

EXPLAINABLE MACHINE LEARNING FOR PATIENT SATISFACTION PREDICTION IN HEALTHCARE: A REGRESSION AND CLASSIFICATION STUDY

Milica Tufegdžić¹, Đorđe Mirjanić², Predrag Dašić^{3,*} and Nevena Tufegdžić⁴

¹Academy of Professional Studies, Information Technology Department
Kragujevac, Serbia

²University of Banja Luka, Faculty of Medicine, Department of Stomatology
Banja Luka, Bosnia and Herzegovina

³Engineering Academy of Serbia
Belgrade, Serbia

⁴SaTCIP Publisher Ltd.
Vrnjačka Banja, Serbia

DOI: [10.7906/indexs.24.6.2](https://doi.org/10.7906/indexs.24.6.2)
Regular article

Received: 16 June 2026.
Accepted: 20 July 2026.

ABSTRACT

This study evaluated regression and classification models for predicting Artificial Intelligence diagnostic confidence and patient satisfaction in healthcare using a synthetic dataset of 5 000 patients with 23 variables generated through clinically informed rules. Linear Regression, Support Vector Regression, Random Forest, HistGradientBoosting, eXtreme Gradient Boosting, Light Gradient Boosting Machine, and CatBoost were tested for regression to estimate the target variable AI_Diagnosis_Confidence. For multiclass classification of Patient_Satisfaction_Class (Low, Medium, High), models evaluated were Logistic Regression, Support Vector Machine, RF, and HGBost. Model performance was assessed using cross-validation with a coefficient of determination R^2 , mean absolute error, and root mean squared error for regression, and accuracy, balanced accuracy, F1 macro, precision macro, and recall macro for classification. SHapley Additive exPlanations (SHAP) was used to interpret predictions and quantify feature contributions, with Linear Regression achieving the best regression performance ($R^2 = 0.9577$) with Diagnosis Complexity, Comorbidity Count, and Lab Abnormality Score as the most influential predictors. For classification, LogR delivered the most balanced performance, capturing minority classes, while tree-based models were biased toward the majority class. SHAP highlighted Recovery Time, Diagnosis Complexity, and Total Risk Score as influential features guiding class-specific predictions. These findings show that interpretable models with SHAP can accurately and transparently predict Artificial Intelligence diagnostic confidence and patient satisfaction, with simpler models excelling when relationships are linear.

KEY WORDS

patient satisfaction, predictive modeling, regression, classification, interpretable artificial intelligence

CLASSIFICATION

JEL: C53, C55, I10, I11

*Corresponding author, *η*: dasicp58@gmail.com; -;
Engineering Academy of Serbia, Bulevar vojvode Mišića 37, II sprat, RS – 11 000 Beograd, Serbia

INTRODUCTION

Artificial intelligence (AI) has emerged as a transformative force across numerous sectors, owing to its capacity to automate complex tasks, enhance operational efficiency, and support informed decision-making [1]. In the healthcare sector, AI is increasingly recognized for its significant impact on diagnostics, treatment planning, personalized medicine, and administrative workflows. Through the analysis of large-scale datasets, the identification of intricate patterns, and the generation of predictive insights, AI contribute to improved diagnostic accuracy, cost reduction, and broader access to healthcare services. AI supports many aspects of healthcare decision-making and can enhance service quality, efficiency, and effectiveness by providing accurate, timely, and personalized information [2]. The integration of AI into clinical decision-making can improve diagnostic accuracy, streamline workflows, reduce errors, and lower costs, with long-term benefits that outweigh the initial investment [3]. However, the adoption of AI-driven Decision Support Systems remains limited due to low trust, poor transparency, and misaligned decision-making processes [4].

The healthcare process involves interactions among multiple stakeholders, including patients, providers, insurers, suppliers, and governmental or nonprofit organizations, across a wide range of settings such as hospitals, clinics, pharmacies, laboratories, and insurance networks [5]. This interconnected structure makes healthcare systems highly complex, as the actions of each stakeholder can influence multiple levels and subsystems, generating uncertainty that current AI models cannot fully capture [6]. In this context, human-in-the-loop AI helps bridge this gap by combining machine learning (ML) capabilities with the expertise and judgment of healthcare professionals. It creates a feedback mechanism in which human contributions, such as data annotation, model validation, and decision support, improve system performance, safety, quality, and efficiency while reducing the risks associated with full automation [7, 8]. The effectiveness of AI in healthcare depends largely on how well it interacts with human users; however, its impact remains constrained by the absence of evaluation frameworks that adequately reflect the complexity and contextual realities of real-world clinical workflows [9].

Applications of AI in diagnostics, decision support, patient monitoring, and research have demonstrated benefits such as increased accuracy, reduced errors, and faster processes. However, successful implementation requires careful attention to system integration, clinical workflows, user training, ethical considerations, and regulatory compliance [7]. AI also contributes to hospital operations and health policy by automating routine tasks, optimizing resource allocation, and supporting data-driven decision-making, although challenges remain, including potential clinical errors, model overfitting, ethical concerns, and adherence to regulations [10]. Virtual healthcare, as a specialized domain, leverages AI to enhance patient engagement, real-time health monitoring, and diagnostic performance through applications such as AI-driven diagnostic tools, predictive analytics, and teleconsultation platforms. These technologies are applied across areas including cardiac care, diabetes management, mental health, and dermatology, while simultaneously addressing ethical and regulatory issues such as bias, data privacy, and accountability [11].

Healthcare is inherently data-rich, and the combination of large datasets, advanced computing power, and AI capabilities has created significant opportunities [12]. Predictive analytics, powered by AI, can improve patient outcomes and reduce costs by analyzing electronic health records (EHRs), claims data, and other health information using statistical models and ML to forecast future health risks [13]. Furthermore, AI is playing an increasingly important role in diagnosis and prognosis by analyzing extensive medical data to identify patterns, enhance accuracy over time, and support early intervention. This is particularly valuable for chronic disease management, where AI can anticipate complications and assist clinicians in planning preventive strategies [14].

ML methods have been widely applied to predict chronic diseases, employing feature selection techniques such as least absolute shrinkage and selection operator (LASSO), minimal-redundancy-maximal-relevance, and RELIEF, along with models including Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and Naïve Bayes (NB) [15]. For instance, ML was used to predict heart disease by training five models, RF, SVM, Logistic Regression (LogR), eXtreme Gradient Boosting (XGBoost), and Light Gradient-Boosting Machine (LGBM), with the best-performing model subsequently deployed in a web application for reliable predictions based on user input [16]. ML approaches, including SVM and LogR, have also been applied to predict multiple diseases, such as diabetes, heart disease, lung cancer, Parkinson's disease, and breast cancer [17]. Several models, including Feedforward Neural Network, XGBoost, RF, and SVM, were evaluated across four datasets encompassing biomarkers, patient history, symptoms, and syndromic surveillance, revealing that prediction performance depends on the type of data, modeling approach, and clinical context, highlighting the need for data-specific solutions [18]. Some systems integrate Convolutional Neural Networks (CNN) for feature extraction with K-Nearest Neighbors (KNN) for data matching to predict diseases using datasets containing symptoms, lifestyle factors, and medical records, with performance compared against NB, DT, and LogR models [19]. ML can also enable continuous image-learning systems that automatically detect common disease patterns through online analysis, maintaining large image collections with data organized using isotropic positioning to improve accuracy and processing efficiency [20]. Furthermore, Articulate Medical Intelligence Explorer, a large language model-based AI system, was introduced for diagnostic dialogue. It uses a self-play simulated environment with automated feedback to scale learning across diseases and specialties, and its performance has been evaluated on history-taking, diagnostic accuracy, clinical management, communication, and empathy [21].

Explainable artificial intelligence (XAI) makes AI decisions interpretable and accountable, promoting health care quality. It enhances the six health care quality pillars of safety, effectiveness, patient centredness, timeliness, efficiency and equity by making medical decisions more transparent and informed [22]. XAI enables users to better understand and interpret AI predictions, which is particularly important in healthcare. It also supports clinicians in identifying the factors underlying AI-generated outputs, especially in disease diagnosis [23]. XAI improves the transparency and interpretability of AI systems through model-agnostic methods such as Local Interpretable Model-agnostic Explanations (LIME) and SHAP, and model-specific methods such as Feature Visualization, Decision Trees, and Rule-based Models. By providing clear explanations, these methods help build trust and support the adoption of AI in decision-making [24]. Combining multiple data sources improves diagnosis, reduces errors, supports disease monitoring, and helps XAI detect health problems earlier, enabling timely treatment [25].

AI can serve as an initial point of contact for patients, providing preliminary diagnoses and allowing clinicians to focus on more complex cases. Its effectiveness depends not only on prediction accuracy but also on the ability to offer clear explanations that can enhance patient satisfaction [26]. Trust in AI is critical in healthcare and is influenced by user-related factors such as education, experience, and cognitive biases, as well as system-related factors including transparency, complexity, and reliability. Outputs that are unreliable or biased due to insufficient or subjective data can compromise performance and reduce user trust and acceptance [27]. Multiple factors shape trust in AI systems, including AI literacy, which involves understanding how AI inputs and processes work to build confidence; AI psychology, which reflects the influence of demographic characteristics and prior experience; and AI utility, which considers perceived usefulness, efficiency, and integration into clinical workflows. Other important considerations include anthropomorphism, data privacy and security, and mechanisms to restore trust following errors, particularly in high-risk clinical environments [28].

A focus group of 17 adults was used to examine patient perspectives on five AI clinical scenarios, revealing the highest comfort with digital scribe and radiology review applications, and the lowest with virtual human interactions. Key themes included validation, usability, transparency, opportunity, and privacy, with participants valuing their involvement in the co-design process [29]. A survey conducted in 2019 found that while most respondents viewed AI positively in diagnosis and treatment, concerns persisted regarding misdiagnosis, privacy, reduced clinician interaction, and increased costs, especially among racial and ethnic minority groups. Participants preferred to be informed when AI was used and expressed greater comfort with its use in certain clinical settings [30]. Among 2039 respondents, overall trust in the healthcare system was moderate, but confidence in AI being used responsibly or safely was low. Trust was higher among individuals with general confidence in the system and lower among females and those with prior experiences of discrimination, while health literacy and AI knowledge had little effect [31].

In a study of 223 physicians and nurses analyzing 1338 wound-maceration diagnoses, accurate AI recommendations significantly improved diagnostic accuracy, making correct decisions ten times more likely, whereas incorrect recommendations reduced accuracy, highlighting potential overreliance on AI [32]. Attitudes toward AI in healthcare were further examined in a survey of 13 806 patients, with 57,6% expressing a favorable view. Attitudes varied according to demographics, health status, and technological literacy: females and individuals in poorer health tended to have fewer positive perceptions, while higher AI knowledge and technology use increased positivity. Trust in AI remained limited, particularly regarding treatment accuracy (41,8%), and most participants preferred Explainable AI (70,2%) and physician-led decision-making (72,9%), even at the cost of a slight reduction in accuracy [33].

This study compares regression and classification approaches for predicting AI diagnostic confidence and evaluates whether SHAP explanations remain consistent across continuous and categorical representations of the target. A synthetic dataset was created to support the analysis, and the study was conducted through several key stages. First, data preprocessing and regression modeling was performed, followed by classification modeling. The performance of the regression models was then evaluated, and the structure of the synthetic data was validated using SHAP. Comparative performance of classification models was assessed, including confusion matrix analysis across models to examine predictive accuracy. Finally, SHAP analysis was conducted to interpret model outputs and assess the consistency and interpretability of feature contributions across different modeling approaches.

DATA

SYNTHETIC DATASET DESIGN PRINCIPLES

The generation of comprehensive and diverse datasets is crucial for enhancing both data collaboration and the reliability of healthcare decisions. Many existing datasets, however, lack sufficient diversity, which can result in biased algorithms and reduced accuracy for certain patient groups, highlighting the need for representative and inclusive data to ensure algorithmic safety [34]. Effectively managing healthcare data as big data requires proper preprocessing and normalization of unstructured and heterogeneous sources, such as EHRs and Internet of Things (IoT) devices, because well-designed data models are essential for accurate analysis and to prevent misleading conclusions [35]. The “AI in Healthcare” dataset [36] that was initially considered for model development was deemed unsuitable due to implausible clinical associations, unrealistic measurement precision, missing statistical dependencies, demographic inconsistencies, and an arbitrary AI confidence metric, undermining both its statistical validity and clinical realism.

Given the practical, ethical, and legal challenges of using real-world clinical data, a synthetic healthcare dataset was generated through a multi-stage pipeline. Each observation represents a distinct patient. First, baseline patient profiles were created using randomized sampling. Next, diagnoses, treatments, and complexity levels were assigned with clinically informed rules, followed by simulation of vital signs and laboratory measurements. Finally, derived features and outcome variables, including risk scores and `AI_Diagnosis_Confidence`, were computed using formula-driven feature engineering. This layered approach produced an internally consistent dataset suitable for ML tasks such as regression, classification, and interpretability analyses like SHAP, enabling examination of feature-level contributions to model predictions.

Baseline variables included Age (18 to 90 years), Gender (Male, Female), Insurance Type (Medicare, Private, Uninsured), and Diagnosis, randomly assigned from six categories: Healthy, Hypertension, Diabetes, Cancer, Heart Disease, and Influenza. Core physiological and laboratory measurements, including `Blood_Pressure`, `Heart_Rate`, `Body_Temperature`, and `Lab_Test_Results`, were sampled from clinically plausible distributions. Structured dependencies were included to reflect realistic clinical interactions, with diagnosis severity influencing treatment, recovery time, and diagnostic complexity. Derived variables such as `Age_Group`, `Patient_Satisfaction_Class`, `Comorbidity_Count`, `Lab_Abnormality_Score`, `Urgency_Flag`, and `Total_Risk_Score` were engineered to capture clinically meaningful constructs and support downstream predictive modeling.

The primary target variable, `AI_Diagnosis_Confidence`, was synthetically generated using a weighted formula (1) designed to mimic an AI diagnostic system under varying case difficulty:

$$\begin{aligned} \text{AI_Diagnosis_Confidence} = & 0,95 - (0,45 \cdot \text{Complexity}) - (0,08 \cdot \text{Comorbidities}) \\ & - (0,04 \cdot \text{Lab_Abnormality}) + \varepsilon, \end{aligned} \quad (1)$$

where ε represents stochastic noise. The constant 0,95 reflects baseline confidence under ideal conditions, while confidence decreases with higher case complexity, comorbidity burden, and laboratory abnormalities. Complexity carries the greatest weight, reflecting its dominant influence on diagnostic uncertainty. The noise term introduces variability to approximate unobserved factors in real-world clinical settings.

DATASET CHARACTERISTICS AND DESCRIPTIVE STATISTICS

The study dataset included 5 000 patient records with 23 variables: 13 numerical and 10 categorical. Numerical features covered demographics, physiological and laboratory measurements, treatments, and outcomes, while categorical features captured demographics, diagnoses, medications, insurance status, imaging results, surgical interventions, allergies, family history, and grouped satisfaction classes. All numerical variables were complete, with no missing values, indicating high data integrity. Descriptive statistics for the numerical features are presented in Table 1.

The dataset demonstrated heterogeneous yet clinically plausible patterns across both numerical and categorical variables. Outcome-related variables, such as `Recovery_Time`, showed substantial variability, reflecting diverse recovery trajectories, while `Patient_Satisfaction` scores were generally high, with the High category predominating ($n = 4\,447$). `AI_Diagnosis_Confidence` values were relatively elevated, and `Diagnosis_Complexity` exhibited moderate variability.

Among categorical variables, several were fully complete, including Gender, Diagnosis, Insurance_Type, X-ray_Results, Family_History, Patient_Satisfaction_Class, and Age_Group. Gender had two categories, with Female most frequent ($n = 2\,542$). Diagnosis comprised six categories, dominated by Hypertension ($n = 1\,125$). Insurance_Type included three categories, with Uninsured most prevalent ($n = 1\,682$). X-ray_Results were predominantly Normal ($n = 4\,043$),

Table 1. Descriptive statistics for numerical variables.

Variable	Minimum	Maximum	Mean	Standard Deviation (SD)	Median	Interquartile Range (IQR)
Age (years)	18	90	53.78	20.92	54.00	36-72
Blood_Pressure (mmHg)	83.45	186.92	126.75	16.04	123.65	115.47-135.95
Heart_Rate (bpm)	39.85	122.22	76.52	10.86	75.66	68.81-83.21
Temperature (°F)	96.93	101.86	98.93	0.84	98.71	98.39-99.10
Treatment_Duration (days)	1	30	15.46	8.60	15.00	8.00-23.00
Lab_Test_Results	35.97	208.23	109.09	29.25	102.35	89.49-20.17
Recovery_Time (days)	0	39	6.25	8.66	4	2-6
Patient_Satisfaction	1	5	4.49	0.76	5	4-5
AI_Diagnosis_Confidence	0.511	0.980	0.772	0.098	0.786	0.696-0.851
Diagnosis_Complexity	0.05	0.80	0.374	0.245	0.30	0.20-0.70
Comorbidity_Count	0.00	3.00	1.48	0.76	2.00	1.00-2.00
Lab_Abnormality_Score	0.00	2.16	0.44	0.43	0.29	0.12-0.58
Total_Risk_Score	0.24	2.61	1.23	0.48	1.28	0.81-1.62

Family_History most commonly indicated Heart Disease ($n = 1698$), and Age_Group included four categories, with Adult as the largest subgroup ($n = 1714$).

Some categorical variables contained clinically meaningful missing values. Surgery_Type had the highest missingness (85,2%), reflecting patients who did not undergo surgery, and among recorded entries, angioplasty was most frequent ($n = 427$). Allergies were missing in 33,8% of cases, indicating no known allergies, with penicillin most frequently reported ($n = 1664$). Medication data had 17,2% missingness, reflecting patients not receiving medications, and Lisinopril was the most common documented drug ($n = 1125$).

Overall, the dataset combines complete quantitative measurements with informative categorical variables and outcome indicators. The variability across physiological, laboratory, and recovery measures, along with structured missingness in treatment-related variables, supports its suitability for descriptive statistics, predictive modeling, and clinical decision-support research.

EXPLORATORY DATA PROFILING AND FEATURE ASSESSMENT

The dataset was profiled by classifying numerical and categorical variables and assessing missing values to evaluate completeness. Distributions of primary outcomes, AI_Diagnosis_Confidence and Patient_Satisfaction_Class, were examined for skewness and class imbalance. Outliers in clinical features were identified using the Interquartile Range (IQR) method. Correlation analysis, visualized in a heatmap (Figure 1), guided the selection of the top 15 features most strongly associated with AI confidence, after removing non-informative columns.

The dataset reveals several strong correlations between patient factors and AI performance. AI_Diagnosis_Confidence is highly negatively correlated with Diagnosis_Complexity ($-0,89$), indicating confidence drops as cases become more complex. Recovery_Time is inversely related to Patient_Satisfaction ($-0,76$), while Age strongly drives Total_Risk_Score ($0,74$), and Lab_Test_Results align closely with Lab_Abnormality_Score ($0,72$). Higher AI confidence correlates positively with Patient_Satisfaction ($0,61$) but decreases as Comorbidity_Count rise ($-0,60$). Complexity also contributes to risk assessment, with more difficult cases linked to higher Total_Risk_Score ($0,61$). Several variables, like Treatment_Duration, that is largely independent of other factors show minimal associations. Vital signs such as Heart_Rate and Temperature exhibit weak correlations with Total_Risk_Score ($0,05$ and $-0,33$). Recovery_Time moderately increases with Diagnosis_Complexity ($0,56$), reinforcing the negative impact of complex cases on patient satisfaction.

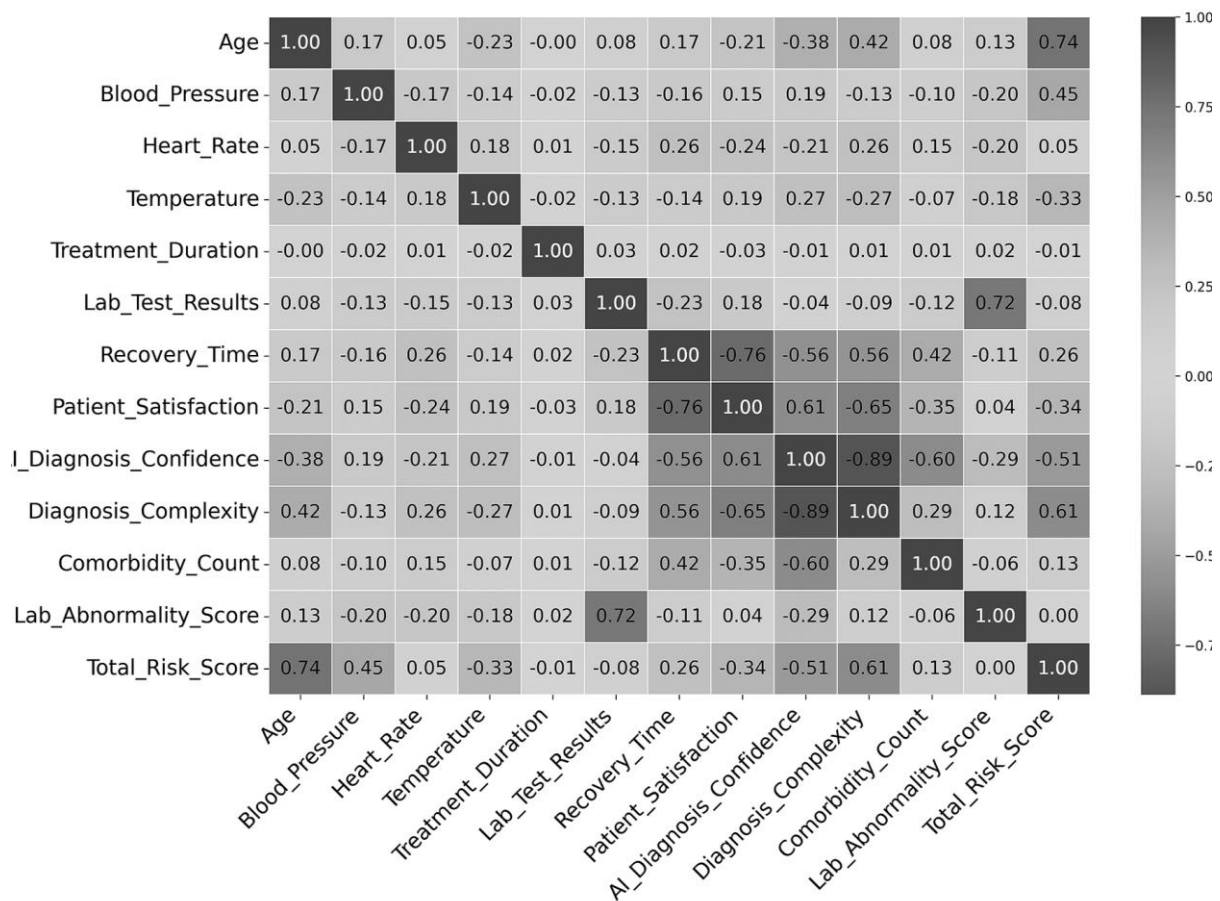


Figure 1. Correlation heatmap.

Scatter plots and categorical boxplots were used to examine how key numerical features and patient demographics influence AI diagnostic performance, Figure 2. For example, boxplots comparing AI_Diagnosis_Confidence across age groups and diagnoses reveal clear patterns. Age affects AI certainty: Minors and Young Adults exhibit higher confidence ($\sim 0,85$), while Seniors show lower confidence ($\sim 0,73$) and greater variability, reflecting the higher prevalence of comorbidities in older patients. Some Young Adult cases appear as outliers with low confidence, indicating that even in groups where AI performs well overall, certain cases remain challenging. Confidence is highest for well-defined conditions such as Healthy (median $\sim 0,88$) and Influenza ($\sim 0,84$), and lowest for complex conditions like Cancer and Heart Disease (medians $\sim 0,62$ - $0,66$), with Diabetes and Hypertension in the middle range.

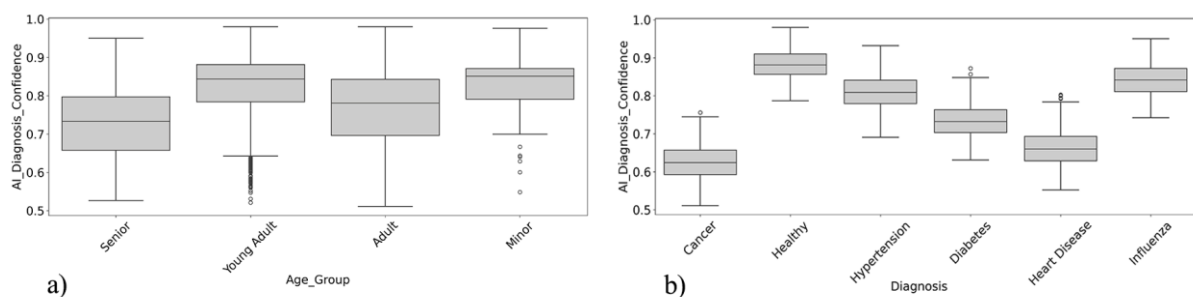


Figure 2. Boxplots: a) AI_Diagnosis_Confidence by Age_Group; b) AI_Diagnosis_Confidence by Diagnosis.

METHODOLOGY

REGRESSION MODELING AND DATA PREPROCESSING

The target variable for the regression pipeline is `AI_Diagnosis_Confidence`, which the model aims to predict using the selected set of predictor features. Feature engineering was applied to create clinically meaningful binary indicators, including elevated blood pressure, elevated heart rate, and fever status. Numerical variables were imputed using the median and standardized using z-score normalization, while categorical variables were imputed using the most frequent category and transformed using one-hot encoding. These preprocessing steps were implemented using a `ColumnTransformer` within a unified scikit-learn pipeline to ensure consistent transformation during model training and evaluation.

The dataset was split into training (80%) and testing (20%) subsets. Multiple regression models were evaluated, including Linear Regression (LR), SVR, RF Regressor, HistGradientBoosting (HGB) Regressor, XGBoost, LGBM, and CatBoost. Model performance was assessed using coefficient of determination (R^2), mean absolute error (MAE), root mean squared error (RMSE) on the test set, while 5-fold cross-validated RMSE (CV_RMSE) on the training set was used for model comparison and ranking. The best-performing model, based on the lowest CV_RMSE, was selected for further analysis, including SHAP-based explainability analysis. For additional model evaluation, predicted values for `AI_Diagnosis_Confidence` were calculated for all cases and for each regression model, along with the corresponding relative errors that were calculated as the ratio of the absolute errors (differences between observed and predicted values) to the observed values. To facilitate a more efficient comparison of model performance, the mean values of these errors were computed as mean relative errors (MREs).

CLASSIFICATION MODELING AND DATA PREPROCESSING

A supervised multiclass classification framework was developed to predict patient satisfaction categories Low, Middle, and High using `Patient_Satisfaction_Class` as the target variable. To prevent target leakage, the original continuous satisfaction score was excluded, and derived binary features including `High_BP` (blood pressure greater than 140 mmHg), `High_HR` (heart rate greater than 100 bpm), and `Fever` (temperature greater than 37,5 °C) were introduced to capture clinically relevant physiological thresholds.

The dataset was split into training and test subsets using a stratified 80:20 split to preserve class distributions. A unified preprocessing pipeline ensured consistency and avoided information leakage. Numerical features were median-imputed and standardized, while categorical features underwent the most frequent imputation and one-hot encoding. Class imbalance was addressed with the Synthetic Minority Oversampling Technique (SMOTE) applied only to the training data.

Four ML models, RF, HGB, LogR, and SVM with a radial basis function (RBF) kernel, were evaluated to compare linear and nonlinear approaches. Performance was assessed on the holdout test set using accuracy, balanced accuracy, F1 macro, precision macro, and recall macro, complemented by five-fold stratified cross-validation on the training data, with mean F1 macro as the primary internal validation metric. Confusion matrices were generated to identify class-specific prediction patterns.

Model interpretability was enhanced through feature importance analysis and SHAP. SHAP values highlighted class-specific feature contributions and the factors most strongly influencing predictions for each satisfaction category. Predicted class labels and their corresponding membership probabilities were stored for subsequent probability-based analyses, and the predicted probabilities for the true class were extracted to assess model confidence. Boxplots of these probabilities, grouped by model and actual class, enabled a comparative assessment of

confidence and consistency across categories. Higher median probabilities indicated stronger support for the correct class, while wider distributions reflected greater variability in model certainty.

This workflow provides a rigorous, transparent, and methodologically robust approach for multiclass patient satisfaction prediction, integrating preprocessing, class imbalance handling, model evaluation, and interpretability in a unified framework.

RESULTS AND DISCUSSION

REGRESSION

Regression Model Performance

To evaluate the predictive performance of the proposed regression models, multiple algorithms were trained to estimate the continuous target variable *AI_Diagnosis_Confidence*. Model performance was assessed using R^2 , MAE, RMSE, and CV_RMSE . The results for all algorithms are visualized in graphs and presented in Figure 3. The results indicate LR achieved the best overall performance among all evaluated models, obtaining the highest R^2 (0,9577) and the lowest MAE (0,0156), RMSE (0,0195), and CV_MSE (0,0201). These findings suggest that the model was able to explain approximately 95,77% of the variance in the AI diagnostic confidence score while maintaining a very low prediction error.

Among the nonlinear and ensemble-based approaches, CatBoost Regressor performed competitively, achieving an R^2 of 0,9568, which was only marginally lower than that of LR. Other tree-based ensemble methods, including RF, HGB, LGBM, and XGBoost, also demonstrated strong predictive capability, with R^2 values above 0,953. However, none of these more complex models surpassed the simpler linear baseline. SVR exhibited the weakest performance in the comparison, with the lowest R^2 and highest error values, indicating comparatively reduced suitability for the present synthetic data structure.

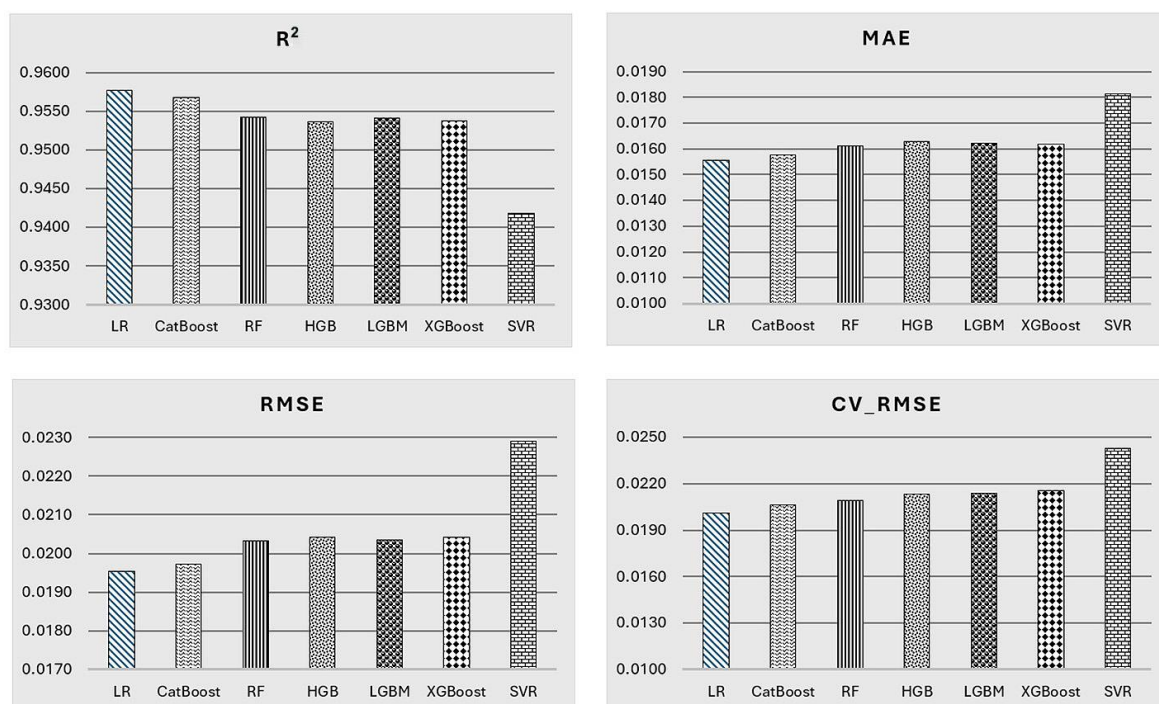


Figure 3. Regression performance comparison for the selected ML algorithms.

The comparative analysis of regression model performance based MREs shows that LR achieved the best overall results. It recorded the lowest MRE (0,020, or approximately 2,00%),

indicating the highest predictive accuracy among the evaluated models. CatBoost Regressor ranked second, with only slightly higher value ($MRE = 0,021$). Other ensemble-based models, including RF Regressor, XGBoost, LGBM, and HGBost, exhibited very similar performance, whereas SVR showed the weakest performance, with the highest MRE (0,024).

Descriptive statistics of absolute and relative errors further confirmed the superiority of LR. In addition to the lowest average error, it also demonstrated one of the lowest standard deviations of absolute error (approximately 0,012), indicating stable and consistent predictions across the dataset. Moreover, it produced the lowest maximum absolute error (0,066), suggesting better robustness and fewer extreme prediction failures compared to models such as SVR and RF Regressor, which showed considerably larger maximum deviations. For all models, the median absolute error was lower than the mean, indicating a right-skewed error distribution in which most predictions were highly accurate, while a small number of larger errors increased the average.

A particularly important result is that the simplest model, LR, outperformed more complex ML algorithms. This finding suggests that the relationship between the input features and the target variable is predominantly linear, making additional model complexity unnecessary for this dataset. Although CatBoost Regressor proved to be a strong nonlinear alternative, the overall results indicate that LR provides the best balance of accuracy, stability, computational efficiency, and interpretability. Therefore, it can be considered the most appropriate model for the final application in this study.

Interpretation of Regression Performance

The strong performance of LR is not only statistically significant but also methodologically consistent with the synthetic data generation process. The target variable, `AI_Diagnosis_Confidence`, was generated using a weighted linear function. Because the underlying relationship between the principal predictors and the target variable is predominantly linear, a LR model is naturally well aligned with the data-generating mechanism. As a result, it is expected that LR would recover this structure more efficiently than nonlinear models, particularly when the added noise is relatively small and the embedded relationships remain stable. This finding should not be interpreted as a limitation of the study, but rather as evidence that the predictive framework is internally coherent and that the model comparison is successfully capturing the known synthetic signal. In fact, the strong performance of the linear baseline reinforces the validity of the simulation design and confirms that the synthetic target variable behaves according to the intended clinical logic.

Although advanced ensemble models such as CatBoost, RF, LGBM, and XGBoost are generally advantageous in complex real-world nonlinear settings, their marginally lower performance in this study suggests that the synthetic confidence mechanism does not require high-capacity nonlinear approximation under the current formulation. The close performance of CatBoost to LR nevertheless indicates that the relationship can still be captured robustly by more flexible learners, which is valuable for comparative analysis and explainability benchmarking.

SHAP-Based Explainability Analysis of the Best Model

To investigate model interpretability and validate whether the learned relationships aligned with the synthetic data-generating process, SHAP were applied to the best-performing model, LR. The SHAP summary plot (Figure 4) provided both global feature importance and directional interpretation of how each feature influenced the predicted `AI_Diagnosis_Confidence` values.

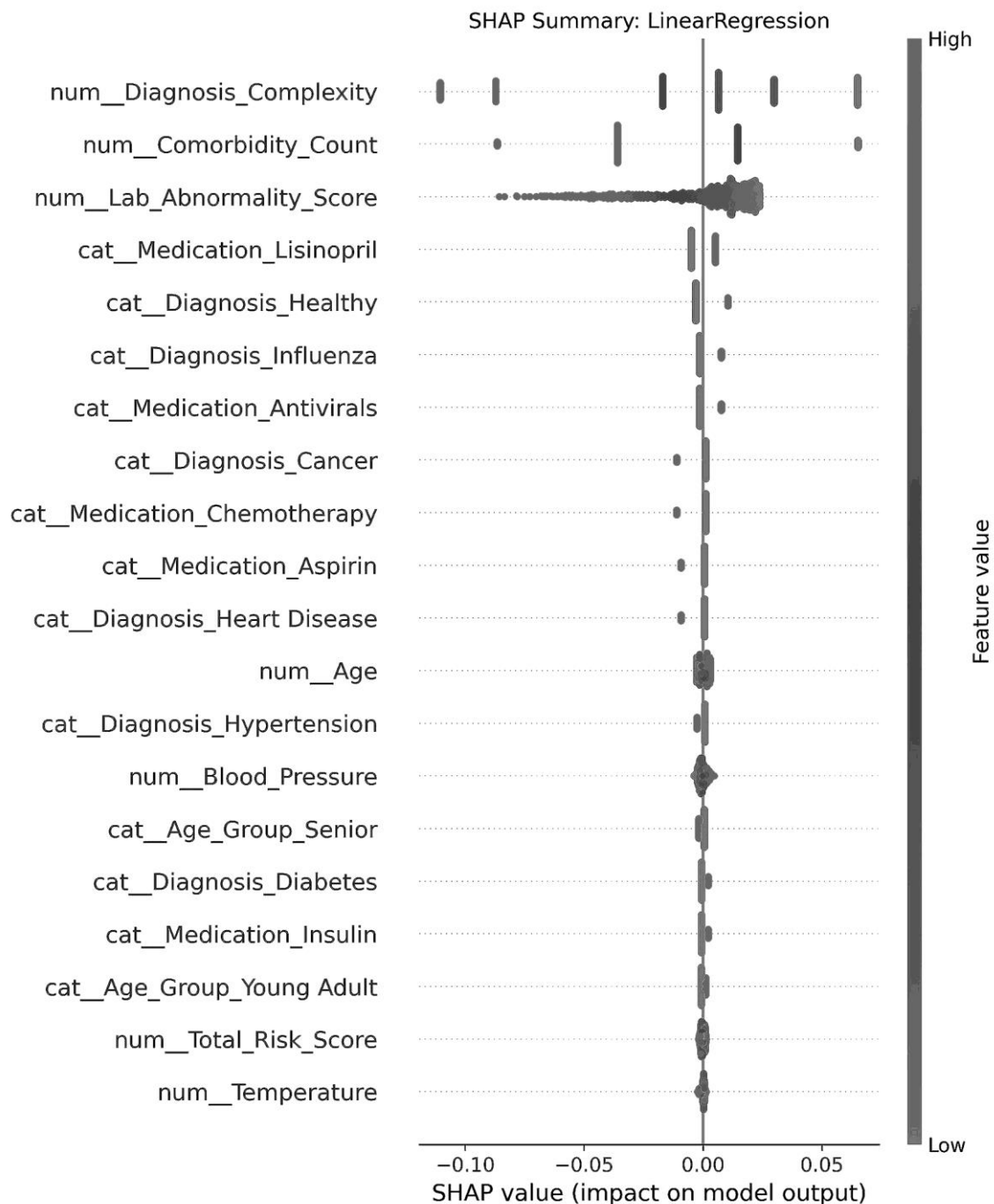


Figure 4. SHAP summary plot for LR.

The SHAP analysis revealed that the three most influential predictors were Diagnosis_Complexity (num_Diagnosis_Complexity), Comorbidity_Count (num_Comorbidity_Count), and Lab_Abnormality_Score (num_Lab_Abnormality_Score). For Diagnosis_Complexity, the higher values were associated with lower model predictions, suggesting an inverse relationship; for Comorbidity_Count, more comorbidities decreased the predicted outcome contrary to typical expectations; and for Lab_Abnormality_Score, lower scores increased predictions while higher scores reduced them, indicating the model treats abnormal labs as negative contributors to the target. This ranking is highly consistent with the synthetic target-generation formula, in which these variables were explicitly assigned descending negative weights. Specifically, Diagnosis_Complexity carried the largest penalty coefficient, followed by comorbidities and laboratory abnormality, making them the expected dominant determinants of AI confidence.

Other notable features include Medication Lisinopril (cat_Medication_Lisinopril), whose presence increases the prediction, and Diagnosis_Healthy (cat_Diagnosis_Healthy), where patients labeled "Healthy" show higher predicted values. Features with minimal impact, like Total_Risk_Score (num_Total_Risk_Score) and Temperature (num_Temperature), have SHAP values tightly clustered around zero.

Features indicating higher illness severity, such as complexity or comorbidities, are negatively correlated with the prediction, which suggests that the target might represent something like a "Patient Health Score" where higher values are better, rather than risk or cost. Categorical variables, including diagnoses or medications, form distinct clusters, and the presence or absence of these features clearly influences the model, with a Cancer diagnosis, for example, slightly reducing the predicted value. Since this is a LR model, the SHAP values closely mirror the model coefficients, making each feature's effect additive and consistently directional.

Validation of the Synthetic Data Structure Through Shap

One of the most important findings of this study is that the SHAP analysis successfully recovered the known synthetic ground-truth relationships embedded in the target-generation process, providing internal validation for both the predictive model and the explainability framework. Because the target variable was intentionally constructed using a predefined weighted relationship, the SHAP analysis serves as a benchmark to assess whether the trained model correctly learned the feature importance hierarchy and the direction of feature influence.

The observed SHAP results confirm both: Diagnosis_Complexity was identified as the strongest negative contributor, Comorbidity_Count was identified as the second-strongest negative contributor, and Lab_Abnormality_Score was identified as the third-strongest negative contributor. This close agreement between the designed synthetic formula and the SHAP-derived explanations demonstrates that the LR and SHAP pipeline is capable of accurately recovering the intended signal structure. This is a particularly important contribution in the context of Explainable AI (XAI), as it shows that the interpretability method is not merely producing visually plausible outputs, but is faithfully reflecting the true synthetic mechanism that governs the target variable.

Practical Implications and Limitations

Although the present study is based on synthetic data, the findings offer several practical insights for the design of trustworthy healthcare AI systems. The results illustrate that AI confidence estimation can be modeled effectively using structured patient and clinical burden variables. In real-world applications, this could support systems that not only produce diagnostic predictions but also provide calibrated confidence estimates to clinicians. The strong interpretability of the LR and SHAP framework suggests that simpler models may be preferable in healthcare contexts where transparency and auditability are essential. In settings where the relationships are approximately linear or intentionally structured, simpler models may offer a better balance between performance and interpretability than more complex black-box alternatives. Finally, the ability of SHAP to recover clinically meaningful feature contributions highlights its value for model auditing, trust building, and human-AI collaboration. For example, a clinician may be more willing to trust a low-confidence AI recommendation if the system clearly explains that the reduction in confidence is due to high case complexity, multiple comorbidities, and severe laboratory abnormalities. Thus, even within a synthetic framework, the study demonstrates how explainable predictive modeling can support the development of safer and more interpretable clinical decision-support systems.

Despite the strong performance and coherent explainability findings, several limitations should be acknowledged. Because the target variable was generated using a predominantly linear

formula, linear models such as LR are naturally favored. As a result, the superior performance of LR should not be interpreted as evidence that linear models will always outperform nonlinear methods in real-world healthcare data. The embedded relationships reflect assumptions made during the simulation process. Although these assumptions were designed to be clinically plausible, they may not fully capture the complexity, heterogeneity, and noise characteristics of real clinical environments. Because the target-generation mechanism did not explicitly incorporate strong nonlinear interactions, the advantage of advanced ensemble methods may be understated. In real healthcare datasets, interactions among features may be far more complex, potentially increasing the utility of models such as XGBoost, CatBoost, or RF.

CLASSIFICATION

Comparative Performance of Classification Models

The comparative evaluation of the multiclass classification models revealed notable differences between overall accuracy and class-balanced performance. Based on the obtained results, the models were primarily compared using balanced accuracy, which is a more informative metric than raw accuracy in the presence of class imbalance. Although all models achieved relatively high overall accuracy values, ranging from 0,900 to 0,924, their balanced accuracy values were substantially lower, ranging from 0,598 to 0,726, Figure 5. This discrepancy indicates that the models generally performed well on the majority class but showed reduced effectiveness when predicting minority classes.

Among the evaluated models, LogR demonstrated the best overall performance when class balance was taken into account. Despite achieving the lowest raw accuracy (0,900), it obtained the highest balanced accuracy (0,726) and the highest macro F1 score (0,649), indicating the most consistent performance across all three patient satisfaction categories. SVM achieved a slightly higher raw accuracy (0,910) but lower balanced accuracy (0,691) and a nearly identical macro F1 score (0,645), suggesting that it represents a competitive but slightly less balanced alternative. RF and HGB (denoted as HGB at Figure 5 for simplicity) achieved the highest raw accuracy values (0,915 and 0,924, respectively), but their substantially lower balanced accuracy scores (0,629 and 0,598) indicate that these models were more strongly influenced by the majority class and less effective in distinguishing minority categories.

The contrast between accuracy and balanced accuracy highlights the presence of a misleading accuracy effect in this multiclass setting. In particular, HGB achieved the highest overall accuracy, yet it produced the lowest balanced accuracy among all evaluated models. This pattern suggests that the model was highly effective at identifying the dominant class but considerably less successful in recognizing the less frequent classes. By contrast, LogR sacrificed a small amount of raw accuracy in exchange for markedly better class-balanced performance, which is particularly important in healthcare-related applications where minority class detection may carry substantial practical significance.

The cross-validation results provided additional insight into model stability and generalizability. The highest cross-validated macro F1 score was observed for HGB (0,685), which exceeded its holdout test macro F1 score (0,605). This discrepancy suggests that the model may be sensitive to the particular train-test split used in the holdout evaluation and may potentially benefit from further hyperparameter optimization or repeated resampling. In contrast, LogR and SVM showed close agreement between their holdout and cross-validated macro F1-scores, indicating more stable and consistent performance across different subsets of the data. RF also showed relatively stable cross-validated performance, although its lower class-specific sensitivity limited its practical effectiveness.

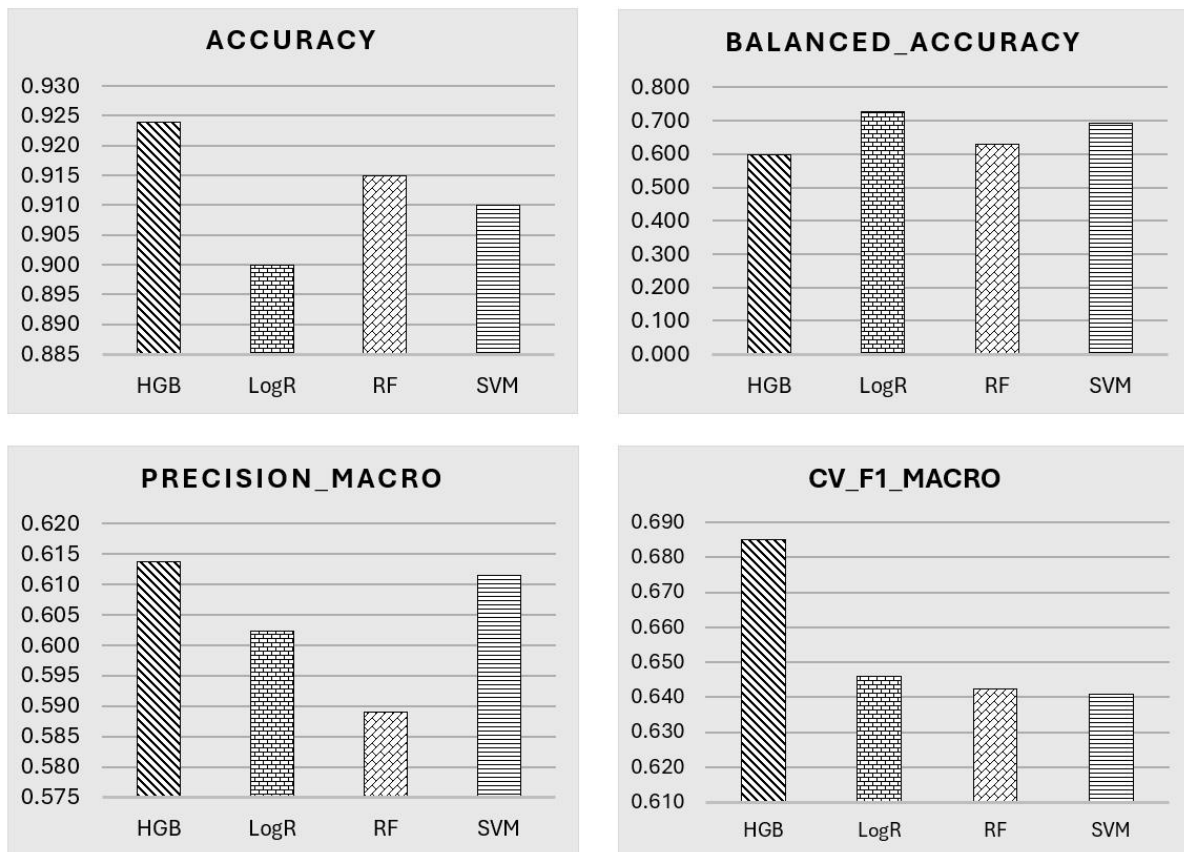


Figure 5. Classification models' performances.

From a practical perspective, LogR emerged as the most suitable model for the current dataset because it provided the best compromise between overall predictive performance and equitable treatment of all classes. While SVM represented a strong alternative with comparable macro-level performance, RF and HGB appeared to be more strongly biased toward the majority class. Therefore, when balanced multiclass prediction is prioritized over raw accuracy, LogR should be considered the preferred model.

Confusion Matrix Analysis Across Models

The multiclass classification models were evaluated on the Low, Medium, and High patient satisfaction classes, revealing consistent patterns of performance and misclassification across all four algorithms (presented in Figure 6). The High class was the most reliably predicted category across all models due to its dominance in the dataset. LogR correctly identified 828 out of 889 High cases, achieving 93.1% recall, while HGB improved this further by detecting 868 out of 889 cases, corresponding to 97.6% recall. RF correctly predicted 842 High cases, yielding 94.7% recall, and SVM performed similarly, with 844 High cases correctly classified at 94.9% recall. Precision was consistently high across models, indicating that nearly all predicted High cases were correct. Misclassifications from High mostly occurred into the adjacent Medium class, with very few instances predicted as Low, highlighting that the models primarily confuse neighboring satisfaction levels rather than extremes.

The Medium class proved to be the most challenging to predict consistently. LogR correctly identified 57 out of 85 Medium cases, resulting in 67.1% recall, but misclassifications were significant, with 57 Medium cases predicted as High and 11 as Low. HGB struggled similarly, correctly predicting 50 Medium cases, achieving 58.8% recall, with misclassifications equally split between High and Low. RF improved recall for Medium to 82.4%, correctly classifying 70 cases, although nearly half of the predicted Medium cases were

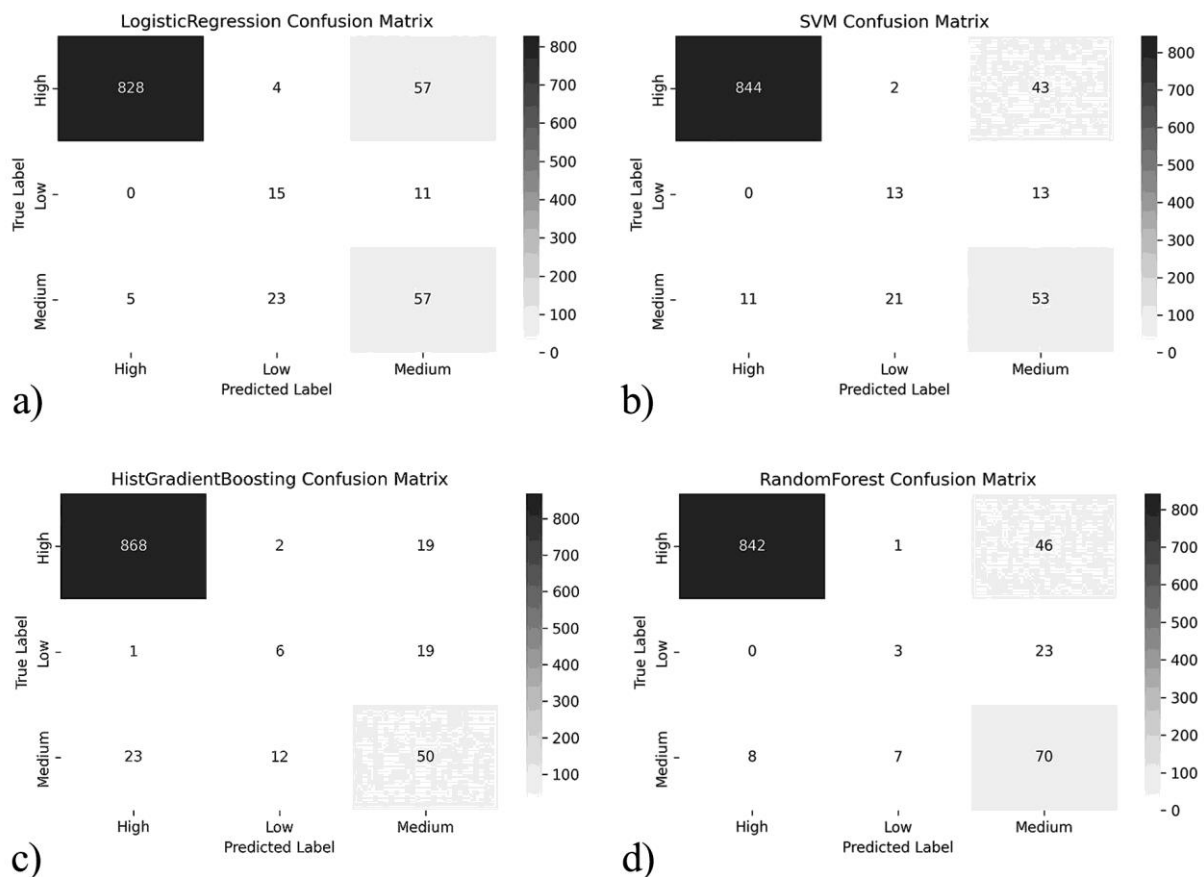


Figure 6. Confusion matrices: a) LogR; b) SVM; c) HGBBoost; d) RF.

actually High or Low, indicating a high noise level in this category. SVM achieved a moderate recall of 62,4%, with errors distributed more broadly; Medium cases were misclassified as both High and Low, showing the model's conservative approach to this middle category.

The Low class remained the most difficult for all models due to extreme class imbalance, with only 26 samples available. LogR predicted 15 Low cases correctly, reaching 57,7% recall, whereas HGBBoost performed poorly with only six correct predictions, achieving 23,1% recall. RF struggled the most, identifying just three Low cases, corresponding to 11,5% recall, effectively ignoring this minority class. SVM showed marked improvement for Low, detecting 13 cases, resulting in 50% recall, demonstrating its ability to handle minority clusters more effectively through margin-based classification. Across all models, precision for the Low class was low, ranging from 27% to 36%, reflecting frequent false positives primarily from Medium predictions.

Several consistent trends emerge from the confusion matrices. Misclassifications mostly occurred between neighboring classes, High and Medium or Medium and Low, rather than between extreme categories, indicating overlapping feature spaces for adjacent satisfaction levels. The Low class consistently suffered the lowest recall across models, suggesting that data scarcity rather than algorithmic choice is the primary limitation. LogR offered a balanced approach, achieving the highest recall for the Low class among simpler models while maintaining reliable High class predictions. HGBBoost achieved the highest overall accuracy and recall for the High class but severely underperformed on the Low class. RF provided the best recall for the Medium class but failed almost completely on Low. SVM was the most balanced model, offering the best compromise for detecting Low and Medium cases without sacrificing High class performance.

Given these patterns, the choice of model depends on the priority of class-specific detection. If the primary goal is minority detection, especially for the Low class, SVM or LogR are

preferable. For majority detection, HGBost provides the highest recall for the High class. RF achieves the highest recall for the Medium class, but with higher misclassification noise.

SHAP Analysis and Model Interpretability

The SHAP analyses reveal the internal logic behind each model's predictions, explaining why the Low and Medium classes are the most difficult to classify. Recovery time emerged as a central feature across all models, acting as a critical discriminator between High and Low predictions. High recovery times push predictions toward Low, while low recovery times push predictions toward High. This explains the models' strong performance on the High class but their struggles with the Medium class, which lacks a clear discriminative feature.

For the Low class, SHAP analysis across LogR, SVM, HGBost and RF (Figure 7) consistently identified Recovery_Time as the most influential predictor.

Additional recurrent contributors included AI_Diagnosis_Confidence, Age, Total_Risk_Score, and Comorbidity_Count, indicating that recovery dynamics and overall clinical burden are central to Low-class classification. LogR emphasized continuous numerical predictors with broad linear effects, whereas HGBost revealed stronger nonlinear and categorical interactions, particularly involving insurance type and allergy history. RF highlighted diagnostic and procedural features such as X-ray findings and surgery type, suggesting subgroup-based decision

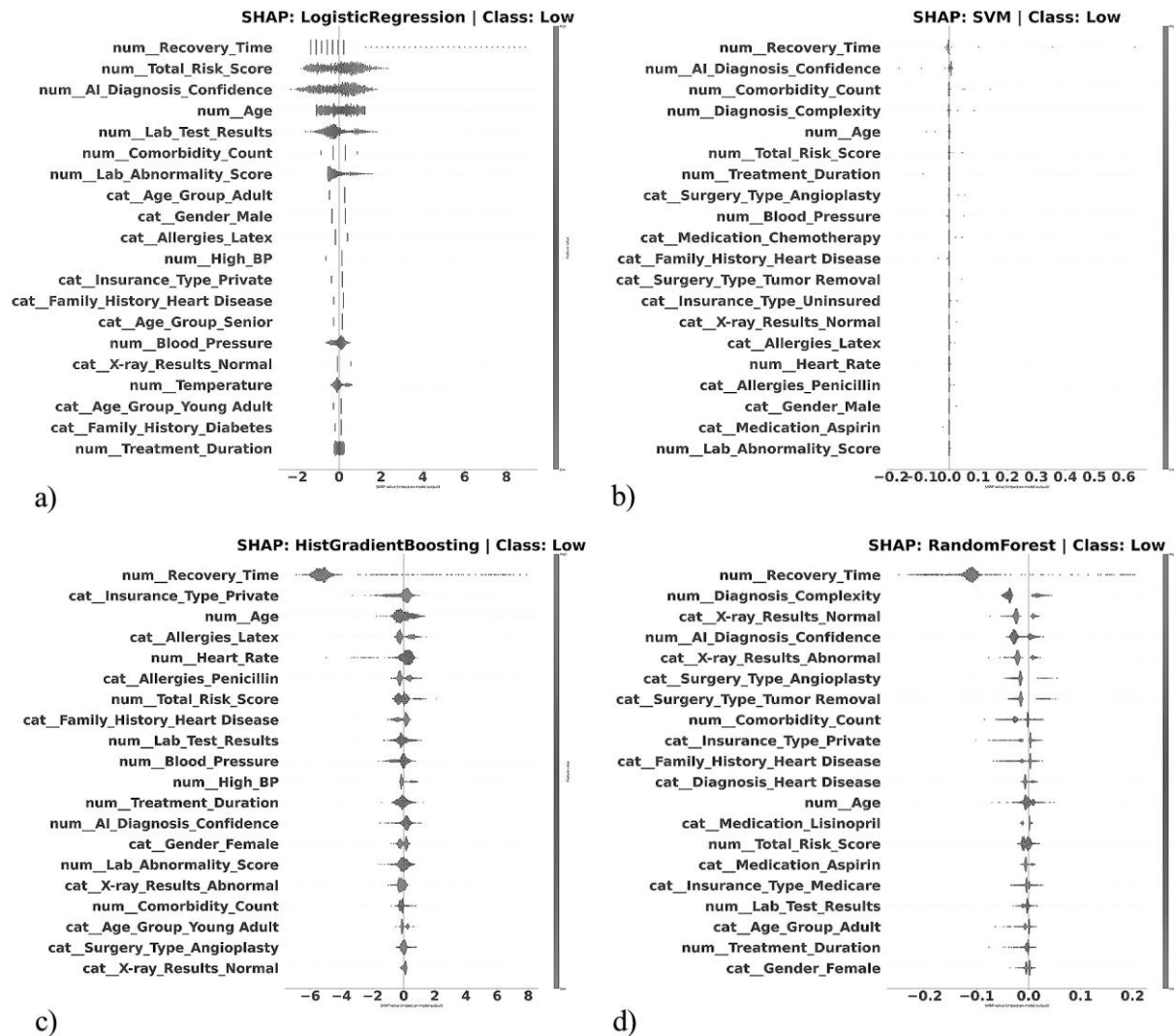


Figure 7. SHAP feature contributions for Low Patient Satisfaction Class predictions: (a) LogR; (b) SVM; (c) HGBost; and (d) RF.

patterns. In contrast, SVM exhibited highly compressed SHAP values centered near zero, indicating comparatively weaker feature attribution strength and reduced interpretability for this class. Overall, HGBost provided the most informative and clinically nuanced explanation for the Low class, while LogR offered the clearest linear interpretability.

For the Medium Patient_Satisfaction_Class, the SHAP plots (Figure 8) indicate that predictions are less dominated by a single feature than in the Low class, although Recovery_Time and Diagnosis_Complexity consistently emerge as major contributors across multiple models.

Compared with the Low class, the Medium class appears to be influenced by a broader combination of recovery-related, diagnostic, physiological, and laboratory variables, which is expected because the Medium category often represents an intermediate or overlapping decision region between the extreme classes. Across the four models, LogR primarily emphasizes numeric clinical severity variables with a moderate spread of SHAP values, suggesting that the model captures broad linear relationships among patient condition indicators. In contrast, SVM again shows compressed SHAP values, indicating weaker or less expressive feature attribution despite identifying similar clinically relevant predictors. Among the models, HGBost provides the richest and most balanced explanatory structure, capturing a wider range of influential features and reflecting more complex nonlinear interactions. Meanwhile, RF highlights the importance of recovery-related variables, diagnosis complexity, and insurance status, reflecting its ensemble nature.

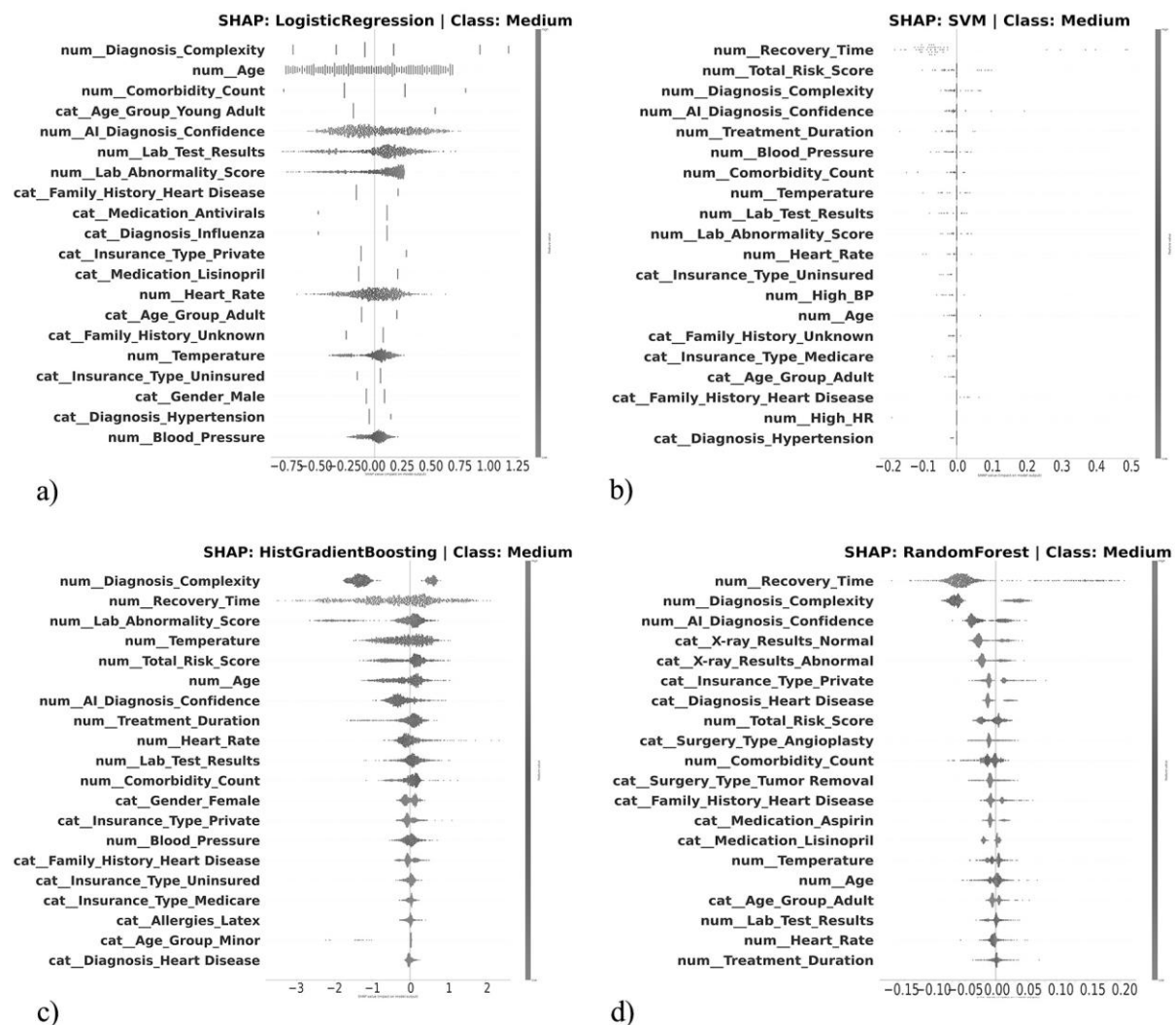


Figure 8. SHAP feature contributions for Medium Patient Satisfaction Class predictions: (a) LogR; (b) SVM; (c) HGBost; and (d) RF.

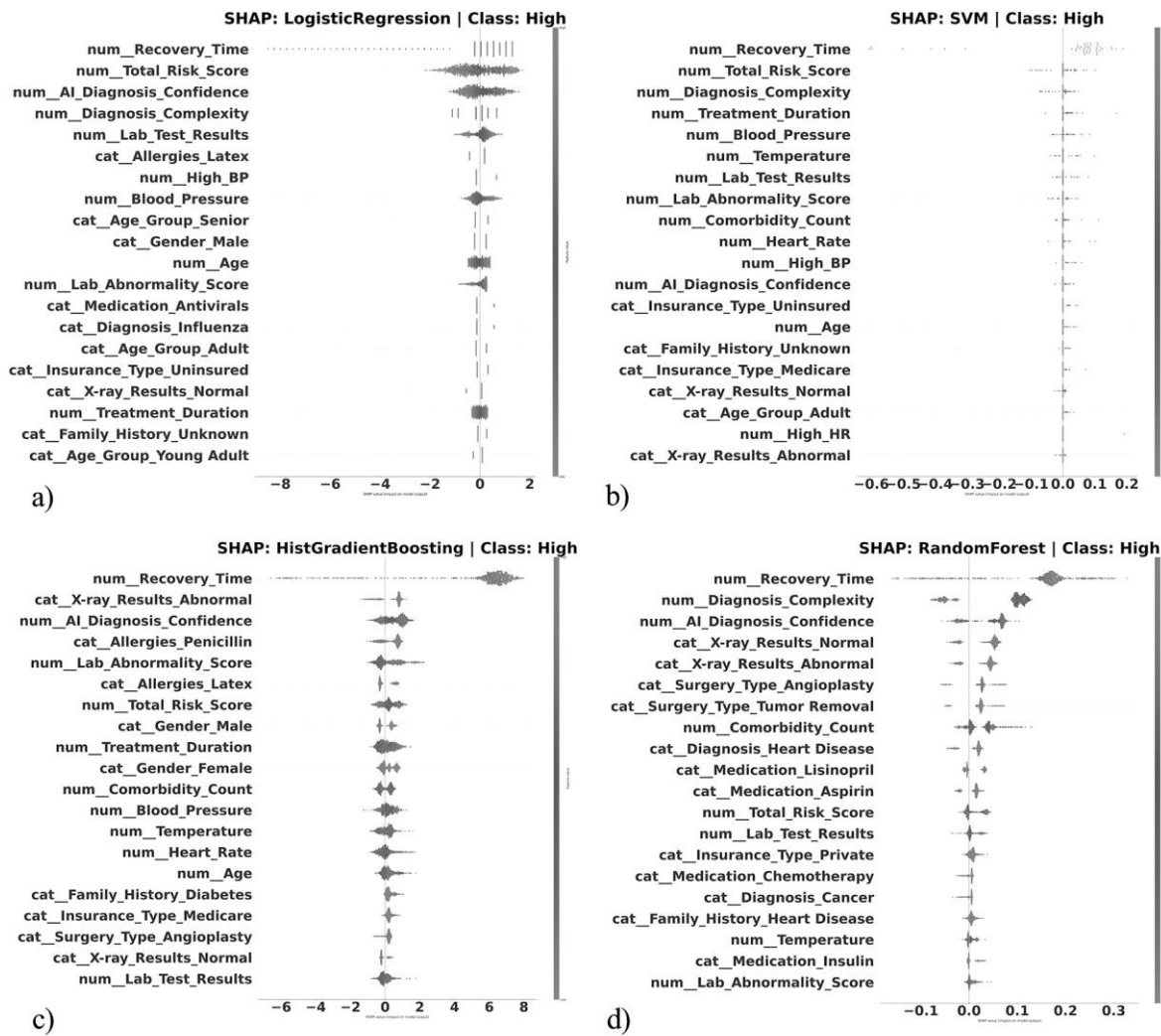


Figure 9. SHAP feature contributions for High Patient Satisfaction Class predictions: (a) LogR; (b) SVM; (c) HGB; and (d) RF.

imaging findings, and AI confidence measures, although its feature contributions appear more localized and distributed across decision paths than those observed in HGB.

For the High class, SHAP analysis across LogR, SVM, HGB, and RF (Figure 9) consistently identified `num_Recovery_Time` as the most influential predictor. Additional recurrent contributors included `num_Total_Risk_Score`, `num_AI_Diagnosis_Confidence`, and `num_Diagnosis_Complexity`, indicating that clinical severity and diagnostic certainty are central to High-class classification.

LogR emphasized continuous numerical predictors with broad linear effects, particularly highlighting the impact of total risk scores and AI confidence levels in a symmetrical distribution. In contrast, HGB revealed significantly stronger nonlinear and categorical interactions, displaying the highest SHAP magnitudes for recovery time and identifying `cat_X-ray_Results_Abnormal` as a critical driver for the High class. RF highlighted diagnostic and procedural complexity, prioritizing features like `num_Diagnosis_Complexity` and `cat_X-ray_Results_Abnormal`, which suggests a decision-making pattern based on clinical subgroups and specific procedural markers. In contrast, SVM exhibited highly compressed SHAP values centered near zero, indicating comparatively weaker feature attribution strength and reduced interpretability for this class. Overall, HGB provided the most informative and clinically nuanced explanation for the High class by capturing extreme predictive values, while LogR offered the clearest linear interpretability of the underlying risk factors.

Patient satisfaction classification is primarily driven by Recovery_Time along with AI_Diagnosis_Confidence, Total_Risk_Score, Diagnosis_Complexity, Age, and Comorbidity_Count, highlighting the central role of recovery dynamics and clinical severity. Among models, HGBost provided the most nuanced and informative explanations, especially for the High class, while LogR offered clear linear interpretability. RF captured Medium cases effectively through diverse feature interactions, and SVM provided comparatively fair detection of the Low class despite weaker feature attribution. Challenges remain in separating Low and Medium classes due to feature overlap and class imbalance. Targeted oversampling, feature engineering such as a combined Severity Index, and probability calibration can further improve performance and interpretability across models. Finally, if the goal is to accurately capture minority cases (Low), SVM is the most reliable choice. RF can serve as a secondary option when distinguishing Medium cases is critical. HGBost remains ideal for maximizing High class accuracy but should be combined with oversampling or class weighting to avoid ignoring Low cases. LogR offers interpretability and fairness but struggles with nonlinear overlaps. Overall, SHAP analyses clarify why these models behave differently and guide the choice of models and feature engineering strategies to optimize multiclass patient satisfaction predictions.

Class Separation and Model Confidence Patterns

Examining the confidence distributions via the confidence distribution boxplot (Figure 10) further clarifies model behavior.

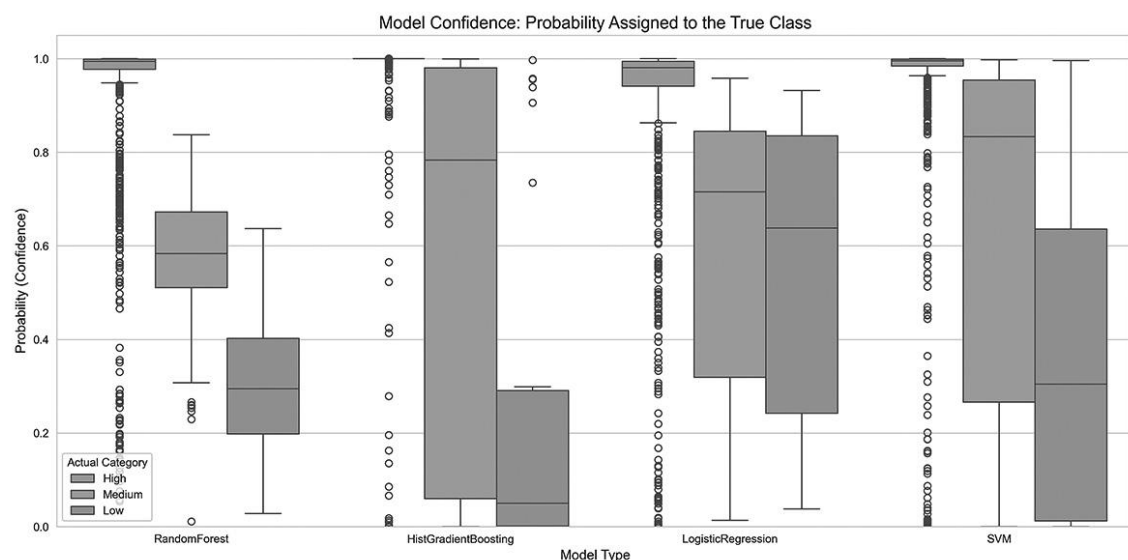


Figure 10. Model confidence probability.

The confidence distributions shown in Figure 10 reinforce the interpretability findings observed in the SHAP analyses. The RF model shows a very distinct stratification between its predictions. For the High class, it is exceptionally confident, with almost all data points clustered near a probability of 1,0, though it does produce a long tail of outliers reaching down to zero. The Medium and Low classes show a clear step-down in confidence, the median probability for Medium sits around 0,6, while the Low class median drops to approximately 0,3. This suggests that while the model is decisive for high-certainty cases, it maintains a structured, lower-confidence threshold for less certain classifications. HGBost displays much higher variance and less stability compared to RF. While the High class still maintains a median near 1.0, it features a significantly wider spread and a dense collection of outliers across the entire probability range. Interestingly, the Medium class is extremely stretched, covering nearly the entire y-axis from 0,0 to 1,0, indicating the model often struggles to differentiate medium-confidence cases from extremes. The Low class shows very low confidence, with the bulk of

the distribution and the median resting near 0,05. The LogR model appears highly “overconfident” or polarized. The High and Medium classes show median probabilities very close to 1,0 and 0,8, though the Medium class has a much larger interquartile range (IQR), stretching down toward 0,3. This indicates that the model frequently assigns high probabilities even to samples that may not strictly belong to the High class. The Low class similarly shows high variance, with a median around 0,65, suggesting this model generally struggles to produce low-probability scores for true classes, leading to less granular separation. The SVM exhibits a behavior similar to LogR but with even more pronounced extremes. The High class is pinned almost exclusively at 1,0 with very few outliers. However, the Medium class shows a massive spread, with a median near 0,8 but whiskers and a box that extend nearly to the bottom of the scale. The Low class for SVM is also widely distributed, with a median around 0,3. Like the gradient boosting model, the SVM seems to struggle with the Medium middle ground, often swinging between high certainty and total uncertainty. The results indicate that model interpretability and predictive confidence do not always align directly, and both should be considered when evaluating classifier performance in multi-class patient satisfaction prediction.

Strategic Implications for Model Choice

Based on performance metrics, confusion matrices, SHAP analyses, and predicted probabilities, LogR is the preferred model for this multiclass patient satisfaction task. While HGBost achieves the highest overall accuracy, it is heavily biased toward the High class, struggles with Low and Medium cases, and assigns near-zero confidence to many Low instances, demonstrating poor reliability for minority classes. RF performs better on Medium cases but still underperforms on Low, and its confidence distributions show moderate overlap between classes.

LogR provides the most balanced performance, achieving the highest balanced accuracy and F1 macro score while maintaining calibrated confidence across all classes. It effectively captures Low and Medium cases and is interpretable via SHAP, with smooth linear transitions between classes. SVM is a close second, showing higher sensitivity for Low cases, but overall metrics and calibration slightly favor LogR.

The superior performance of LogR and SVM is explained by the use of SMOTE, which creates synthetic minority samples. Linear and kernel-based models can learn optimal decision boundaries in this augmented space, whereas tree-based models overfit to synthetic points and fail to generalize. LogR’s balanced predictions, reliability, and interpretability make it the optimal choice for fair and robust multiclass classification.

CONCLUSION

LR proved to be the most accurate, stable, and interpretable model for predicting AI_diagnostic_confidence in this synthetic dataset, outperforming more complex nonlinear and ensemble based models such as CatBoost, RF, LGBM, and XGBoost. Its strong performance aligns with the linear structure of the data generating process, while SHAP analyses confirmed that the most influential features, Diagnosis_Complexity, Comorbidity_Count, and Lab_Abnormality_Score, affected predictions as expected. These results demonstrate that a simple, interpretable model can provide both highly predictive accuracy and clear feature-level insights in confidence aware AI applications. The experimental results indicate that the proposed regression framework effectively models AI_diagnostic_confidence with high accuracy. SHAP-based explainability analyses offer transparent and clinically intuitive insights into model behavior, demonstrating a strong alignment between the learned model structure and the known synthetic target-generation mechanism. This alignment supports the methodological validity of the study and underscores the potential of XAI for evaluating confidence-aware AI systems in healthcare.

For multiclass patient satisfaction prediction, LogR proves to be the most robust and fair model. It balances overall accuracy with performance on minority classes, effectively utilizes the SMOTE-augmented dataset, and offers interpretable predictions via SHAP. SVM serve as a viable alternative when accurately identifying rare Low cases is critical. In contrast, HGBost and RF models are less reliable without additional adjustments. Enhancing feature engineering and optimizing classification thresholds could improve recognition of Medium and Low classes while maintaining fairness and interpretability. Overall, the evaluation shows that balanced, interpretable models with calibrated probability outputs outperform less transparent or biased methods in imbalanced multiclass settings. Combining LogR with carefully engineered features and probability-based monitoring provides the most reliable and actionable approach for deployment.

Although these findings are derived from a synthetic dataset, they provide a valuable proof-of-concept for future research on interpretable clinical AI and confidence-informed decision support. Future studies should extend this approach to real-world clinical datasets or more sophisticated simulation scenarios to assess generalizability.

REFERENCES

- [1] Marjanovic, U.; Mester, G. and Milic Marjanovic B.: *Assessing the success of artificial intelligence tools: An evaluation of ChatGPT using the information system success model*. Interdisciplinary Description of Complex Systems **22**(3), 266-275, 2024, <http://dx.doi.org/10.7906/indecs.22.3.3>,
- [2] Khosravi, M.; Zare, Z.; Mojtabaian, S.M. and Izadi, R.: *Artificial Intelligence and Decision-Making in Healthcare: A Thematic Analysis of a Systematic Review of Reviews*. Health Services Research and Managerial Epidemiology **11**, 2024, <http://dx.doi.org/10.1177/23333928241234863>,
- [3] Nouis, S.C.E.; Uren, V. and Jariwala, S.: *Evaluating accountability, transparency, and bias in AI-assisted healthcare decision-making: a qualitative study of healthcare professionals' perspectives in the UK*. BMC Medical Ethics **26**(1), No. 89, 2025, <http://dx.doi.org/10.1186/s12910-025-01243-z>,
- [4] Yu, Y., et al.: *Enhancing clinician trust in AI diagnostics: A dynamic framework for confidence calibration and transparency*. Diagnostics **15**(17), No. 2204, 2025, <http://dx.doi.org/10.3390/diagnostics15172204>,
- [5] Tan, J.; Wen, H.J. and Awad, N.: *Health care and services delivery systems as complex adaptive systems*. Communications of the ACM **48**(5), 36-44, 2005, <http://dx.doi.org/10.1145/1060710.1060737>,
- [6] Gillner, S.: *We're implementing AI now, so why not ask us what to do? - How AI providers perceive and navigate the spread of diagnostic AI in complex healthcare systems*. Social Science & Medicine **340**, No. 116442, 2024, <http://dx.doi.org/10.1016/j.socscimed.2023.116442>,
- [7] Olawade, D.B., et al.: *Human in the loop artificial intelligence in healthcare: applications, outcomes, and implementation challenges*. International Journal of Medical Informatics **213**, No. 106362, 2026, <http://dx.doi.org/10.1016/j.ijmedinf.2026.106362>,
- [8] Jiang, T.; Sun, Z.; Fu, S. and Lv, Y.: *Human-AI interaction research agenda: A user-centered perspective*. Data and Information Management **8**(4), No. 100078, 2024, <http://dx.doi.org/10.1016/j.dim.2024.100078>,

- [9] Wekenborg, M.K.; Gilbert, S. and Kather, J.N.: *Examining human-AI interaction in real-world healthcare beyond the laboratory*. npj Digital Medicine **8**, No. 169, 2025, <http://dx.doi.org/10.1038/s41746-025-01559-5>,
- [10] Nasef, D., et al.: *Integrating artificial intelligence in clinical practice, hospital management, and health policy: Literature review*. Journal of Hospital Management and Health Policy **9**, No. 20, 2025, <http://dx.doi.org/10.21037/jhmhp-24-138>,
- [11] Chaturvedi, D.; Baghel Chauhan, S. and Singh, I.: *The impact of artificial intelligence on remote healthcare: Enhancing patient engagement, connectivity, and overcoming challenges*. Intelligent Pharmacy **3**(5), 323-329, 2025, <http://dx.doi.org/10.1016/j.ipha.2024.12.003>,
- [12] Apell, P. and Eriksson, H.: *Artificial intelligence (AI) healthcare technology innovations: the current state and challenges from a life science industry perspective*. Technology Analysis & Strategic Management **35**(2), 179-193, 2023, <http://dx.doi.org/10.1080/09537325.2021.1971188>,
- [13] Olawade, D.B. et al.: *Artificial intelligence in healthcare delivery: Prospects and pitfalls*. Journal of Medicine, Surgery, and Public Health **3**, No. 100108, 2024, <http://dx.doi.org/10.1016/j.glmedi.2024.100108>,
- [14] Maleki Varnosfaderani, S. and Forouzanfar, M.: *The Role of AI in Hospitals and Clinics: Transforming Healthcare in the 21st Century*. Bioengineering **11**(4), No. 337, 2024, <http://dx.doi.org/10.3390/bioengineering11040337>,
- [15] Islam, R.; Sultana, A. and Islam, M.R.: *A comprehensive review for chronic disease prediction using machine learning algorithms*. Journal of Electrical Systems and Information Technology **11**, No. 27, 2024, <http://dx.doi.org/10.1186/s43067-024-00150-4>,
- [16] Gabriel, J.: *A Machine learning-based Web application for heart disease prediction*. Intelligent Control and Automation **15**(1), 9-27, 2024, <http://dx.doi.org/10.4236/ica.2024.151002>,
- [17] Bagul, V., et al.: *Multiple disease prediction using machine learning*. International Journal of Innovative Science and Research Technology **9**(4), 1155-1158, 2024, <http://dx.doi.org/10.38124/ijisrt/IJSRT24APR1453>,
- [18] Mohamed, A.; Abdelrehim, M. and Al-Barazie, R.: *Context matters in machine learning based disease prediction with insights from diverse clinical and symptom data*. Scientific Reports **15**, No. 42669, 2025, <http://dx.doi.org/10.1038/s41598-025-26855-8>,
- [19] Alanazi, R.: *Identification and prediction of chronic diseases using machine learning approach*. Journal of Healthcare Engineering **2022**, No. 2826127, 2022, <http://dx.doi.org/10.1155/2022/2826127>,
- [20] Rajesh, E.; et al.: *Machine learning for online automatic prediction of common disease attributes using never-ending image learner*. Diagnostics **13**(1), No. 95, 2023, <http://dx.doi.org/10.3390/diagnostics13010095>,
- [21] Tu, T., et al.: *Towards conversational diagnostic artificial intelligence*. Nature **642**, 442-450, 2025, <http://dx.doi.org/10.1038/s41586-025-08866-7>,
- [22] Johannssen, A. and Chukhrova, N.: *The crucial role of explainable artificial intelligence (XAI) in improving health care management*. Health Care Management Science **28**(3), 565-570, 2025, <http://dx.doi.org/10.1007/s10729-025-09720-y>,
- [23] Sadeghi, Z., et al.: *A review of Explainable Artificial Intelligence in healthcare*. Computers and Electrical Engineering **118**(Part A), No. 109370, 2024, <http://dx.doi.org/10.1016/j.compeleceng.2024.109370>,

- [24] Adeniran, A.A.; Onebunne, A.P. and William, P.: *Explainable AI (XAI) in healthcare: Enhancing trust and transparency in critical decision-making*. World Journal of Advanced Research and Reviews **23**(03), 2647-2658, 2024, <http://dx.doi.org/10.30574/wjarr.2024.23.3.2936>,
- [25] Al-Antari, M.A.: *Artificial intelligence for medical diagnostics-existing and future AI technology!* Diagnostics **13**(4), No. 688, 2023, <http://dx.doi.org/10.3390/diagnostics13040688>,
- [26] Alam, L. and Mueller, S.: *Examining the effect of explanation on satisfaction and trust in AI diagnostic systems*. BMC Medical Informatics and Decision Making **21**(1), No. 178, 2021, <http://dx.doi.org/10.1186/s12911-021-01542-6>,
- [27] Asan, O.; Bayrak, A.E. and Choudhury, A.: *Artificial intelligence and human trust in healthcare: focus on clinicians*. Journal of Medical Internet Research **22**(6), No. e15154, 2020, <http://dx.doi.org/10.2196/15154>,
- [28] Mertz, M., et al.: *Exploring trust factors in AI-healthcare integration: a rapid review*. Frontiers in Artificial Intelligence **8**, No. 1658510, 2025, <http://dx.doi.org/10.3389/frai.2025.1658510>,
- [29] Foresman, G., et al.: *Patient perspectives on artificial intelligence in health care: Focus group study for diagnostic communication and tool implementation*. Journal of Participatory Medicine **17**, No. e69564, 2025, <http://dx.doi.org/10.2196/69564>,
- [30] Khullar, D., et al.: *Perspectives of patients about artificial intelligence in health care*. JAMA Network Open **5**(5), No. e2210309, 2022, <http://dx.doi.org/10.1001/jamanetworkopen.2022.10309>,
- [31] Nong, P. and Platt, J.: *Patients' Trust in Health Systems to use Artificial Intelligence*. JAMA Network Open **8**(2), No. e2460628, 2025, <http://dx.doi.org/10.1001/jamanetworkopen.2024.60628>,
- [32] Kücking, F., et al.: *Impact of AI recommendation correctness on diagnostic accuracy in clinical decision-making*. International Journal of Medical Informatics **207**, No. 106223, 2026, <http://dx.doi.org/10.1016/j.ijmedinf.2025.106223>,
- [33] Busch, F., et al.: *Multinational attitudes toward AI in health care and diagnostics among hospital patients*. JAMA Network Open **8**(6), No. e2514452, 2025, <http://dx.doi.org/10.1001/jamanetworkopen.2025.14452>,
- [34] Arora, A., et al.: *The value of standards for health datasets in artificial intelligence-based applications*. Nature Medicine **29**(11), 2929-2938, 2023, <http://dx.doi.org/10.1038/s41591-023-02608-w>,
- [35] Chatterjee, P.; Cymberknop, L.J. and Armentano, R.L.: *Nonlinear Systems in Healthcare towards Intelligent Disease Prediction*. In: Legnani, W. and Moschandreou, T.E., eds.: *Nonlinear Systems - Theoretical Aspects and Recent Applications*. IntechOpen, 2020, <http://dx.doi.org/10.5772/intechopen.88163>,
- [36] -: *Synthetic dataset for AI in healthcare*. <https://www.kaggle.com/datasets/smmmmmmmmmmmm/synthetic-dataset-for-ai-in-healthcare>.