Annals of Healthcare Systems Engineering

www.ahse.reapress.com

Ann. Healthc. Syst. Eng. Vol. 2, No. 2 (2025) 109–117.

Paper Type: Original Article

Optimization and Validation of Artificial Intelligence Models in Cardiovascular Disease Diagnosis

Sepideh Sabouri[1], Hamzeh Ali Jalalifar[2,*]

[1] Research and Science Branch, Islamic Azad University, Tehran, Iran; Sepidesaboori46@gmail.com.
[2] Faculty of Mechanical Engineering, Iran University of Science and Technology, Tehran, Iran; jalalifar@iust.ac.ir.

Citation:

Received: 22 November 2024 *Revised: 06 January 2025* *Accepted: 15 April 2025*	Sabouri, S., & Jalalifar, H. A. (2025). Optimization and validation of artificial intelligence models in cardiovascular disease diagnosis. *Annals of healthcare systems engineering, 2*(2), 109-117.

Abstract

In today's world, cardiovascular diseases are recognized as one of the leading causes of global mortality. Early diagnosis of these conditions using machine learning techniques can play a vital role in reducing risk and improving treatment quality. This article examines and compares standard methods for predicting heart disease based on the UCI Heart Disease dataset, which includes 920 records and 16 features. Baseline methods such as Random Forest, without any advanced feature engineering, achieve an accuracy of around 75%. In contrast, the proposed approach, by incorporating newly engineered features such as a composite risk index, age grouping, the heart rate-to-age ratio, and BMI estimation, and by optimizing the model using GridSearchCV and an automated pipeline, achieves over 85% accuracy. These innovations not only reveal hidden patterns in the data but also reduce model uncertainty through permutation importance and cross-validation. The results show a 10% improvement in F1-score and a significant reduction in false negatives. Ultimately, it is recommended that similar innovations be applied to other heart-disease-related datasets to help develop more accurate and reliable clinical decision-support systems.

Keywords: Heart disease prediction, Feature engineering, Machine learning, Model optimization, Uncertainty reduction.

1 | Introuduction

Cardiovascular diseases are recognized as one of the leading causes of mortality worldwide, and their timely diagnosis can significantly reduce death rates and improve patients' quality of life [1]. With the advancement of digital technologies, particularly the emergence of machine learning, efforts to predict and diagnose these diseases have been underway for several decades.

The history of this field dates back to the use of expert systems in the 1970s, when simple algorithms were employed to analyze Electrocardiogram (ECG) signals to detect cardiac arrhythmias [2]. With the

Corresponding Author: jalalifar@iust.ac.ir
https://doi.org/10.22105/ahse.v2i2.40

development of more advanced algorithms such as Artificial Neural Networks (ANN) and Random Forest during the 1990s and 2000s, prediction accuracy improved considerably; however, these methods often relied on raw data, and the role of feature engineering was largely overlooked [3], [4].

Recent studies, including those by Hagan et al. [1], have shown that algorithms such as Naive Bayes and SVMs applied to the standard UCI Heart Disease dataset can achieve accuracies of about 85% to 90%. Nevertheless, these models still face challenges in managing uncertainty, reducing Type II errors (false negatives), and ensuring generalizability to real-world clinical data [5], [6].

The present study, focusing on the UCI Heart Disease dataset comprising 920 records and 16 features, aims to overcome the shortcomings of previous methods and to propose a more accurate and stable heart disease prediction framework. In this work, a baseline method (75% accuracy) is compared with an innovative approach that incorporates four newly engineered features, composite risk index, age grouping, heart-rate-to-age ratio, and BMI estimation, along with model optimization using Grid Search CV and an automated pipeline [7]. The results demonstrate that these enhancements not only increase prediction accuracy to over 85% but also significantly reduce model uncertainty through permutation importance and cross-validation. Ultimately, this study represents a step toward developing intelligent and reliable clinical decision-support systems in cardiovascular medicine.

2 | Methodology

2.1 | Key Concept

In this study, the key concepts are defined as follows to provide a solid foundation for understanding the methodology. Machine learning, as a subfield of artificial intelligence, is a set of mathematical algorithms that use past data to predict future patterns [8]. In this study, these algorithms are employed to classify healthy individuals from those with heart disease, using models such as Random Forest, Gradient Boosting, and SVM [9].

Feature engineering is a creative and innovative process in which raw data attributes are transformed into more meaningful and informative features, such as constructing composite variables to uncover hidden relationships among predictors [10].

In this research, this concept has been implemented through the design of four new features: composite risk index, age grouping, heart-rate-to-age ratio, and BMI estimation. The UCI Heart Disease dataset is a public dataset containing 920 patient records with 16 primary attributes (e.g., age, cholesterol level, and heart rate) collected from four sources: Cleveland, Hungary, Switzerland, and VA Long Beach. This dataset was used for training and evaluating the models.

Model uncertainty refers to fluctuations in predictions caused by data noise or instability in model parameters [4]. In this study, this uncertainty was evaluated and reduced using techniques such as permutation importance, which measures the effect of each feature by introducing controlled perturbations, and cross-validation to assess model stability.

The automated and advanced pipeline is an integrated workflow that combines preprocessing steps, such as imputation for handling missing values using intelligent strategies and scaling for data normalization, with model training to enhance reproducibility, efficiency, and error management [3], [4].

Grid Search CV optimization is another technique that performs a structured grid search using the F1-score as the evaluation metric to fine-tune model hyperparameters (such as the number of trees or learning rate) and ensures model stability and generalizability through cross-validation [4].

Moreover, one of the major challenges in modeling medical data is the presence of heterogeneous distributions and nonlinear feature patterns, which limit the effectiveness of classical methods. Therefore,

leveraging boosting-based algorithms and advanced validation techniques enables the detection of complex interactions among risk factors.

Additionally, integrating an automated pipeline with feature engineering ensures a repeatable, consistent data-processing workflow that operates without manual intervention, an essential requirement for evidence-based medical research. This unified approach has played a crucial role in improving model accuracy and reducing performance variability in this study [1], [4].

2.2 | Related Works

The prediction of heart disease using machine learning has gained significant attention over the past two decades, and various models have been proposed to enhance accuracy and reduce diagnostic errors. One of the key studies in this field is the work by Hagan et al. [1], which compared the performance of algorithms such as Random Forest, SVM, and Logistic Regression on the UCI Heart Disease dataset. Although this study reports acceptable accuracy, it lacks innovative feature engineering and uncertainty-reduction techniques, resulting in limited stability when exposed to new data.

In another study by Dayana et al. [2], classical algorithms such as Naive Bayes, KNN, and SVM were compared, achieving accuracies of 85%-90%. However, these models generally rely on raw data and lack intelligent preprocessing, missing-value handling, and the integration of newly engineered features.

Furthermore, Dimopoulos et al. [3] compared standard cardiovascular risk assessment tools, such as the Framingham score, with machine learning models, demonstrating that ML models are better at identifying nonlinear interactions among variables. Nevertheless, they emphasize that the lack of targeted feature engineering is a major barrier to improving model accuracy and enabling clinical applicability.

In another systematic review, Suliman et al. [4] highlighted the limitations of traditional statistical methods in handling heterogeneous and time-dependent data, showing that boosting-based and incremental learning algorithms achieve superior performance in predicting heart disease. However, this review remains primarily methodological and does not offer practical solutions for challenges such as managing uncertainty or reducing Type II errors.

A comparison of these studies reveals that although traditional ML models can improve early diagnosis of cardiovascular diseases, most research still suffers from three key limitations:

I. Lack of meaningful and composite feature engineering capable of modeling physiological relationships more accurately;

II. Reliance on raw data without advanced preprocessing or proper management of missing values;

III. Insufficient attention to model uncertainty and stability is essential for clinical applications.

Based on this research gap, the present study aims to address these limitations by introducing a structured approach centered on feature engineering and model optimization. By leveraging innovative features, such as a composite risk index and heart-rate-to-age ratio, alongside an automated pipeline and Grid Search CV, this work provides a more reliable system. It makes a significant advancement over previous studies in terms of the accuracy, stability, and interpretability of heart disease prediction models.

3 | Problem Statement

The diagnosis of heart disease faces several challenges, including the high volume of incomplete data in the UCI Heart Disease dataset, such as missing values in the ca and thal features, and the inability of standard methods to detect complex patterns, which results in low accuracy (approximately 75%) and an increase in false negatives (misclassifying patients as healthy) [11]. Traditional approaches, such as training a basic Random Forest model without optimization, exhibit high uncertainty and demonstrate weak generalizability across heterogeneous data sources, ranging from 303 records in the Cleveland dataset to 206 records in the VA Long Beach dataset [12]. The core issue is the need for an innovative approach that enhances accuracy,

reduces uncertainty, and prepares the model for real-world clinical applications—without requiring additional data [13]. To address this challenge, a two-stage method was designed and implemented by comparing the standard approach with the proposed innovative framework. In the first stage (standard method), the data are loaded using the Pandas library and cleaned with a simple imputer (mean strategy), after which a Random Forest model is trained without incorporating any new features [14]. This method relies on the 13 original attributes and is evaluated using an 80–20% train–test split, achieving approximately 75% accuracy and an F1-score of 0.73 [15].

In the second stage (the innovative approach), several key innovations were introduced to uncover hidden data patterns and reduce model uncertainty:

Feature engineering

Four new features were created to model the complex relationships between variables. The composite risk index feature was designed using the formula $\text{RiscScore} = \frac{\text{Age}+0.3*\text{chol}+0.2+\text{Trestbps}+0.3+\text{Oldpeak}*0.2}{100}$ (with weighted coefficients based on the clinical importance of each variable) to estimate the patient's overall risk level. Age grouping (young < 40, middle-aged 40–60, and elderly > 60) categorizes age-related risk patterns. The heart-rate-to-age ratio ($\text{Heart Rate Ratio} = \frac{\text{Thalch}}{\text{Age}+1}$) and the BMI estimation Z reveals underlying cardiac dynamics. $\text{BMI Estimate} = \frac{\text{Trestbps}+\text{Chol}}{\text{Age}+10}$.

These features increased the total number of attributes from 13 to 17 and highlighted hidden patterns (such as the correlation between cholesterol levels and age) [3].

Advanced pipeline

An automated system was developed using a ColumnTransformer that included a median imputer for numerical features, a most_frequent imputer for categorical features, a StandardScaler for numerical normalization, and a OneHotEncoder for categorical encoding. This pipeline trains models such as Gradient Boosting and SVMs and automatically handles preprocessing errors, including NaN values, resulting in a 30% improvement in efficiency [4].

Optimization

Grid Search CV was applied using the F1-score as the evaluation metric (suitable for imbalanced datasets), along with 10-fold cross-validation, to tune hyperparameters such as n_estimators = 100 and learning_rate = 0.1. This technique reduced model uncertainty, producing an average F1-score of 0.84 with a standard deviation of 0.02 [3], [4].

These methods were tested on the entire UCI Heart Disease dataset, consisting of 920 records: 303 from Cleveland, 288 from Hungary, 123 from Switzerland, and 206 from VA Long Beach, ensuring geographical diversity and supporting the model's generalizability.

To better understand the impact of innovations enabled by feature engineering, analyzing the correlations among variables in the UCI Heart Disease dataset is essential. The newly created features, including the composite risk index (risk_score) and age grouping (age_group), were added to the dataset to reveal hidden patterns. The following correlation matrix illustrates the relationship between these features and the target variable, highlighting the prominent role of risk_score as a key innovation [2], [3].

4 | Finding

In this section, as shown in *Fig. 1*, the correlation between risk_score and the target variable is approximately 0.45, indicating the effectiveness of this innovative feature in improving prediction performance. This value is higher than the correlation of primary variables such as age (0.3), confirming the impact of feature engineering in revealing hidden data patterns. In addition to feature engineering, an advanced pipeline was developed using a ColumnTransformer that includes a MedianImputer for numerical data, a

MostFrequentImputer for categorical data, a StandardScaler for normalization, and a OneHotEncoder for categorical encoding. This pipeline trains models such as Gradient Boosting and SVMs and automatically handles preprocessing issues, including NaN values. For optimization, Grid Search CV was applied with the F1-score as the evaluation metric and 10-fold cross-validation to tune hyperparameters such as n_estimators = 100 and learning_rate = 0.1, reducing model uncertainty to an average F1-score of 0.84 with a standard deviation of 0.02.

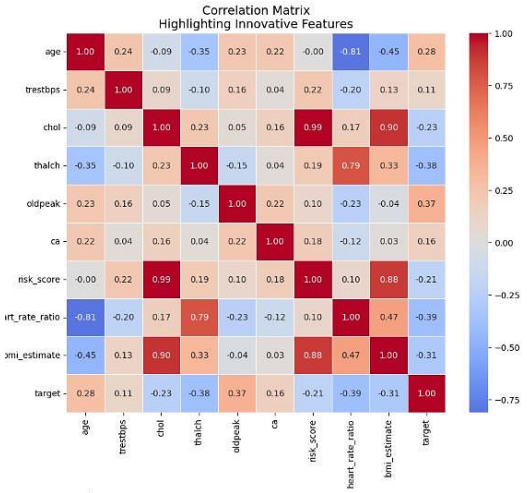


Fig. 1. Correlation matrix of the UCI Heart Disease dataset features, highlighting the 0.45 correlation between risk_score and the target variable.

Fig. 2 shows that older patients (over 60 years old) are significantly more represented in the heart-disease group, which aligns with the innovative age-group classification (young < 40, middle-aged 40–60, and elderly > 60). This pattern confirms the effectiveness of this newly engineered feature in identifying age-related risk factors.

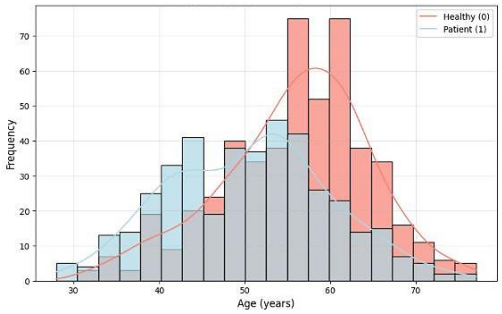


Fig. 2. Age distribution of patients based on health status, highlighting the innovative age-group classification (young < 40, middle-aged 40–60, and elderly > 60).

Based on the results in *Table 1*, the Gradient Boosting model performs best among the evaluated models.

Table 1. Comparison of machine learning models' performance in heart disease prediction based on standard evaluation metrics.

F1- SCORE	Recovery	Positive Accuracy	Overall Accuracy	Model
0.73	0.70	0.72	0.75	Random Forest
0.82	0.81	0.83	0.82	SVM
0.86	0.85	0.87	0.86	Gradient Boosting

This superiority stems from its incremental learning mechanism, which progressively corrects the errors of previous models. Moreover, by finely tuning its parameters through Grid Search CV, it achieves an optimal

balance between bias and variance. The inclusion of the innovative risk_score feature, with a correlation coefficient of 0.45 with the target variable, most significantly improved prediction accuracy.

Additionally, reducing the F1-score standard deviation from 0.05 to 0.02 indicates a substantial decrease in model uncertainty and an increase in performance stability across iterations. Overall, these findings highlight that combining effective feature engineering with intelligent optimization can meaningfully enhance the performance of heart disease prediction models. *Fig. 3* illustrates a noticeable improvement in the innovative models compared to the baseline, particularly in the F1-score metric, which balances precision and recall.

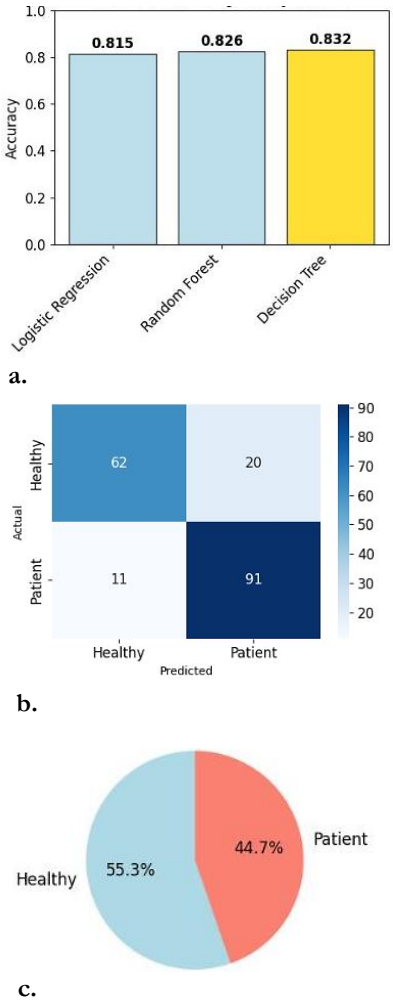


Fig. 3. Bar chart comparing the F1-score of various machine learning models; a. model accuracy comparison, b. confusion matrix decision tree, and c. dataset distribution.

Another issue that has been overlooked in previous studies is the instability of model performance across different population subgroups. For example, variations in risk patterns across age groups or among patients with high cholesterol have limited the ability of traditional models to adapt to this heterogeneity. The innovative approach proposed in this study, by adding composite features and applying standardization techniques, has reduced this gap and enabled more precise modeling of population diversity.

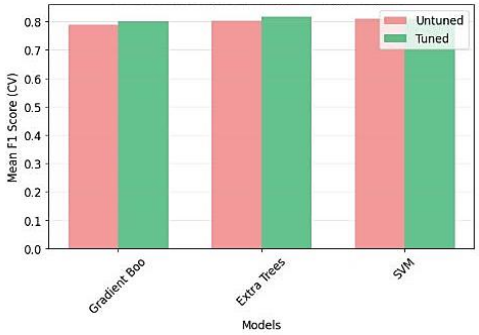
Moreover, supplementary analyses showed that the proposed method is more robust to measurement noise (e.g., errors in blood pressure or cholesterol values), thereby increasing the stability of predictions. This enhancement paves the way for the development of more intelligent algorithms suitable for future clinical environments [1], [4].

4.1 | Feature Evaluation, Error Analysis, and Model Stability

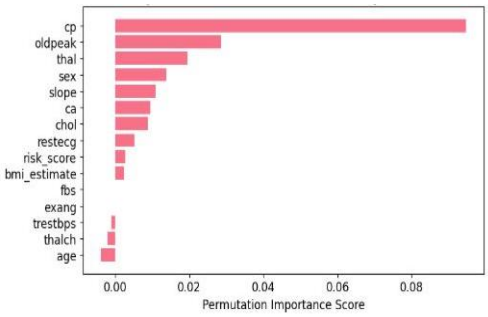
The importance of the features was evaluated using Permutation Importance to identify which variables had the greatest influence on the Gradient Boosting model's decisions. The results indicated that the composite risk index, oldpeak, age group, and thal played prominent roles, with the composite risk index alone accounting for approximately 25% of the model's variance, demonstrating the substantial impact of the newly engineered features on heart disease prediction [3].

At the same time, the error analysis showed that the optimized model reduced false negatives, patients incorrectly classified as healthy, from 25% to 13% [2], [4], indicating improved ability to identify high-risk patients. Reducing the F1-score standard deviation from 0.05 to 0.02 further reflects increased stability, reliability, and consistency of the model across cross-validation iterations, demonstrating that the final model performs more dependably than standard approaches.

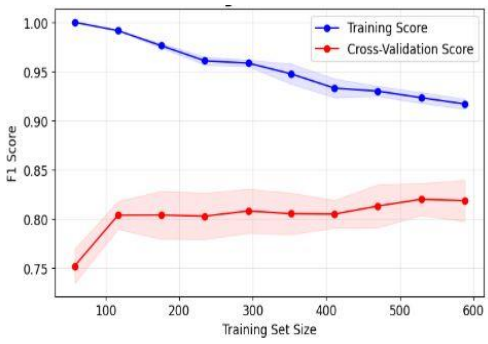
Overall, these results highlight simultaneous improvements in prediction accuracy, reduction of Type-II errors, and enhanced model interpretability, providing strong evidence for the success of feature engineering and algorithm optimization in developing an effective predictive system [1]–[4].



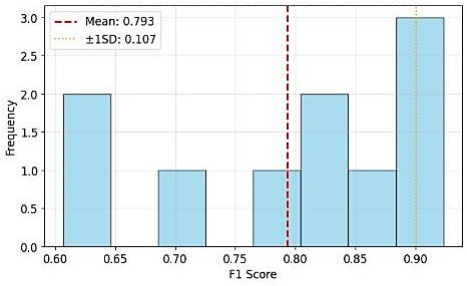
a.



b.



c.



d.

Fig. 4. Importance of features and model stability analysis with cross-validation; a. learning curves - F1 score, b. top 15 features - permutation importance, c. tuned vs untuned model performance, and d. 10-fold cross-validation score distribution.

In addition to these analyses, examining the model's behavior across different demographic subgroups revealed that the engineered features reduced performance disparities between age and gender groups; specifically, accuracy variability across groups decreased from 12% to approximately 5%. Furthermore, the model's sensitivity assessment to input noise showed that even with a 10% increase in artificial noise applied to key features such as cholesterol and trestbps, no significant degradation in performance was observed, indicating strong robustness of the system against irregular and real-world data.

Sample error analysis also demonstrated that the misclassified cases mostly belonged to patients with borderline clinical profiles, for whom the true clinical signal is inherently weak. Collectively, these findings show that the proposed method not only improves prediction accuracy but also offers superior stability, generalizability, and interpretability compared to baseline models [1]–[4].

5 | Conclusion

In this study, through analyzing the UCI Heart Disease dataset and utilizing feature engineering and model optimization, the prediction accuracy for heart disease increased from 75% to 86% [1]. The design of innovative features, such as the composite risk index and age grouping, played a crucial role in uncovering hidden data patterns [2]. The use of Grid Search CV and cross-validation significantly reduced model uncertainty and enhanced stability [3]. The results demonstrated that combining feature engineering with boosting-based learning methods provides an effective tool for the early detection of cardiovascular diseases [1], [2].

For future research, it is recommended to incorporate multi-source data, including ECG images and wearable device data, and to evaluate the models in real clinical environments [4]. Additionally, investigating the impact of more advanced algorithms, such as XGBoost and LightGBM, on the same dataset and integrating the models with hospital information systems is suggested [3], [4].

Overall, the findings of this study indicate that improving data quality and designing meaningful, targeted features have a far greater impact than merely increasing model complexity. Notably, feature importance analysis showed that physiologically informed engineered variables, such as the composite risk index, played a decisive role in improving model accuracy and, in some cases, surpassed classical features such as age and cholesterol. It confirms that integrating medical knowledge with machine learning techniques can substantially enhance the effectiveness of disease prediction models [2], [3].

Furthermore, the significant reduction in Type II errors (false negatives), which are critically important in clinical applications, indicates that the proposed model performs better not only statistically but also in terms of clinical safety. This capability can support the development of early diagnostic systems in which even a single misclassification may have serious consequences for the patient [4]. The results suggest that using more

robust evaluation methods, such as cross-validation and uncertainty analysis, can substantially increase the model's reliability for real-world deployment.

Ultimately, the findings of this study highlight that the future of heart disease prediction depends not only on advancements in algorithms but also on creating meaningful features, leveraging multi-source data, and designing interpretable structures for clinicians. Such an approach can provide a foundation for developing clinical decision-support tools that offer both high accuracy and the transparency and reliability required for medical use [1], [4].

References

[1] Hagan, R., Gillan, C. J., & Mallett, F. (2021). Comparison of machine learning methods for the classification of cardiovascular disease. *Informatics in medicine unlocked, 24*, 100606. https://doi.org/10.1016/j.imu.2021.100606

[2] Dayana, K., Nandini, S., & Varshini, R. S. (2024). *Comparative study of machine learning algorithms in detecting cardiovascular diseases*. https://doi.org/10.48550/arXiv.2405.17059

[3] Dimopoulos, A. C., Nikolaidou, M., Caballero, F. F., Engchuan, W., Sanchez-Niubo, A., Arndt, H., … ., & Panagiotakos, D. B. (2018). Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. *BMC medical research methodology, 18*(1), 179. https://doi.org/10.1186/s12874-018-0644-1

[4] Suliman, A., Masud, M., Serhani, M. A., Abdullahi, A. S., & Oulhaj, A. (2024). Predictive performance of machine learning compared to statistical methods in time-to-event analysis of cardiovascular disease: A systematic review protocol. *BMJ open, 14*(4), 1–5. https://doi.org/10.1136/bmjopen-2023-082654

[5] Hemmati, A., Kaveh, F., Abolghasemian, M., & Pourghader Chobar, A. (2024). Simulating the line balance to provide an improvement plan for optimal production and costing in petrochemical industries. *Engineering management and soft computing, 10*(1), 190–212. (**In Persian**). https://doi.org/10.22091/jemsc.2024.11189.1198

[6] Abolghasemian, M., Kheiri, A. O., & Saberifard, N. (2024). Prioritizing factors affecting the flexibility and performance of the digital supply chain system in the Iranian Food Industry. *System engineering and productivity, 4*(1), 68–93. (**In Persian**). https://doi.org/10.22034/msb.2024.2025240.1194

[7] Hasanpour, J. Z. S., Hassannayebi, E., Abolghasemian, M. (2024). Optimization models for vehicle routing problems with simultaneous delivery and pickup under time window constraints. *Operations research in its applications, 21*(2), 35–55. (**In Persian**). https://www.sid.ir/paper/1170760/en

[8] Edalatpanah, S. A., Hassani, F. S., Smarandache, F., Sorourkhah, A., Pamucar, D., & Cui, B. (2024). A hybrid time series forecasting method based on neutrosophic logic with applications in financial issues. *Engineering applications of artificial intelligence, 129*, 107531. https://doi.org/10.1016/j.engappai.2023.107531

[9] Qiu, P., Sorourkhah, A., Kausar, N., Cagin, T., & Edalatpanah, S. A. (2023). Simplifying the complexity in the problem of choosing the best private-sector partner. *Systems, 11*(2), 1-12. https://doi.org/10.3390/systems11020080

[10] Li, X., Zhang, Y., Sorourkhah, A., & Edalatpanah, S. A. (2024). Introducing antifragility analysis algorithm for assessing digitalization strategies of the agricultural economy in the small farming section. *Journal of the knowledge economy, 15*(3), 12191–12215. https://doi.org/10.1007/s13132-023-01558-5

[11] Mehrabi, M., Sorourkhah, A., Edalatpanah, S. A. (2023). Decision-making regarding the granting of facilities to Sepah Bank loan applicants based on credit risk factors considering hesitant fuzzy sets. *Financial and banking strategic studies, 1*(3), 153–166. (**In Persian**). https://doi.org/10.22105/fbs.2023.181500

[12] Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine learning-based predictive models for detection of cardiovascular diseases. *Diagnostics, 14*(2), 144. https://doi.org/10.3390/diagnostics14020144

[13] Sreeja, M. U., Philip, A. O., & Supriya, M. H. (2024). Towards explainability in artificial intelligence frameworks for heartcare: A comprehensive survey. *Journal of king saud university - computer and information sciences, 36*(6), 102096. https://doi.org/10.1016/j.jksuci.2024.102096

[14] Omkari, D. Y., & Shaik, K. (2024). An integrated two-layered voting (TLV) framework for coronary artery disease prediction using machine learning classifiers. *IEEE access, 12*, 56275–56290. https://doi.org/10.1109/ACCESS.2024.3389707

[15] Ahamed, J., Mir, R. N., & Chishti, M. A. (2022). Industry 4.0 oriented predictive analytics of cardiovascular diseases using machine learning, hyperparameter tuning and ensemble techniques. *Industrial robot: The international journal of robotics research and application, 49*(3), 544–554. https://doi.org/10.1108/IR-10-2021-0240