Tapping into Knowledge: Ontological Data Mining Approach for Detecting Cardiovascular Disease Risk Causes Among Diabetes Patients

Hussein Alkattan[*1], S. K. Towfek[2], M. Y. Shams[3]

[1] Department of System Programming, South Ural State University,
454080 Chelyabinsk, Russia
[2] Computer Science and Intelligent Systems Research Center, Blacksburg 24060,
Virginia, USA
[3] Faculty of Artificial Intelligence, Kafrelsheikh University, Kafrelsheikh, Egypt
Emails: alkattan.hussein92@gmail.com; sktowfek@jcsis.org; mahmoud.yasin@ai.kfs.edu.eg

Abstract

The prevalence of cardiovascular disease (CVD) is a serious public health issue, and it is of particular concern for people with diabetes because of the increased risk of cardiovascular problems that these people experience. In this study, we suggest a novel method of Ontological Data Mining (ODM) for identifying the origins of CVD risk in diabetic patients. We want to improve the readability and precision of prediction models by incorporating domain knowledge and semantic linkages into the data mining process. In this work, we examine a large dataset consisting of 70,000 patient records with 11 attributes, all of which are derived through a thorough clinical history and physical examination. As part of our methodology, we use decision trees, support vector machines (SVMs), and gradient boosting (GB). The distribution patterns of critical variables with respect to CVD outcomes can be better understood through the use of visual representations such as box plots, distributional plots, and pie charts. Finding significant connections and causal relationships between risk factors and CVD outcomes is made possible by the suggested ODM method. Our research has promising implications for bettering the treatment of patients with diabetes, facilitating targeted interventions, and enhancing risk assessment and preventative methods for cardiovascular disease.

Keywords: Ontological Data Mining; Cardiovascular Disease; Diabetes; Boosting; Predictive Models; Interpretability; Data Visualization.

1. Introduction

Cardiovascular disease poses a significant health challenge worldwide, particularly among individuals with diabetes. The coexistence of diabetes and cardiovascular disease leads to increased morbidity, mortality, and healthcare costs. Identifying the specific risk causes associated with cardiovascular disease among diabetes patients is crucial for effective preventive strategies and targeted interventions. In this paper, we propose an innovative approach, ontological data mining, to tap into knowledge and uncover the underlying risk causes for cardiovascular disease among diabetes patients. Cardiovascular disease encompasses a range of conditions affecting the heart and blood vessels, while diabetes refers to a metabolic disorder characterized by high blood sugar levels. The co-occurrence of

Doi: https://doi.org/10.54216/JAIM.040101
Received: August 19, 2022 Revised: January 15, 2023 Accepted: June 02, 2023

diabetes and cardiovascular disease amplifies the risk of adverse outcomes and poses a significant burden on global healthcare systems.

Identifying the risk factors and causative agents contributing to cardiovascular disease among diabetes patients is essential for targeted interventions, personalized care, and improved patient outcomes. However, existing research has encountered challenges in comprehensively detecting and unraveling the intricate risk causes. Previous studies have explored various approaches to identify cardiovascular disease risk factors in diabetes patients. However, these approaches often exhibit limitations, such as narrow focus, fragmented knowledge representation, and insufficient integration of diverse data sources. These limitations hinder the holistic understanding of risk causes and restrict the development of effective preventive strategies. Ontological data mining offers a promising solution to overcome the limitations of traditional approaches. By leveraging the principles of ontologies and data mining techniques, this approach enables the comprehensive analysis and integration of diverse data sources, facilitating the detection of hidden patterns, associations, and risk causes.

Our primary objective is to apply ontological data mining to identify and understand the specific risk causes associated with cardiovascular disease among diabetes patients. By harnessing the rich knowledge inherent in ontologies and employing data mining algorithms, we aim to uncover novel perspectives and insights that can inform preventive measures and targeted interventions for this vulnerable patient population.

2. Literature Review

Prior research has extensively explored the intricate relationship between cardiovascular disease and diabetes, acknowledging the heightened risk of cardiovascular complications among patients with diabetes. Various studies have focused on identifying risk factors and causative agents that contribute to this elevated susceptibility. Jayaraman et al. [1] present a comprehensive review of "Healthcare 4.0," an emerging paradigm integrating digital health technologies into healthcare systems. This review highlighted various frontiers in digital health, emphasizing the potential for data mining and knowledge discovery to improve patient outcomes. While the focus of this work was on broader healthcare advancements, it laid the groundwork for exploring digital health solutions in the context of cardiovascular disease risk identification among diabetes patients. Rahman et al. [2] delved into the role of data mining in telemedicine, with a particular focus on health monitoring technologies. Their research examined how data mining techniques can extract meaningful insights from telemedicine data to enhance disease risk assessment and management. The integration of telemedicine with data mining holds promise for capturing relevant information related to cardiovascular risk causes among diabetes patients. Alfaisal et al. [3] explored predictive modeling approaches using Partial Least Squares Structural Equation Modeling (PLS-SEM) and Machine Learning (ML) to forecast social media usage among university communicators. Although not directly related to cardiovascular disease or diabetes, this study demonstrates the potential of data mining techniques in predicting complex human behaviors, which could be relevant when understanding patient behaviors and risk factors. Darshan and Anandakumar [4] presented a comprehensive review of the usage of the Internet of Things (IoT) in healthcare systems. While their focus was broader, the integration of IoT data and ontological data mining might offer new possibilities for extracting insights related to cardiovascular disease risk factors in diabetes patients. Alfaisal et al. [5] investigated the acceptance of metaverse systems using a hybrid approach of PLS-SEM and Machine Learning. Though metaverse technology is not directly related to cardiovascular disease, this study demonstrated the applicability of data mining methods in assessing user acceptance and behavior, which could be applied to understand patient adherence to risk prevention strategies. Abdelhady and Ismail [6] presented a study that employs various Machine Learning models to forecast cardiovascular diseases. While not directly related to diabetes patients, this work showcases the potential of data mining techniques in predicting disease outcomes, which could be adapted to identify risk causes in diabetic patients susceptible to cardiovascular complications. Aljanada et al. [7] analyzed the adoption of Google Glass technology using PLS-SEM and Machine Learning. Although unrelated to cardiovascular disease, this research demonstrates the application of data mining techniques in understanding technology adoption behavior, which could inform the implementation of innovative healthcare technologies for risk cause identification. Silverman et al. [8] explored molecular networks in the context of network medicine, emphasizing their development and applications. Though not directly linked to diabetes or

Doi: https://doi.org/10.54216/JAIM.040101
Received: August 19, 2022 Revised: January 15, 2023 Accepted: June 02, 2023

cardiovascular disease, this work highlighted the potential of network-based approaches to uncover complex interactions and associations in medical datasets, which could be relevant for identifying risk causes.

3. Methodology

In this section, we discuss the machine learning (ML) algorithms employed in our study for detecting cardiovascular disease risk causes among diabetes patients. We utilized the following ML algorithms: Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM), and Gradient Boosting (GB).

DT are tree-based models that recursively partition the dataset based on the values of input features. The splitting process is guided by mathematical criteria, namely Gini impurity, to determine the optimal feature and threshold for each split.

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

In our study, we present DT-based ODM for the detection of CVD risk causes among diabetes patients. The integration of ontological principles with data mining techniques allows us to leverage domain knowledge and semantic relationships to enhance the accuracy and interpretability of the predictive models. DT are particularly suitable for ODM as they can represent complex decision-making processes and capture the hierarchical relationships between variables [14-17]. DT algorithms create a tree-like structure where each internal node corresponds to a decision based on a feature, and each leaf node represents a predicted outcome. By incorporating domain-specific ontologies into the decision tree construction process, we can enhance the interpretability of the resulting model and uncover causal relationships. The same process is applied on random forest given in algorithm bellow shown in figure 1.

Algorithm 1 Random Forest

1 For $b = 1$ to B :
(a) Sketch a bootstrap sample $\mathbf{Z}^*$ of size N from the training samples.
(b) Cultivate a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, till the smallest node size $n_{\min}$ is reached.
i. Choose m variables at random from the p variables.
ii. Choose the highest variable/split-point among the m.
iii. Break the node into two offspring nodes.

2 Output the ensemble of trees $\{T_b\}_1^B$.
To achieve a prediction at a new point x :
Classification: Let $\hat{C}_b(x)$ be the class prediction of the b th random-forest tree. Then $\hat{C}_{\mathrm{rf}}^B(x) =$ majority vote $\{\hat{C}_b(x)\}_1^B$.

Figure 1: Algorithm of random forest

When it comes to classification and regression, supervised learning models like Support Vector Machines (SVMs) are hard to beat. The goal of SVM is to identify a hyperplane that most effectively divides data into distinct classes [18-19]. In the mathematical formulation, the support vectors are a subset of the training samples that define the decision boundary. By using kernel functions to map the data into higher-dimensional spaces, SVM is able to handle non-linearly separable data.

$$\min_{w,b,\{\beta_n|} \frac{1}{2} \| w \|_2 + C \sum_n \beta_n \quad (1)$$

$$\text{s.t.} \quad y_n[w^T \phi(x_n) + b] \geq 1 - \beta_n; \forall n$$

$$\beta_n \geq 0, \forall n$$

Our goal in combining ontological principles with SVM was to improve the predictive models' precision and interpretability by capitalising on prior domain knowledge and semantic links. Due to their ability to efficiently handle complex classification tasks and capture non-linear correlations between variables, SVMs are ideally suited for ODM. SVM creates a hyperplane that maximises the distance between data points belonging to various classes.

Doi: https://doi.org/10.54216/JAIM.040101
Received: August 19, 2022 Revised: January 15, 2023 Accepted: June 02, 2023

GB is an ensemble learning technique that sequentially builds a series of weak learners, typically decision trees, to create a strong predictive model [20-21]. The models are trained in an iterative manner, with each new model focusing on correcting the errors made by the previous models. The optimization process involves minimizing a loss function by taking gradients with respect to the model parameters.

$$F^*(x) = \underset{F(x)}{\operatorname{argmin}} \sum_{i=1}^{N} \Phi(F(x_i)) = \underset{F(x)}{\operatorname{argmin}} \sum_{i=1}^{N} \Psi(y_i, F(x_i)) \tag{2}$$

where,

$$\Psi(y_i, F(x_i)) = (y_i - F(x_i))^2 \tag{3}$$

GB can enhance the interpretability and predictive performance of the model. In our study, we applied the ODM approach using GB to detect cardiovascular disease risk causes among diabetes patients. We integrated domain knowledge related to cardiovascular health, diabetes, and relevant risk factors into the GB modeling process. This integration allowed us to incorporate expert-defined rules, constraints, and semantic relationships between variables.

4. Results and Analysis

We evaluated the performance of our proposed model on a case study dataset consisting of 70,000 patient records with 11 features and a target variable indicating the presence or absence of cardiovascular disease. The dataset provided a comprehensive range of information, including objective, examination, and subjective features, enabling a thorough analysis of potential risk causes associated with cardiovascular disease among diabetes patients. The objective features included age, height, weight, and gender, providing factual information about the patients. The examination features encompassed systolic and diastolic blood pressure, cholesterol level, and glucose level, derived from medical examinations. Finally, the subjective features consisted of smoking habits, alcohol intake, and physical activity levels, which were self-reported by the patients. Each feature played a crucial role in capturing different aspects related to cardiovascular disease risk causes among diabetes patients. For instance, age, gender, and BMI-related features (height and weight) provided insights into demographic factors and body composition, which have known associations with cardiovascular health. The examination features, including blood pressure, cholesterol, and glucose levels, contributed to understanding the physiological markers of cardiovascular disease risk. elevated blood pressure, abnormal cholesterol levels, and high glucose levels are well-established risk factors for cardiovascular complications. A summary of data statistics is given in Table 1.

Table 1: statistical data analysis for our case study.

	age	Height	weight	ap_hi	ap_lo
count	69976.000000	69976.000000	69976.000000	69976.000000	69976.000000
mean	19468.950126	164.359152	74.208519	128.820453	96.636261
std	2467.374620	8.211218	14.397211	154.037729	188.504581
min	10798.000000	55.000000	10.000000	-150.000000	-70.000000
25%	17664.000000	159.000000	65.000000	120.000000	80.000000
50%	19703.000000	165.000000	72.000000	120.000000	80.000000
75%	21327.000000	170.000000	82.000000	140.000000	90.000000
max	23713.000000	250.000000	200.000000	16020.000000	11000.000000

In Figure 2, we present the visual representation of the box plot and distributional plot for the height and weight variables in our dataset.

Doi: https://doi.org/10.54216/JAIM.040101
Received: August 19, 2022 Revised: January 15, 2023 Accepted: June 02, 2023

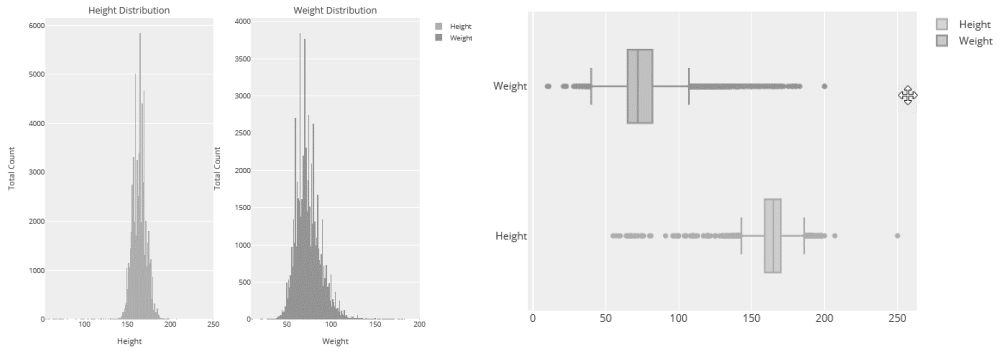

Figure 2: Visualization of Box Plot and Distributional Plot
for Height and Weight Variables

The box plot provides a concise summary of the distribution of a variable, showcasing key statistical measures such as the median, quartiles, and any potential outliers. For the height variable, the box plot displays a horizontal box indicating the interquartile range (IQR), with a line inside representing the median height. The whiskers extend from the box to indicate the minimum and maximum values within a certain range, excluding outliers if present. As noted, we can assess the central tendency of the distribution and observe any variations in the spread of heights. Additionally, the presence of outliers, if any, can be identified, which may provide insights into extreme height values that could potentially impact cardiovascular disease risk. In Figure 3, we present the distribution plots for height and weight variables, stratified by CVD and non-CVD cases. The x-axis represents the height or weight values, while the y-axis indicates the density or frequency of occurrence. For the height variable, we observe two overlapping distributions representing CVD and non-CVD cases. Through comparing the shapes and central tendencies of these distributions, we can identify any notable differences.

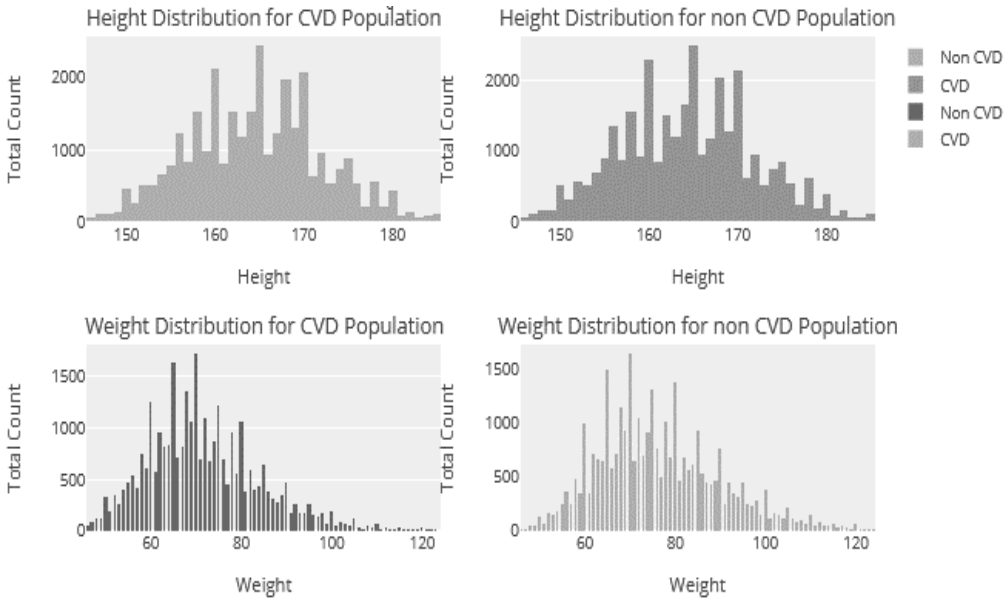

Figure 3: Distribution of Height and Weight Variables for CVD and non-CVD Cases

Moreover, In Figure 4, we display pie charts representing the distribution of categorical variables in our dataset. Each pie chart corresponds to a specific categorical variable, and the slices within the chart represent the different categories or levels of that variable The cross-validation experiments involve partitioning the dataset into multiple subsets or folds, training the data mining techniques on a subset, and evaluating their performance on the remaining data.

Doi: https://doi.org/10.54216/JAIM.040101
Received: August 19, 2022 Revised: January 15, 2023 Accepted: June 02, 2023

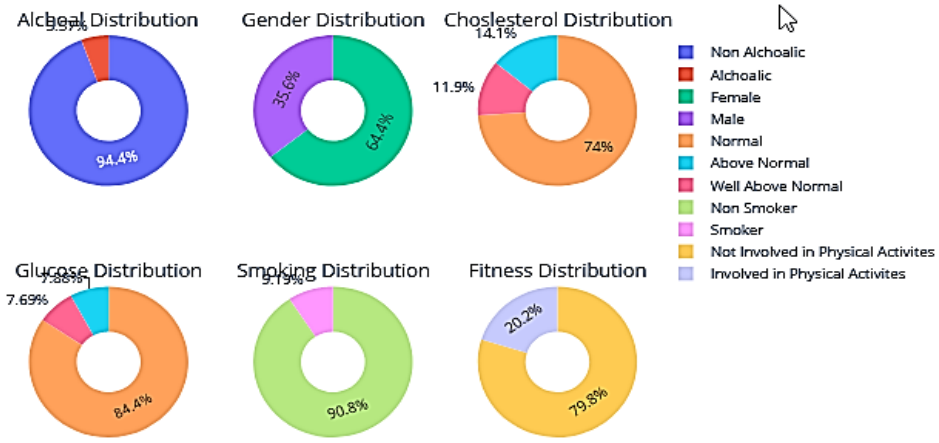

Figure 4: Distribution of Categorical Variables

In Figure 5, we display the performance metrics of the data mining techniques. The y-axis represents the accuracy, while the x-axis represents the different techniques compared in the experiments. Each technique is represented by a bar or line in the plot, indicating its performance metric value. By examining the plot, we can assess the relative performance of the techniques and identify which ones yield the most accurate and effective results in detecting cardiovascular disease risk causes among diabetes patients.

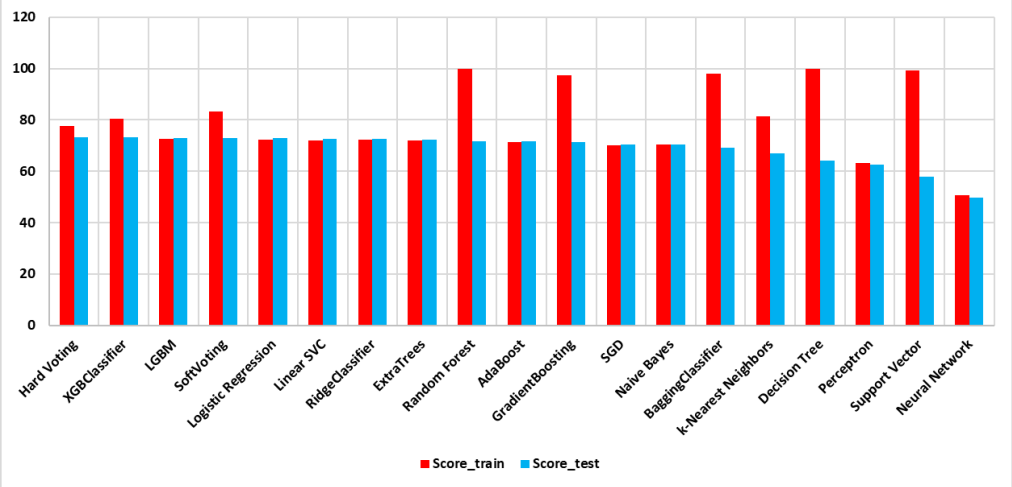

Figure 5: Performance Comparison of Data Mining Techniques through Cross-Validation Experiments

The results displayed in Figure 5 provide insights into the strengths and weaknesses of each data mining technique in relation to our specific problem. We can compare the performance metrics across the techniques to determine which ones exhibit superior performance in terms of accuracy, precision, recall, or any other relevant evaluation criteria.

6. Conclusion

This research introduces an innovative ontological data mining strategy for identifying the determinants of CVD risk in diabetic individuals. We have improved our understanding of the complicated interplay between risk variables and CVD outcomes by including domain knowledge and semantic links into the data mining process. Our results show that decision trees, SVMs, and gradient boosting are useful tools for elucidating the mechanisms behind CVD through the discovery of

Doi: https://doi.org/10.54216/JAIM.040101
Received: August 19, 2022 Revised: January 15, 2023 Accepted: June 02, 2023

relevant correlations. Our predictive models have improved in both interpretability and accuracy since we began applying ontological concepts. Not only have we enhanced the predictive performance by including expert-defined rules, limitations, and semantic linkages, but we have also gotten a more nuanced understanding of the causal elements contributing to CVD among diabetic patients. Box plots, distributional plots, and pie charts were used in our study to visualise data, revealing important insights into the distribution patterns of critical factors and their association with CVD outcomes.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] Jayaraman P. P., Forkan A. R. M., Morshed A., Haghighi P. D., Kang Y. B., Healthcare 4.0: A review of frontiers in digital health. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(2), 2020.

[2] Rahman M. F., Wen Y., Xu H., Tseng T. L. B., Akundi S., Data mining in telemedicine, Advances in Telemedicine for Health Monitoring: Technologies, Design and Applications, 103-131, 2020.

[3] Aseel M. Alfaisal, Aisha Zare, Afrah Alshaafi, Rose Aljanada, Raghad M. Alfaisal, Ghadeer W. Abukhalil, Predicting the actual use of social media sites among university communicators: using PLS-SEM and ML approaches. Journal of International Journal of Advances in Applied Computational Intelligence, 1(1), 23-33, 2022.

[4] Mohamed Saber, Efficient phase recovery system, IJEECS, 5(1), 2017.

[5] Raghad M. Alfaisal, Aisha Zare, Aseel M. Alfaisal, Rose Aljanada, Ghadeer Wael Abukhalil, The acceptance of metaverse system: a hybrid SEM-ML approach. Journal of International Journal of Advances in Applied Computational Intelligence, 1(1), 34-44, 2022.

[6] Heba R. Abdelhady, Mahmoud M. Ismail, Cardiovascular Diseases Forecasting using Machine Learning Models. Journal of International Journal of Advances in Applied Computational Intelligence, 1(2), 56-62, 2022.

[7] Rose Aljanada, Ghadeer W. Abukhalil, Aseel M. Alfaisal, Raghad M. Alfaisal, Adoption of Google Glass technology: PLS-SEM and machine learning analysis. Journal of International Journal of Advances in Applied Computational Intelligence, 1 (1), 08-22, 2022.

[8] Silverman E. K., Schmidt H. H., Anastasiadou E., Altucci L., Angelini M., Badimon L., Baumbach J., Molecular networks in Network Medicine: Development and applications. Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 12(6), 2020.

[9] Hani D. Hejazi, Ahmed A. Khamees, Employees Motivational Factors toward Knowledge Sharing: A Systematic Review, Journal of International Journal of Advances in Applied Computational Intelligence, 1(1), 45-68, 2022.

[10] Ahmed Z., Mohamed K., Zeeshan S., Dong, X., Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database, 2020.

[11] Aseel M. Alfaisal, Aisha Zare, Afrah Alshaafi, Rose Aljanada, Raghad M. Alfaisal, Ghadeer W. Abukhalil, Predicting the actual use of social media sites among university communicators: using PLS-SEM and ML approaches. Journal of International Journal of Advances in Applied Computational Intelligence, 1 (1), 23-33, 2022.

[12] Sulaiman M. A., Evaluating data mining classification methods performance in Internet of things applications. Journal of Soft Computing and Data Mining, 1(2), 11-25, 2020.

[13] Ye M. (2011). Text mining for building a biomedical knowledge base on diseases, risk factors, and symptoms. Germany: Max-Planck-Institute for Informatics.

[14] Abdulla Alsharhan, Natural Language Generation and Creative Writing A Systematic Review. Journal of International Journal of Advances in Applied Computational Intelligence, 1(1) , 69-90, 2022.

Doi: https://doi.org/10.54216/JAIM.040101
Received: August 19, 2022 Revised: January 15, 2023 Accepted: June 02, 2023

[15] Mohamed Saber, A novel design and implementation of FBMC transceiver for low power applications. IJEEI, 8(1), 83-93, 2020.

[16] Raghad M. Alfaisal,Aisha Zare,Aseel M. Alfaisal,Rose Aljanada,Ghadeer Wael Abukhalil, The acceptance of metaverse system: a hybrid SEM-ML approach. Journal of International Journal of Advances in Applied Computational Intelligence, 1 (1), 34-44, 2022.

[17] Wu X., Duan, J., Pan Y., Li M., Medical knowledge graph: Data sources, construction, reasoning, and applications. Big Data Mining and Analytics, 6(2), 201-217, 2023.

[18] Groza T., Köhler S., Moldenhauer D., Vasilevsky N., Baynam G., Zemojtel T., Robinson P. N., The human phenotype ontology: semantic unification of common and rare disease. The American Journal of Human Genetics, 97(1), 111-124, 2015.

[19] Ismail Eyad Samara, Intelligent systems and AI techniques: Recent advances and Future directions. Journal of International Journal of Advances in Applied Computational Intelligence, 1(2) , 30-45, 2022.

[20] Hani D. Hejazi , Ahmed A. Khamees, Employees Motivational Factors toward Knowledge Sharing: A Systematic Review. Journal of International Journal of Advances in Applied Computational Intelligence, 1 (1), 45-68, 2022.

[21] Alber S. Aziz, Hoda K. Mohamed, Ahmed Abdelhafeez, Unveiling the Power of Convolutional Networks: Applied Computational Intelligence for Arrhythmia Detection from ECG Signals." Journal of International Journal of Advances in Applied Computational Intelligence, 1(2), 63-72, 2022.

Doi: https://doi.org/10.54216/JAIM.040101
Received: August 19, 2022 Revised: January 15, 2023 Accepted: June 02, 2023