ICIC Express Letters
Part B: Applications
Volume **15**, Number **11**, November **2024**
ICIC International ©2024 ISSN 2185-2766
pp. **1175–1182**

COVID-19 CASES CLASSIFICATION USING A SEMANTIC DECISION FOREST METHOD

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

[1]Faculty of Science and Agricultural Technology
Rajamangala University of Technology Lanna Phitsanulok
52 Moo 7, Ban Krang, Mueang Phitsanulok District, Phitsanulok 65000, Thailand
sirichanyac@rmutl.ac.th

[2]Department of Computer Science and Information Technology
Faculty of Science
Naresuan University
99 Moo 9, Thapo Sub-District, Muang District, Phitsanulok 65000, Thailand
*Corresponding author: kraisakk@nu.ac.th

Received February 2024; accepted April 2024

Abstract. *The COVID-19 pandemic has been a global health crisis, impacting millions of people worldwide. In such critical circumstances, early detection of infected patients is crucial to minimize fatalities. To facilitate rapid diagnosis of COVID-19, we have developed the 'Semantic Decision Forest (SDF)' – a tree-based ensemble method that leverages an ontology model to classify infected patients. During model construction, the ontology knowledge assists in selecting important attributes for decision tree nodes. In this study, we evaluate the effectiveness of our proposed method using the COVID-19 dataset from Brazil. By employing advanced tree-based ensemble algorithms such as AdaBoost, Gradient Boosting, and Random Forest, we comprehensively assess the performance of our approach. The experimental results unequivocally demonstrate that our proposed method outperforms other algorithms, achieving the highest classification accuracy and sensitivity.*

Keywords: Classification, Ensemble, Semantic, Knowledge-based, Decision tree, Semantic decision forest

1. **Introduction.** The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has led to a wide range of respiratory symptoms, varying from mild to severe, which often require intensive hospitalization [1]. Timely diagnosis and vaccination play a crucial role in preventing infection and reducing the risk of death.

In various domains, Machine Learning (ML) has emerged as a powerful technology for classification and prediction, enabling the identification of patterns and hidden knowledge within collected data. However, traditional ML algorithms often overlook the semantics and relationships embedded in the data [2]. To address this limitation, researchers have integrated knowledge to enhance the analytical process. For example, Idoudi et al. [3] employed a mammographic ontology to evaluate the significance and similarity of association rule items, leading to the development of valuable rules.

Classification, a common ML task, involves assigning labels to new observations based on training data. Decision trees, known for their interpretability, are frequently utilized for classifying unknown data into specific categories. This algorithm has been utilized for disease diagnosis tasks [4] due to its interpretable and comprehensible results. Nonetheless, decision tree algorithms face challenges such as overfitting, biased attribute selection, and instability issues [5]. To overcome these limitations, various methods have been proposed, including the generation of novel splitting conditions [6,7] and the utilization of

DOI: 10.24507/icicelb.15.11.1175

ensemble methods [8]. In our previous study [7], we incorporated an ontology to establish a new splitting criterion, effectively addressing the issue of biased attribute selection.

In this paper, we present a novel method that leverages ensemble learning and domain knowledge to enhance the effectiveness of the Decision Forest (DF) algorithm in classifying COVID-19 patients. This approach differs from our previous research [9], which primarily focused on developing a single model relying solely on domain knowledge. By employing an ensemble method, our purpose was to enhance the performance of the individual decision tree model and effectively address issues associated with overfitting and instability within the DF. As well, we utilize knowledge from an ontology as a crucial factor in guiding the algorithm during the node selection process. Given these factors, we are confident that our research offers an innovative approach that integrates preexisting knowledge with tree-based ensemble learning for accurate classification of COVID-19 cases. The experimental results demonstrate that our proposed approach consistently achieves higher levels of accuracy and sensitivity than conventional tree-based ensemble methods.

The structure of the paper is organized as follows: Section 2 provides a comprehensive overview of related studies in the field, Section 3 meticulously describes the proposed approach, Section 4 presents the outcomes and facilitates an in-depth discussion, and finally, Section 5 concludes the paper by outlining future directions for further research and improvement in this domain.

2. **Literature Review.**

Random Forest (RF) [9] is widely recognized as a machine learning algorithm renowned for its robustness and accuracy as a classifier across diverse domains [10]. RF constructs its model exclusively based on the training data, abstaining from leveraging prior knowledge. Nevertheless, researchers have acknowledged the potential benefits arising from the integration of domain knowledge into the RF algorithm throughout the data preparation phase [11] and the modeling phase. This integration is driven by the underlying belief that incorporating domain knowledge can augment the performance of conventional ML algorithms by shedding light on decision patterns and curtailing the search space [12].

During the model construction process, external knowledge was utilized to assist the RF algorithm in achieving high performance. To illustrate, Wang et al. [13] proposed a two-stage machine learning model for predicting the survival time of cancer patients. In the first stage, a classification tree was used to classify the patient's survival case, followed by the prediction of survival time using a regression tree. The importance of features, determined through the regression tree using observed data, served as prior knowledge to aid in feature selection during model construction. The proposed approach generated multiple feature subsets and assessed the importance of each subset during tree construction. The feature subset with the highest total importance was selected to construct the individual tree. The experimental results convincingly demonstrate that the proposed method exhibited the lowest prediction error of all other approaches.

Similarly, Cantor et al. [14] developed a knowledge-slanted RF algorithm that utilized a protein-protein interaction network to enhance RF. The knowledge derived from the network was used to rank genes associated with calcific aortic valve stenosis, enabling the selection of the gene with the highest priority in the early steps of the algorithm. Their findings showed that this knowledge improved classification performance by distinguishing between three predefined classes.

Sun et al. [15] introduced a knowledge-based RF known as the Multiplicative Integration Weighted-Random Forest (MIW-RF), which incorporated control process knowledge. This knowledge acted as the controlling input for generating multiple forests, and the results from these forests were averaged to obtain the final outputs. The results indicated that the proposed approach demonstrated improved performance over that of benchmark methods.

Previous research has consistently demonstrated that integrating knowledge with ensemble methods can significantly enhance the overall performance of individual methods. In our study, to develop an effective classification model for COVID-19 cases, we integrated an ontology [16], which serves as a formal descriptive and shared body of knowledge, into a tree-based ensemble. Section 3 provides a detailed discussion of the methodology for integrating an ontology with the tree-based ensemble approach.

3. **Materials and Methods.** In this section, we present a comprehensive explanation of the framework used in our approach, as illustrated in Figure 1. The framework consists of three main components: 1) data preparation, 2) construction of the Semantic Decision Forest (SDF), and 3) model evaluation. Each process is described in detail below.

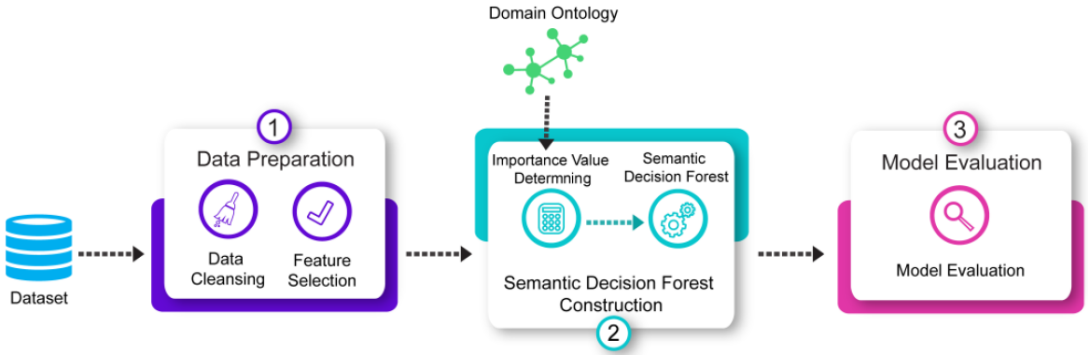


FIGURE 1. A proposed framework that integrates knowledge into a tree-based ensemble method

3.1. **Dataset and ontology.** We evaluated our methodology using the Brazilian COVID-19 dataset, available in the Mendeley Data repository [17]. This dataset consists of 3,128 records with 11 attributes, including patient demographics and COVID-19 symptoms, enabling the classification of COVID-19 and non-COVID-19 cases. To improve the classification process, we integrated a COVID-19 ontology [18] that includes patient demographics and covers all important aspects of the dataset. The COVID-19 ontology contains 2,276 concepts and 11 types of relationships.

3.2. **Data preparation.** Data preparation is a crucial step in data analysis that ensures the integrity of the data and facilitates accurate results. To address missing values, we applied the list-wise deletion method [19], eliminating instances with missing values from the dataset. For feature selection, we employed the Chi-square test [20] to identify the most informative features for model construction. A p-value of the Chi-square test that exceeds 0.05 indicates that the test attribute is uncorrelated with the target class. In such cases, the uncorrelated attributes are removed from the dataset. Subsequently, we divided the resulting cleansed dataset into a 70 : 30 ratio for training and testing data, respectively, in the subsequent steps.

3.3. **Semantic decision forest construction.** This section introduces the SDF approach, which enhances the analysis process through the integration of knowledge. SDF is a tree-based ensemble method that combines the ID3 algorithm with ontology-based knowledge. By incorporating ontology knowledge, importance values can be assigned to attributes, guiding the algorithm in selecting significant attributes during the node selection process. The construction procedure of SDF is presented in Figure 2 and described in detail below.

In the SDF construction procedure, we leverage the information from the domain ontology to determine the importance value of each attribute. This is achieved by utilizing the Weighted Semantic PageRank method [21], which assigns importance values to ontology

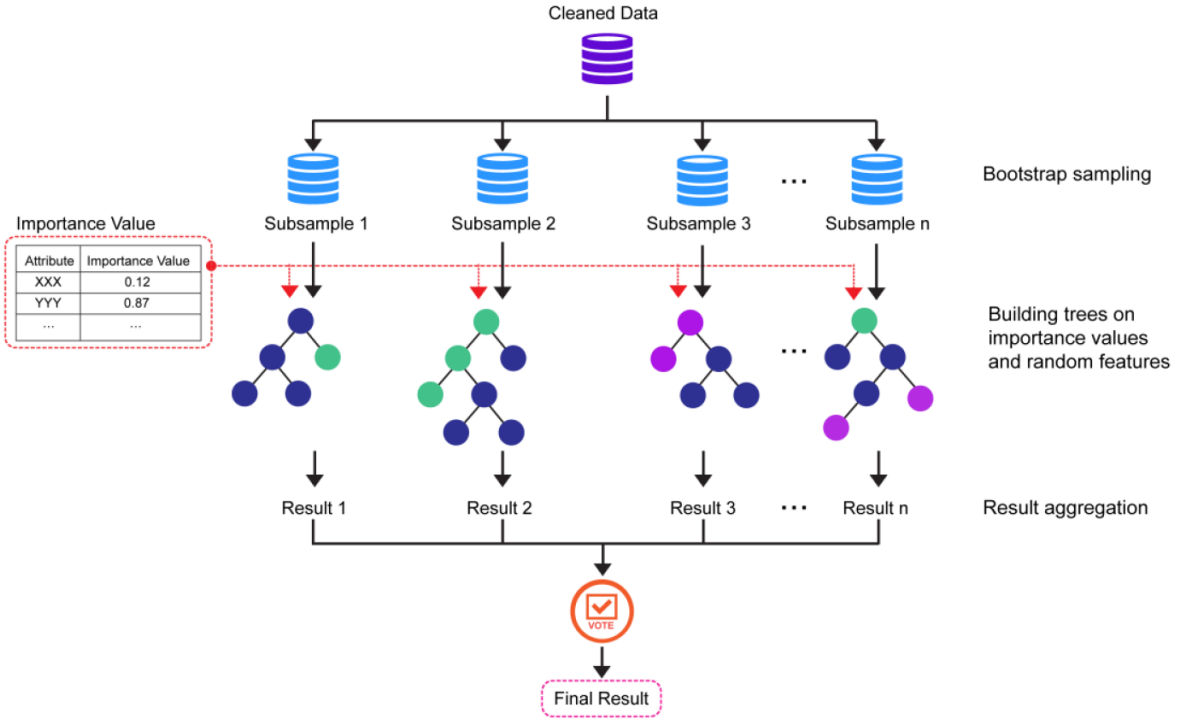


FIGURE 2. A semantic decision forest construction process

concepts based on their relationships. The resulting importance values serve as the basis for assigning importance values to the corresponding attributes during the construction of individual decision trees within the SDF. To promote diversity and minimize generalization errors, we created multiple subsamples of the same size as the original training data during the SDF construction procedure. These subsamples are generated by randomly selecting instances from the original training data with replacement, resulting in an average presentation rate of 63.2% [22]. Each subsample serves as training data for developing a decision tree within the SDF.

The random subspace method was employed to select a random subset of features for the node selection process in each decision tree. For each feature, we calculate the information gain, which acts as an indicator for the ID3 splitting criterion. To address the bias in attribute selection that favors attributes with a large number of values, we incorporate the importance values derived from the ontology to modify the information gain. This ensures that attributes with significant importance values have a higher likelihood of being selected as decision tree nodes compared to attributes with numerous distinct values. The computation for the revised information gain is defined by Equations (1) through (3).

$$Entropy(D) = -\sum_{i=1}^{m} p(i|D) \log_2 p(i|D) \tag{1}$$

where $Entropy(D)$ is the entropy of the dataset D which comprises m classes, and $p(i|D)$ is the probability that the instances in the dataset D are categorized under class i.

$$Entropy(A) = \sum_{j=1}^{k} \frac{n_j}{n} Entropy(j) \tag{2}$$

where $Entropy(A)$ represents the average entropy of the attribute A, which encompasses k unique values. The term n refers to the total number of instances of the dataset D and n_j refers to the number of instances that have class j of the attribute A [22].

$$RGAIN(A) = IValue(A) \times (Entropy(D) - Entropy(A)) \tag{3}$$

where $RGAIN(A)$ refers to the revised information gain of the attribute A and $IValue(A)$ refers to the importance value of the attribute A, which is determined based on the ontology.

Each decision tree is fully grown without pruning. Finally, the result of each decision tree will be aggregated, and a majority vote will be applied to obtain the final classification result of the SDF.

3.4. **Model evaluation.** Accuracy and sensitivity, which are the classification evaluation metrics, are used to measure the performance of the SDF. The accuracy is used to measure the correctly classified data, and sensitivity represents the percentage of the COVID-19 patients that are correctly classified. The accuracy and sensitivity are defined as follows [22].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

where TP represents the number of correctly predicted positive instances that belong to the positive class, whereas TN denotes the number of accurately labeled negative instances that belong to the negative class. FP indicates the mislabeling of negative instances as positive, while FN represents the count of positive instances wrongly predicted as negative.

4. **Results and Discussion.** The purpose of this section is to assess the performance of the SDF. The COVID-19 dataset is applied to examine the proposed approach. The findings are presented in the subsequent sections.

4.1. **Determining the optimal number of decision trees.** The goal of this experiment was to identify the optimal number of decision trees for the SDF that achieves the highest classification performance. The number of decision trees used in the forest is a critical parameter: having too many trees can result in slower algorithm execution. We experimented by varying the number of decision trees, including 10, 50, 100, 500, 800, and 1,000, to construct the SDF. The classification results of the proposed approach with the different numbers of decision trees are presented in Table 1.

Table 1. The classification result of the SDF with different numbers of decision trees

Number of decision trees	Average accuracy
10	88.44%
50	88.98%
100	***89.25%***
500	89.15%
800	89.11%
1,000	89.15%

The study determined that the optimal number of decision trees for achieving the highest accuracy was 100, resulting in an accuracy rate of 89.25%. However, when the number of decision trees exceeded 100 in the construction of the SDF, the classification performance exhibited fluctuations, and the accuracy dropped to 89.15% with 1,000 decision trees. It is important to note that the COVID-19 dataset used in this experiment has low dimensionality, consisting of only ten predictor attributes. Consequently, an excessive

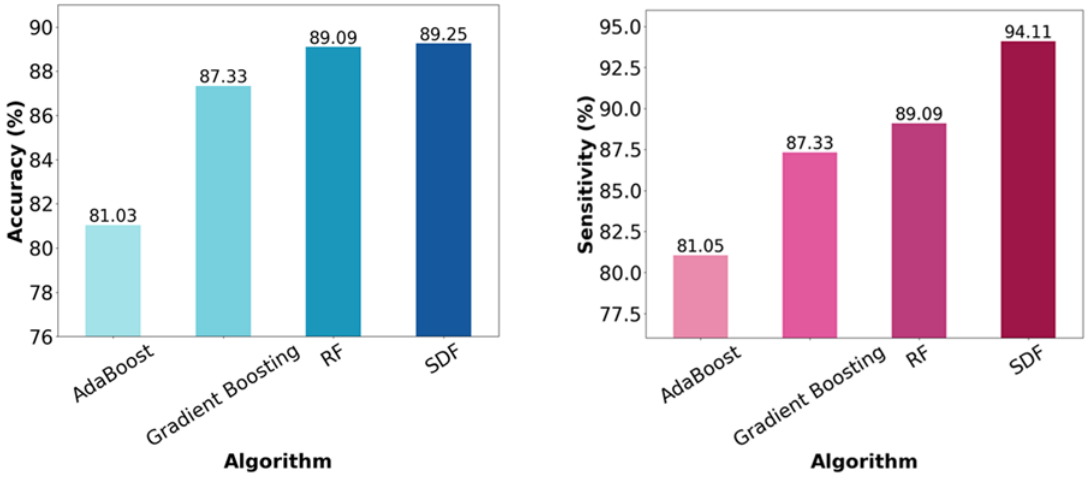
number of decision trees may generate redundancies. Furthermore, if each individual decision tree produces identical classification results, the effectiveness of classification may not significantly improve.

These findings emphasize the influence of the number of decision trees on the precision of the SDF. However, merely increasing the number of decision trees does not guarantee a significant enhancement in classification effectiveness. Thus, the careful adjustment of the number of decision trees employed in the SDF becomes a critical parameter that necessitates thorough consideration to achieve satisfactory classification performance.

4.2. **Comparing the SDF and other tree-based ensemble methods.** The purpose of this investigation was to assess the classification performance of the proposed method by comparing it with established ensemble algorithms, namely AdaBoost, Gradient Boosting, and RF. To ensure a fair comparison, the number of decision trees for all algorithms was set to 100, based on the findings of a previous experiment. A total of thirty independent experiments were conducted, and the resulting accuracies and sensitivities were averaged to derive the outcomes.

Figure 3(a) presents the classification accuracies of each algorithm. The proposed method demonstrated superior precision compared to all other algorithms. The SDF achieved the highest accuracy of 89.25%, slightly surpassing RF with an accuracy of 89.09%. When applied to the dataset, AdaBoost and Gradient Boosting yielded classification accuracies of 81.03% and 87.33%, respectively.

In the medical context, misclassifying patients infected with COVID-19 can have severe implications. Therefore, sensitivity is employed to assess the ability of each algorithm to accurately classify infected patients. Figure 3(b) illustrates the sensitivity of the investigated algorithms. The proposed approach outperformed the other algorithms in terms of sensitivity, achieving a maximum sensitivity of 94.11%. Utilizing RF to classify the COVID-19 dataset resulted in a decrease in sensitivity of 5.02% to 89.09%. Among the algorithms, AdaBoost exhibited the lowest sensitivity, with a value of 81.05% when classifying the dataset.



(a) (b)

FIGURE 3. Comparison of classification performance of the SDF and other tree-based ensemble algorithms

These findings underscore the significance of ontology-based knowledge in achieving high sensitivity. The attribute importance values extracted from the COVID-19 ontology guide the algorithm in selecting the most significant and relevant attributes for the decision tree nodes. By incorporating pertinent attributes in the model construction, improved classification performance is attained [23].

The findings of this experiment revealed that the proposed approach, the SDF, is capable of producing a model with superior classification performance. The developed model precisely categorizes infected patients, thereby reducing the risk of patient mortality and preventing a pandemic.

5. **Conclusions.** To enhance the classification of COVID-19 patients, the SDF method was introduced as an ensemble approach based on decision trees, which incorporates ontology information. Significance values for each concept are calculated using a PageRank-based algorithm, taking account of the interrelationships among the concepts within the ontology. These values play a crucial role in guiding the selection of attributes within individual decision trees, resulting in improved classification accuracy. The ensemble method exhibits remarkable performance in classification tasks. However, it is imperative to emphasize the significance of ensuring the quality of the employed ontology. Insufficient ontology knowledge may lead to inaccurate significance values and, subsequently, inaccurate classifications. The responsibility of maintaining high-quality ontology information lies with human decision-making. Moving forward, our future research endeavors will be dedicated to refining techniques aimed at increasing decision tree diversity and ultimately enhancing the overall performance of the classification system.

Acknowledgement. This research was supported by Naresuan University (NU) and National Science, Research, and Innovation Fund (NSRF), Grant No. R2567B017. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation. We wish to thank Mr. Roy I. Morien of the Naresuan University Graduate School for his editing of the grammar, syntax and general English expression in this document.

REFERENCES

[1] G. Pascarella, A. Strumia, C. Piliego, F. Bruno, R. D. Buono, F. Costa, S. Scarlata and F. E. Agrò, COVID-19 diagnosis and management: A comprehensive review, *J. Intern. Med.*, vol.288, no.2, pp.192-206, 2020.

[2] C. Sirichanya and K. Kraisak, Semantic data mining in the information age: A systematic review, *Int. J. Intell. Syst.*, vol.36, no.8, pp.3880-3916, 2021.

[3] R. Idoudi, K. S. Ettabaa, B. Solaiman and K. Hamrouni, Ontology knowledge mining based association rules ranking, *Procedia Comput. Sci.*, vol.96, pp.345-354, 2016.

[4] H. Ilyas, S. Ali, M. Ponum, O. Hasan, M. T. Mahmood, M. Iftikhar and M. H. Malik, Chronic kidney disease diagnosis using decision tree algorithms, *BMC Nephrology*, vol.22, no.1, p.273, 2021.

[5] R. Rivera-Lopez, J. Canul-Reich, E. Mezura-Montes and M. A. Cruz-Chávez, Induction of decision trees as classification models through metaheuristics, *Swarm Evol. Comput.*, vol.69, 101006, 2022.

[6] F. Es-Sabery, K. Es-Sabery, J. Qadir, B. Sainz-De-Abajo, A. Hair, B. García-Zapirain and I. De La Torre-Díez, A MapReduce opinion mining for COVID-19-related tweets classification using enhanced ID3 decision tree classifier, *IEEE Access*, vol.9, pp.58706-58739, 2021.

[7] S. Chanmee and K. Kesorn, Exploiting a knowledge base for intelligent decision tree construction to enhance classification power, *Eng. Appl. Sci. Res.*, vol.49, no.4, 2022.

[8] O. Sagi and L. Rokach, Ensemble learning: A survey, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol.8, no.4, e1249, 2018.

[9] L. Breiman, Random forests, *Mach. Learn.*, vol.45, no.1, pp.5-32, 2001.

[10] M. S. A. Rahaman, P. Vasant, I. ul Haq, E. Padmanabhan, S. R. Jufar, R. Sokkalingam and J. Watada, Evaluation of tree-based ensemble learning algorithms to estimate total organic carbon from wireline logs, *International Journal of Innovative Computing, Information and Control*, vol.17, no.3, pp.807-829, 2021.

[11] K. Gao, Y. Yang, T. Zhang, A. Li and X. Qu, Extrapolation-enhanced model for travel decision making: An ensemble machine learning approach considering behavioral theory, *Knowl. Based Syst.*, vol.218, 106882, 2021.

[12] S. S. Anand, D. A. Bell and J. G. Hughes, The role of domain knowledge in data mining, *Proc. of the 4th International Conference on Information and Knowledge Management*, Baltimore, MD, USA, pp.37-43, 1995.

[13] Y. Wang, D. Wang, X. Ye, Y. Wang, Y. Yin and Y. Jin, A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction, *Inf. Sci.*, vol.474, pp.106-124, 2019.
[14] E. Cantor, R. Salas, H. Rosas and S. Guauque-Olarte, Biological knowledge-slanted random forest approach for the classification of calcified aortic valve stenosis, *BioData Mining*, vol.14, no.1, p.35, 2021.
[15] L. Sun, Y. Ji, X. Zhu and T. Peng, Process knowledge-based random forest regression for model predictive control on a nonlinear production process with multiple working conditions, *Adv. Eng. Inform.*, vol.52, 101561, 2022.
[16] S. Staab and R. Studer, *Handbook on Ontologies*, 2nd Edition, Springer Science & Business Media, Heidelberg, 2009.
[17] Í. V. dos S. Santana, A. C. M. da Silveira, A. Sobrinho, L. C. e Silva, L. D. da Silva, D. F. de S. Santos, E. Candeia and A. Perkusich, A Brazilian dataset of symptomatic patients for screening the risk of COVID-19, *Mendeley Data*, vol.5, 2021.
[18] A. T. Kodamullil, COVID-19 ontology, *BioPortal*, https://bioportal.bioontology.org/ontologies/COVID-19, 2021.
[19] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago and O. Tabona, A survey on missing data in machine learning, *J. Big Data*, vol.8, no.1, p.140, 2021.
[20] M. L. McHugh, The Chi-square test of independence, *Biochem. Med.*, vol.23, no.2, pp.143-149, 2013.
[21] H.-G. Jun, D.-H. Im and H.-J. Kim, An RDF metadata-based Weighted Semantic PageRank algorithm, *Int. J. Semant. Web Inf. Syst.*, vol.7, pp.11-24, 2016.
[22] J. Han, J. Pei and M. Kamber, Data mining: Concepts and techniques, 3rd Edition, In *The Morgan Kaufmann Series in Data Management Systems*, Elsevier Science, Burlington, MA, 2011.
[23] R. Spencer, F. Thabtah, N. Abdelhamid and M. Thompson, Exploring feature selection and classification methods for predicting heart disease, *Digit. Health*, vol.6, 2055207620914777, 2020.