A Decision Tree Knowledge-based System for Reviewing Research Ethics Protocol

Ratih Nur Esti Anggraini
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ratih_nea@if.its.ac.id

Nurul Fajrin Ariyani
Computer Science and Informatics
Cardiff University
Surabaya, Indonesia
ariyaninf@cardiff.ac.uk

Abdullah F. Septiyanto
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
abdullahfaqih161@gmail.com

Riyanarto Sarno
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

Zahra D. Meilani
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
zahradyahmeilani@gmail.com

Triono Soendoro
Komite Etik Penelitian dan Pelayanan KEPPIN
Jakarta, Indonesia
tsoendor12@gmail.com

2022 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT) | 978-1-6654-6030-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/COMNETSAT56033.2022.9994414

***Abstract*— Knowledge base systems have undergone many developments in providing similar results to experts. This can help increase time effectiveness in determining decisions and analysis results. Several methods have given good results in determining decisions based on knowledge, one of which is using a decision tree. In this study, the researchers applied decision tree modeling to determine the results of the review on ethical research protocols. Our target is to classify ethical protocols into one of three decisions: Exempted, Expedited, or Full Board. Three decision tree models are used in this research to evaluate the best results that can predict the ethical review protocol results according to the expert's dataset. The experiments showed that all models showed the same result, with an accuracy value of 0,91, precision of 0,93, and recall of 0,91. However, manual checking showed that the second model with Gini criteria parameters and class weight balance resulted in 10 data correctly predicted based on the dataset used.**

Keywords—knowledge-based system, decision tree, ethics protocol

I. INTRODUCTION

Ethics has to be applied in research involving humans, especially in medical research. Research ethics is a guideline of norms, values, and morality that must be followed during scientific research to protect the research participants [1]. The ethics of research needs to be stated in the research ethics protocol. Thus, this document must clearly explain ethical issues that possibly occur during research, such as approval, trust, protection of personal health data, equality, accessibility, and human empowerment as research objects [2]. Moreover, it must ensure that three ethical principles (respect for person, beneficence, and justice) are implemented in the research with humans as the participants [3].

Based on World Health Organization (WHO), an ethics protocol has to go through a review system and follow seven ethical standards [4]. Moreover, it must also follow the twenty-five guidelines created by the Council for International Organizations of Medical Sciences (CIOMS) [5].

Ethical protocol reviewers must pay attention to the many guidelines, causing the review process to take longer. Thus, a rule is needed to determine if the research has no ethical issues. As a result, it is necessary to build a system that can determine whether the protocols have no ethical issues using a Knowledge-Based System (KBS).

KBS is commonly used to represent the relationship between the cause and effect of input or processed data [6]. In applying KBS, developers experience many challenges, one of which is extracting knowledge into the Field [7], since experts intuitively make decisions based on their experiences. Therefore, it can cause some differences of opinion if the reference comes from more than one expert. For the KBS to be more straightforward in determining decisions, the Rule Base System (RBS) is applied. RBS can model clinical knowledge into several rules in the form of IF (condition) - THEN (conclusion) statements. The IF-THEN form is one of the popular methods used to present knowledge [8] because the application is simple and can give relevant results if the rules do not have complex values. One method that can be applied is the Decision Tree (DT).

A DT is a method that can provide intuitive clarity to state which set of attributes can lead to different prediction results [9]. Using a tree structure to build a decision model from the knowledge of expert teams, the DT model is a natural mechanism for dealing with problems in decision-making with predictive performance, and computational effort is relatively small and easy to understand [10]. DT modeling is distinguished from constructing a tree based on the train data. The split criteria process forms a decision tree by repeatedly optimizing the function and data partitioning on the root and internal nodes until the termination conditions are met [11]. Some of the models that are often used are ID3 [12] which has an extension model of C4.5 [13], and the famous tree model used is the Classification And Regression Tree (CART) [14].

CART is popularly used because CART is based on the Gini index, in contrast to ID3 and C4,5, which uses information gain (Entropy). Thus, CART modeling can also use entropy. Decision tree modeling using CART can be applied using the sci-kit-learn library in Python, which focuses on classification with nominal data because the library does not support categorical data. Although some previous methods are different, the modelling method is not classified as an absolute scheme [15].

In this research, to solve the problem of requests for reviewing ethical protocol documents that increased and could not be handled by the existing expert team resources, the researchers contributed by building KBS using DT modeling. Knowledge from the expert team is extracted into several attributes to describe how the expert team evaluates all points and sub-points to determine whether the research is allowed to be carried out. The dataset is extracted from some of the attributes obtained to provide a data train used in modeling the DT. The target is to categorize it into three classes, namely Exempted (E1), Expedited (E2), and Full Board (FB). Thus, based on the decision classification target, the attributes from the knowledge extraction results of the expert team are used as a reference to determine the success rate of DT modeling.

With the total attribute and classification results obtained from the knowledge extraction of the expert team, the possibility of a decision tree being made is very large. Therefore, an important factor that needs to be considered in decision tree modeling is the complexity of the resulting model [10]. The most popular model used in DT modeling is CART, with the data split method using Gini (CART-Gini) and Entropy (CART-Entropy). The DT modelling was carried out in several scenarios to get the simplest model with the most concise tree and leaf formation. Each of these scenarios will compare the values of precision, recall, and F-Measure to determine which model is the best in providing the simplest DT results and with the highest precision, recall, and F-Measure.

II. LITERATURE REVIEW

A. Research Ethics Protocol

The research ethics protocol is a scientific process that must be carried out by institutions and researchers who carry out research involving humans as the participants, especially in medical research. Those protocols have to incorporate seven ethical principles and twenty-five guidelines defined by WHO to be executed during the research. Subsequently, it has to be submitted to the ethical review committee to evaluate any ethical issues. If there is no issue, the committee will issue ethical clearance so the researchers can conduct their study.

The results of the ethical protocol review process can be classified into three categories, namely Exempted (E1), Expedited (E2), and Full Board (FB), as shown in Fig. 1. E1 is when the research has no ethical issues; thus, ethical clearance can be issued, and the research can be directly carried out. In E2, there is a minor ethical issue or different opinions between reviewers, so the review process must be repeated. In FB, the protocol contains major ethical issues, so the ethical committee must review them in more detail.

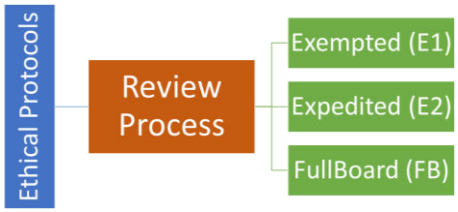

Fig. 1. Protocol Review Results

Therefore, researchers must ensure that the proposed research is scientific research with an adequate knowledge base and can produce valuable information [16].

B. Decision Tree

Quinlan [12] developed the simplest classification algorithm called Decision Tree. The decision tree is one of the approaches used for predictive modelling in machine learning and data mining [17]. This method can be used to predict decisions independently or in combination with other methods. The functionality of the decision tree algorithm is based on a data set of input variables to create a predictive model based on the values of the dependent variable [12]. The decision tree construction results are built from the feature set into a hierarchical structure by showing the relationship between features. Decision Trees have been widely used to handle categorical and numeric feature types. However, the Decision Tree algorithm has shortcomings in building models because of its efficiency in determining predictions [12]. The splitting measures commonly used in the decision tree are Gini-index and information gain (Entropy). In this research, equations to calculate Gini-index as in Eq. (1) [18] and information gain based on entropy as in Eq. (2) [19].

In DT, the model called ID3 [12] uses a split criterion based on gain (entropy) information which has several shortcomings in handling noise and miss value data [20]. C4.5 is a manufactured version of ID3 to overcome the weaknesses of ID3 by using the trimming method to reduce over-installation and using the information gain ratio as a splitting criterion to get a smoother partition [10]. Then another DT algorithm was developed by Breiman et al. [21], namely CART. The split criteria on CART use the Gini index and create a binary tree with discrete and continuous attributes using the complexity pruning method to reduce overfitting [10]. In addition to these three models, there are many other models that previous researchers have developed to handle various problems in building a knowledge base as a decision-making system and automatic classification.

Vanfretti and Arava [22] applied a decision tree as a method to classify the condition of voltage stability of the electric power system. The voltage condition is divided into four classes: unstable, outside grid limits, marginally stable, and unstable. This method used a dataset to build a decision tree model. The results of these decisions are used to assist operators in choosing preventive measures to keep the system stable and not violate operational requirements w.r.t voltage stability. The study's results showed the average accuracy value reached 99.06%.

In addition, [23] builds a decision tree with a supervised learning technique to determine the best decision on a complex production system optimization problem. The decision tree models are the Logistic Model Tree (LMT) and Hoeffding Tree. The specified classification consists of 4 classes: do nothing, produce, maintain, and recycle. The results of this study indicate that the LMT model is better than the Hoeffding Tree, with an average acquisition of precision, recall, and F-Measure of 0.99%.

In other research, [24] a decision tree is used to classify module errors in software. Various tree models are used to

get the most optimal tree model. In general, the decision tree models used are Gini-Index, CHID, Gain Ratio, and Information Gain. Still, some of these tree models have advantages and disadvantages, so a new tree model combines Gini-Index and Entropy called EGIA. The model was tested on several datasets and compared with CART, C4.5, Logistic Regression and Naive Bayesian. Results of comparison show that the model could predict faulty modules in the software better than the existing approach. The proposed node separation method can be modified to suit the needs and can be integrated with different optimization methods to be more dynamic than the previous method.

Recently, [9] developed a new evolutionary algorithm for fuzzy decision trees in identifying patients after endovascular treatment with poor outcomes and predictive results from the decision tree model. It was made so that nurses and patients could easily understand it. In this study, a comparison of decision tree modelling with the proposed method with the CART model was carried out on classification accuracy and interpretation ability. The CART model used is CART with Gini index and CART with Entropy. The results showed that fuzzy decision trees outperformed CART with an average accuracy of 72% with fewer nodes than CART. Based on some of this research, the CART modelling algorithm is still widely used to model decision trees in various tests. However, its implementation is still inferior to some decision tree modelling algorithms. Still, the many studies using this modelling show that CART is feasible to be used as the first test in modelling the dataset to get a concise decision tree model with high accuracy.

C. Performance Evaluation

Performance evaluation is used to measure the model that has been made. In this work, we use accuracy, precision, and recall. Accuracy is calculated to measure the percentage of correct predictions with total data instances, precision is used to measure how accurate our model is out of predicted faulty modules [24], and recall is employed to measure how much our model can predict correctly from the true data. In this research, equations to calculate accuracy as in Eq. (3), precision as in Eq. (4), and F-recall as in Eq. (5).

$$Gini = 1 - \sum_{i-1}^{n}(p_i)^2 \quad (1)$$

where:

p_i = probability of an object being classified to a particular class

$$E(S) = \sum_{i-1}^{n}(p_i \log_2 p_i) \quad (2)$$

where:

p = probability distribution

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

where:

TP = true positive
TN = true negative
FP = false positive
FN = false negative

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

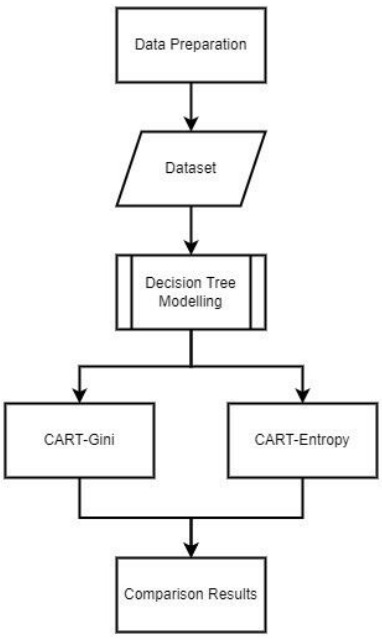

Fig. 2. Proposed Method

III. PROPOSED METHOD

This section explains how to extract knowledge from the expert team from the data preparation stage to compare the DT formation model using CART. The proposed method is shown in Fig. 2.

A. Data Preparation

At this stage, the data preparation process begins with extracting knowledge from the expert. Information mining is carried out by conducting discussions with experts to assess each point in the ethical protocol document. Six attributes were obtained based on all the points and sub-points with the value for each attribute to determine the results of the expert's decision. On the other hand, the expert assigned a value to each attribute along with the decision results. This data is used as a reference for training data to create a knowledge base using DT modeling. The six attributes and values for each attribute are shown in TABLE I.

In TABLE I, a part of the dataset obtained from the expert, the dataset consists of six attributes based on the points assessed by the expert team to determine that the research is by the research ethics protocol, which is grouped into three classes, namely Exempted (E1), Expedited (E2), and Full Board (FB).

B. Dataset

At this stage, the results of the knowledge extraction process from the expert into categorical data are converted into numerical data that will be used as a dataset in DT modeling. The total data obtained from expert knowledge is 204 with six attributes and classes, and the complete data for each class is shown in Fig. 3. In Fig. 3, the total data is from the FB class with a total of 108 data, followed by E2 with a total of 72, and last is E1 with a total of 24. Based on this dataset, it is an unbalanced dataset where each class does not have the same total data, which can also affect the training results in DT modeling.

TABLE I. PART OF DATASET FROM EXPERT

No	Do the subjects include vulnerable groups?	Secret guaranteed	Potential secrets revealed	Information consent	Persuasion / compression	Risk more than minimum	Label
1	Yes	Guaranted	No Potential	No	Not Forced	No	E1
2	No	Guaranted	Potential	Yes	Forced	Yes	E2
3	Yes	Guaranted	No Potential	Yes	Forced	Yes	FB

TABLE II. THE EXAMPLE OF DATASET IN NUMERICAL FEATURES

No	X1	X2	X3	X4	X5	X6	Label
1	1	1	1	1	1	1	E1
2	0	1	0	1	0	0	E2
3	0	1	1	1	0	0	FB

Dataset conversion is carried out to use the dataset for DT CART modeling because the CART model can only be applied to numerical datasets. The results of converting the dataset to numerical are shown in TABLE II.

In TABLE II, each categorical value is converted to a numerical value between 1 and 0 based on each attribute's meaning value (positive and negative) that affects the expert's decision. If the meaning is positive, it has a value of 1; if the meaning is negative, it has a value of 0. As in the attribute "Do subjects include vulnerable groups?" in row 1, if the value is Yes, it is worth 0. If the subject is a vulnerable group, then the research has the potential to be discontinued, so it is negative. Then the attribute "Secret is guaranteed" in row 1 has a guaranteed value, and then it is worth 1. If the confidentiality of the subject is guaranteed, then the research has the potential to be continued, so it has a positive meaning.

C. Decision Tree Modeling

At this stage, DT modeling is carried out based on the dataset prepared in the previous stage. Based on several previous studies, DT modeling still uses the CART model to compare the confusion matrix values. This research experiment focuses on proving the best modeling with the dataset used using the CART model with different split criteria, namely Gini and Entropy. Several scenarios were carried out to test the split criteria used to model the simplest DT with the smallest depth value a nd the highest accuracy value. This is an important factor that needs to be considered in DT modeling, namely the complexity of the resulting model. Thus, the more complex the resulting DT model, the more difficult it is to understand and it requires a longer computational time [10]. DT Modelling is done using the sci-kit-learn library in Python.

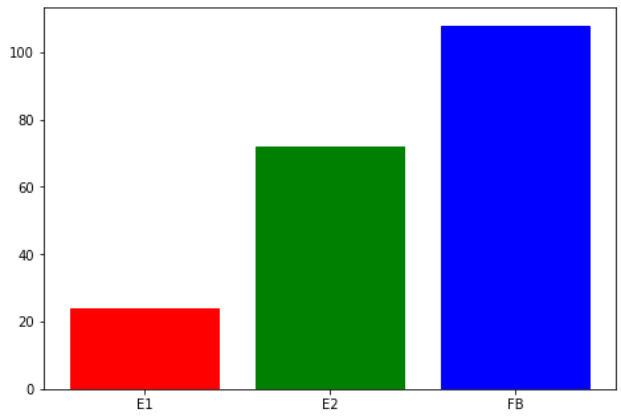

Fig. 3. Total dataset

IV. EXPERIMENT RESULT

A. The Result of Decision Tree Model

The DT model was successfully created at this stage using CART-Gini and CART-Entropy. Several DT modelling was conducted to determine whether the parameters used when building a DT CART model could affect the accuracy, precision, and recall values obtained. The purpose of this modelling is to find the most concise DT model with the highest accuracy value so that the decision tree model can be easily understood by experts and does not require a heavy level of computation to be able to provide decision results. The parameters to be tested are the maximum depth from 2 to get the maximum level of accuracy. The CART-Gini modelling is shown in Fig. 4, and the CART-Entropy modelling is shown in Fig. 5.

In Fig. 4, the results of the CART-Gini modeling have been successfully carried out. The results of the CART-Gini modeling differ from those of the CART-Entropy modeling shown in Fig. 5 when the max_depth parameter value is 4. There is no difference when modeling using the max_depth parameter with values 2 and 3, so the experiment continues and stops at the modeling experiment with parameter value max_depth 4. Differences are marked with a red box on CART-Gini classified as E2, while on CART-Entropy, the value is FB.

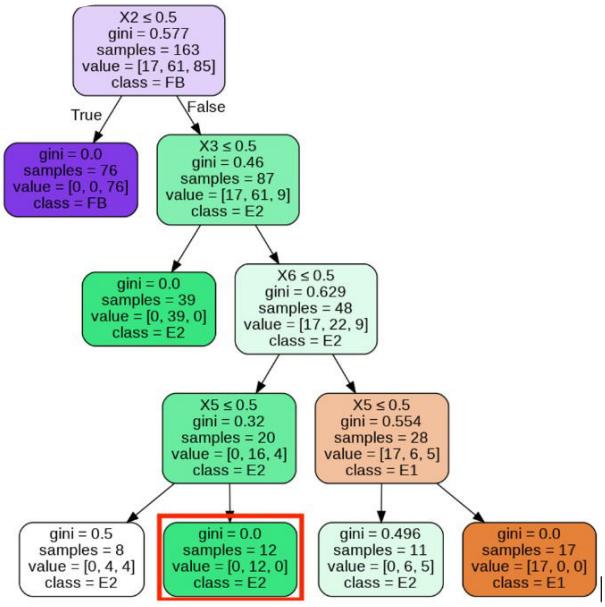

Fig. 4. CART-Gini

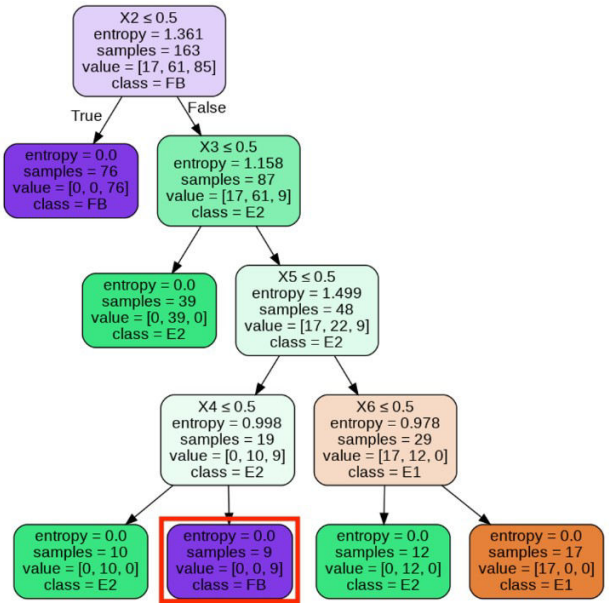

Fig. 5. CART-Entropy

From these differences, the evaluation results obtained are different, so a comparison is carried out later to find out the most optimal model in modeling DT with the dataset used.

B. Comparison Result

A decision tree model comparison is made. The results of the performance evaluation are shown in TABLE III.

In TABLE III, three experiments are carried out with the accuracy, precision, and recall values obtained in each model, namely CART-Gini and CART-Entropy. In experiments using the max_depth 2 and 3 parameter values, both modeling methods get the same accuracy, precision, and recall values. However, in the following experiment with a max_depth value of 4, CART-Entropy achieved the highest accuracy, precision, and recall values of 1. The CART-Gini model got an accuracy value of 0.93, precision 0.94, and recall 0.93. This shows that the DT model made with CART-Entropy and only a max_depth value of 4 can provide a concise DT model with the most optimal accuracy value. While for the CART-Gini model, there is no difference from the experiment with max_depth values of 3 and 4.

TABLE III. PERFORMANCE EVALUATION

M	MD	CART-Gini			CART-Entropy		
		AC	PR	RC	AC	PR	RC
1	2	0.76	0.70	0.76	0.76	0.70	0.76
2	3	0.93	0.94	0.93	0.93	0.94	0.93
3	4	0.93	0.94	0.93	**1**	**1**	**1**

where :
M : Model DT
MD : Maximal Depth Tree
AC : Accuracy
PR : Precision
RC : Recall

V. CONCLUSIONS

This research focuses on creating a knowledge base to assist the expert team in reviewing the research ethics protocol document by extracting knowledge from the expert to obtain a dataset that will be used as a reference for modeling the knowledge-based system. Using a knowledge base using DT can help the performance of the expert team in reviewing research ethics protocols that have limited resources with the increasing level of documented research ethics protocols being reviewed.

The proposed method results prove that the dataset can be modeled using DT with CART's most popular model. This modelling aims to get the most concise model tree with the most optimal accuracy value. The tree model that has been successfully created can be easily understood and helps the expert team complete the review of research ethics protocol documents that continue to increase. The comparison results show that the CART-Entropy model can provide accuracy, precision, and recall values of 1 with a compact tree model. In contrast, the CART-Gini can only provide an accuracy value of 0.93, precision 0.94, and recall 0.93.

Future work in this research is that a more in-depth review of the research ethics protocols can be carried out with all existing points without having to do extraction as in this study. Thus, the complexity of the reviews carried out by the expert team can be implemented as a whole into the knowledge-based model. Therefore, the results will also be more complex and flexible than using only six attributes, so a suitable DT modeling is needed to handle the complexity of the dataset used.

ACKNOWLEDGEMENT

This research was funded by Institut Teknologi Sepuluh Nopember (ITS), Surabaya through the Consortium research grant scheme (2021) and Department research grant (2022)

REFERENCES

[1] M. Guillemin, L. Gillam, D. Rosenthal, and A. Bolitho, "Human research ethics committees: Examining their roles and practices," *Journal of Empirical Research on Human Research Ethics*, vol. 7, no. 3., pp.38-49, 2012, doi: 10.1525/jer.2012.7.3.38.

[2] A. K. L. Jumelle *et al.*, "Ethical assessment in e-Health," *2014 IEEE 16th International Conference on e-Health Networking, Applications and Services, Healthcom 2014*, pp. 262–268, 2014, doi: 10.1109/HealthCom.2014.7001852.

[3] J. M. Sims, "A brief review of the Belmont report," *Dimensions of Critical Care Nursing*, vol. 29, no. 4, pp. 173-174, 2010, doi: 10.1097/DCC.0b013e3181de9ec5.

[4] World Health Organization, "Standards and Operational Guidance for Ethics Review of Health-Related Research with Human Participants," *Appropriate Technology*, 2011.

[5] C. for International Organizations of Medical Sciences and others, "International ethical guidelines for health-related research involving humans.," *International ethical guidelines for health-related research involving humans.*, 2017.

[6] T. Dudek and B. Śmiałkowska, "Integrated quality assessment of services in an adaptive expert system with a rule-based knowledge base," *Transportation Research Procedia*, vol. 39, pp. 34–41, 2019, doi: 10.1016/j.trpro.2019.06.005.

[7] T. V. Avdeenko and E. S. Makarova, "Integration of Case-based and Rule-based Reasoning Through Fuzzy Inference in Decision Support Systems," *Procedia Computer Science*, vol. 103, no. October 2016, pp. 447–453, 2017, doi: 10.1016/j.procs.2017.01.016.

[8] M. Dehghani Soufi, T. Samad-Soltani, S. Shams Vahdati, and P. Rezaei-Hachesu, "Decision support system for triage management: A hybrid approach using rule-based reasoning and fuzzy logic," *International Journal of Medical Informatics*, vol. 114, no. May 2017, pp. 35–44, 2018, doi: 10.1016/j.ijmedinf.2018.03.008.

[9] N. Kappelhof *et al.*, "Evolutionary algorithms and decision trees for predicting poor outcome after endovascular treatment for acute ischemic stroke," *Computers in Biology and Medicine*, vol. 133, no. November 2020, pp. 104414, 2021, doi: 10.1016/j.compbiomed.2021.104414.

[10] X. Guan, J. Liang, Y. Qian, and J. Pang, "A multi-view OVA model based on decision tree for multi-classification tasks," *Knowledge-Based Systems*, vol. 138, pp. 208–219, 2017, doi: 10.1016/j.knosys.2017.10.004.

[11] J. Yan, Z. Zhang, K. Lin, F. Yang, and X. Luo, "A hybrid scheme-based one-vs-all decision trees for multi-class classification tasks," *Knowledge-Based Systems*, vol. 198, pp. 105922, 2020, doi: 10.1016/j.knosys.2020.105922.

[12] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1023/A:1022643204877.

[13] S. L. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," *Machine Learning*, vol. 16, no. 3, pp. 235-240, 1994, doi: 10.1007/bf00993309.

[14] W. Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14-23, 2011, doi: 10.1002/widm.8.

[15] M. Mehta, R. Agrawal, and J. Rissanen, "SLIQ: A fast scalable classifier for data mining," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 1996, vol. 1057 LNCS, doi: 10.1007/bfb0014141.

[16] CIOMs &WHO, International Ethical Guidelines for Health-related Research Involving Humans, no. February. 2016.

[17] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008, doi: 10.1007/s10115-007-0114-2.

[18] S. Samet and A. Miri, "Privacy preserving ID3 using gini index over horizontally partitioned data," in *AICCSA 08 - 6th IEEE/ACS International Conference on Computer Systems and Applications*, 2008, pp. 645–651, doi: 10.1109/AICCSA.2008.4493598.

[19] B. HSSINA, A. MERBOUHA, H. EZZIKOURI, and M. ERRITALI, "A comparative study of decision tree ID3 and C4.5," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 2, pp. 13–19, 2014, doi: 10.14569/specialissue.2014.040203.

[20] C. Jin, F. Li, and Y. Li, "A generalized fuzzy ID3 algorithm using generalized information entropy," *Knowledge-Based Systems*, vol. 64, pp 13-21, 2014, doi: 10.1016/j.knosys.2014.03.014.

[21] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.

[22] L. Vanfretti and V. S. N. Arava, "Decision tree-based classification of multiple operating conditions for power system voltage stability assessment," *International Journal of Electrical Power and Energy Systems*, vol. 123, no. April, pp. 106251, 2020, doi: 10.1016/j.ijepes.2020.106251.

[23] G. Koulinas, P. Paraschos, and D. Koulouriotis, "A decision trees-based knowledge mining approach for controlling a complex production system," *Procedia Manufacturing*, vol. 51, no. 2019, pp. 1439–1445, 2020, doi: 10.1016/j.promfg.2020.10.200.

[24] M. Singh and J. Kumar Chhabra, "EGIA: A new node splitting method for decision tree generation: Special application in software fault prediction," *Materials Today: Proceedings*, 2021, doi: https://doi.org/10.1016/j.matpr.2021.05.325.