International Journal of Innovative Computing 16(1–2) 279–285

UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

# Machine Learning–Driven Ontology System for Coronary Heart Disease Diagnosis

Glory Ita Okon*
Department of Computer Science, University of Uyo
UNIUYO
Uyo, Nigeria
Email: gloryita2095@gmail.com

Moses Effiong Ekpenyong
Department of Computer Science, University of Uyo
STEM Centre, University of Uyo, and Centre for Research, University of Uyo
Email: mosesekpenyong@uniuyo.edu.ng

Patience U. Usip
UNIUYO
Uyo, Nigeria
Email: patienceusip@uniuyo.edu.ng

Submitted: 15/9/2025. Revised edition: 28/11/2025. Accepted: 5/1/2026. Published online: 28/7/2026
DOI: https://doi.org/10.11113/ijic.v16n1-2.702

***Abstract*—Coronary heart disease is a disease that affects the larger coronary arteries on the surface of the heart and has been seen to be one of the major factors contributing to morbidity and death globally. This project presents a machine learning-driven ontology system for diagnosing coronary heart disease (CHD) based on patient data and clinical parameters. The system integrates a trained machine learning model, developed with the Framingham Heart Study dataset on the Weka platform, to predict the risk of CHD risk by extracting classification rules and embedding them within an ontology framework built in Protégé. The ontology models patient demographics, symptoms, and risk factors, providing a structured and semantically rich representation to support accurate diagnosis and clinical decision-making. The decision tree machine learning model achieved an 82.8 percent accuracy, and the ontology model was evaluated to be coherent and consistent. The system was further deployed to the web, enabling user-friendly interaction. By combining machine learning with ontology, this project not only leverages data-driven predictive power but also enhances interpretability, transparency, and adaptability in diagnosis. This synergy bridges the gap between statistical prediction and structured medical knowledge, ultimately contributing to more effective CHD risk assessment and improved patient outcomes in clinical practice.**

***Keywords*—Coronary heart disease, Machine learning driven prediction, Ontology model, Diagnosis, Knowledge representation, Intelligent system**

## I. INTRODUCTION

The heart is one of the most vital organs in the human body. It is a muscle that uses a network of arteries to pump blood throughout the body. The heart's importance cannot be overstated, as human life may be unsustainable if the heart ceases to function. Cardiovascular disease (CVD) is among the leading causes of morbidity and mortality worldwide, representing a significant public health burden. The World Health Organisation (WHO) recorded an estimate of 17.9 million people who died from CVDs in 2019, representing 32% of all global deaths [1].

Cardiovascular diseases (CVDs) encompass coronary heart disease (CHD), Cerebrovascular Disease, Peripheral Arterial Disease, Rheumatic Heart Disease, Congenital Heart Disease etc. Coronary heart disease (CHD) is usually caused by atherosclerosis (thickening or hardening of the arteries) that results in decreased blood flow through the vessel. This leads to decreased oxygen supply to the heart muscle, resulting in impaired heart function or damage to heart muscle cells (myocardial infarction or heart attack).

Despite advancements in medical technology, current methods for CHD diagnosis often fall short in terms of promptness and effectiveness. The traditional diagnostic approaches frequently rely on late-stage symptoms or invasive procedures, delaying the identification of at-risk individuals. The critical gap lies in the absence of rapid, accurate, and noninvasive diagnostic tools for early identification of individuals predisposed to CHD. Prompt identification is essential for executing prompt treatments, lifestyle adjustments, and individualized strategies for treatment.

Ontology formalization combined with machine learning (ML) offers a promising pathway to bridge this gap. By harnessing the predictive power of ML algorithms alongside the structured representation of domain knowledge provided by ontology, innovative diagnostic tools can be developed to analyse diverse patient data swiftly and accurately. Such tools have the potential to revolutionize CHD diagnosis, equipping healthcare professionals with efficient means of risk assessment and enabling the initiation of preventive measures [1].

## II. Related Works

Chandrasekhar *et al.* [2] explored improving the accuracy of heart disease prediction by applying six machine learning models: Random Forest, K-Nearest Neighbour, Logistic Regression, Naïve Bayes, Gradient Boosting, and AdaBoost. Their experiments were conducted using two benchmark datasets, Cleveland and IEEE Dataport. To enhance performance, they incorporated hyperparameter tuning with GridSearchCV and validated results using a five-fold cross-validation strategy. On the Cleveland dataset, Logistic Regression achieved the highest accuracy at 90.16%, whereas AdaBoost produced the best outcome on the IEEE Dataport dataset with 90% accuracy. When the six classifiers were combined into a soft voting ensemble, performance improved further, yielding 93.44% accuracy for the Cleveland dataset and 95% for the IEEE Dataport dataset. These results outperformed the strongest individual models. The study also assessed model robustness by examining accuracy loss across folds and evaluated outcomes using both accuracy and negative log loss metrics. Overall, the ensemble method demonstrated clear improvements on both datasets and showed superior results compared with previous heart disease prediction studies.

Esposito [3] proposed an automated approach for the examination of cardiovascular systems that could support cardiologists. The study proposed a structured, formal model based on SNOMED terminology to depict the cardiovascular system's architecture. The model defined either the anatomy of the cardiovascular system in normal patients or the anatomy characterised by malformations and abnormalities in congenital heart disease patients. The model was formalised in OWL ontologies and SWRL rules and then a logic reasoner was used to identify either congenital heart disease patients or other heart abnormalities and malformations they are affected by.

Knowledge representation of unstructured health data using ontology yielded efficient spatial knowledge retrieval [4]. Massari *et al.* [5] proposed an approach to predict cardiovascular disease (CVD) using machine learning techniques and ontology to build an efficient ontology-based model able to accurately predict the presence of cardiac disease and establish early diagnosis.

The approach consisted of using the decision tree algorithm's rules to distinguish between people who have cardiovascular disease and those who do not and then implementing these rules in the ontology reasoner using Semantic Web Rule Language (SWRL). The ontology model result reached a classification accuracy of 75% compared to the decision tree method.

Massari *et al*. [6] concentrated on developing an intelligent system based on ontology to predict diabetes patient diagnoses using decision tree model outcomes. The work introduced a novel method for using Semantic Web Rule Language in medicine by combining machine learning with semantic web ontology language to create an ontology-based model for diabetes patients. The rules that differentiate a positive patient from a negative patient were also extracted using the decision tree technique. Semantic Web Rule Language was then used to include these rules in the ontology. Therefore, in terms of diabetes prediction, the study verified the outcome produced by the ontology model using the outcomes of the decision tree model.

Prior studies have made significant contributions by improving predictive accuracy with machine learning techniques and advancing the use of ontologies for medical knowledge representation. However, most works either emphasize predictive modelling without strong semantic integration or focus on ontology without leveraging machine learning's data-driven insights. Additionally, practical deployment and interpretability remain underexplored. This study seeks to bridge these gaps by integrating decision tree–based machine learning with ontology modelling to enhance both accuracy and interpretability in coronary heart disease diagnosis.

## III. Methodology

The intelligent system for coronary heart disease diagnosis developed in this work comprises two components: the front end and the back end.

***The Front end:*** The front-end is the user-facing component of the system. It provides a user-friendly platform for healthcare professionals (like doctors) and patients to interact with the system, input data, and receive results in a user-friendly manner. The front end will be accessible through a web application.

***The Back end:*** As illustrated in Fig. 1, the back end constitutes the core of the system architecture, encompassing the data collection process, classification using the machine learning model, rule extraction, integration of rules into the ontology, and result evaluation. This layer hosts the machine learning-driven ontology model and performs the computational tasks of data processing, ontology reasoning, and diagnosis generation. The evaluated results are then transmitted to the front end, where users can interact with the system in a simplified and user-friendly manner.

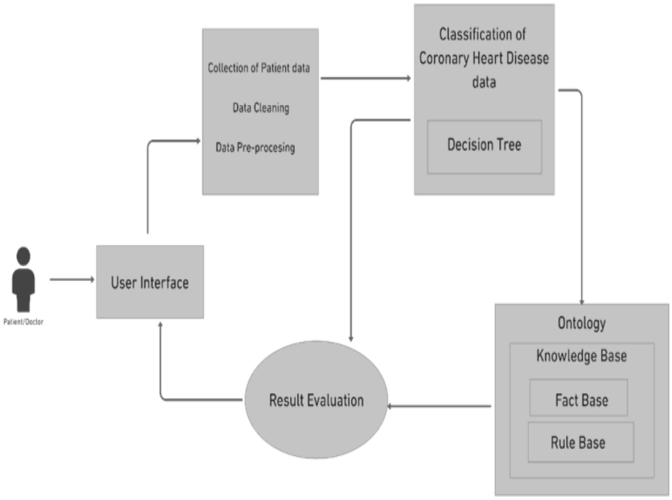



Fig. 1. Intelligent System for Coronary Heart Disease Architecture

### 3.1 User Interface

The User Interface functions as the principal point of interaction between the users (such as doctors, healthcare professionals, and patients) and the system. It is designed to be intuitive and user-friendly, allowing users to input data, view results, and interact with the system's features without understanding the underlying complexities. Its key features include:

a. It facilitates the entry of patient information, including medical history, symptoms, and risk factors.

b. Provides validation to ensure data is complete and accurate before submission.

c. Displays diagnostic results, including risk scores and predictions, in a clear and understandable format.

d. It also includes secure login and authentication features to protect patient data and restrict access to authorised personnel.

e. It offers different views and functionalities based on the user's role (e.g., doctor, patient, administrator).

### 3.2 Data Collection and Preparation

The data used for the model training containing information like clinical history, demographic details, and behavioural details collected from medical tests and questioning was gotten from an online repository Kaggle.com. The dataset consists of patient information obtained from the heart study in Framingham, Massachusetts. The dataset has 4,240 records, 15 attributes – gender, age, educational level, currentsmoker, cigsperday, BPmeds, prevalentstroke, prevalentHYP, diabetes, totchol, sysBP, diaBP, BMI, heartrate, glucose and a target variable - TenYearCHD where each record corresponds to a new patient.

The collected data was cleaned to remove missing rows, duplicates, and outliers using filters from the Weka tool. Relevant features were also extracted, and the data normalised to ensure consistency and accuracy of results.

### 3.3 Model Training

Model Training is the process where the machine learning model learns from historical data. This component is crucial as it determines the accuracy and reliability of the system in diagnosing coronary heart disease. Its key features include:

a. The collection of large datasets of patient information, including diagnosed cases of coronary heart disease and corresponding symptoms, risk factors, and outcomes.

b. The data is divided into training and testing sets, with 20% put aside for testing and the remaining 80% used to train the model.

c. Model development, in which the training dataset is used to train the decision tree model. The decision tree was chosen because it is interpretable, handles both categorical and numerical data, and allows direct extraction of rules for integration into the ontology.

d. Model validation utilises techniques like cross-validation to verify the model is generalisable and performs better with unfamiliar data.

### 3.4 Ontology Model

The Ontology Model provides a structured representation of knowledge related to coronary heart disease. It connects medical concepts (e.g., symptoms, risk factors) and their relationships, enriching the data before it is processed by the machine learning model. It also maps patient data (such as symptoms and risk factors) to corresponding concepts within the ontology. Its key feature is the knowledge base which comprises the fact base holding all defined and organised relevant coronary heart disease-related concepts and their interrelationships and the rule base containing the classification rules extracted from the machine learning algorithms. Fig. 2 shows the schema of the ontology for CHD patient.

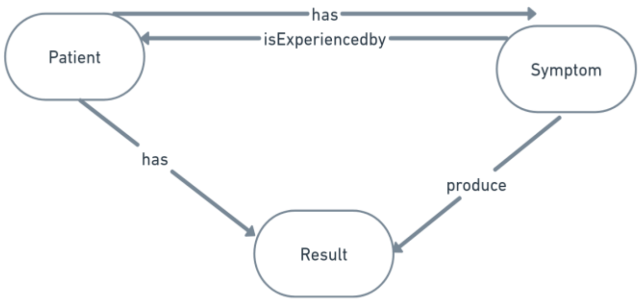



Fig. 2. Schema of CHD patient Ontology

## IV. Implementation and Results

The realisation of the proposed system and the expected outcome from implementing the various design objectives are as follows:

### 4.1 Model Development with WEKA

The dataset was uploaded to Weka and pre-processed using the available filters, after which a predictive decision tree model was built using the J48 classifier. J48 was chosen over other decision tree models in Weka because it is a well-established implementation of the C4.5 algorithm, capable of handling both categorical and numerical attributes, managing missing values, and producing clear, interpretable rules suitable for ontology integration. The model was trained with default settings and validated using 10-fold cross-validation. Fig. 3 presents the decision tree classification result, while Fig. 4 shows a snippet of the extracted classification rules.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3030               82.8321 %
Incorrectly Classified Instances       628               17.1679 %
Kappa statistic                          0.1129
Mean absolute error                      0.2378
Root mean squared error                  0.382
Relative absolute error                 92.0748 %
Root relative squared error            106.3102 %
Total Number of Instances             3658

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.954    0.869    0.859      0.954    0.904      0.129    0.616     0.867     0
                 0.131    0.046    0.336      0.131    0.189      0.129    0.616     0.243     1
Weighted Avg.    0.828    0.744    0.780      0.828    0.795      0.129    0.616     0.772

=== Confusion Matrix ===

    a    b   <-- classified as
 2957  144 |    a = 0
  484   73 |    b = 1
```

Fig. 3. Decision Tree Classification Result

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

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

age <= 48
|   diabetes = 0: 0 (1780.0/133.0)
|   diabetes = 1
|   |   glucose <= 225
|   |   |   cigsPerDay <= 20: 0 (15.0)
|   |   |   cigsPerDay > 20: 1 (3.0/1.0)
|   |   glucose > 225: 1 (5.0/1.0)
age > 48
|   glucose <= 122
|   |   sysBP <= 143.5
|   |   |   BPMeds = 0
|   |   |   |   prevalentStroke = 0
|   |   |   |   |   male = 0: 0 (576.0/71.0)
|   |   |   |   |   male = 1
|   |   |   |   |   |   age <= 57: 0 (330.0/54.0)
|   |   |   |   |   |   age > 57
|   |   |   |   |   |   |   prevalentHyp = 0
|   |   |   |   |   |   |   |   currentSmoker = 0
|   |   |   |   |   |   |   |   |   glucose <= 74: 0 (23.0/1.0)
|   |   |   |   |   |   |   |   |   glucose > 74
|   |   |   |   |   |   |   |   |   |   diaBP <= 81.5
|   |   |   |   |   |   |   |   |   |   |   sysBP <= 128.5
|   |   |   |   |   |   |   |   |   |   |   |   sysBP <= 119.5: 0 (20.0/2.0)
```

Fig. 4. Snippet of Decision Tree Rules

### 4.2 Ontology Development with Protégé

After consulting the SNOMED-CT and NHLBI documentation, information related to coronary heart disease was obtained and used to create a simple ontology consisting of classes, object properties, and data properties; axioms were also imported from the same dataset used to train the ML model in Weka (Fig. 5). The decision tree output was then systematically transformed into SWRL rules. The conditions at each decision node were mapped to ontology data properties, with numeric thresholds expressed using SWRL built-ins. The terminal leaf nodes of the decision tree were translated into class assertions representing the diagnostic outcome. This process ensured that the decision logic captured by the model was preserved in a machine-interpretable form. The generated rules were then integrated into the ontology model for reasoning, as illustrated in Fig. 6.

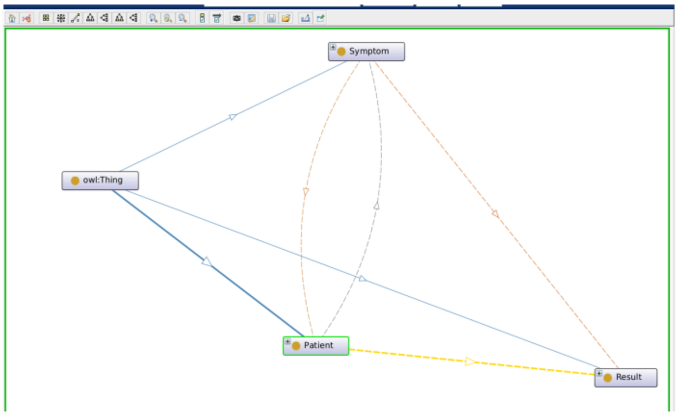



Fig. 5. Graphical Representation of the Ontology

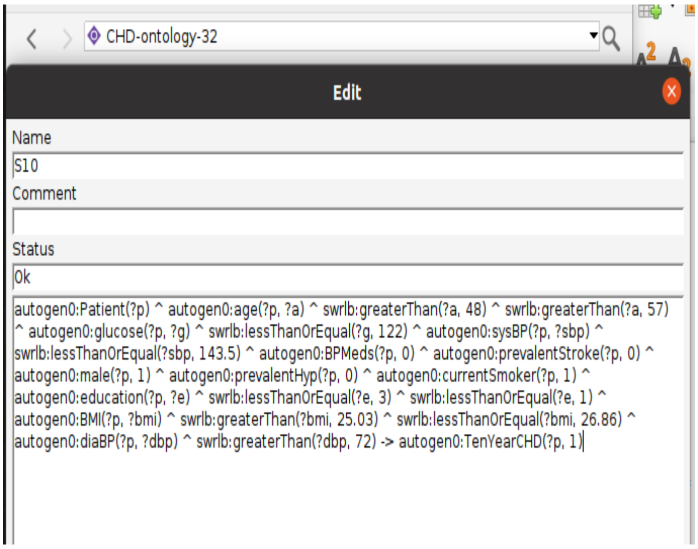



Fig. 6. Snippet of SWRL Rules

### 4.3 System Integration

Users of the system will need to be able to interface with the system, input data, and get real-time feedback and diagnoses based on the ontology's reasoning capabilities.

A JavaScript framework - React JS was used to create the user interface that allows for interaction with the system; using Express JS, an API (Application programming interface) endpoint was created and provided as a back end to process the data from the front-end, apply the ontology and return the result. Fig. 7 shows the user interface and the output diagnosis.

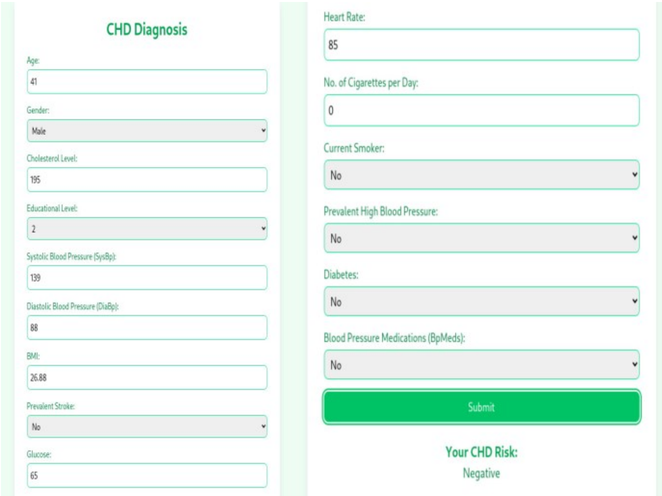



Fig. 7. User Interface and Diagnosis

## 4.4 Evaluation of model/Results

The decision tree model was accessed using metrics such as accuracy, precision, recall, F1-score, and the confusion matrix. The results show the classification metric precision, Recall and F1 score per class. These metrics are calculated using the True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) as defined in equations (1)-(3).

$$Precision = \frac{TP}{(TP+FP)} \quad \text{- - - - (1)}$$

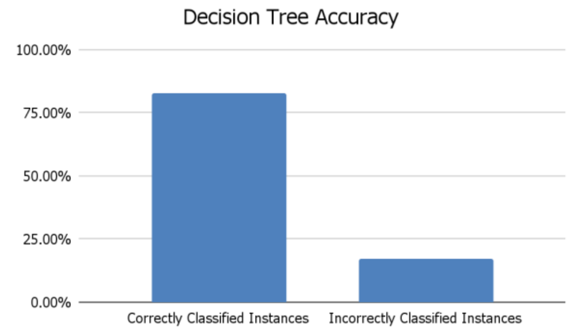
Recall is the ability of the classifier to find all positive instances. It is given by

$$Recall = \frac{TP}{(TP+FN)} \quad \text{- - - - (2)}$$

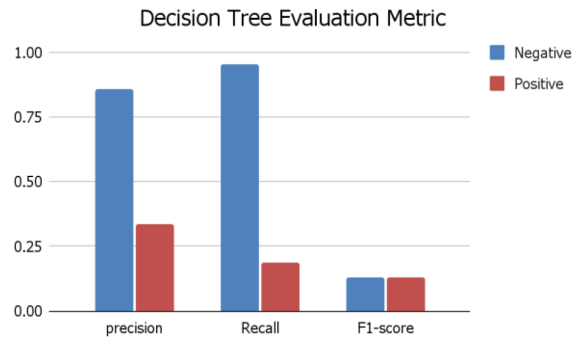
The F1 score is a weighted harmonic mean of the precision and recall. It is given as shown in equation 3

$$F1\ score = \frac{2*(Recall*Precision)}{(Recall+Precision)} \quad \text{- - - - (3)}$$

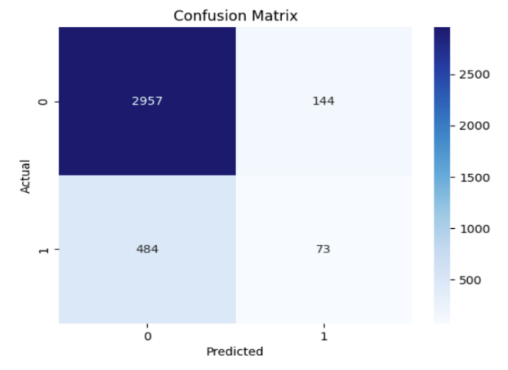
Ultimately, the model's accuracy was 82.8%, with per-class metrics showing its ability to differentiate between CHD-positive and CHD-negative cases. The confusion matrix further highlighted the distribution of correct and incorrect classifications. For ontology evaluation, coherence and consistency were first confirmed with the Pellet reasoner in Protégé (Fig. 9). Queries executed on the Apache Jena-Fuseki server (Fig. 10) successfully retrieved patient details including age, symptoms, and CHD status, demonstrating that rules integrated from the decision tree produced meaningful inferences. OntoMetrics (Fig. 11) was chosen for additional evaluation because it provides automated structural and quality metrics (e.g., class richness, relationship richness, inheritance depth) that complement logical reasoning results. The OntoMetrics results confirmed that ontology is well-structured and sufficiently rich to support knowledge representation and retrieval.




(a)




(b)




(c)

Fig. 8. Decision Tree Evaluation Results: (a) Decision Tree Accuracy; (b) Decision Tree Evaluation metric; (c) Confusion Matrix

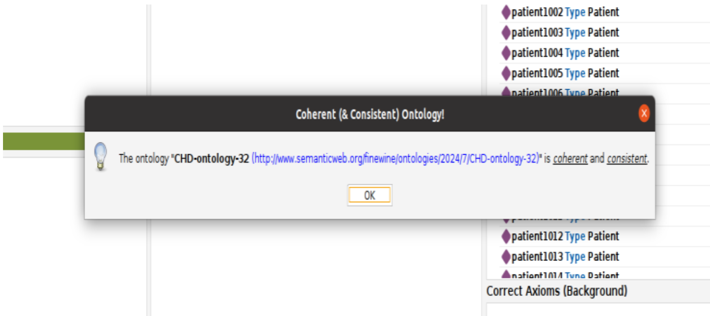



Fig. 9. Pellet Reasoner Result

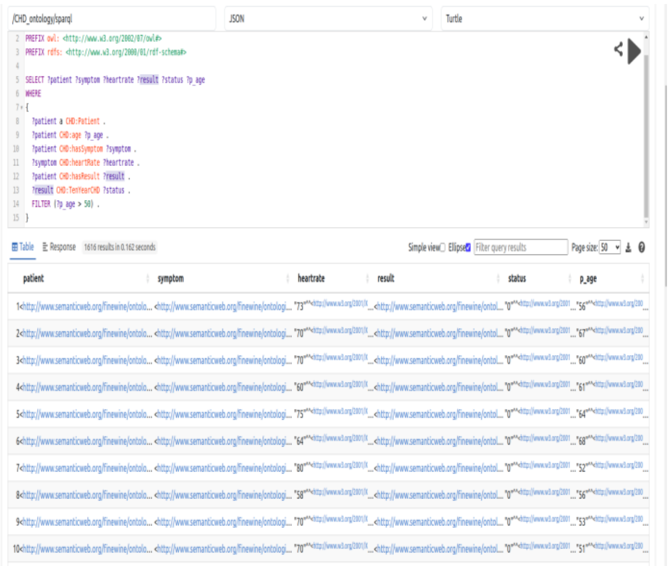



Fig. 10. Sample SPARQL query and query result from Apache Jena-Fuseki server

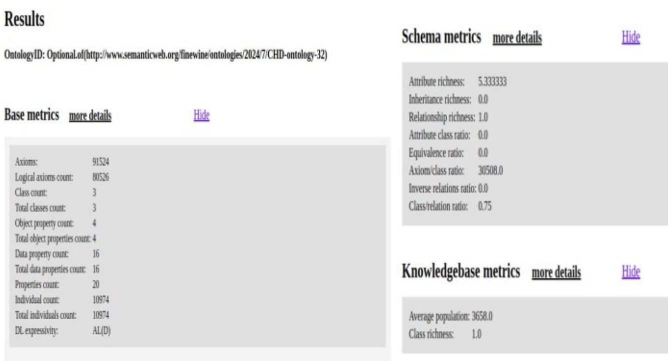



Fig. 11. Ontology Evaluation result from OntoMetrics

## V. CONCLUSION

This work developed a machine learning-driven ontology model for diagnosing coronary heart disease (CHD). The model integrates domain knowledge from medical ontologies with machine learning algorithms to enhance the accuracy and interpretability of CHD diagnosis. A comprehensive ontology model was created incorporating medical concepts, relationships, and rules related to CHD using the Protege software. Classification rules extracted from the decision tree machine learning algorithm using Weka were integrated with the ontology model to analyse patient records and identify patterns indicative of CHD. The model was later queried on the Apache Jena-Fuseki server and deployed to the web. The decision tree machine learning model achieved an accuracy of 82.8% in diagnosing CHD and the ontology model was confirmed to be coherent and consistent. The model demonstrated robustness and potential generalizability across different patient scenarios. However, the dataset used is a benchmark dataset and may not fully cover the variability and complexity of real-world patient populations, potentially affecting the model's external validity. In addition, the implementation of the ontology with machine learning rules was performed manually, which may limit scalability. Despite these limitations, the project demonstrates the potential of machine learning-driven ontology models in improving CHD diagnosis. By integrating domain knowledge and machine learning, we can develop more accurate, interpretable, and clinically relevant diagnostic systems.

The outcomes of this research can contribute to enhancing the CHD diagnosis process and strengthening decision support in healthcare. Future work could focus on automating the rule-extraction and ontology integration process, expanding the approach to other cardiovascular diseases, and exploring real-world clinical applications. Integrating the model with electronic health records (EHRs) would allow seamless access to real-time patient data, thereby improving usability, clinical adoption, and impact in routine healthcare practice.

## ACKNOWLEDGMENT

The authors would like to acknowledge the support of Tertiary Education Trust Fund (TETFund), Nigeria for the financial support on dissemination of this research.

## CONFLICT OF INTEREST STATEMENT

The authors agree that this research was conducted in the absence of any self-benefits, commercial or financial conflicts and declare the absence of conflicting interests with the funders.

## AUTHORS' CONTRIBUTIONS

Glory Ita carried out the research, wrote and revised the article. Patience Usip conceptualised the central research idea and provided the theoretical framework. Moses Ekpenyong supervised the writing of this manuscript for dissemination and approved the article submission.

## References


[1] Massari, E. H., Noreddine, G., Sajida, M., Hamza, G., Mohamed, B., & Naqvi, M. R. (2022). The impact of ontology on the prediction of cardiovascular disease compared to machine learning algorithms. *International Journal of Online and Biomedical Engineering (iJOE), 18*(11), 143–157. https://doi.org/10.3991/ijoe.v18i11.32647

[2] Chandrasekhar, N., & Peddakrishna, S. (2023). *Enhancing heart disease prediction accuracy through machine learning techniques and optimization*. School of Electronics Engineering, VIT-AP University, Amaravati, India.

[3] Esposito, M. (2008). Congenital heart disease: An ontology-based approach for the examination of the cardiovascular system. In I. Lovrek, R. J. Howlett, & L. C. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems (KES 2008)* (Vol. 5177, pp. 618–625). *Lecture Notes in Computer Science*. Springer. https://doi.org/10.1007/978-3-540-85563-7_65

[4] Usip, P. U., Ekpenyong, M. E., Ijebu, F. F., Usang, K. J., & Udo, I. J. (2021). PeNLP parser: An extraction and visualization tool for precise maternal, neonatal and child healthcare geo-locations from unstructured data. In *Deep Learning in Biomedical and Health Informatics*.

[5] Massari, E. H., Noreddine, G., Sajida, M., Zineb, S., & Hamza, G. (2022). Ontology-based decision tree model for prediction of cardiovascular disease. *Indian Journal of Computer Science and Engineering, 13*(3), 851–859. https://doi.org/10.21817/indjcse/2022/v13i3/221303143

[6] Massari, E. H., Sajida, M., Zineb, S., & Noreddine, G. (2022). Ontology-based machine learning to predict diabetes patients. In *Proceedings of the International Conference* (pp. 437–445). https://doi.org/10.1007/978-3-030-91738-8_40


.