a10040124

[13] Pattama Charoenporn, Reservoir Inflow Forecasting Using ID3 and C4.5 Decision Tree Model, 2017 IEEE 3rd International Conference on Control Science and Systems Engineering, 978-1-5386-0484-7/17/$31.00 ©2017 IEEE, 698-701. https://doi.org/10.1109/ccsse.2017.8088023

[14] Sudrajat, I. Irianingsih, D. Krisnawan, Analysis of data mining classification by comparison of C4.5 and ID algorithms,, IOP Conf. Series: Materials Science and Engineering 166: 1-9, 2017. https://doi.org/10.1088/1757-899x/166/1/012031

[15] Joko Azhari Suyatno, Fhira Nhita, Aniq Atiqi Rohmawati, Rainfall Forecasting in Bandung Regency using C4.5 Algorithm, 2018 6th International Conference on Information and Communication Technology (ICoICT), ISBN: 978-1-5386-4571-0 (c) 2018 IEEE, 324-328. https://doi.org/10.1109/icoict.2018.8528725

[16] X. Wanga, C. Zhoua, X. Xub, Application of C4.5 decision tree for scholarship evaluations, The 10th International Conference on Ambient Systems, Networks and Technologies (ANT), The Authors. Published by Elsevier Ltd. ScienceDirect, Procedia Computer Science 151 (2019) 179–184. https://doi.org/10.1016/j.procs.2019.04.027

[17] Daniel Aletaha, Tuhina Neogi, Alan J. Silman, Julia Funovits, David T. Felson, Clifton O. Bingham., … Gillian Hawker(2010). “2010 Rheumatoid Arthritis Classification Criteria,” Arthritis & Rheumatism, (62)9: 2569-2581, 2010. https://doi.org/10.1002/art.27584

[18] Lakshmi.B.N, Dr.Indumathi.T.S, Dr.Nandini Ravi, A study on C.5 Decision Tree Classification Algorithm for Risk Predictions during Pregnancy, International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015), Elsevier Ltd., ScienceDirect, Procedia Technology 24: 1542–1549, 2016. https://doi.org/10.1016/j.protcy.2016.05.128

[19] Yanwei Xing, Jie Wang and Zhihong Zhao, Yonghong Gao, Combination data mining methods with new medical data to predicting outcome of coronary heart disease, 2007 International Conference on Convergence Information Technology,204:868-872,2007. https://doi.org/10.1109/iccit.2007.204

[20] Begona Garcia-Zapirian, Yolanda Garcia-Chimeno, and Heather Rogers(2015), ”Machine Learning Techniques for Automatic Classification of Patients with Fibromyalgia and Arthritis,” International Journal of Computer Trends and Technology, 25(3):2231-2803,2015. https://doi.org/10.14445/22312803/ijctt-v25p129

[21] Begum Cigsar and Deniz Unal, Comparison of Data Mining Classification Algorithms Determining the Default Risk, Hindawi Scientific Programming Feb. 2019, Article ID 8706505.

[22] Nikita Jain, Vishal Srivastava, Data Mining Techniques: A Survey Paper, International Journal of Research in Engineering and Technology, 2(11), eISSN: 2319-1163, pISSN: 2321-7308, 2013. https://doi.org/10.15623/ijret.2013.0211019

[23] S. Umadevi, and K. S. Jeen Marseline, A Survey on Data Mining Classification algorithms, International Conference on Signal Processing and Communication (ICSPC’ 17), July 2017. https://doi.org/10.1109/cspc.2017.8305851

[24] Shanmugam, S., & Preethi, J., “Design of Rheumatoid Arthritis Predictor Model Using Machine Learning Algorithms,” SpringerBriefs in Applied Sciences and Technology, https://doi.org/10.1007/978-981-10-6698-6_7

[25] Vaishali S. Parsania, Krunal Kamani, and Gautam J Kamani, “Comparative Analysis of Data Mining Algorithms on EHR of Rheumatoid Arthritis of Multiple Systems of Medicine International,” Journal of Engineering Research and General Science, 3 (1): 344-350, 2015.

[26] Beau Norgeot, MS; Benjamin S. Glicksberg, Laura Trupin, Dmytro Lituiev, Milena Gianfrancesco, Boris Oskotsky, Gabriela Schmajuk, Jinoos Yazdany, and Atul J. Butte, Assessment of a Deep Learning Model Based on Electronic Health Record Data to Forecast Clinical Outcomes in Patients with Rheumatoid Arthritis, JAMA Network Open. 2019,2(3). https://doi.org/10.1001/jamanetworkopen.2019.0606

[27] Jihyung Yoo, Mi Kyoung Lim, Chunhwa Ihm, Eun Soo Choi and Min Soo Kang (2017). A Study on Prediction of Rheumatoid Arthritis Using Machine Learning. International Journal of Applied Engineering Research, 12 (20). ISSN 0973-4562, pp. 9858-9862, 2017.

[28] Catia Sofia Tadeu Botas (2017), “Feature analysis to predict treatment outcome in Rheumatoid Arthritis,” Instituto Superior Tecnico, Lisboa, Portugal. pp. 1-10, 2017.

[29] Elena Myasoedova, John M. Davis Ill, Eric L.Matteson, Sara J. Achenbach, Soko Setoguchi, Shannon M. Dunlay, Veronique L. Roger, Sherine E. Gabriel, and Cynthia S. Crowson, ” Increased hospitalization rates following heart failure diagnosis in rheumatoid arthritis as compared to the general population,” Seminars in Arthritis and Rheumatism, 50:25-29, 2020 © Elsevier Inc. https://doi.org/10.1016/j.semarthrit.2019.07.006

[30] Cynthia S Crowson, Katherine P Liao, John M Davis, Daniel H Solomon, Eric L Matteson, Keith L Knutson, Mark A Hlatky, and Sherine E Gabriel, Rheumatoid Arthritis and Cardiovascular Disease, NIH Public Access Am Heart J. 166(4):622–628, 2013. https://doi.org/10.1016/j.ahj.2013.07.010

[31] Usman Khalid, Alexander Egeberg, Ole Ahlehoff, Deirdre Lane, Gunnar H. Gislason, Gregory Y. H. Lip and Peter R. Hansen, Incident Heart Failure in Patients with Rheumatoid Arthritis: A Nationwide Cohort Study, Journal of the American Heart Association 2018. https://doi.org/10.1161/jaha.117.007227

[32] Khalid Raza, Application of Data Mining in Bioinformatics, Indian Journal of Computer Science and Engineering 1(2):114-118, 2012, ISSN: 0976-5166.

[33] Xing-Ming Zhao, Data Mining in Systems Biology, IEEE/ACM Transactions on Computational Biology and Bioinformatics 2016, vol 13(6), 1003-1003. https://doi.org/10.1109/tcbb.2016.2617698

[34] Zijing Wang, Yo Liu and Li Liu,” A New Way to Choose Splitting Attribute in ID3 Algorithm,” 978-1-5090-6414-4/17 ©2017 IEEE. https://doi.org/10.1109/itnec.2017.8284813

[35] Audu Musa Mabu, Rajesh Prasad, Raghav Yadav, Suleiman S Jauro,” A Review of Data Mining Methods in Bioinformatics,” Recent Advances of Engineering, Technology and Computational Sciences, 978-1-5386-1686-4/18 ©2018 IEEE. https://doi.org/10.1109/raetcs.2018.8443785

[36] He Zhang and Runjing Zhou,” The Analysis and Optimization of Decision Tree Based on ID3 Algorithm,” The 9th International Conference on Modeling, identification and Control, 2017. https://doi.org/10.1109/icmic.2017.8321588

[37] Jorn Lotsch, Lars Alfredsson, Jon Lampa, “Machine-Learning-based knowledge discovery in rheumatoid arthritis-related registry data to identify predictors of persistent pain,” The International Association for the Study of Pain Research Paper – PAIN, 161:114-126, 2020. https://doi.org/10.1097/j.pain.0000000000001693

[38] Tiffany D. Pan, Beth A. Mueller, Carin E. Dugowson, Michael L. Richardson, and J. Lee Nelson, Disease progression in relation to pre-onset parity among women with rheumatoid arthritis, Seminars in Arthritis and Rheumatism 2019, 0049-0172/© 2019 Elsevier Inc. https://doi.org/10.1016/j.semarthrit.2019.06.011

[39] Ho Sharon, I. Elamvazuthi, CK. Lu, S. Parasuraman and Elango Natarajan, Classification of Rheumatoid Arthritis using Machine Learning Algorithms, IEEE Student Conference on Research and Development (SCOReD) :345-350, 2019. https://doi.org/10.1109/scored.2019.8896344

[40] Ho Sharon, Irraivan Elamvazuthi, Cheng-Kai Lu, S. Parasuraman and Elango Natarajan, Development of Rheumatoid Arthritis Classification From Electronic Image Sensor Using Ensemble Method, Sensors 2020, 20, 167. https://doi.org/10.3390/s20010167

[41] Fautrel, B. Guillemin, F. Meyer, O. Bant, M.D. et al., Choice of Second-Line Diases-Modifying Antirheumatic Drugs After Failure of Methotrexate Therapy for Rheumatoid Arthritis: A Decision Tree for Clinical Practice Based on Rheumatologists’ Preferences. Arthritis Rheumatol. 61:425–434, 2009. https://doi.org/10.1002/art.24588

[42] Keyser, F.D. Choice of biologic therapy for patients with rheumatoid arthritis: The infection perspective. Curr. Rheumatol. Rev., 7: 77-87, 2011. https://doi.org/10.2174/157339711794474620

[43] Wu, C.-T.; Lo, C.-L.; Tung, C.-H.; Cheng, H.-L. Applying Data Mining Techniques for Predicting Prognosis in Patients with Rheumatoid Arthritis. Healthcare,8:85,2020. https://doi.org/10.3390/healthcare8020085

[44] Li Y, Sun X, Zhang X, Liu Y, Yang Y, Li R, Liu X, Jia R, Li Z. Establishment of a decision tree model for diagnosis of early rheumatoid arthritis by proteomic fingerprinting. Int J Rheum Dis. 18(8):835-41, 2015. PMID: 26249836. https://doi.org/10.1111/1756-185x.12595

[45] Ma Dan, Liang Nana, Zhang Liyun. Establishing Classification Tree Models in Rheumatoid Arthritis Using Combination of Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry and Magnetic Beads. Frontiers in Medicine. 2021; 8:190. ISSN:2296-858X. https://doi.org/10.3389/fmed.2021.609773

[46] Hans Ulrich Scherer, Thomas Häupl, Gerd R. Burmester,The etiology of rheumatoid arthritis, Journal of Autoimmunity, Volume 110,2020,102400, ISSN 0896-8411, https://doi.org/10.1016/j.jaut.2019.102400

[47] Maria Hugle, Patrick Omoumi, Jacob M. van Laar, Joschka Boedecker and Thomas Hugle. Applied machine learning and artificial intelligence in

rheumatology.Rheumatology Advances in Practice 20;0:1–10,2020. https://doi.org/10.1093/rap/rkaa005

[48] Shanmugam, S., Preethi, J. Improved feature selection and classification for rheumatoid arthritis disease using weighted decision tree approach (REACT). J Supercomput 75, 5507–5519 (2019). https://doi.org/10.1007/s11227-019-02800-1

[49] Guo Q, Wang Y, Xu D, Nossent J, Pavlos NJ, Xu J. Rheumatoid arthritis: pathological mechanisms and modern pharmacologic therapies. Bone Res. 6:15, 2018. PMID: 29736302; PMCID: PMC5920070. https://doi.org/10.1038/s41413-018-0016-9

[50] Maria Kourilovitch, Claudio Galarza-Maldonado, Esteban Ortiz-Prado. Diagnosis and classification of rheumatoid arthritis. Journal of autoimmunity, 2014. https://doi.org/10.1016/j.jaut.2014.01.027

[51] Vodencarevic, A., Tascilar, K., Hartmann, F. et al. Advanced machine learning for predicting individual risk of flares in rheumatoid arthritis patients tapering biologic drugs, Arthritis Res Ther, 23, 67 (2021). https://doi.org/10.1186/s13075-021-02439-5

[52] Rehberg, M., Giegerich, C., Praestgaard, A. et al., Identification of a Rule to Predict Response to Sarilumab in Patients with Rheumatoid Arthritis Using Machine Learning and Clinical Trial Data, Rheumatol Ther 8, 1661–1675 (2021), https://doi.org/10.1007/s40744-021-00361-5

[53] Liu, J., Chen, N. A 9 mRNAs-based diagnostic signature for rheumatoid arthritis by integrating bioinformatic analysis and machine-learning, J Orthop Surg Res 16, 44 (2021). https://doi.org/10.1186/s13018-020-02180-w

[54] Huang Jie, Fu Xuekun, Chen Xinxin, Li Zheng, Huang Yuhong, Liang Chao, Promising Therapeutic Targets for Treatment of Rheumatoid Arthritis, Frontiers in Immunology 2021, Vol.12, ISSN:1664-3224, https://doi.org/10.3389/fimmu.2021.686155

[55] Maarseveen TD, Meinderink T, Reinders MJT, Knitza J, Huizinga TWJ, Kleyer A, Simon D, van den Akker EB, Knevel R, Machine Learning Electronic Health Record Identification of Patients with Rheumatoid Arthritis: Algorithm Pipeline Development and Validation Study, JMIR Med Inform, 8(11):2020, e23930, PMID: 33252349, PMCID: 7735897, https://doi.org/10.2196/23930

[56] Koo, B.S., Eun, S., Shin, K. et al. Machine learning model for identifying important clinical features for predicting remission in patients with rheumatoid arthritis treated with biologics. Arthritis Res Ther 23, 178, 2021. https://doi:10.1186/s13075-021-02567-y

[57] Fraenkel L, Bathon JM, England BR, et al. 2021 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis, Arthritis Rheumatol,73(7):1108-1123,2021. https://doi:10.1002/art.41752

[58] nptelhrd. (2008, October 16). Lecture – 35 Rule Induction and Decision Trees – I. Retrieved from https://www.youtube.com/watch?v=WfsRaLmh8js&t=547s.

[59] Tom M. Mitchell. Machine Learning McGraw-Hill Science/Engineering/Math. 1997; pp. 52-76.

[60] Wikipedia Website. [Online]. Available: https://en.wikipedia.org/wiki/Information_gain_ratio

[61] Seema Sharma, Jitendra Agrawal, Sanjeev Sharma, Classification Through Machine Learning Technique: C4.5 Algorithm based on Various Entropies, International Journal of Computer Applications, 82(16):0975-8887, 2013. https://doi.org/10.5120/14249-2444

[62] R. Sudrajat, I. Irianingsih, D. Krisnawan, Analysis of data mining classification by comparison of C4.5 and ID algorithms, IOP Conf. Series: Materials Science and Engineering 166, 2017, 012031, https://doi.org/10.1088/1757-899x/166/1/012031

[63] Wikipedia Website. [Online]. Available: https://en.wikipedia.org/wiki/ID3_algorithm.

[64] Poonkuzhali. S, Saravanakumar. C. Data Warehousing & Data Mining. Charulatha Publications 2008. 1st Ed. pp: 6.12.