ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

AN EXPLAINABLE CLINICAL VARIANT RISK ASSESSMENT FRAMEWORK FOR GENOMIC DECISION SUPPORT

[1]Fahmi Izhari✉, [2]Hanna Willa Dhany, [1]Syarif Hidayat Matondang, [1]Rizki Khairani Nasution
[1]UIN Syekh Ali Hasan Ahmad Addary, Padangsidimpuan, Indonesia
[2]Universitas Pembangunan Panca Budi, Medan, Indonesia
Email: fahmi_izhari@uinsyahada.ac.id

DOI: https://doi.org/10.46880/jmika.Vol10No1.pp350-356

ABSTRACT

Clinical variant interpretation is essential for precision medicine; however, conventional machine learning approaches often focus on prediction accuracy without providing sufficient interpretability and decision-support capabilities. This study proposes a hybrid framework integrating Light Gradient Boosting Machine (LightGBM), SHapley Additive exPlanations (SHAP), and a Fuzzy Decision Support System (FDSS) for clinical variant risk assessment using the ClinVar dataset. A stratified sample of 200,000 genetic variants was utilized for model development and evaluation. LightGBM was employed to predict variant pathogenicity, while SHAP was applied to identify feature contributions and improve model transparency. The resulting prediction probabilities were subsequently processed through fuzzy inference to generate interpretable risk categories and recommendation-oriented outputs. Experimental results showed that the proposed framework achieved an Accuracy of 95.89%, Precision of 95.58%, Recall of 82.97%, F1-Score of 88.83%, and ROC-AUC of 98.73%. Explainability analysis revealed that variant-type representation was the most influential predictor of pathogenicity. The proposed framework extends conventional classification by transforming predictive outputs into actionable risk assessments, thereby enhancing transparency and supporting informed genomic decision-making. These findings demonstrate the potential of integrating predictive analytics, explainable artificial intelligence, and fuzzy reasoning for clinical variant assessment in precision medicine.

Keyword: Clinical Variant Classification, LightGBM, SHAP, Fuzzy Decision Support System, Precision Medicine.

INTRODUCTION

The rapid advancement of next-generation sequencing (NGS) technologies has significantly increased the availability of genomic data for clinical and biomedical applications (Satam et al., 2023). Public genomic repositories such as ClinVar provide extensive information regarding genetic variants and their clinical significance, supporting the development of precision medicine and personalized healthcare strategies (Méndez-Vidal et al., 2025). However, the continuous growth of genomic databases has created substantial challenges in variant interpretation, particularly in distinguishing pathogenic variants from benign variants accurately and efficiently (Barbitoff et al., 2024).

The classification of genetic variants is a critical task in clinical decision-making because inaccurate interpretations may affect diagnosis, treatment planning, and genetic counseling. Traditional manual assessment performed by experts is often time-consuming and difficult to scale as the volume of genomic data continues to increase. Consequently, intelligent information systems capable of supporting automated variant interpretation have become increasingly important in healthcare environments (Amiri, 2024).

Machine learning techniques have demonstrated strong capabilities in analyzing large-scale biomedical datasets and identifying complex patterns that are difficult to detect through conventional approaches (Dhanka et al., 2026). Among various machine learning methods, gradient boosting algorithms have attracted considerable attention due to their predictive performance, computational efficiency, and scalability when processing large datasets (Alshboul et al., 2022). Nevertheless, many existing studies primarily focus on improving classification accuracy while providing limited support for interpretability and decision-making processes (Kruschel et al., 2026; Tursunalieva et al., 2024).

The lack of transparency in machine learning models remains a major challenge in healthcare applications. Clinical experts often require explanations regarding the factors influencing model

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

predictions before adopting artificial intelligence systems in practice. Explainable Artificial Intelligence (XAI) has therefore emerged as an important research area, with SHapley Additive exPlanations (SHAP) becoming one of the most widely used approaches for interpreting machine learning models and quantifying feature contributions (Makumbura et al., 2024).

Another limitation of current pathogenicity prediction systems is their dependence on binary classification outputs. In practical clinical settings, decision-makers frequently require risk-oriented assessments rather than simple categorical predictions. Decision Support Systems (DSS) combined with fuzzy logic have been widely recognized as effective tools for handling uncertainty and generating interpretable recommendations in complex decision environments (Kostopoulos et al., 2024; Srivastava et al., 2025).

Despite recent advances in machine learning-based genomic analysis, limited studies have integrated predictive modeling, explainability, and decision-support mechanisms within a unified framework. Most existing approaches focus on classification performance while overlooking the need for transparent and actionable decision support. This gap highlights the necessity of developing an intelligent framework capable of transforming predictive outputs into meaningful risk assessments.

Therefore, this study proposes a hybrid framework that integrates LightGBM, SHapley Additive exPlanations (SHAP), a Pathogenicity Index (PI), and a Fuzzy Decision Support System (FDSS) for clinical variant assessment using large-scale ClinVar data. The proposed framework aims not only to improve pathogenicity prediction but also to enhance interpretability and support risk-oriented decision-making through the integration of machine learning, explainable artificial intelligence, and fuzzy reasoning.

The contributions of this study are threefold. First, it develops a scalable machine learning model for large-scale genomic variant classification. Second, it introduces a Pathogenicity Index as an interpretable indicator for evaluating variant-associated risk. Third, it establishes a fuzzy decision-support mechanism that converts predictive outputs into meaningful risk categories and recommendation-oriented assessments, thereby supporting more transparent and effective decision-making in precision medicine.

RELATED WORK

Machine Learning For Clinical Variant Classification

The rapid growth of genomic repositories has encouraged the development of machine learning approaches for clinical variant classification. Previous studies have employed various algorithms, including Support Vector Machines, Random Forests, Gradient Boosting Machines, and Deep Learning models, to distinguish pathogenic variants from benign variants (Bahmane et al., 2026; Divya et al., 2025; Izhari & Meiyanti, 2025). These approaches have demonstrated promising predictive performance and have significantly improved the efficiency of genomic data interpretation. Among them, gradient boosting methods have gained considerable attention due to their ability to handle high-dimensional data, capture nonlinear relationships, and maintain strong predictive performance across large-scale datasets (Qiuqian et al., 2025).

Recent studies have further shown that boosting-based algorithms are particularly effective in biomedical data analysis because of their robustness and scalability. Nevertheless, most existing studies focus primarily on optimizing classification accuracy and evaluation metrics. As a result, the generated predictions often remain difficult to interpret and provide limited support for practical decision-making in clinical environments.

Explainable Artificial Intelligence In Healthcare

The increasing adoption of machine learning in healthcare has raised concerns regarding transparency and trustworthiness. Clinical practitioners require not only accurate predictions but also understandable explanations regarding how predictive decisions are generated. Consequently, Explainable Artificial Intelligence (XAI) has emerged as an important research area aimed at improving transparency and accountability in intelligent systems.

Among various XAI approaches, SHapley Additive exPlanations (SHAP) has become one of the most widely adopted methods because it provides a theoretically grounded framework for quantifying feature contributions and explaining model behavior (Santos et al., 2024). Previous studies have demonstrated that SHAP can improve the interpretability of healthcare prediction models by identifying influential variables and revealing their impact on prediction outcomes. However, existing implementations of SHAP generally focus on post-hoc explanation and rarely extend explanatory information into decision-support mechanisms capable of generating actionable recommendations.

Fuzzy Decision Support Systems In Healthcare

Decision Support Systems (DSS) have been extensively applied in healthcare to assist diagnosis,

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

risk assessment, treatment planning, and resource allocation. Within DSS research, fuzzy logic has been recognized as an effective approach for handling uncertainty because it enables the representation of linguistic reasoning and gradual transitions between decision categories (Aregbesola et al., 2025). Unlike conventional binary systems, fuzzy inference systems can provide more flexible and interpretable assessments that better reflect real-world decision-making processes.

Several studies have integrated machine learning and fuzzy logic to improve decision-support capabilities. In these hybrid approaches, machine learning models are typically used to generate predictive outputs, while fuzzy inference systems transform those outputs into risk levels or recommendation-oriented results (Jiang et al., 2024). Although such frameworks have demonstrated advantages in terms of interpretability and usability, most existing studies focus on disease diagnosis, patient monitoring, or healthcare risk assessment rather than genomic variant interpretation using large-scale clinical repositories.

Research Gap and Research Position

Despite significant advances in machine learning, explainable artificial intelligence, and fuzzy decision-support systems, important limitations remain. Existing studies generally emphasize predictive performance, model interpretability, or decision-support capability as separate objectives. Machine learning approaches have demonstrated strong classification performance but often lack transparency, whereas explainable artificial intelligence methods improve model interpretability without directly supporting risk-oriented decision-making. Similarly, fuzzy decision-support systems provide interpretable recommendations but are rarely integrated with explainable genomic prediction models. Consequently, most pathogenicity prediction frameworks still generate classification outputs without transforming predictive information into actionable risk assessments. To address these limitations, this study proposes a hybrid framework that integrates LightGBM, SHAP, a Pathogenicity Index (PI), and a Fuzzy Decision Support System (FDSS) for clinical variant assessment using large-scale ClinVar data. Unlike previous studies that focus primarily on prediction or interpretation, the proposed framework combines classification, explainability, risk quantification, and recommendation-oriented assessment within a unified decision-support architecture. Through this integration, the framework aims to provide more transparent, interpretable, and actionable support for genomic decision-making in precision medicine.

RESEARCH METHOD

As illustrated in Figure 1, the proposed framework integrates machine learning, explainable artificial intelligence, and fuzzy decision support techniques for clinical variant risk assessment. The framework consists of four sequential stages: predictive modeling, explainability analysis, risk index construction, and decision support.

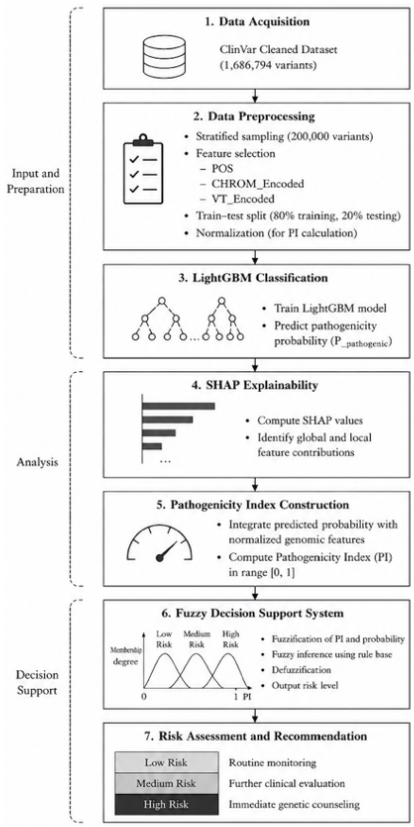

Fig 1. Clinical Variant Risk Assessment Methodology

The study utilized the ClinVar Cleaned Dataset containing 1,686,794 clinically annotated genetic variants. To improve computational efficiency while preserving the original class distribution, a stratified sample of 200,000 variants was selected for model development and evaluation. Three genomic attributes were employed as predictive features, namely genomic position (POS), chromosome representation (CHROM_Encoded), and variant-type representation (VT_Encoded), while pathogenicity status was used as the target variable.

The feature selection process was guided by data completeness, biological relevance, and interpretability considerations. Genomic position

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

represents the physical location of a variant within the genome, chromosome representation captures chromosomal context, and variant-type representation reflects mutation characteristics. These attributes were selected because they provide essential genomic information while maintaining model simplicity and interpretability.

The dataset was partitioned using a stratified training–testing scheme to preserve the original class distribution. Variant pathogenicity probabilities were subsequently estimated using a LightGBM classifier. To enhance model transparency, SHAP analysis was employed to quantify feature contributions and identify the relative importance of genomic attributes.

To provide a continuous representation of pathogenic risk, a Pathogenicity Index (PI) was constructed by integrating prediction probability with normalized genomic attributes:

$$PI = 0.60P + 0.25VT_{norm} + 0.15CHROM_{norm}$$

where P denotes the pathogenicity probability predicted by LightGBM, VT_{norm} represents the normalized variant-type feature, and $CHROM_{norm}$ denotes the normalized chromosome feature. The weighting scheme prioritizes model confidence while preserving relevant genomic information.

The resulting prediction probability and PI values were incorporated into a Fuzzy Decision Support System (FDSS). Probability, PI, and Risk Score were represented using Low, Medium, and High fuzzy sets. A Mamdani inference mechanism was employed to generate risk assessments, while centroid defuzzification was applied to obtain the final risk score. The fuzzy rule base used in the proposed framework is presented in table 1.

Table 1. Fuzzy Rule Base

Rule	Condition	Output
R1	Probability High AND PI High	High Risk
R2	Probability Medium OR PI Medium	Medium Risk
R3	Probability Low AND PI Low	Low Risk

RESULTS AND DISCUSSION

Classification Performance

Table 2 presents the classification performance of the proposed LightGBM model. The model achieved an Accuracy of 95.89%, Precision of 95.58%, Recall of 82.97%, F1-Score of 88.83%, and ROC-AUC of 98.73%, indicating excellent discriminative capability in distinguishing pathogenic and benign variants.

Table 2. Overall Classification Performance

Metric	Value
Accuracy	95.89%
Precision	95.58%
Recall	82.97%
F1-Score	88.83%
ROC-AUC	98.73%

Furthermore, Table 3 presents the detailed classification report for each class. The model achieved near-perfect performance for benign variants while maintaining strong performance for pathogenic variants.

Table 3. Classification Report

Class	Precision	Recall	F1-Score	Support
Benign	0.96	0.99	0.97	32,122
Pathogenic	0.96	0.83	0.89	7,878

Explainability Analysis

To improve model transparency, SHAP analysis was employed to quantify feature contributions. Figure 2 illustrates the SHAP summary plot, while Table 4 summarizes the mean absolute SHAP values of each feature.

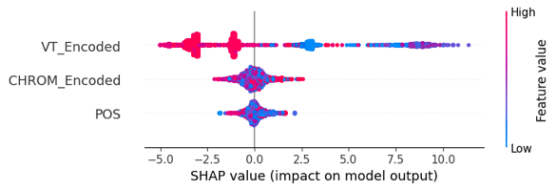

Fig 2. SHAP Summary Plot of Feature Contributions

The SHAP results indicate that VT_Encoded, representing variant-type information, contributes substantially more to prediction outcomes than chromosome-related and positional genomic attributes. This finding suggests that variant-type characteristics provide stronger predictive signals for pathogenicity assessment.

Table 4. SHAP Feature Importance

Feature	Mean Absolute SHAP
VT_Encoded	3.3687
CHROM_Encoded	0.4270
POS	0.3046

The dominance of VT_Encoded suggests that variant-type information contains stronger predictive signals than chromosome representation or genomic position.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Comparison of Feature Importance Methods

To further investigate feature relevance, a comparison between LightGBM feature importance and SHAP importance was conducted.

Table 5. Comparison of Feature Ranking

Rank	LightGBM Importance	Score
1	CHROM_Encoded	14,718
2	VT_Encoded	11,583
3	POS	4,533
Rank	**SHAP Importance**	**Mean Abs SHAP**
1	VT_Encoded	3.3687
2	CHROM_Encoded	0.4270
3	POS	0.3046

Figure 3 presents the LightGBM feature importance ranking. Although LightGBM identifies CHROM_Encoded as the most frequently utilized feature, SHAP reveals that variant-type representation (VT_Encoded) has the greatest influence on prediction outcomes, whereas chromosome representation contributes primarily to tree-splitting decisions within the LightGBM model.

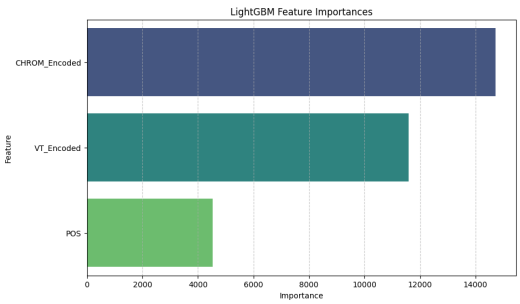

Fig 3. LightGBM Feature Importance Ranking

Pathogenicity Index Analysis

The proposed Pathogenicity Index (PI) was developed to transform probabilistic predictions into a continuous risk representation. Table 6 summarizes the descriptive statistics of PI values.

Table 6. Descriptive Statistics of Pathogenicity Index

Statistic	Value
Mean	0.3859
Standard Deviation	0.1869
Minimum	0.1465
25th Percentile	0.2750
Median	0.3167
75th Percentile	0.3626
Maximum	0.9746

The wide distribution of PI values indicates that the proposed index effectively captures varying pathogenicity levels and provides richer information than conventional binary classification outputs.

Fuzzy Risk Assessment

Figure 4 illustrates the distribution of risk levels generated by the Fuzzy DSS. Most variants were categorized as Low Risk or Medium Risk, whereas a smaller proportion was classified as High Risk.

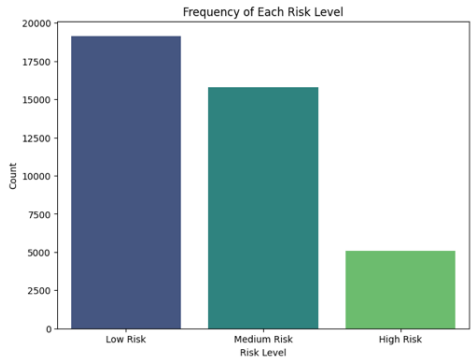

Fig 4. Distribution of Risk Levels Generated by the Fuzzy DSS

The distribution demonstrates the ability of the proposed framework to stratify genetic variants into clinically meaningful categories.

Figure 5 presents the violin plot of risk scores across risk categories. Clear separation among Low, Medium, and High-Risk groups can be observed, indicating effective fuzzy inference and defuzzification processes.

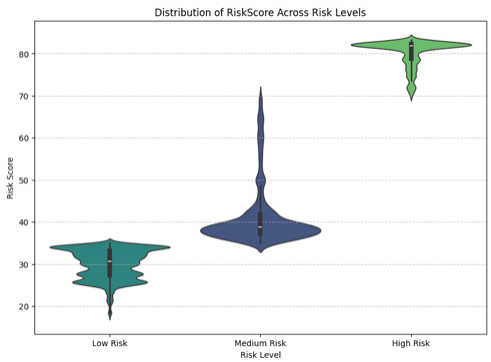

Fig 5. Distribution of Risk Scores Across Risk Categories

Low Risk variants were concentrated around scores of 20–35, Medium Risk variants around 35–50, and High-Risk variants above 70.

Figure 6 illustrates the relationship between pathogenicity probability and risk score. A strong positive correlation is observed, confirming that higher

pathogenicity probabilities consistently produce higher risk scores.

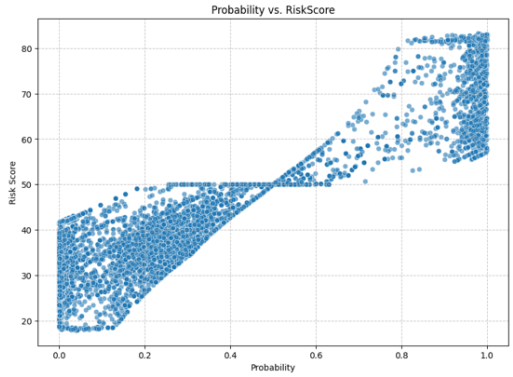

Fig 6. Relationship Between Prediction Probability and Risk Score

Discussion

The findings demonstrate that the proposed framework extends conventional genomic variant classification by integrating predictive modeling, explainability, and risk-oriented decision support within a unified architecture. Previous studies have shown the effectiveness of machine learning approaches for genomic variant classification; however, the proposed framework further incorporates explainability, risk quantification, and decision-support mechanisms to improve the interpretability and practical utility of prediction outcomes.

The SHAP analysis identified variant-type representation as the most influential predictor of pathogenicity, indicating that mutation characteristics contribute more substantially to risk assessment than chromosomal and positional genomic attributes. This observation supports the relevance of variant-type information in genomic interpretation and demonstrates the value of explainable artificial intelligence in identifying factors that influence model predictions.

Furthermore, the proposed Pathogenicity Index and FDSS transformed probabilistic outputs into interpretable risk categories, providing a more informative assessment than conventional binary classification. By incorporating risk stratification into the assessment process, the framework offers additional support for variant prioritization and genomic risk evaluation.

Despite these promising findings, the study was conducted using a ClinVar-derived dataset and did not include external validation using independent genomic repositories. Therefore, further studies involving robustness testing and cross-dataset validation are required to assess the generalizability of the proposed framework across diverse genomic datasets and clinical contexts.

CONCLUSION

This study proposed a hybrid framework that integrates LightGBM, SHAP, a Pathogenicity Index, and a Fuzzy Decision Support System for clinical variant risk assessment. The proposed framework extends conventional pathogenicity classification by combining predictive modeling, explainability, risk quantification, and decision-support mechanisms within a unified architecture. By transforming probabilistic predictions into interpretable risk categories, the framework provides a more transparent and actionable approach to genomic risk evaluation. These findings highlight the potential of integrating machine learning, explainable artificial intelligence, and fuzzy reasoning to support clinical variant interpretation and precision medicine applications.

DAFTAR PUSTAKA

Alshboul, O., Shehadeh, A., Almasabha, G., & Almuflih, A. S. (2022). Extreme gradient boosting-based machine learning approach for green building cost prediction. *Sustainability*, *14*(11), 6651.

Amiri, Z. (2024). Leveraging AI-enabled information systems for healthcare management. *Journal of Computer Information Systems*, 1–28.

Aregbesola, G. D., Asghar, I., Akbar, S., & Ullah, R. (2025). Fuzzy Logic Model for Informed Decision-Making in Risk Assessment During Software Design. *Systems*, *13*(9), 825.

Bahmane, K., Bhattacharya, S., & Kassem, M. A. (2026). PathoPredictor: A Machine Learning Framework for Predicting Pathogenic Missense Variants in the Human Genome. *Journal of Genome Biotechnology and Genetics*, *1*(1), 3.

Barbitoff, Y. A., Ushakov, M. O., Lazareva, T. E., Nasykhova, Y. A., Glotov, A. S., & Predeus, A. V. (2024). Bioinformatics of germline variant discovery for rare disease diagnostics: current approaches and remaining challenges. *Briefings in Bioinformatics*, *25*(2), bbad508.

Dhanka, S., Sharma, A., Kumar, A., Maini, S., & Vundavilli, H. (2026). Advancements in hybrid machine learning models for biomedical disease classification using integration of hyperparameter-tuning and feature selection methodologies: A comprehensive review. *Archives of Computational Methods in Engineering*, *33*(1), 289–324.

Divya, N., Desai, B. S., Prem, D. S., Bindu, C., Vybhavi, G. Y., Sundaray, M., & Hussein, L. (2025). Advancing genetic variant classification: Integrating machine learning with population genetics and clinical interpretation. *AIP Conference Proceedings*, *3361*(1), 050059.

Izhari, F. I., & Meiyanti, R. (2025). Hybrid Ensemble Learning to Improve Prediction Disease Kidney Chronic. *JADEN: Journal of Algorithmic Digital Engineering and Networks*, *1*(1), 39–46.

Jiang, P. H. W., Wang, W. Y. C., Goh, T., & Hsieh, C.-C. (2024). System integration framework for implementing a machine learning-driven clinical decision support system in emergency departments. *Proceedings of the 2024 8th International Conference on Medical and Health Informatics*, 120–126.

Kostopoulos, G., Davrazos, G., & Kotsiantis, S. (2024). Explainable artificial intelligence-based decision support systems: A recent review. *Electronics*, *13*(14), 2842.

Kruschel, S., Hambauer, N., Weinzierl, S., Zilker, S., Kraus, M., & Zschech, P. (2026). Challenging the performance-interpretability trade-off: an evaluation of interpretable machine learning models. *Business & Information Systems Engineering*, *68*(1), 159–183.

Makumbura, R. K., Mampitiya, L., Rathnayake, N., Meddage, D. P. P., Henna, S., Dang, T. L., Hoshino, Y., & Rathnayake, U. (2024). Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence (XAI) techniques like shapley additive explanations (SHAP) for interpreting the black-box nature. *Results in Engineering*, *23*, 102831.

Méndez-Vidal, C., Bravo-Gil, N., Pérez-Florido, J., Marcos-Luque, I., Fernández, R. M., Fernández-Rueda, J. L., González-del Pozo, M., Martín-Sánchez, M., Fernández-Suárez, E., & Mena, M. (2025). A genomic strategy for precision medicine in rare diseases: integrating customized algorithms into clinical practice. *Journal of Translational Medicine*, *23*(1), 86.

Qiuqian, W., GaoMin, KeZhu, Z., & Chenchen. (2025). A light gradient boosting machine learning-based approach for predicting clinical data breast cancer. *Multiscale and Multidisciplinary Modeling, Experiments and Design*, *8*(1), 75.

Santos, M. R., Guedes, A., & Sanchez-Gendriz, I. (2024). SHapley additive explanations (SHAP) for efficient feature selection in rolling bearing fault diagnosis. *Machine Learning and Knowledge Extraction*, *6*(1), 316–341.

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., & Das, G. (2023). Next-generation sequencing technology: current trends and advancements. *Biology*, *12*(7), 997.

Srivastava, A., Bhanot, D., Jasim, L. H., Varshney, N., & Patil, V. (2025). Advancements in fuzzy logic applications for diagnostic decision support systems in healthcare. *Fuzzy Information and Engineering*, *17*(3), 284–297.

Tursunalieva, A., Alexander, D. L. J., Dunne, R., Li, J., Riera, L., & Zhao, Y. (2024). Making sense of machine learning: A review of interpretation techniques and their applications. *Applied Sciences*, *14*(2), 496.

