ONTOLOGY-BASED DECISION TREE MODEL FOR PREDICTION OF CARDIOVASCULAR DISEASE

Hakim El Massari

National School of Applied Sciences, Sultan Moulay Slimane University,
Lasti Laboratory, Khouribga, Morocco
h.elmassari@usms.ma

Noreddine Gherabi

National School of Applied Sciences, Sultan Moulay Slimane University,
Lasti Laboratory, Khouribga, Morocco
n.gherabi@usms.ma

Sajida Mhammedi

National School of Applied Sciences, Sultan Moulay Slimane University,
Lasti Laboratory, Khouribga, Morocco
sajida.mhammedi@usms.ac.ma

Zineb Sabouri

National School of Applied Sciences, Sultan Moulay Slimane University,
Lasti Laboratory, Khouribga, Morocco
zineb.sabouri@usms.ac.ma

Hamza Ghandi

National School of Applied Sciences, Sultan Moulay Slimane University,
Lasti Laboratory, Khouribga, Morocco
hamza.ghandi@usms.ac.ma

Abstract

Nowadays, cardiovascular diseases (CVD) are one of the most critical reasons for death. Thus, CVD prediction is a crucial challenge in the field of clinical data analysis. Researchers are using a variety of statistical and machine learning methods to assess immense amounts of complex medical data, to help doctors predict heart disease. In this paper, we proposed a new approach to predict CVD using ML techniques and Ontology to build an efficient ontology-based model able to predict accurately the presence of cardiac disease and establish an early diagnosis. the approach consists of extracting rules from the Decision Tree algorithm that differentiate the patients with or without cardiovascular disease then implementing these rules in the ontology reasoner using Semantic Web Rule Language (SWRL). The ontology model result reach high classification accuracy of 75% compared to the decision tree model. The approach can be employed in the medical field for the prediction of cardiovascular diseases.

***Keywords*: Ontology; SWRL; Machine learning; Cardiovascular disease; Decision tree.**

1. Introduction

The World Health Organization (WHO), has published a report on CVD, the report said that the main reason for CVD, is cause by mortality globally which is estimated around 17.9M people died of it in 2019, number were at the report that 32% of all death were worldwide. 85% suffered heart attack or stroke. In 2019, the too early deaths occurred before the age of 70 represent 38 percent of the aforementioned 17million are attributable for noncommunicable diseases. Given these statistical numbers, it is important to reveal cardiovascular disease as early as possible so that management with assistance and medicines can begin.

In addition, the report has come with that we can prevent most cardiovascular disease by addressing behavioral risk factors such as alcohol and tobacco use, unhealthy diet and obesity, physical inactivity, age, gender, and history of the family. Several studies have been accomplished to predict cardiovascular diseases and numerous machine learning models have been deployed, for the purpose of classifying and predicting heart disease diagnoses

[Saranya and Pravin 2020]. For instance, this study [Ali et al. 2021] aimed to determine the best machine learning methods with the highest accuracy for predicting patients with CVD. For this purpose, various supervised machine-learning algorithms were applied and compared for performance such as K-nearest neighbor, Random Forest, Decision tree, AdaboostM1, Logistic regression, and Multilayer perceptron [AL-Taie et al. 2021].

In this paper, we introduced a new approach consisting of merging Machine Learning and Semantic Web. On the one hand, ML algorithms autonomously learn to perform a task or make predictions from data and improve their performance over time, while the Semantic Web provides several formats for displaying data and ontological background knowledge. Merging the two together allowed us to build an ontology-based model able of predicting CVD with high accuracy. We have established a model ontological of knowledge representation through the OWL language, SWRL rules, and reasoner. To do so, we generate the rules from the decision tree algorithm, then we implemented them into the ontology by using Semantic Web Rule Language (SWRL).

The rest of the paper is structured as follows: we present related work in Section 2, then we describe the methodology in section 3. Discussion and Result in section 4. Finally, conclusion and future works in section 5.

2. Literature Review

Recently, CVDs detection and prediction has received much attention in the last years, and different approaches have addressed this problem. Machine learning and the semantic web are the main focus currently lie on.

In this survey [Swathy and Saruladha 2021], authors present a comparison of several classification and predictions for CVD. Such as, Data Mining Techniques, Machine Learning Models, and Deep Learning Models. [Mohd Faizal et al. 2021] A review of conventional and artificial intelligence (AI) risk prediction models in CVD, the authors found that while conventional risk prediction models are still the current gold standard and are generally used today, AI approaches such as machine learning and deep learning have been demonstrated to be practical to the analysis in meaningful ways in terms of the quantity of input and time reduction as compared to the conventional approach.

The ML approach was discovered to be objective and helpful. Various methods, like Random Forest (RF) Algorithm, Support Vector Machines (SVM), Genetic algorithm (GA), and Artificial Neural Networks (ANN), Particle Swarm Optimization (PSO), Decision Trees (DT), Naive Bayes (NB) and K-Nearest Neighbor (KNN) [Mohan et al. 2019; Shah et al. 2020; Kondababu et al. 2021]. In another research, [Faieq and Mijwil 2022] the authors introduce two machine learning algorithms SVM and ANN to Early diagnose heart disease, and the high-accuracy prediction results go to support vector machine. In [Lim et al. 2021], authors create a new hybrid model using support vector machine and artificial neural network to early detection of breast cancer disease, high accuracy of 98% was attained by the model.

Moreover, on another side, Deep learning is used to learn complex prediction models and it has been successfully applied to several issues in healthcare in particular. Among them, [Bensenane et al. 2022] authors proposed a model with high performance in terms of accuracy, the model is about a two-stage deep learning LSTM neural network to classify arrhythmias. In [Krishnan et al. 2021], the authors introduce a predictive deep learning model for heart disease prediction, by implementing RNN, GRU, and LSTM. high-value accuracy of 98% was achieved.

In another side, Semantic technology [Seeliger et al. 2019] uses formal semantics to assist AI techniques to understand language and processing information the way humans do. Also Semantic Web Technologies are used to support digital healthcare services [Barisevičius et al. 2018]. The use of ontologies and semantic web technologies [Manika et al. 2018; Mhammedi et al. 2022] like RDF, OWL and SPARQL can furnish the essential semantics for a variety of medical domains and, besides, can help as tools for making creative solutions technology to current issues. Moreover, ontology with big data can play an important role in healthcare issues [Mezghani et al. 2015; Irfan et al. 2019; El Massari et al. 2022].

Research has begun to publish regarding the integration of ML technologies and Semantic Web technology. [Khan et al. 2019] A machine-learning prediction model for a manufacturing network that can assist a project manager in allocating a newly received order among its suppliers, this model based on decision tree rules, ontology, and SWRL rules. Thus according to [JABARDI and Hadi 2020], they used Ontology and SWRL to create an effective model for detecting and classifying fake Twitter accounts. In [EL Massari et al. 2022; El Massari et al. 2022b], the authors developed and tested an ontology-based model that can predict diabetes patients using an ontology classifier based on a decision tree algorithm.

3. Methodology

The methodology adopted in this research is divided into three sections as shown in Fig. 1. The first step is data preprocessing which consist of transforming raw data into a useful format ready to be used in the classification process. In the second step, we used decision tree algorithm as a classifier capable to differentiate all possible rules to classify patients with or without CVD. In the third step, ontology engineering is employed for constructing the ontology and knowledge representation, after that we implemented the decision tree rules by switching or

converting them to rules-based reasoner of semantic web rule language which is used for detecting the absence or presence of cardiovascular disease.

Fig. 1. Methodology steps.

3.1. *Data preprocessing*

The cardiovascular Disease dataset is used in this paper. The dataset consists of 70 000 records of patients' data, 11 features plus a target. Table 1 gives a detailed explanation of the attributes. The dataset can be downloaded online [a] in CSV format, and should be converted to ARFF format to be ready for use by Weka software.

Attribute	Description
1- age	Age (days)
2- height	Height (cm)
3- weight	Weight (Kg)
4- gender	Gender (Male/Female)
5- ap_hi	Systolic blood pressure
6- ap_lo	Diastolic blood pressure
7- cholesterol	Cholesterol (1: normal, 2: above normal, 3: well above normal)
8- gluc	Glucose (1: normal, 2: above normal, 3: well above normal)
9- smoke	Smoking (binary)
10- alco	Alcohol intake (binary)
11- active	Physical activity (binary)
12- cardio	Target Variable (0 or 1). Presence or absence of cardiovascular disease

Table 1. Dataset feature's information.

3.2. *Decision tree algorithm*

The Decision Tree algorithm is a member of the supervised learning algorithm family used in statistics, data mining, and machine learning [Sabouri et al. 2022]. The decision tree approach, unlike other supervised learning algorithms, may also be utilized to solve regression and classification issues. We have selected the decision tree algorithm for many reasons, the result of the classification tree is easier to understand and interpret, and it supports multiple data types such as numeric, nominal, categorical, etc. The objective of employing a Decision Tree is to build a training model that can predict the class or value of the target variable by learning basic decision rules from past data (training data).
WEKA Software [Srivastava 2014] was used to create a predictive decision tree using the J48 classifier algorithm. The training set was validated using 10-fold cross-validation. Furthermore, all J4.8 settings were kept at their default values.
Fig. 2 illustrates the decision tree classification result, and Fig. 3 provides a snippet of the decision tree output (we got 580 leaves) that will be employed to generate SWRL rules which will be used in the ontology model.

[a] https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

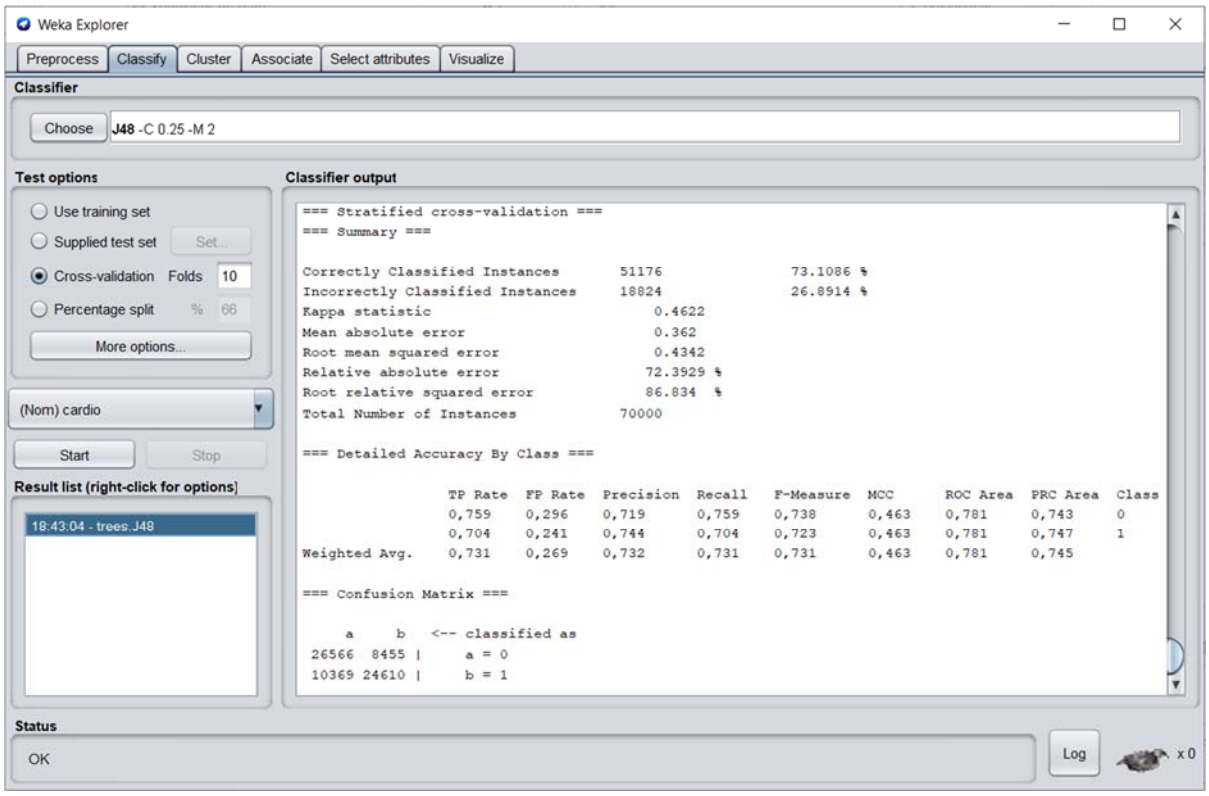

Fig. 2. Decision Tree classification result.

```
J48 pruned tree
------------------

ap_hi <= 129
|   age <= 19931
|   |   cholesterol = 1
|   |   |   ap_hi <= 118
|   |   |   |   ap_hi <= 24
|   |   |   |   |   ap_hi <= 12
|   |   |   |   |   |   height <= 167: 0 (41.0/5.0)
|   |   |   |   |   |   height > 167
      • • •
|   |   ap_lo <= 68
|   |   |   ap_hi <= 240
|   |   |   |   smoke <= 0: 1 (118.0/32.0)
|   |   |   |   smoke > 0
|   |   |   |   |   alco <= 0: 0 (8.0/1.0)
|   |   |   |   |   alco > 0: 1 (3.0)
|   |   |   ap_hi > 240
|   |   |   |   weight <= 79: 0 (10.0)
|   |   |   |   weight > 79: 1 (2.0)
|   |   ap_lo > 68: 1 (19270.0/3093.0)

Number of Leaves  :     580

Size of the tree :      1113
```

Fig. 3. Snippet of the decision tree output.

3.3. *Ontology construction*

Protégé is used to build the ontology, it's an open-source platform that offers a suite of tools to a growing user community for building domain models and knowledge-based applications with ontologies. Fig. 4 illustrates the graphical representation of the ontology. Two main classes "Patient" and "Diagnostic", two subclasses (absence and presence) of the Diagnostic class, and four subclasses (TP, TN, FN, FP) of "PatientEvaluationMetrics" class which is a subclass of "Patient". Moreover, the data properties are the same as the dataset attributes explained in Table 1.

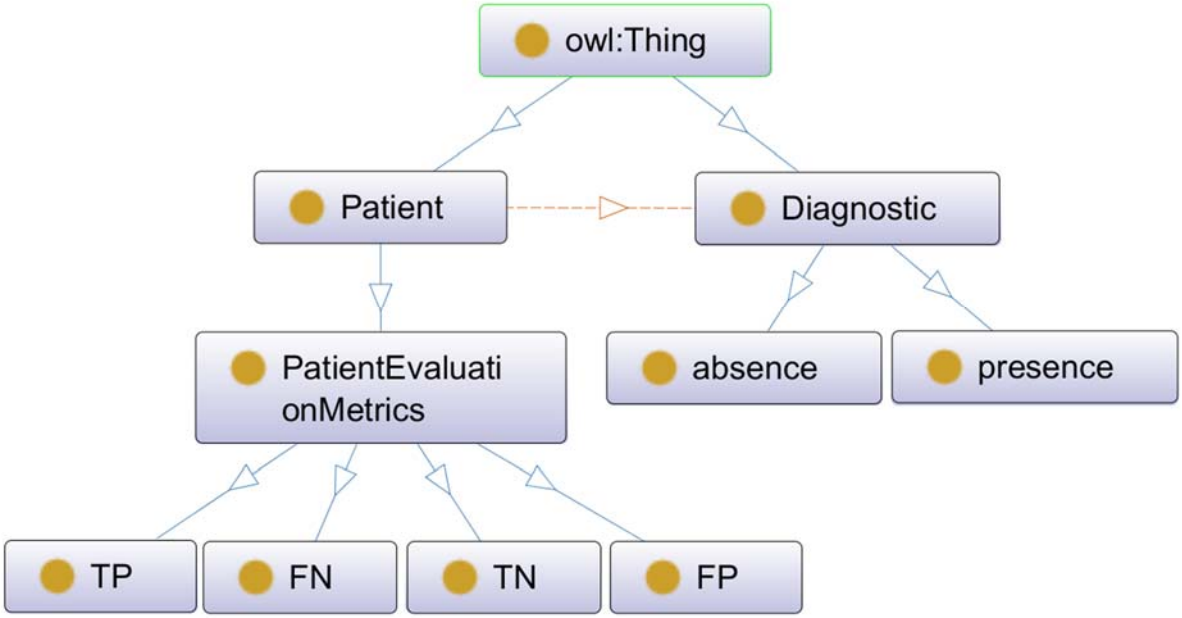

Fig. 4. The ontology graph.

After creating our ontology, we import the same cardiovascular disease dataset instances to it using the Cellfie plugin of Protégé software.

Semantic Web Language Rules

After creating and filling the ontology with the dataset, in this step we ought to determine the Semantic Web Language rules for reasoning. For that purpose, a java programing language is used to convert the extracted rules from the DT algorithm as illustrated in Fig. 3. Each leaf of the tree was extracted as a single SWRL rule using the java program. For instance, consider the SWRL rule taken from the first line of Fig. 3.

First leaf of DT algorithm

If `ap_hi` ≤ 129 && `age` ≤ 19931 && `cholesterol` = 1 && `ap_hi` ≤ 118 && `ap_hi` ≤ 24 && `ap_hi` ≤ 12 && `height` ≤ 167 *THEN* put the patient in `absence`

SWRL obtained

`Patient`(?pt) ^ `ap_hi`(?pt, ?AP) ^ *swrlb:lessThanOrEqual*(?AP, '129'^^*xsd:decimal*) ^ `age`(?pt, ?AG) ^ *swrlb:lessThanOrEqual*(?AG, '19931'^^*xsd:decimal*) ^ `cholesterol`(?pt, ?CH) ^ *swrlb:equal*(?CH, '1'^^*xsd:decimal*) ^ `ap_hi`(?pt, ?AP) ^ *swrlb:lessThanOrEqual*(?AP, '118'^^*xsd:decimal*) ^ `ap_hi`(?pt, ?AP) ^ *swrlb:lessThanOrEqual*(?AP, '24'^^*xsd:decimal*) ^ `ap_hi`(?pt, ?AP) ^ *swrlb:lessThanOrEqual*(?AP, '12'^^*xsd:decimal*) ^ `height`(?pt, ?H) ^ *swrlb:lessThanOrEqual*(?H, '167'^^*xsd:decimal*) → `absence`

The program java generated a total of 580 rules which is the identical number of leaves extracted from the DT algorithm, then the SWRL tab plugin was used to import these rules into Protégé. For executing SWRL rules and inferring new ontology axioms, we need an OWL reasoner such as Pellet [Khamparia and Pandey 2017], it has better functionality to work with OWL and SWRL rules, particularly because it allows to define custom SWRL rules. The Pellet reasoner uses CVD data and SWRL rules to produce assumption and delivers the final decision which is absence or presence of CVD. Fig. 5 illustrates the results of individuals inferred by type based on SWRL rules.

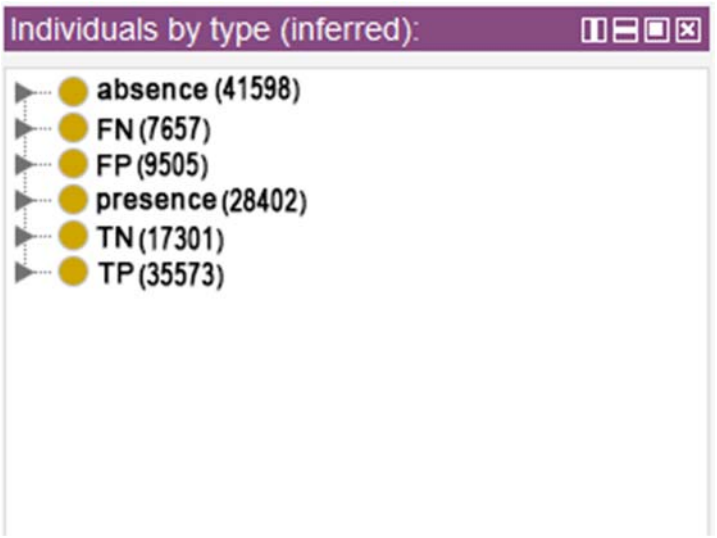

Fig. 5. Result of individuals inferred.

4. Result and discussion

To evaluate the results obtained from the decision tree algorithm and the ontology model, we used the confusion matrix and various performance measures derived from it, such as Accuracy, Precision, Recall, F-Measure, etc.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F-Measure} = 2 * \frac{PREC * REC}{PREC + REC} \quad (4)$$

Based on the calculations of the confusion matrix, Table 2 and Fig. 6 summarizes the results of ML and ontology classifiers used in this approach.

	Accuracy	Precision	Recall	F-Measure
Decision Tree	0.731	0.719	0.759	0.738
Ontology Model	0.755	0.789	0.823	0.805

Table 1. Evaluation metrics of the decision tree and ontology.

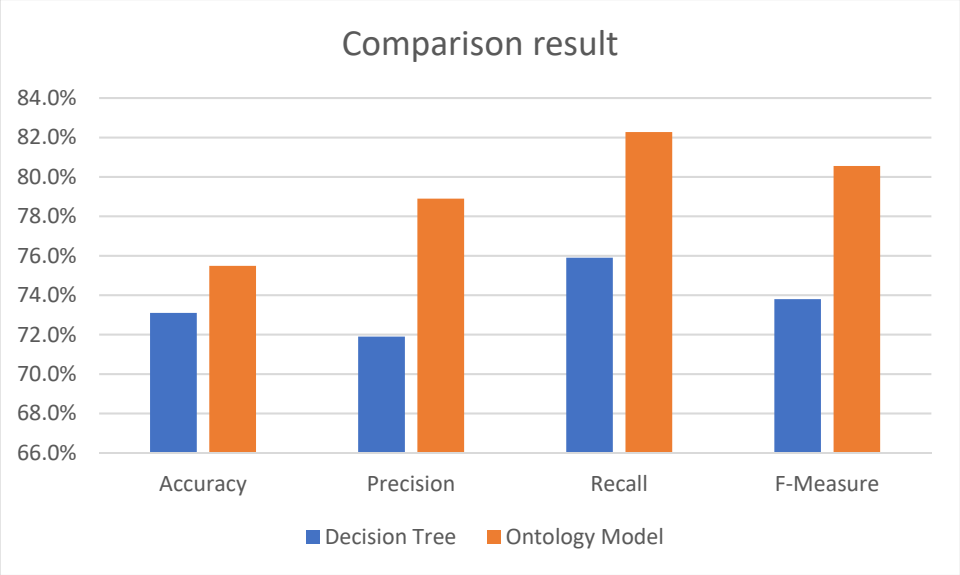

Fig. 6. Comparison result of the decision tree and ontology.

Based on results illustrated in Fig. 6 and Table 2, we observed that the ontology classifier is considered the best with high accuracy 75.5% compared to the decision tree classifier. We conclude that integrating machine learning and ontological reasoning may provide successful results. Furthermore, these comparative results demonstrate how OWL ontology's knowledge representation and reasoning capabilities may bring benefits other than classification. Furthermore, because the ontology classifier is an interpretable model, it may offer information on how the process reaches the decision. The ontology classifier produces equivalent and comparable results to machine learning classifiers. The results may also be interpreted by humans, and the rules can be modified or changed as needed.

5. Conclusion

In this paper, a new approach has been suggested to identify and classify patients with or without Cardiovascular disease, using Decision Tree algorithm, ontology, reasoner, and SWRL rules. This purpose was achieved by creating an ontology model based on a decision tree method and mapping it into an ontology using SWRL rules.

The ontology classifier is considered the best with high accuracy of 75.5% compared to the decision tree classifier, this proves the efficacy of this approach and can be employed in the medical field for the prediction of cardiovascular diseases particularly or other diseases generally.

As future work, we intend to automate the whole work by using the java APIs of all Materials used in this approach. In addition, we will compare our approach with other machi learning algorithms.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] ALI, M.M., PAUL, B.K., AHMED, K., BUI, F.M., QUINN, J.M.W., AND MONI, M.A. 2021. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Computers in Biology and Medicine 136, 104672.
[2] AL-TAIE, R.R.K., SALEH, B.J., SAEDI, A.Y.F., AND SALMAN, L.A. 2021. Analysis of WEKA data mining algorithms Bayes net, random forest, MLP and SMO for heart disease prediction system: A case study in Iraq. International Journal of Electrical and Computer Engineering (IJECE) 11, 6, 5229–5239.
[3] BARISEVIČIUS, G., COSTE, M., GELETA, D., ET AL. 2018. Supporting Digital Healthcare Services Using Semantic Web Technologies. The Semantic Web – ISWC 2018, Springer International Publishing, 291–306.
[4] BENSENANE, H., AKSA, D., OMARI, F.W., AND RAHMOUN, A. 2022. A deep learning-based cardio-vascular disease diagnosis system. Indonesian Journal of Electrical Engineering and Computer Science 25, 2, 963–971.
[5] EL MASSARI, H., MHAMMEDI, S., GHERABI, N., AND NASRI, M. 2022. Virtual OBDA Mechanism Ontop for Answering SPARQL Queries Over Couchbase. Advanced Technologies for Humanity, Springer International Publishing, 193–205.
[6] EL MASSARI, H., MHAMMEDI, S., SABOURI, Z., AND GHERABI, N. 2022. Ontology-Based Machine Learning to Predict Diabetes Patients. Advances in Information, Communication and Cybersecurity, Springer International Publishing, 437–445.
[7] EL MASSARI, H., SABOURI, Z., MHAMMEDI, S., AND GHERABI, N. 2022b. Diabetes Prediction Using Machine Learning Algorithms and Ontology. Journal of ICT Standardization, 319–338.
[8] FAIEQ, A.K. AND MIJWIL, M.M. 2022. Prediction of of heart diseases utilising support vector machine and artificial neural network. Indonesian Journal of Electrical Engineering and Computer Science 26, 1, 374–380.

[9] IRFAN, R., REHMAN, Z., ABRO, A., CHIRA, C., AND ANWAR, W. 2019. Ontology Learning in Text Mining for Handling Big Data in Healthcare Systems. Journal of Medical Imaging and Health Informatics 9, 4, 649–661.
[10] JABARDI, M. AND HADI, A. 2020. Twitter Fake Account Detection and Classification using Ontological Engineering and Semantic Web Rule Language. Karbala International Journal of Modern Science 6, 4.
[11] KHAMPARIA, A. AND PANDEY, B. 2017. Comprehensive analysis of semantic web reasoners and tools: a survey. Education and Information Technologies 22, 6, 3121–3145.
[12] KHAN, Z.M.A., SAEIDLOU, S., AND SAADAT, M. 2019. Ontology-based decision tree model for prediction in a manufacturing network. Production & Manufacturing Research 7, 1, 335–349.
[13] KONDABABU, A., SIDDHARTHA, V., KUMAR, BHK.B., AND PENUMUTCHI, B. 2021. A comparative study on machine learning based heart disease prediction. Materials Today: Proceedings.
[14] KRISHNAN, S., MAGALINGAM, P., AND IBRAHIM, R. 2021. Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction. International Journal of Electrical and Computer Engineering (IJECE) 11, 6, 5467–5476.
[15] LIM, T.S., TAY, K.G., HUONG, A., AND LIM, X.Y. 2021. Breast cancer diagnosis system using hybrid support vector machine-artificial neural network. International Journal of Electrical and Computer Engineering (IJECE) 11, 4, 3059–3069.
[16] MANIKA, P., XHUMARI, E., KTONA, A., AND DEMIRI, A. 2018. Application of Ontologies and Semantic Web Technologies in the Field of Medicine. RTA-CSIT.
[17] MEZGHANI, E., EXPOSITO, E., DRIRA, K., DA SILVEIRA, M., AND PRUSKI, C. 2015. A Semantic Big Data Platform for Integrating Heterogeneous Wearable Data in Healthcare. Journal of Medical Systems 39, 12, 185.
[18] MHAMMEDI, S., EL MASSARI, H., AND GHERABI, N. 2022. Composition of Large Modular Ontologies Based on Structure. Advances in Information, Communication and Cybersecurity, Springer International Publishing, 144–154.
[19] MOHAN, S., THIRUMALAI, C., AND SRIVASTAVA, G. 2019. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access 7, 81542–81554.
[20] MOHD FAIZAL, A.S., THEVARAJAH, T.M., KHOR, S.M., AND CHANG, S.-W. 2021. A review of risk prediction models in cardiovascular disease: conventional approach vs. artificial intelligent approach. Computer Methods and Programs in Biomedicine 207, 106190.
[21] SABOURI, Z., MALEH, Y., AND GHERABI, N. 2022. Benchmarking Classification Algorithms for Measuring the Performance on Maintainable Applications. Advances in Information, Communication and Cybersecurity, Springer International Publishing, 173–179.
[22] SARANYA, G. AND PRAVIN, A. 2020. A comprehensive study on disease risk predictions in machine learning. International Journal of Electrical and Computer Engineering (IJECE) 10, 4, 4217–4225.
[23] SEELIGER, A., PFAFF, M., AND KRCMAR, H. 2019. Semantic Web Technologies for Explainable Machine Learning Models: A Literature Review. PROFILES/SEMEX@ISWC.
[24] SHAH, D., PATEL, S., AND BHARTI, S.K. 2020. Heart Disease Prediction using Machine Learning Techniques. SN Computer Science 1, 6, 345.
[25] SRIVASTAVA, S. 2014. Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining. International Journal of Computer Applications 88, 10, 26–29.
[26] SWATHY, M. AND SARULADHA, K. 2021. A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques. ICT Express.

Authors Profile

Noreddine Gherabi, is a professor of computer science with industrial and academic experience. He holds a doctorate degree in computer science. In 2013, he worked as a professor of computer science at Mohamed Ben Abdellah University and since 2015 has worked as a research professor at Sultan Moulay Slimane University, Morocco. Member of the International Association of Engineers (IAENG). Professor Gherabi having several contributions in information systems namely: big data, semantic web, pattern recognition, intelligent systems…His research areas include Machine Learning, Deep Learning, Big Data, Semantic Web, and Ontology. He can be contacted at email: n.gherabi@usms.ma.

Hakim El Massari, received his master degree from Normal Superior School of Abdelmalek Essaadi University, Tétouan, Morocco, in 2014. Currently, he is preparing his Ph.D. in computer science at the National School of Applied Sciences, Sultan Moulay Slimane University, Khouribga, Morocco. His research areas include Machine Learning, Deep Learning, Big Data, Semantic Web, and Ontology. He can be contacted at email: h.elmassari@usms.ma.

Sajida Mhammedi, received her Ms Degree in Computer Engineering from Faculty of Science and Technologie, Beni Mellal Morocco, she worked as a visiting researcher at the Sultane Moulay Slimane University, her research interests include Machine Learning, Semantic Web, recommendation systems, Ontology, and Big Data.

Zineb Sabouri, is pursuing her Ph.D. in Computer Engineering from the National School of Applied Sciences of Khouribga, her area of interest is Machine Learning, Intelligent Systems, Deep Learning, and Big Data.

Hamza Ghandi, is pursuing her Ph.D. in Computer Engineering from the National School of Applied Sciences of Khouribga, her area of interest is Machine Learning, Intelligent Systems, Deep Learning, and Big Data.