science, engineering and health studies

Research Article

*Science, Engineering and Health Studies*
*https://lio1.tci-thaijo.org/index.php/sehs*
*ISSN (Online): 2630-0087*

# A smarter forest: Enhancing cardiovascular risk prediction using a knowledge-based random forest

*Sirichanya Chanmee*[1] *and Kraisak Kesorn*[2*]

[1] Faculty of Science and Agricultural Technology, Rajamangala University of Technology Lanna Phitsanulok, Phitsanulok 65000, Thailand
[2] Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok 65000, Thailand

***Corresponding author**:
*Kraisak Kesorn*
*kraisakk@nu.ac.th*

***Received***: *3 December 2024*
***Revised***: *22 February 2025*
***Accepted***: *3 March 2025*
***Published***: *25 December 2025*

***Citation***:
*Chanmee, S., & Kesorn, K. (2025). A smarter forest: Enhancing cardiovascular risk prediction using a knowledge-based random forest. Science, Engineering and Health Studies, 19, 25020007.*

## ABSTRACT

Predicting heart disease and other cardiovascular issues accurately is critical for enabling early intervention and improving patient outcomes. This study proposed the semantic random forest (SRF) framework, which enhances the classification performance of conventional random forest (RF) algorithms for heart disease prediction. The conventional RF framework is augmented through the integration of knowledge from a formal ontology model that encapsulates domain-specific medical knowledge, thereby providing a structured representation of concepts, relationships, and axioms. The SRF framework utilizes this ontology during the classification process to yield more precise predictions. The effectiveness of the proposed SRF framework was evaluated against the conventional RF, AdaBoost, and gradient boosting algorithms, with a focus on their ability to classify heart disease instances accurately. Experimental results demonstrate that the proposed SRF framework outperformed the baseline algorithms on two datasets, achieving the highest accuracy and Matthews correlation coefficient values of 0.8296 and 0.6589 on the University of California at Irvine dataset and 0.9856 and 0.9706 on Mendeley dataset, respectively. The results demonstrate that ontology-based structured knowledge significantly improves the classification power of traditional RF algorithms, which highlights this knowledge-driven approach's potential to predict heart disease risks in computer-aided medical diagnoses.

**Keywords:** random forest; heart disease; machine learning; knowledge base; classification; ontology

## 1. INTRODUCTION

Heart disease is a prevalent cause of morbidity and mortality globally, with approximately 17.9 million deaths reported in 2019 (AlGhanem et al., 2020; Ed-daoudy et al., 2023), and cardiovascular diseases, including heart disease and stroke, comprise approximately 31% of all deaths globally. The risk factors for heart disease include nonmodifiable demographic factors, e.g., age, gender, and family history, and modifiable lifestyle factors, e.g., smoking, physical inactivity, obesity, diabetes, and hypertension (Ishak et al., 2020). Predicting an individual's cardiovascular disease risk accurately is essential to facilitate timely preventive measures and mitigate health burdens. Conventional risk prediction models utilize multivariate regression techniques based on a limited set of established risk factors (Kwon et al., 2020); however, these models frequently demonstrate suboptimal predictive performance due to restrictive assumptions and their inability to capture the complex, nonlinear relationships among risk factors (Shouman et al., 2011). Recent advancements in machine learning (ML)

present opportunities to improve cardiovascular risk prediction accuracy. ML algorithms, notably the random forest (RF) algorithm, have exhibited promising results in various medical applications, including cardiovascular disease prediction and diagnosis (Hossain et al., 2023; Shanmugasundaram et al., 2018).

The proposed approach combines the advantages of RFs with domain-specific knowledge about cardiovascular risk factors to enhance the predictive accuracy and clinical utility of the traditional RF model. Typically, conventional risk prediction models for cardiovascular disease rely on multivariate regression techniques, e.g., logistic regression and Cox proportional hazard models (Harrell, 2001), which use a limited set of well-established risk factors to estimate an individual's risk of developing cardiovascular disease. However, these models apply restrictive modeling assumptions, they cannot detect the complex, nonlinear relationships between risk factors, and they provide suboptimal predictive performance (Tripoliti et al., 2017).

RF comprises ensemble-based ML algorithms that combine multiple decision trees (DTs) to make predictions. Unlike traditional regression models, RFs can handle the complex, nonlinear relationships between variables and are less susceptible to overfitting (Shouman et al., 2011). Several studies have demonstrated the potential of RFs and other ML algorithms in cardiovascular disease prediction and diagnosis (Dinh et al., 2019). These models accurately identify individuals at high risk of developing cardiovascular disease and detect specific cardiovascular conditions, e.g., heart failure. The RF method has seen significant advancements through the introduction of the WildWood (WW) algorithm, which was proposed by Gaïffas et al. (2023). This novel ensemble algorithm enhances traditional RF methods by innovatively utilizing out-of-bag samples for prediction aggregation. Conventional RF techniques rely solely on leaf nodes for predictions; however, WW's distinguishing feature is its ability to aggregate predictions from every possible subtree within each decision tree. This comprehensive approach employs an exponential weighting mechanism that incorporates the complete tree structure, which results in more refined decision boundaries. Through context tree weighting and histogram-based split optimization, WW delivers exceptional computational performance while producing results that match or surpass current state-of-the-art algorithms. This combination of improved accuracy and interpretability establishes WW as a significant advancement in ML ensemble methods. Incorporating knowledge-based approaches into ML algorithms is an important area of research. These methods leverage domain-specific knowledge to improve ML models' performance and interpretability (Miraftabzadeh et al., 2021). A representative example is theory-guided data science, which emphasizes the integration of domain knowledge into the data analysis and model development process (Karpatne et al., 2017; Chanmee & Kesorn, 2021). This approach has been examined in various fields and offers promising avenues for scientific discovery and supporting decision-making processes. However, most existing ML-based approaches treat risk factor relationships as purely data-driven patterns without incorporating domain-specific knowledge, which can result in models that lack interpretability and clinical relevance. In addition, ML models trained solely on historical data frequently fail to generalize well to new populations because they do not explicitly integrate expert-driven insights into cardiovascular disease mechanisms. To address these gaps, this study proposed a semantic RF (SRF) model that integrates domain expertise into the learning process to improve the performance of cardiovascular risk prediction. Unlike traditional RF models, which rely solely on statistical correlations, the proposed SRF method enhances decision-making processes by embedding structured knowledge about cardiovascular risk factors, thereby improving both interpretability and predictive accuracy.

Chanmee and Kesorn (2021) introduced the "semantic data mining" concept after surveying the use of domain ontologies and overcame the limitations of traditional methods. A knowledge-based approach enables a deeper understanding of data that goes beyond statistical patterns to uncover meaningful insights. Chanmee and Kesorn (2023) also proposed the semantic DT (SDT) method to incorporate a knowledge base into a DT algorithm, which improves the traditional Iterative Dichotomiser 3 (ID3) algorithm by leveraging domain expertise. This integration allows the SDT method to exploit structured knowledge, potentially resulting in more accurate and interpretable DTs. Inspired by this approach, the proposed ensemble-based SRF method integrates expert knowledge into the DT induction process. Similar knowledge-based enhancements have been applied successfully in other domains and ML techniques, e.g., the ARIMAXS model (Juraphanthong & Kesorn, 2024, 2025) for COVID-19 incidence prediction, highlighting the value of integrating structured semantic information into predictive modeling. The primary contributions of this study are threefold. First, it proposes the SRF framework, which incorporates domain-specific knowledge into the RF model development process. Second, it evaluates the predictive performance of the proposed SRF framework by comparing it with traditional RF methods and other ML algorithms on cardiovascular risk prediction tasks. Finally, the findings demonstrate that incorporating domain knowledge can enhance the accuracy and clinical utility of ML models for cardiovascular risk prediction.

This study builds on our previous research on semantic data mining (Chanmee & Kesorn, 2021) and SDTs (Chanmee & Kesorn, 2023, 2024), which demonstrated that incorporating domain knowledge into tree-based algorithms enhances their reasoning capability. The main difference between the current study and our previous work (Chanmee & Kesorn, 2024) is the application of various sampling methods to determine which method is the most suitable for collaborating with ontology knowledge to enhance the performance of the tree-based ensemble approach. By incorporating structured cardiovascular knowledge into the RF modeling process, the proposed SRF framework aimed to achieve higher predictive accuracy by integrating expert-driven risk factor relationships, improved interpretability for clinical decision-making, thereby enabling medical professionals to better understand and trust model predictions, and greater generalizability to diverse populations, thereby reducing the bias inherent in purely data-driven approaches.

## 2. MATERIALS AND METHODS

The architecture of the proposed SRF framework is shown in Figure 1. As can be seen, the SRF framework comprises three primary components, i.e., data preparation, SRF construction, and evaluation. Each component is described in detail in the following subsections, including the materials and processes utilized in this study.

### 2.1 Materials

To evaluate the proposed SRF framework, we utilized two public cardiovascular disease datasets. The first dataset, sourced from the University of California, Irvine (UCI) data repository (Andras Janosi, 1988), contains 303 records with 14 attributes. The second dataset, found in the Mendeley data repository (Maghdid & Rashid, 2022), is an unbalanced dataset containing 1,319 samples and nine attributes. Note that both datasets include patient demographics, disease signs, and physical test results, e.g., age, gender, heart rate, and blood pressure. The details of these datasets are given in Table 1. The main distinction between these datasets lies in their attribute types. The first dataset comprises both nominal and numerical attributes, while most of the attributes in the second dataset are numerical. The presence of numerous nominal attributes in the first dataset raises concerns about potential bias in the attribute selection for DT nodes, which could impact classification performance because attributes with a larger number of values are more likely to be selected. In contrast, the second dataset, with just one nominal attribute, is less susceptible to this issue.

To address the bias selection problem and enhance the classification performance, we integrated the heart failure ontology (Wang, 2015) into the SRF construction process. This ontology encompasses 1,652 classes of heart failure information, covering disease signs, symptoms, and causes, as well as diagnostic tests. The process of designing and selecting an ontology involves several key steps. First, the scope of the ontology is defined. Then, existing ontologies are assessed for their relevance to the study area. Subsequently, the concepts and their interrelationships are examined using the Protégé (Knublauch et al., 2004) ontology editing tool. Finally, the ontology that is most relevant to the examined datasets is selected for use in the proposed SRF. To update and maintain an ontology effectively, several critical tasks must be completed, e.g., adding new concepts, refining existing definitions, merging duplicates, and removing outdated entities, to ensure the ontology's accuracy and relevance. Regular reviews and audits are also required to maintain consistency and alignment with the latest domain knowledge. We also handle change management and ensure efficient ontology evolution by implementing versioning and, in certain instances, automated updates using ML.

### 2.2 Data preparation

In this study, we employed the list-wise deletion method (Emmanuel et al., 2021) to address missing data in the datasets, where samples with missing values were removed. In addition, we utilized the interquartile range (IQR) measure (Smiti, 2020) to identify and remove outliers. For the numerical attributes, each attribute's values were divided into four equal parts, and the IQR was calculated using the first quartile (Q1) and the third quartile (Q3). Any value that falls below $Q1 - 1.5 \times IQR$ or exceeds $Q3 + 1.5 \times IQR$ was identified as an outlier and removed from the dataset. Statistical methodologies, including the Chi-squared test (McHugh, 2013) and point-biserial correlation (Verma, 2019) methods, were employed in the feature selection process to assess the relationship between the attributes and the target class. Here, attributes that elicited a $p$-value of less than 0.05 from these statistical tests were viewed as correlated with the target classes and were included in the construction of the proposed SRF. In addition, the synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002) was employed to address class imbalance in the cardiovascular dataset obtained from the Mendeley data repository. Finally, the cleaned datasets were divided into training and testing data at a ratio of 70:30 for the subsequent processes. The number of samples in the datasets after the data preparation process is presented in Table 2, and examples of each dataset are shown in Figure 2.

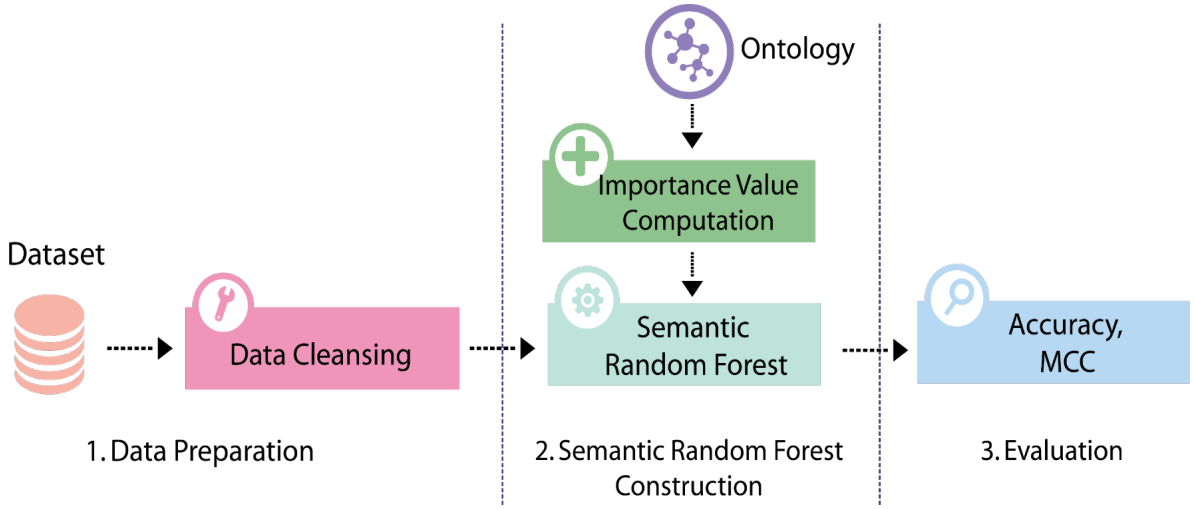



**Figure 1.** Architectural framework of the proposed SRF model

**Table 1.** Datasets used to evaluate the performance of the proposed SRF framework

| Attribute | Data type | Description |
|---|---|---|
| **Heart disease (UCI)** | | |
| age | Numerical | Age of patient |
| sex | Nominal | Sex of patient |
| cp | Nominal | Type of chest pain |
| Thresbps | Nominal | Resting blood pressure |
| chol | Numerical | Serum cholesterol |
| fbs | Nominal | Fasting blood sugar levels above 120 mg/dL. |
| restecg | Nominal | Resting electrocardiographic results |
| thalach | Numerical | Maximum heart rate attained |
| exang | Nominal | Exercise-induced angina |
| oldpeak | Numerical | ST depression caused by exercise compared with rest |
| slope | Nominal | Slope of the peak exercise ST segment |
| ca | Numerical | Number of major vessels (ranging from 0–3) identified by fluoroscopy |
| thal | Numerical | A blood disorder known as thalassemia |
| num | Numerical | diagnosis of heart disease |
| **Heart disease (Mendeley)** | | |
| age | Numerical | Age of patient |
| gender | Nominal | Sex of patient |
| impulse | Numerical | Heart rate |
| pressurehigh | Numerical | Systolic blood pressure |
| pressurelow | Numerical | Diastolic blood pressure |
| glucose | Numerical | Blood sugar |
| kcm | Numerical | Creatine kinase MB |
| troponin | Numerical | Troponin test |
| class | Nominal | diagnosis of heart disease |

**Table 2**. Number of samples in each dataset after data preparation

| Dataset | Number of samples | | Total number of samples |
|---|---|---|---|
| | Positive class (presence of disease) | Negative class (absence of disease) | |
| Heart disease (UCI) | 137 | 160 | 297 |
| Heart disease (Mendeley) | 319 | 319 | 638 |

**Heart disease (UCI)**

| age | sex | cp | Thresbps | restecg | thalach | exang | oldpeak | slope | ca | thal | num |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 51 | male | Non-Anginal_Pain | 125 | Left_V_Hypertrophy | 166 | No | 2.4 | Flat | 0 | Normal | Absence_of_Disease |
| 56 | male | Asymtomatic | 130 | Left_V_Hypertrophy | 103 | Yes | 1.6 | Downsloping | 0 | Reversable_Defect | Presence_of_Disease |
| 65 | male | Asymtomatic | 120 | Normal | 140 | No | 0.4 | Upsloping | 0 | Reversable_Defect | Absence_of_Disease |
| 41 | male | Atypical_Angina | 120 | Normal | 182 | No | 0 | Upsloping | 0 | Normal | Absence_of_Disease |
| 67 | male | Asymtomatic | 120 | Normal | 71 | No | 1 | Flat | 0 | Normal | Presence_of_Disease |
| 47 | male | Asymtomatic | 110 | Left_V_Hypertrophy | 118 | Yes | 1 | Flat | 1 | Normal | Presence_of_Disease |

**Heart disease (Mendeley)**

| age | gender | kcm | troponin | class |
|---|---|---|---|---|
| 68 | 1 | 3.76 | 0.012 | negative |
| 70 | 1 | 5.02 | 0.016 | positive |
| 72 | 0 | 1.79 | 0.008 | negative |
| 46 | 1 | 3.43 | 0.008 | negative |
| 56 | 0 | 6.38 | 0.006 | positive |
| 65 | 0 | 6.78 | 0.197 | positive |

**Figure 2.** Representative samples from the UCI and Mendeley datasets

## 2.3 SRF mantic random forest construction

This section outlines the process of leveraging knowledge to enhance classification performance. Based on DTs, the proposed SRF is an ensemble method that utilizes the ID3 algorithm, which is integrated with knowledge from the ontology. This knowledge was used to assess each attribute's significance, thereby helping the algorithm identify the key attributes for building each DT in the SRF.

In this study, the weighted semantic PageRank method, as used by Chanmee and Kesorn (2023), was employed to evaluate each attribute's significance based on the concepts and relationships in the ontology. The computation of the

attribute importance value is described in the Appendix (Table S1). The process of constructing the SRF is shown in Figure 3. Here, multiple subsamples were generated to enhance diversity and reduce the ensemble classifier's generalization error. Each subsample served as training data to construct an individual DT within the SRF. The random subspace technique (Ho, 1998) was also employed to increase diversity by randomly selecting subsets of features rather than using the entire set of features. In constructing the DT, we calculated the information gain (IG), which served as the splitting criterion in the ID3 algorithm. To counteract bias favoring attributes with many values, we adjusted each attribute's IG using the importance values obtained from the ontology to ensure that significant attributes with fewer values were more likely to be selected as nodes in the DT. The altered IG was calculated using Equations (1)–(3).

$$Entropy(D) = \sum_{i=1}^{j} p_i \, log_2 p_i \quad (1)$$

where, *Entropy (D)* represents the entropy of dataset *D*, which comprises *j* classes, and $p_i$ denotes the probability of the samples being classified into class *i*.

$$Entropy(A) = \sum_{v=1}^{m} \frac{|D_v|}{|D|} \, Entropy(D_v) \quad (2)$$

where, *Entropy (A)* represents the average entropy of attribute *A*, which possesses *m* unique values, $|D|$denotes the total number of samples in dataset *D*, and $|D_v|$ denotes the number of samples for attribute *A* with value *v*. In addition, *Entropy* ($D_v$) represents the entropy of attribute *A* for the specific value *v* (Han et al., 2011).

$$AIG(A) = \big(Entropy(D) - Entropy(A)\big) + Ip(A) \quad (3)$$

where, AIG(A) represents the altered IG for attribute A, and Ip(A) denotes the importance value of attribute A.

Note that each DT was constructed to its maximum depth without undergoing pruning. To derive the final classification result, the outcomes from each DT were aggregated, and a majority vote was employed to determine the overall classification.

Figure 4 shows the fundamental difference between how the proposed SRF and the conventional method approach the DT construction process when analyzing the same dataset. The proposed SRF incorporates domain knowledge through a heart failure ontology, which provides crucial context about the relative importance of different attributes in the diagnosis of cardiovascular disease. In the conventional methods, when constructing individual DTs, attributes like *Age* and *Chest Pain* are initially considered based on random selection. Then, the algorithm calculates the IG for these attributes to determine their splitting effectiveness. Although the *Chest Pain* attribute is a direct clinical indicator of cardiovascular disease, the traditional DT method selects the *Age* attribute as the root node simply because it shows higher IG. This selection pushes the more diagnostically relevant *Chest Pain* attribute further down the tree, which results in a more convoluted structure that may compromise the diagnostic accuracy of the model. The proposed SRF addresses this limitation by synthesizing two key factors, i.e., the attribute importance values derived from the heart failure ontology and the calculated IG. This integrated approach ensures that attributes are evaluated based on both their statistical properties and their clinical significance. Thus, the *Chest Pain* attribute is selected as the root node, thereby creating a DT structure that better reflects medical knowledge and potentially enhances the model's ability to classify heart failure accurately.

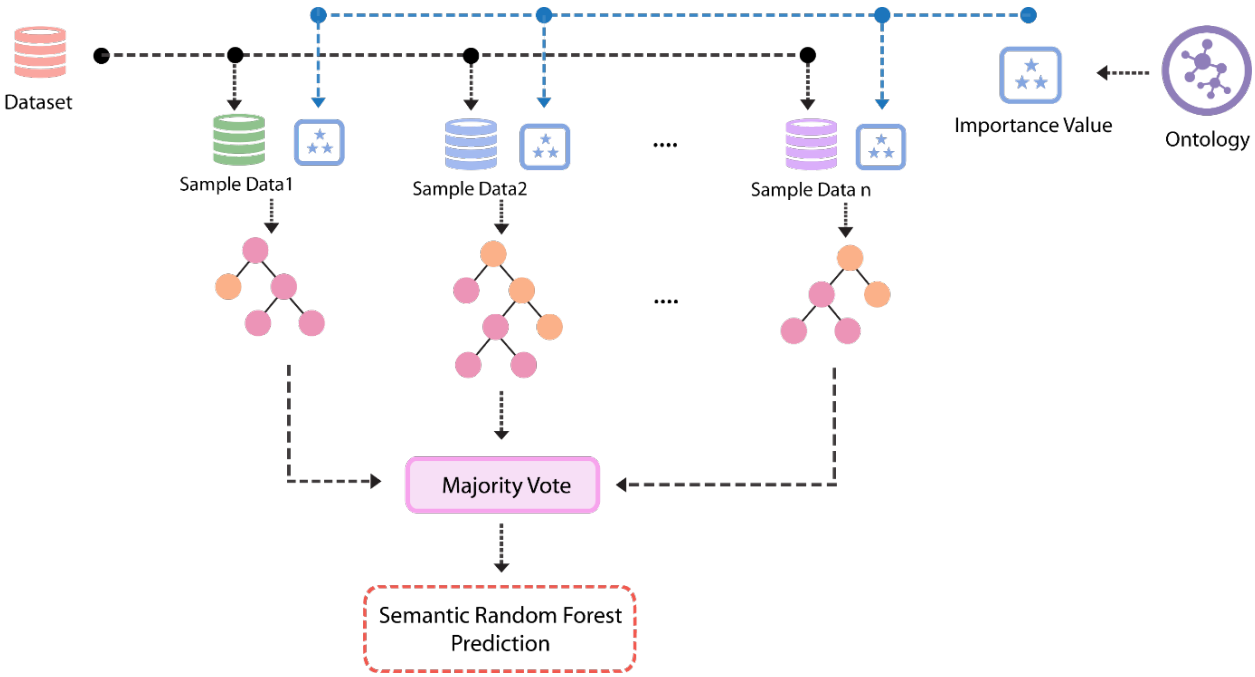
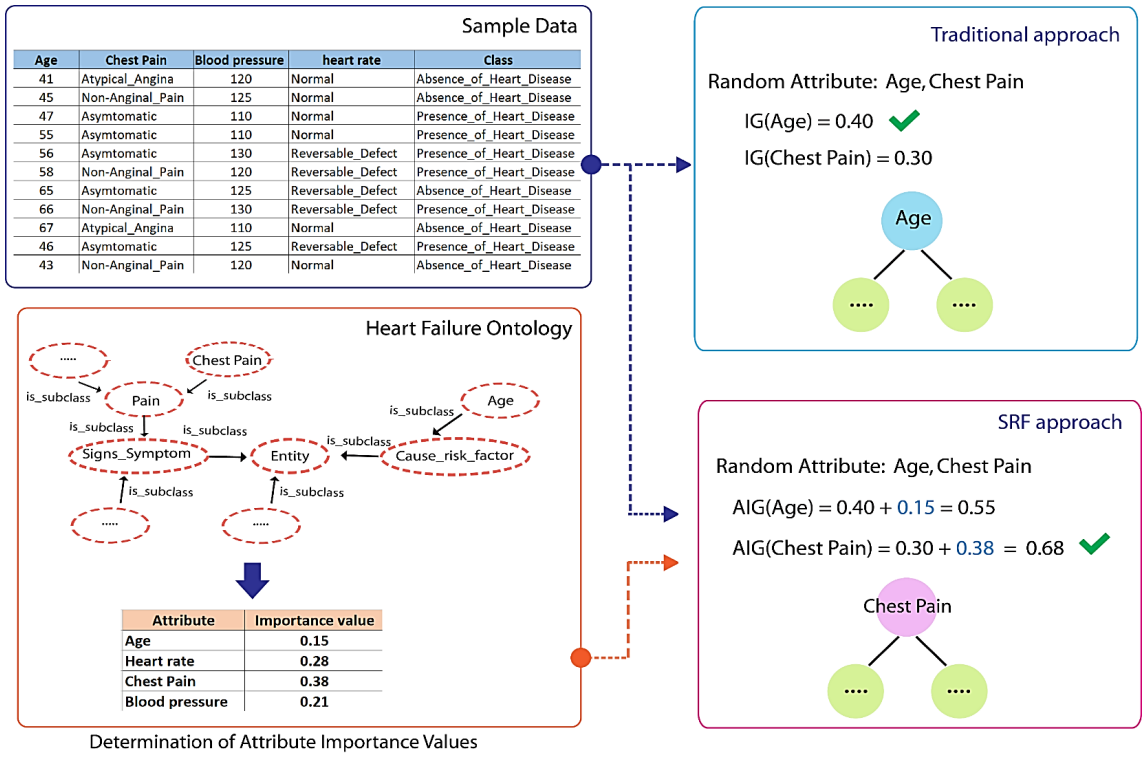



**Figure 3.** Systematic construction and implementation of SRF




| Age | Chest Pain | Blood pressure | heart rate | Class |
|---|---|---|---|---|
| 41 | Atypical_Angina | 120 | Normal | Absence_of_Heart_Disease |
| 45 | Non-Anginal_Pain | 125 | Normal | Absence_of_Heart_Disease |
| 47 | Asymtomatic | 110 | Normal | Presence_of_Heart_Disease |
| 55 | Asymtomatic | 110 | Normal | Presence_of_Heart_Disease |
| 56 | Asymtomatic | 130 | Reversable_Defect | Presence_of_Heart_Disease |
| 58 | Non-Anginal_Pain | 120 | Reversable_Defect | Presence_of_Heart_Disease |
| 65 | Asymtomatic | 125 | Reversable_Defect | Absence_of_Heart_Disease |
| 66 | Non-Anginal_Pain | 130 | Reversable_Defect | Presence_of_Heart_Disease |
| 67 | Atypical_Angina | 110 | Normal | Absence_of_Heart_Disease |
| 46 | Asymtomatic | 125 | Reversable_Defect | Presence_of_Heart_Disease |
| 43 | Non-Anginal_Pain | 120 | Normal | Absence_of_Heart_Disease |

**Figure 4.** Diagram illustrating the influence of attribute importance values on the node selection mechanism

## 2.4 Evaluation

In this study, standard performance metrics were used to evaluate and validate the proposed SRF framework. Here, accuracy was used to assess the proposed SRF's overall performance, and the Matthews correlation coefficient (MCC) (Chicco & Jurman, 2020) was used as an alternative metric, offering robustness in the presence of unbalanced datasets and maintaining its value when positive and negative classes were interchanged. The MCC value ranges from −1 to 1, where −1 indicates extreme misclassification, 1 indicates perfect classification, and 0 indicates random guessing by the classifier.

The accuracy and MCC are defined in Equations (4) and (5), respectively.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \quad (5)$$

where, *TP* represents the number of instances correctly classified as cardiovascular disease, *TN* represents the number of instances correctly classified as the absence of cardiovascular disease, *FP* represents the number of instances of absent cardiovascular disease incorrectly classified as cardiovascular, and *FN* represents the number of cardiovascular disease instances incorrectly classified as absent.

# 3. RESULTS AND DISCUSSION

After preprocessing, the cleaned UCI dataset comprised 11 attributes and 297 records, and the cleaned Mendeley dataset comprised four attributes and 853 records. In addition, various tree-based ensemble methods were employed to perform a comparative analysis of the proposed SRF.

## 3.1 Feature selection

In this experiment, statistical techniques, i.e., the Chi-squared ($X^2$) test and point-biserial correlation ($R_{pb}$), were utilized to identify the attributes associated with the defined classes in the datasets. Table 3 shows the results of the statistical tests. In this evaluation, an association was indicated when the *p*-value of a statistical test between an attribute and the target class was less than 0.05.

**Table 3.** Results of Chi-squared test and point-biserial correlation analysis

| Attribute | $R_{pb}$ | *p*-value | Attribute | $X^2$ | *p*-value |
|---|---|---|---|---|---|
| **Heart disease (UCI)** | | | | | |
| age | 0.227** | < 0.001 | sex | 23.030** | < 0.001 |
| Thresbps | 0.153** | 0.008 | cp | 77.276** | < 0.001 |
| ***chol*** | ***0.080*** | ***0.168*** | ***fbs*** | ***0.003*** | ***0.956*** |
| thalach | -0.424** | < 0.001 | restecg | 9.576** | 0.008 |
| oldpeak | 0.424** | < 0.001 | exang | 52.730** | < 0.001 |
| ca | 0.463** | < 0.001 | slope | 43.473** | < 0.001 |
| | | | thal | 82.460** | < 0.001 |

**Table 3.** Results of Chi-squared test and point-biserial correlation analysis (continued)

| Attribute | $R_{pb}$ | *p*-value | Attribute | $X^2$ | *p*-value |
|---|---|---|---|---|---|
| **Heart disease (Mendeley)** | | | | | |
| age | 0.238** | < 0.001 | gender | 11.36** | < 0.001 |
| ***impulse*** | ***0.007*** | ***0.801*** | | | |
| ***pressurehigh*** | ***-0.021*** | ***0.449*** | | | |
| ***pressurelow*** | ***-0.010*** | ***0.726*** | | | |
| ***glucose*** | ***-0.033*** | ***0.230*** | | | |
| kcm | 0.218** | < 0.001 | | | |
| troponin | 0.229** | < 0.001 | | | |

In the UCI dataset, the "chol" and "fbs" attributes were unrelated to the target class. However, in the Mendeley dataset, no association was found between the "impulse," "pressurehigh," "pressurelow," and "glucose" attributes and the target class. Thus, these uncorrelated attributes were removed from the datasets. Consequently, eleven attributes from the UCI dataset and four attributes from the Mendeley dataset were used to construct the SRF.

## 3.2 Sampling method examination

In ensemble learning, generating multiple data subsamples to construct uncorrelated individual classifiers is a strategy utilized to enhance diversity, which directly influences the performance of the ensemble classifier (Kuncheva & Whitaker, 2003). Thus, this experiment attempted to identify the most effective sampling method that incorporates the ontological knowledge to generate uncorrelated DTs. In this study, various techniques to generate subsample data to optimize the performance of the proposed SRF were investigated. The first method was bagging (Han et al., 2011), which is a widely used technique in the RF algorithm. The bagging approach randomly selected samples with replacement, comprising 63.2% of the original training data, and the remaining samples were duplicates. The second method involved varying the sampling rate between 60% and 80%, as discussed by Adnan and Islam (2015). Here, the remaining portion of the subsample data was generated by randomly duplicating data from the initially selected 30%. The third method adopted the sampling approach employed in the balanced decision forest (BDF) technique (Adnan et al., 2021), which involves adjusting both the number of samples and the number of random features. Specifically, an inverse relationship exists between the number of selected samples and the number of random features. As the number of samples increases, the number of features decreases (and vice versa). For example, when the proportion of subsample data comprises 37% of the total dataset, the number of features in the random subset used to construct the DT comprises approximately 99% of the total available features. In this experiment, 100 DTs were constructed for the proposed SRF. The classification results obtained by the proposed SRF framework using different sampling methods are shown in Figure 5.

The classification results indicate that using a variable sampling rate between 60% and 80% to generate subsample data yielded the highest performance compared with other approaches tested on both datasets. Figure 5(a) shows the accuracy of the SRF's classification results obtained using different sampling methods. For the UCI dataset, applying the variable sampling rate method to generate subsample data results in an accuracy of 0.8296. However, utilizing the bagging method resulted in an accuracy of 0.8239, which was reduced to 0.8085 with the BDF-based method for subsample data generation. For the Mendeley dataset, the proposed SRF framework with the variable sampling rate method achieved the highest classification accuracy of 0.9856. In comparison, the SRF framework with the bagging method achieved an accuracy of 0.9850, and the BDF-based method resulted in an accuracy of 0.9842. These findings suggest that the BDF-based method is less effective than the varying sampling rate method. In the BDF-based method, if a low sampling rate is selected at random, there is a risk of using nearly all attributes to construct the DT, which can result in significant attributes, identified as important by integrating knowledge into the algorithm, being frequently selected as nodes. This may lead to the creation of redundant DTs.

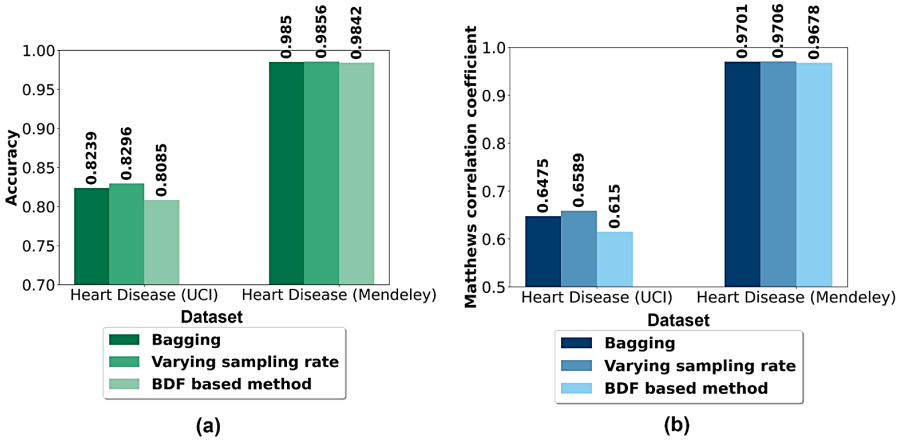



**Figure 5.** Comparison of classification results obtained by the proposed SRF using different sampling techniques

The MCC was also employed to evaluate the effectiveness of each subsampling method (Figure 5-b) aligning with the accuracy measurements. Utilizing varying sampling rates for subsample data demonstrates superior performance compared with the alternative methods. The MCC value for the varying sampling rate method was 0.6589 for the UCI dataset and 0.9706 for the Mendeley dataset. In addition, the precision, recall, and F1-score metrics for each target class are reported in Table 4. These metrics are derived from the confusion matrix, which consists of the following terms:

*Precision* measures the proportion of correctly predicted positive instances among all instances predicted as positive. It is defined as (6):

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

A high precision indicates that when the model predicts a positive class, it is likely to be correct.

Recall (also called Sensitivity or True Positive Rate) measures the proportion of correctly predicted positive instances among all actual positive instances. It is given by (7):

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

A high recall indicates that the model is good at identifying all positive instances.

*F1-score* is the harmonic mean of precision and recall. It balances both metrics and is useful when we need to consider both false positives and false negatives equally. It is calculated as (8):

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

This metric ensures that both precision and recall are maximized, preventing an imbalance between the two. The F1-score provides a single score that considers both precision and recall, making it useful for evaluating models where both false positives and false negatives are significant concerns. There is often a trade-off between precision and recall. Increasing precision may lead to a decrease in recall and vice versa. For example, if a model is very conservative in labeling positive instances, precision may increase, but recall may decrease because fewer true positives are detected.

The results demonstrate that the varying sampling rate method achieved strong performance across all metrics on both datasets. The experimental results indicate that varying the sampling rate improved the classifier performance compared with the traditional techniques, e.g., the bagging method. However, when the BDF-based method was employed to generate multiple subsample datasets, the proposed SRF framework produced unsatisfactory results. These findings emphasize the importance of selecting a sampling method that effectively integrates the available knowledge to realize satisfactory classification performance.

As shown in Figure 6, the dataset includes attributes *A1* through *A5*, with attribute *A3* having the highest importance. When the varying sampling rate method is employed, a maximum of three attributes are used to construct each DT. As a result, attribute *A3* may not be selected as the root node, leading to the creation of DTs with different structures. However, when the BDF-based method is used, most attributes are employed to construct the DT, depending on the sample size. This increases the probability of selecting attribute *A3* as the root node, thereby resulting in the generation of identical DTs. If all DTs in the proposed SRF yield the same classification outcomes, the improvement in classification performance may be minimal. However, when the varying sampling rate method is utilized, the attributes are selected randomly in consistent quantities for the node selection process, which enhances the likelihood of selecting diverse root nodes for the DT, and this enables the generation of distinct DTs in the proposed SRF and consequently improving the classification performance.

### 3.3 Comparison with tree-based ensemble algorithms

We also conducted an experiment to compare the classification performance of the SRF framework with existing ensemble algorithms, e.g., the RF, AdaBoost, and gradient boosting algorithms. Here, each algorithm utilized 100 DTs, and the final outcomes were obtained by averaging the results obtained over 30 experiments. As shown in Table 5, the proposed SRF framework outperformed the other tree-based ensemble algorithms in terms of accuracy and MCC. On the UCI dataset, the SRF framework obtained the highest accuracy of 0.8296. In contrast, the RF, AdaBoost, and gradient boosting algorithms obtained lower classification accuracies of 0.8141, 0.8170, and 0.7978, respectively. Note that similar trends were observed when analyzing the Mendeley dataset. On this dataset, the proposed SRF framework achieved a classification accuracy of 0.9856, which was the highest among the evaluated algorithms. In comparison, the RF, AdaBoost, and gradient boosting algorithms obtained accuracy values of 0.9807, 0.9817, and 0.9794, respectively. In addition, the MCC was used as a validation measure for the proposed SRF framework. On both datasets, the proposed SRF achieved the highest MCC scores, obtaining a value of 0.6589 on the UCI dataset and 0.9706 on the Mendeley dataset.

**Table 4.** Performance metrics (precision, recall, and F1-score) of the SRF with various sampling techniques

| Dataset | Sampling method | Positive class (presence of disease) | | | Negative class (absence of disease) | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1-score | Precision | Recall | F1-score |
| Heart disease (UCI) | Bagging | 0.8241 | 0.7773 | 0.8000 | 0.8232 | 0.8631 | 0.8427 |
| | Varying sampling rate | 0.8343 | 0.7837 | 0.8082 | 0.8284 | 0.8714 | 0.8494 |
| | BDF-based method | 0.8063 | 0.7664 | 0.7859 | 0.8129 | 0.8460 | 0.8291 |
| Heart disease (Mendeley) | Bagging | 0.9870 | 0.9790 | 0.9830 | 0.9840 | 0.9902 | 0.9871 |
| | Varying sampling rate | 0.9873 | 0.9793 | 0.9833 | 0.9843 | 0.9904 | 0.9874 |
| | BDF-based method | 0.9881 | 0.9751 | 0.9816 | 0.9812 | 0.9912 | 0.9861 |

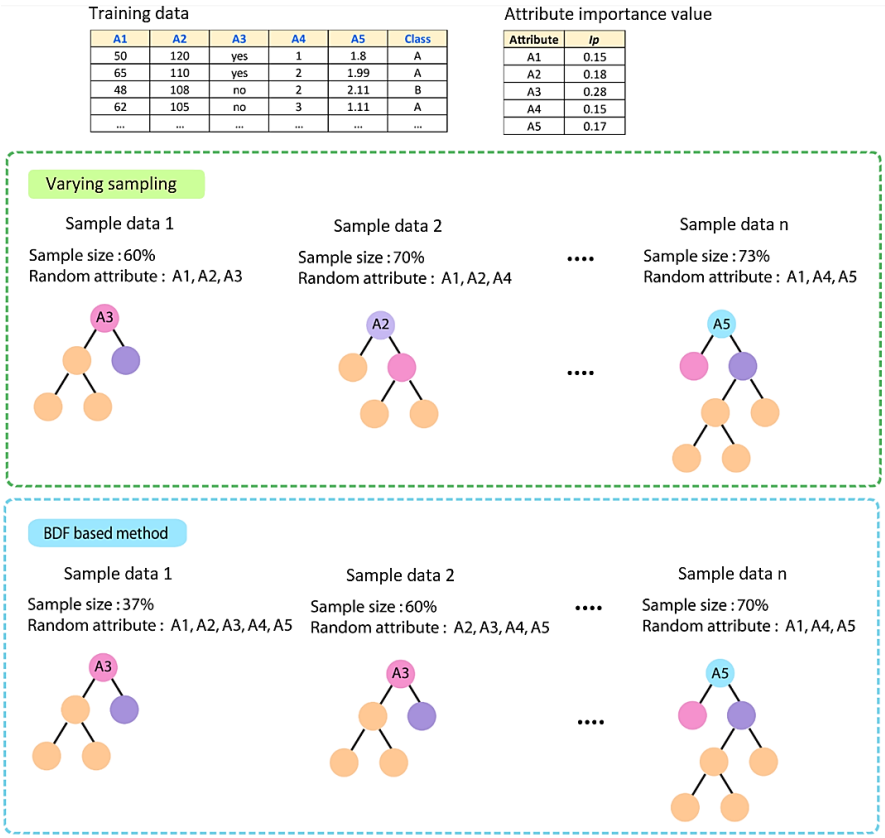



**Figure 6.** Example of the SRF construction process using different sampling methods

**Table 5.** Performance comparison of proposed SRF framework and other ensemble learning algorithms

| Dataset | Algorithms | Accuracy | MCC |
|---|---|---|---|
| **Heart disease (UCI)** | SRF | **0.8296** | **0.6589** |
| | RF | 0.8141 | 0.6319 |
| | AdaBoost | 0.8170 | 0.6367 |
| | Gradient boosting | 0.7978 | 0.5978 |
| **Heart disease (Mendeley)** | SRF | **0.9856** | **0.9706** |
| | RF | 0.9807 | 0.9606 |
| | AdaBoost | 0.9817 | 0.9628 |
| | Gradient boosting | 0.9794 | 0.9580 |

The knowledge derived from the ontology, particularly in terms of the attribute importance values, helps the algorithm identify significant attributes to serve as nodes in each DT of the SRF. Integrating relevant attributes into the model construction process further improves the classification performance (Spencer et al., 2020); however, slight differences were observed in the classification performance in terms of both the accuracy and MCC results and between the SRF and other algorithms on the Mendeley dataset. Utilizing the attribute importance values effectively mitigates the bias in attribute selection, particularly when handling datasets that contain categorical attributes with numerous values. When considering the Mendeley dataset, in which most attributes are numerical, the likelihood of such bias is reduced. Consequently, applying only the proposed SRF yields a modest enhancement in the classification performance. The findings of this experiment demonstrate that the proposed SRF framework can be used to develop a classification model that accurately identifies patients with cardiovascular diseases. As a result, this improved classification capability may help reduce the mortality risk among these patients.

## 3.4 Comparison with nontree-based machine learning models

In our previous study (Chanmee & Kesorn, 2024), the performance of the proposed approach was not compared with that of nontree-based classification algorithms. However, currently, various algorithms, e.g., neural networks and support vector machines (SVMs), are recognized as high-performance classification methods. Thus, the proposed SRF framework was also compared with such nontree-based algorithms. In this experiment, the performance of the proposed SRF framework was compared with that of several nontree-based classification algorithms, including the multilayer perceptron (MLP), a type of artificial neural network (ANN), SVM, and k-nearest neighbors (*k*-NN) methods. Here, the GridSearchCV (Avinash et al., 2022) technique was employed to identify the optimal parameter values that obtained the best performance for each algorithm. Examples of the parameters used to optimize each algorithm are shown in Table 6. In this evaluation, 30 experiments were conducted, and the results, which were averaged to derive the final outcomes, are shown in Table 7.

**Table 6.** Parameter configurations used in different classification algorithms

| Algorithm | Parameter | Parameter range |
|---|---|---|
| MLP | Hidden layer size | {50, 100, 150} |
| | Maximum number of iterations | {50, 100, 150, 200} |
| | Activation function | {"tanh," "relu," "logistic"} |
| SVM | C | {0.1, 1, 10,100} |
| | gamma | {0.001, 0.01, 0.1, 1} |
| | kernel | {"rbf," "linear," "poly," "sigmoid"} |
| k-*NN* | k | {3, 5, 7, 9, 11, 13, 15} |

**Table 7.** Performance evaluation of the SRF against various classification algorithms

| Dataset | Algorithm | Accuracy | MCC |
|---|---|---|---|
| Heart disease (UCI) | SRF | **0.8296** | **0.6589** |
| | MLP | 0.7710 | 0.5589 |
| | SVM | 0.8256 | 0.6569 |
| | k-NN | 0.7202 | 0.4576 |
| Heart disease (Mendeley) | SRF | **0.9856** | **0.9706** |
| | MLP | 0.8324 | 0.6842 |
| | SVM | 0.8053 | 0.6407 |
| | k-NN | 0.7852 | 0.6048 |

The results demonstrate that the proposed SRF framework outperformed the compared classification algorithms in terms of both accuracy and the MCC score. The proposed SRF framework is based on an ensemble learning method that combines multiple models to achieve better predictive performance than a single model (Sagi & Rokach, 2018), thereby leading to superior classification performance. However, the accuracy and MCC score obtained by the SVM on the UCI dataset, i.e., 0.8256 and 0.6569, respectively, were close to the results obtained by the SRF framework, as this balanced dataset with a binary class tends to perform well with SVMs (Zhang et al., 2017). In addition, the ANN algorithm, which has garnered increasing interest from researchers, obtained low classification results on both datasets. The ANN algorithm requires a large amount of training data to achieve high predictive performance (Alwosheel et al., 2018), and the limited size of the examined datasets may have been insufficient to produce accurate classification models. The results suggest that employing an ensemble learning approach alongside ontology knowledge can yield satisfactory outcomes. Notably, the proposed SRF framework demonstrated strong performance even when applied to small datasets.

## 3.5 Computation complexity of proposed SRF framework

Computational complexity is a foundational concept in computer science that examines the intrinsic difficulty of solving computational problems. To assess the computational performance of the proposed SRF framework relative to different baseline models, we evaluated and compared the computational complexity costs of the DT, RF, and SRF models. Typically, constructing a single DT within an RF has a time complexity of O ($n$ * $m$ * log [$m$]), where $n$ denotes the number of samples in the training data, $m$ denotes the number of features, and log ($m$) denotes the average depth of the DT. In a traditional RF, numerous DTs are constructed, and if the RF comprises $k$ trees, the overall time complexity of constructing the RF model is O ($k$ *$n$ *$m$ *log [$m$]).

To make a prediction using a traditional RF, the data point is passed through each DT in the forest, and the final prediction is made by aggregating the predictions of individual trees. Typically, the time complexity of making a prediction with an RF is O ($k$ * log [$m$]), where $k$ denotes the number of trees in the forest, and log ($m$) denotes the average depth of the DT. In this evaluation, two algorithms were employed to construct the SRF model. The first algorithm, detailed in Algorithm 1, calculated the attribute importance values. The second algorithm, outlined in Algorithm 2, focused on the SRF construction process. To evaluate the time complexity of Algorithm 1, we analyzed its performance in the worst-case scenario for size $n$ inputs. In Algorithm 1, lines 1–4 represent simple operations that are conducted once. However, line 3 involves the *Relationship Weighted* function, which is executed $n^2$ times based on the amount of knowledge and their relationship in the ontology. In addition, the FOR loop in lines 6–8 is executed $n$ times. Thus, the total number of executions for lines 1–8 can be determined from Equation (9).

$$T_1 = n^2 + n + 3 \quad (9)$$

The REPEAT loop continues until the stopping condition is satisfied; thus, we assumed that it would be executed $p$ times. Lines 11–14 include the FOR loop, which ran $n$ times, and line 16 represents a single statement that is executed only once. Thus, the execution time for lines 10–16 is obtained from Equation (10).

$$T_2 = (p \times n) + 1 \quad (10)$$

Therefore, the total time complexity of Algorithm 1 is calculated from Equation (11).

$$T_{algorithm1} = T_1 + T_2 = n^2 + (p \times n) + n + 4 \quad (11)$$

Consequently, Algorithm 1's complexity is $O(n^2)$, where $n^2$ denotes the function's highest order of growth. Note that the SRF procedure is similar to that of the RF algorithm; therefore, the time complexity of Algorithm 2 is $O(k * n * m * (\log m))$, which is consistent with the time complexity of the RF algorithm. Finally, the time complexity of the proposed SRF is $O(n^2) + O(k * m * n * (\log m))$. In addition, the prediction procedure for an SRF is similar to that of the traditional RF; thus, the time complexity for the prediction of the SRF is O ($k$ * log [$m$]). These findings suggest that the proposed SRF reinforcement process has the highest complexity cost in this implementation. Therefore, when the ontology contains numerous concepts and relationships, the algorithm may need considerable time to calculate attribute importance values and conduct data classification. Thus, in the future, we plan to conduct algorithmic optimizations to achieve a more acceptable computational complexity cost for the overall process.

## 3.6 Limitations

The purpose of this study was to enhance the classification performance of the traditional RF algorithm when applied

to a cardiovascular dataset, and our main innovation is the incorporation of domain knowledge into the ensemble learning method. However, certain limitations are notable when applying a knowledge-based approach in a clinical setting. The first limitation relates to knowledge maintenance. Medical knowledge evolves over time, which requires the knowledge-based approach to remain updated with new advancements to achieve optimal performance. Thus, the ontology used in the SRF must be adjusted to include the latest knowledge, which is a manual process conducted by experts. However, maintaining the ontology, e.g., adding new information or correcting inaccuracies, is a time-consuming and tedious process. To streamline this task, integrating ontology learning techniques (Khadir et al., 2021) that automatically generate ontologies into the knowledge-based approach can yield satisfactory results. Although the proposed SRF framework has its limitations, it can still offer significant assistance in reducing the mortality rates of cardiovascular patients. To apply the proposed SRF to other illnesses, collaboration with established and trustworthy disease ontologies, e.g., COVID-19 (Sargsyan et al., 2020) and dengue fever (Mitraka et al., 2015), is essential. By incorporating these ontologies into the proposed SRF framework, it will be possible to develop classification models for these specific conditions. In addition, adjustments to the SRF's source code are required to enable it to learn the characteristics of these diseases.

**Algorithm 1: Procedure to determine attribute importance values**

**Input:** Ontology ($O$), Relationships in ontology ($R$)
**Output**: Attribute importance value ($Ip$)
generate an empty set to store the importance values for attributes $Ip$ = {}
$d$ = 0.85
$r\_weight$ = *RelationshipWeighted* (O, R)
$N$ = the quantity of concepts present in the ontology $O$
// set an initial importance value for all concepts
**FOR** each $O_i$ where $O_i \in O$
  $\boldsymbol{IP(O_i) = 1/N}$
**ENDFOR**
//determine importance value
**REPEAT**
**FOR** each concept $O_i$ has connections from the concept $O_j$ and $r \in R$
  $Ip(O_i) = d \times \sum_{j \in outlink(i)} \frac{Ip(O_i) \times r\_weight(r,o_j)}{\sum_{r \in R} r\_weight(r,o_j)} + (1-d)$
  **UPDATE** $Ip$ with $Ip(O_i)$
  **ENDFOR**
**UNTIL** $Ip(O_i)$of all $O_i$ no longer changes
**RETURN** $Ip$

**Algorithm 2: Semantic random forest algorithm**

**Input:** Dataset (D), Target Attribute ($a_{target}$), Number of trees ($N$), Attribute importance values ($Ip$)
**Output:** Semantic Random Forest ($SRF$)
//build the Semantic Random Forest
**SemanticRandomForest** ($D, a_{target}, Ip$)
  **FOR** $i$ = 1 to $N$
    randomly create subsample $D_i$ where $D_i \in D$
    $SRF\_Tree$ = *Build_SDT* ($D, a_{target}, Ip$)
  **ENDFOR**
  // establish the final decision by using a majority vote
  $SRF$ = mode ($SRF\_Tree$)
  **RETURN** $SRF$
// build each decision tree of the Semantic Random Forest
**Build_SDT** ($D, a_{target}, Ip$)
  Creating an empty set designated for the decision tree $SDT\_Tree$ = {}
  **IF** the instances within $D_i$ are uniform in class or alternative stopping criteria are activated
    generate a leaf node that represents the predominant class in $D_i$
  **ENDIF**
  randomly select $m$ attributes from the set of attributes
  **FOR** each attribute $a_j$ where $a_j \in m$ and $a_j \neq a_{target}$
    $AIG(a_j) = (Entropy(D_i) - Entropy(a_i)) + Ip(a_j)$
  **ENDFOR**
  $a_{best}$= attribute that achieves the maximum $AIG(a_j)$
  $SDT\_Tree$ = create a decision tree node based on the attribute $a_{best}$
  $D_{jk}$ = create a sub-dataset from $D_j$ using the attribute $a_{best}$
  **FOR** each attribute $a_j$ where $a_j \in D_{jk}$ and $a_j \neq a_{target}$
  // execute the recursive algorithm
    $SDT\_Tree_j$ = *Build_SDT* ($D_{jk}, a_{target}, Ip$)
    connect $SDT\_Tree_j$ to the appropriate branch of the $SDT\_Tree$
  **ENDFOR**
  **RETURN** $SDT\_Tree$

### 3.7 Extending the SRF model to other medical domains

The current study focused on cardiovascular risk prediction; however, the proposed SRF framework has the potential to be applied to other medical domains. By modifying key components of the proposed framework, e.g., feature selection, ontology design, and model training strategies, the SRF model can be adapted to support a variety of healthcare applications. To ensure successful extension, a structured approach that considers key aspects, e.g., data quality, ontology integration, parameter tuning, and model evaluation, should be adopted. Ensuring dataset quality is fundamental when adapting the SRF model to new medical domains, which involves applying data preparation techniques, e.g., handling missing values and implementing feature selection to identify the most relevant attributes for model training. High-quality data enhance the model's predictive power and reliability across different medical contexts. Another approach to extending the model is domain-specific feature engineering. By identifying and incorporating relevant features from diverse medical datasets, e.g., imaging biomarkers for radiology or genomic data for precision medicine, the model can be tailored to new clinical applications. In addition, transfer learning techniques (Tan et al., 2023) can facilitate adaptation by leveraging knowledge from cardiovascular risk prediction to train models for other diseases with minimal retraining.

Furthermore, a critical component of adaptation is ontology customization and integration. Incorporating well-validated disease ontologies, e.g., those for cancer (Polpinij, 2011), diabetes (Spoladore et al., 2024), or neurological disorders (Jensen et al., 2013), into the knowledge-based component of the SRF model would enable it to capture specialized relationships and improve prediction accuracy for different medical conditions. Ensuring the use of robust ontologies enhances the model's generalizability across medical domains. In addition, parameter tuning is essential in terms of achieving satisfactory performance when extending the model. Selecting an appropriate number of decision trees, optimizing the hyperparameters, and adjusting the model configurations based on the characteristics of the new dataset can enhance the predictive accuracy and efficiency. Finally, to validate the model's adaptability, a cross-domain evaluation should be performed. Applying the proposed SRF model to diverse medical datasets and assessing its performance using key metrics, e.g., accuracy, precision, recall, and the F1-score, will ensure its effectiveness across different healthcare applications, and we can refine the methodology and improve its robustness through rigorous evaluations.

By following this structured approach, i.e., ensuring data quality, utilizing reliable ontologies, tuning parameters, and evaluating results, the SRF model can be extended beyond cardiovascular risk prediction, thereby making it a versatile tool for a wider range of medical applications.

## 4. CONCLUSION

This paper has proposed the SRF framework to improve the classification of cardiovascular disease. The proposed SRF framework combines the traditional RF algorithm with the weighted semantic PageRank method to determine attribute importance. By considering both the statistical significance and contextual relevance of features, the proposed SRF framework selects features that enhance the overall classification accuracy. In addition, incorporating ontological knowledge into the SRF framework provides a structured representation of medical concepts, relationships, and axioms, and this structured knowledge guides each DT in the SRF, which enables more informed and precise predictions about heart disease. Compared with the baseline RF, AdaBoost, and gradient boosting algorithms, the proposed SRF framework outperforms in terms of accuracy and MCC. Furthermore, the use of diverse subsampling techniques enhances the ensemble's ability to generalize across datasets, thereby reducing overfitting and improving the robustness of the predictions.

However, the success of the proposed SRF framework is strongly dependent on the quality of the ontology. Thus, an inadequate or outdated ontology can yield incorrect attribute importance values, which would result in inaccurate classifications. Therefore, ongoing human oversight is required to maintain and update the ontology to ensure sufficient relevance and accuracy. The ontology must evolve alongside medical knowledge advancements to maintain the effectiveness of the proposed SRF framework. We acknowledge that ontology information extraction and processing can be time-consuming tasks, particularly when handling large datasets. To address this limitation, we suggest two potential solutions. One approach involves applying feature selection and dimensionality reduction techniques to minimize the number of features used during the model training process. By reducing the dimensionality of the dataset, we can decrease the computational complexity and improve the processing time without significantly compromising the model's accuracy. Another promising strategy is ontology pruning using DeepOnto (He et al., 2024), which is a Python package designed for ontology engineering. Pruning reduces the size and complexity of the ontology, thereby decreasing the time required for ontology extraction and processing in the SRF model. This optimization can improve computational efficiency significantly while maintaining the semantic integrity of the ontology-based knowledge representation. Thus, in future work, we plan to investigate and implement these strategies to further optimize the method. In addition, future research will also focus on refining the SRF by incorporating additional techniques, e.g., instance weighting, that effectively handle unbalanced data by assigning different weights to instances based on importance, frequency, or by adjusting the influence of knowledge to achieve optimal performance for each dataset. Furthermore, examining the integration of probabilistic ontologies or knowledge graphs is expected to enhance the proposed SRF framework's ability to capture and utilize complex medical knowledge.

To this end, the proposed SRF framework represents a significant advancement in the application of knowledge-based systems to medical diagnosis. Its integration of structured ontological knowledge and advanced ML techniques positions it as a powerful tool to predict heart disease and other complex medical conditions. As the SRF framework continues to evolve, it holds promise for advancing computer-aided diagnosis and improving patient outcomes by realizing more accurate disease prediction models.

## ACKNOWLEDGMENTS

This research was supported by the Thailand National Science, Research, and Innovation (Fundamental Fund-NU: Grant No. R2567B017). The authors would like to thank Enago (www.enago.com) for the English language review. The funder was not involved in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## REFERENCES

Adnan, M. N., Ip, R. H. L., Bewong, M., & Islam, M. Z. (2021). BDF: A new decision forest algorithm. *Information Sciences*, *569*, 687–705. https://doi.org/10.1016/j.ins.2021.05.017

Adnan, M. N., & Islam, M. Z. (2015). Improving the random forest algorithm by randomly varying the size of the bootstrap samples for low dimensional data sets. In *Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (pp. 391–396). ESANN. https://www.esann.org/sites/default/files/proceedings/legacy/es2015-21.pdf

AlGhanem, H., Shanaa, M., Salloum, S., & Shaalan, K. (2020). The role of KM in enhancing AI algorithms and systems. *Advances in Science, Technology and Engineering Systems Journal*, *5*(4), 388–396. https://doi.org/10.25046/aj050445

Alwosheel, A., van Cranenburgh, S., & Chorus, C. G. (2018). Is your dataset large enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modeling*, *28*, 167–182. https://doi.org/10.1016/j.jocm.2018.07.002

Andras Janosi, W. S. (1988). *Heart disease* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X

Avinash, M., Nithya, M., & Aravind, S. (2022). Automated machine learning-algorithm selection with fine-tuned parameters. In *Proceedings of the Sixth International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1175–1180). IEEE. https://doi.org/10.1109/ICICCS53718.2022.9788236

Chanmee, S., & Kesorn, K. (2021). Semantic data mining in the information age: A systematic review. *International Journal of Intelligent Systems*, *36*(8), 3880–3916. https://doi.org/10.1002/int.22443

Chanmee, S., & Kesorn, K. (2023). Semantic decision trees: A new learning system for the ID3-Based algorithm using a knowledge base. *Advanced Engineering Informatics*, *58*, Article 102156. https://doi.org/10.1016/j.aei.2023.102156

Chanmee, S., & Kesorn, K. (2024). COVID-19 cases classification using a semantic decision forest method. *ICIC Express Letters Part B: Applications*, *15*(11), 1175–1182. https://doi.org/10.24507/icicelb.15.11.1175

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), Article 6. https://doi.org/10.1186/s12864-019-6413-7

Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making*, *19*(1), Article 211. https://doi.org/10.1186/s12911-019-0918-5

Ed-daoudy, A., Maalmi, K., & El Ouaazizi, A. (2023). A scalable and real-time system for disease prediction using big data processing. *Multimedia Tools and Applications*, *82*(20), 30405–30434. https://doi.org/10.1007/s11042-023-14562-3

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1), Article 140. https://doi.org/10.1186/s40537-021-00516-9

Gaïffas, S., Merad, I., & Yu, Y. (2023). WildWood: A new random forest algorithm. *IEEE Transactions on Information Theory*, *69*(10), 6586–6604. https://doi.org/10.1109/TIT.2023.3287432

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques* (3rd ed.). Elsevier Science.

Harrell, F. E., Jr. (2001). Cox proportional hazard regression model. In F. E. Harrell, Jr. (Ed.), *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis* (pp. 465–507). Springer. https://doi.org/10.1007/978-1-4757-3462-1_19

He, Y., Chen, J., Dong, H., Horrocks, I., Allocca, C., Kim, T., & Sapkota, B. (2024). DeepOnto: A Python package for ontology engineering with deep learning. *Semantic Web*, *15*(5), 1991–2004. https://doi.org/10.3233/SW-243568

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844. https://doi.org/10.1109/34.709601

Hossain, M. I., Maruf, M. H., Khan, M. A. R., Prity, F. S., Fatema, S., Ejaz, M. S., & Khan, M. A. S. (2023). Heart disease prediction using distinct artificial intelligence techniques: Performance analysis and comparison. *Iran Journal of Computer Science*, *6*(4), 397–417. https://doi.org/10.1007/s42044-023-00148-7

Ishak, A., Ginting, A., Siregar, K., & Junika, C. (2020). Classification of heart disease using decision tree algorithm. *IOP Conference Series: Materials Science and Engineering*, *1003*, Article 012119. https://doi.org/10.1088/1757-899X/1003/1/012119

Jensen, M., Cox, A. P., Chaudhry, N., Ng, M., Sule, D., Duncan, W., Ray, P., Weinstock-Guttman, B., Smith, B., Ruttenberg, A., Szigeti, K., & Diehl, A. D. (2013). The neurological disease ontology. *Journal of Biomedical Semantics*, *4*(1), Article 42. https://doi.org/10.1186/2041-1480-4-42

Juraphanthong, W., & Kesorn, K. (2024). The intelligent approach of auto-regressive integrated moving average with exogenous semantic (ARIMAXS) variables for COVID-19 incidence prediction. *ICIC Express Letters Part B: Applications*, *15*(2), 207–216. https://doi.org/10.24507/icicelb.15.02.207

Juraphanthong, W., & Kesorn, K. (2025). Autoregressive integrated moving average with semantic information: An efficient technique for the intelligent prediction of dengue cases. *Engineering Applications of Artificial Intelligence*, *143*, Article 109985. https://doi.org/10.1016/j.engappai.2024.109985

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., &

Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168

Khadir, A. C., Aliane, H., & Guessoum, A. (2021). Ontology learning: Grand tour and challenges. *Computer Science Review*, *39*, Article 100339. https://doi.org/10.1016/j.cosrev.2020.100339

Knublauch, H., Fergerson, R. W., Noy, N. F., & Musen, M. A. (2004). The protégé OWL plugin: An open development environment for semantic web applications. In S. A. McIlraith, D. Plexousakis, & F. van Harmelen (Eds.), *The semantic Web–ISWC 2004* (pp. 229–243). Springer. https://doi.org/10.1007/978-3-540-30475-3_17

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with ensemble accuracy. *Machine Learning*, *51*(2), 181–207. https://doi.org/10.1023/A:1022859003006

Kwon, O., Na, W., & Kim, Y.-H. (2020). Machine learning: A new opportunity for risk prediction. *Korean Circulation Journal*, *50*(1), 85–87. https://doi.org/10.4070/kcj.2019.0314

Maghdid, S. S., & Rashid, T. A. (2022). *An extensive dataset for the heart disease classification system* [Dataset]. Mendeley Data, V2. https://doi.org/10.17632/65gxgy2nmg.2

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia Medica*, *23*(2), 143–149. https://doi.org/10.11613/BM.2013.018

Miraftabzadeh, S. M., Longo, M., Foiadelli, F., Pasetti, M., & Igual, R. (2021). Advances in the application of machine learning techniques for power system analytics: A survey. *Energies*, *14*(16), Article 4776. https://doi.org/10.3390/en14164776

Mitraka, E., Topalis, P., Dritsou, V., Dialynas, E., & Louis, C. (2015). Describing the breakbone fever: IDODEN, an ontology for dengue fever. *PLoS Neglected Tropical Diseases*, *9*(2), Article e0003479. https://doi.org/10.1371/journal.pntd.0003479

Polpinij, J. (2011). The Cancerology ontology: Designed to support the search of evidence-based oncology from biomedical literature. In *Proceedings of the 24th International Symposium on Computer-Based Medical Systems (CBMS)* (pp. 1–6). IEEE. https://doi.org/10.1109/CBMS.2011.5999168

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4), Article e1249. https://doi.org/10.1002/widm.1249

Sargsyan, A., Kodamullil, A. T., Baksi, S., Darms, J., Madan, S., Gebel, S., Keminer, O., Jose, G. M., Balabin, H., DeLong, L. N., Kohler, M., Jacobs, M., & Hofmann-Apitius, M. (2020). The COVID-19 ontology. *Bioinformatics*, *36*(24), 5703–5705. https://doi.org/10.1093/bioinformatics/btaa1057

Shanmugasundaram, G., Selvam, V. M., Saravanan, R., & Balaji, S. (2018). An investigation of heart disease prediction techniques. In *2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICSCAN.2018.8541165

Shouman, M., Turner, T., & Stocker, R. (2011). Using a decision tree for diagnosing heart disease patients. In P. Vamplew, A. Stranieri, & K.-L. Ong (Eds.), *Proceedings of the Ninth Australasian Data Mining Conference - Volume 121* (pp. 23–30). Australian Computer Society. https://dl.acm.org/doi/10.5555/2483628.2483633

Smiti, A. (2020). A critical overview of outlier detection methods. *Computer Science Review*, *38*, Article 100306. https://doi.org/10.1016/j.cosrev.2020.100306

Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. *Digital Health*, *2020*, Article 6. https://doi.org/10.1177/2055207620914777

Spoladore, D., Tosi, M., & Lorenzini, E. C. (2024). Ontology-based decision support systems for diabetes nutrition therapy: A systematic literature review. *Artificial Intelligence in Medicine*, *151*, Article 102859. https://doi.org/10.1016/j.artmed.2024.102859

Tan, Z., Luo, L., & Zhong, J. (2023). Knowledge transfer in evolutionary multi-task optimization: A survey. *Applied Soft Computing*, *138*, Article 110182. https://doi.org/10.1016/j.asoc.2023.110182

Tripoliti, E. E., Papadopoulos, T. G., Karanasiou, G. S., Naka, K. K., & Fotiadis, D. I. (2017). Heart failure: Diagnosis, severity estimation, and prediction of adverse events using machine learning techniques. *Computational and Structural Biotechnology Journal*, *15*, 26–47. https://doi.org/10.1016/j.csbj.2016.11.001

Verma, J. P. (2019). Non-parametric Correlations. In J. P. Verma (Ed.), *Statistics and research methods in psychology with Excel* (pp. 523–565). Springer. https://doi.org/10.1007/978-981-13-3429-0_13

Wang, L. (2015, December 8). *Heart failure ontology*. BioPortal. https://bioportal.bioontology.org/ontologies/HFO

Zhang, Y., Xin, Y., Li, Q., Ma, J., Li, S., Lv, X., & Lv, W. (2017). Empirical study of seven data mining algorithms on different characteristics of datasets for biomedical classification applications. *BioMedical Engineering OnLine*, *16*(1), Article 125. https://doi.org/10.1186/s12938-017-0416-x

## Supplementary: Weighted Semantic PageRank approach

The weighted semantic PageRank method (Chanmee & Kesorn, 2023) is employed to determine the importance value of each attribute. Here, let $C$ be the set of concepts and $R$ be the set of relationships in the ontology, where $c_a \in C$ represents a concept in the ontology, and $r \in R$ is a relationship linking each concept. First, the weight of each relationship is determined based on the frequency of a relationship for a specific concept (FR) and the inverse value of the FR (IFR). The weight of each relationship is calculated using Equations (12)–(14) (Chanmee & Kesorn, 2023).

$$FR(r, c_a) = \frac{f(r,c_a)}{max\{f(x,c_a):x \in R\}} \quad (12)$$

Where, $FR(r, c_a)$ represents a frequency relationship for concept $c_a$, and $f(r, c_a)$ is the number of relationships $r$ associated with concept $c_a$, which is the starting concept linked to another concept in the ontology. The term $max\{f(x, c_a)\}$ indicates the highest number of relationships associated with that concept.

$$IFR(r, C) = log \frac{|C|}{tr(r)} \quad (13)$$

Where, $|C|$ denotes the total number of concepts in the ontology, and $tc(r)$ denotes the total number of starting concepts of relationship $r$.

$$W(r, c) = FR(r, c_a) \times IFR(r, C) \quad (14)$$

Where, $W(r, c)$ denotes the weight of relationship $r$, which is associated with concept $c_a$.

The next step is determining the concept's importance value, which is calculated as follows (15):

$$IP(C_a) = d \sum_{b \epsilon link} \frac{Ip(c_b) \times W(r,c_b)}{\sum_{r \epsilon R} W(r,c_b)} + (1 - d) \quad (15)$$

where $Ip(c_a)$ denotes the importance value of concept $c_a$, $Ip(c_b)$ denotes the importance value of concept $c_b$, and $W(r, c_b)$ indicates the weight of the relationship r associated with concept $c_b$. In addition, *link(a)* refers to the set of concepts connected to concept $c_a$.

Figure S1, which shows an example ontology for the cardiovascular disease domain, illustrates the process of computing the attribute importance values. The example ontology comprises five concepts, i.e., *Gender, Patient, Chest Pain, Dyspnea, and Chest Pressure*, and three types of relationships, i.e., *has_gender*, *has_symptom*, and *is_subclass*.

As shown in Figure S1, *Patient* is the starting concept of the *has_gender* and *has_symptom* relationships, and *Chest Pressure* is the starting concept of the *is_subclass* relationship. The *Patient* concept is connected to another concept by one *has_gender* link and two *has_symptom* links. As a result, the *FR* of the *has_gender* relationship is $\frac{1}{2}$, and that of the *has_symptom* relationship is $\frac{2}{2}$. The *Chest Pressure* concept is connected to another concept by one link of the *is_subclass* relationship; therefore, the FR is $\frac{1}{1}$.

As shown in Figure S1, the ontology contains a total of five concepts, and each relationship is connected to a single concept. As a result, the IFR of each concept is $log \frac{|5|}{1} = 0.7$.

The weight of each relationship is computed using Equation (14), resulting in the following.

- W(has_gender, Patient) = 0.5×0.7=0.35
- W(has_symptom, Patient) = 1×0.7=0.7
- W(is_subclass, ChestPressure) = 1×0.7=0.7

Computing the importance value for each concept is an iterative process in which the values are updated continuously until convergence is achieved. In the first iteration, the importance value is initialized as $\frac{1}{|c|}$, and then the importance value becomes $\frac{1}{5} = 0.2$. The importance value of the *Chest Pain* concept, which is associated with the *Patient* and *Chest Pressure* concepts, can be computed as follows.

$$Ip(ChestPain) = 0.85 \times \left(\left(\frac{0.2 \times 0.7}{0.35+0.7+0.7}\right) + \left(\frac{0.2 \times 0.7}{0.7}\right)\right) + (1 - 0.85)$$
$$= 0.388$$

For the *Patient* concept, which does not have any other concept linked to it, the importance value can be calculated as follows.

$$Ip(Patient) = 0.85 \times 0 + (1 - 0.85)$$
$$= 0.150$$

Note that the importance values of the remaining concepts can be computed using the same approach. The computed importance values are used in the subsequent iteration and updated iteratively until they remain unchanged. The results of this example are shown in Table S1.

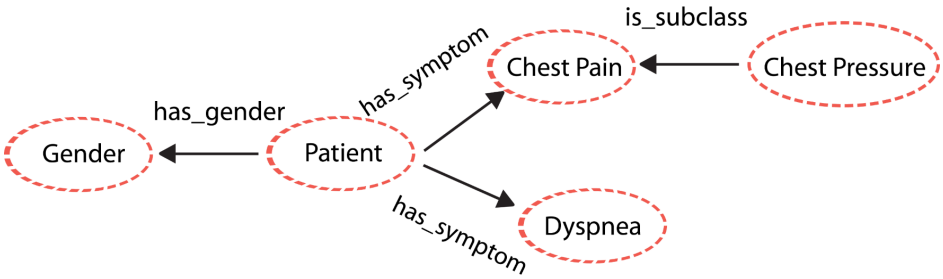



**Figure S1.** Example of an ontology for the cardiovascular disease domain

**Table S1.** Importance value of each concept

| **Concept** | **Importance value ($O_i$)** | | |
|---|---|---|---|
| | **First iteration** | **Second iteration** | **Third iteration** |
| Gender | 0.184 | 0.176 | 0.176 |
| Patient | 0.150 | 0.150 | 0.150 |
| Chest Pain | 0.388 | 0.329 | 0.329 |
| Dyspnea | 0.218 | 0.201 | 0.201 |
| Chest Pressure | 0.150 | 0.150 | 0.150 |