Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00217-5

Original Article

Interpretable Machine Learning for Mortality Risk Stratification in Hantavirus Infection: A Global Study Across Hemorrhagic Fever with Renal Syndrome and Hantavirus Cardiopulmonary Syndrome

Yetunde O. Enigbokan [1] , Akeem A. Abiona [1] , Micheal O. Ajinaja [1, ✉]

1. *Department of Computer Science, Federal Polytechnic Ile Oluji, Ile Oluji, NGA*

Received: May 31, 2026 | Review began: June 28, 2026 | Review ended: July 30, 2026 | Published: August 11, 2026

Abstract

Hantavirus infection remains a rare but potentially fatal zoonosis, and early identification of patients at highest risk of death is essential for timely triage and resource allocation. This study developed an interpretable machine learning (ML) framework for mortality risk stratification using a global hantavirus epidemiology dataset. The dataset included both Hemorrhagic Fever with Renal Syndrome (HFRS) and Hantavirus Cardiopulmonary Syndrome (HCPS) cases. The study conducted a retrospective supervised learning analysis on 10,000 patient records, including demographic, clinical, epidemiological, and treatment variables. The binary outcome was mortality. Preprocessing included identifier removal, missing-value handling, categorical and symptom encoding, feature selection, and class-imbalance correction. Logistic regression, random forest, and extreme gradient boosting (XGBoost) models were trained and compared on a held-out test set using receiver operating characteristic-area under the curve (ROC-AUC), accuracy, precision, recall, F1-score, and Brier score. Model interpretation was planned using SHapley Additive exPlanations-based feature attribution. The cohort included 8,938 recovered and 1,062 deceased cases (mortality rate: 10.62%), comprising 6,460 HFRS cases (64.6%) and 3,540 HCPS cases (35.4%). Logistic regression achieved the highest discrimination, with an ROC-AUC of 0.858 and the highest recall for mortality detection (0.741), but modest precision (0.321) and weaker calibration (Brier score 0.119). Gradient boosting showed comparable discrimination (AUC 0.857) with better precision (0.511) and calibration (Brier score 0.076). Random forest performed less well for mortality detection, with markedly low recall (0.085). The study trained and compared interpretable classifiers using 10,000 global hantavirus cases representing both HFRS (64.6%) and HCPS (35.4%) presentations, and reported pooled performance as well as syndrome-stratified results. Severity, syndrome type, viral load category, age, and geographic setting were the most informative predictors. Mortality risk in hantavirus infection can be modeled using routine clinical and epidemiological features, but clinical deployment should prioritize sensitivity, calibration, and transparency over accuracy alone. These findings support the use of interpretable ML as a practical framework for early hantavirus risk stratification and external validation in independent cohorts.

Categories: AI/ML-based decision support systems, Explainable AI, Machine Learning (ML)

Keywords: hantavirus, mortality prediction, machine learning, interpretability, shap

Introduction

Hantaviruses are emerging zoonotic pathogens of significant global public health concern, collectively responsible for an estimated 150,000 to 200,000 human infections annually [1]. They belong to the family *Hantaviridae*, order *Bunyavirales*, and are broadly classified as Old World hantaviruses, primarily causing Hemorrhagic Fever with Renal Syndrome (HFRS) across Asia and Europe, and New World hantaviruses, responsible for Hantavirus Cardiopulmonary Syndrome (HCPS) in

the Americas [2]. In the historical North American literature, this syndrome has often been referred to as Hantavirus Pulmonary Syndrome (HPS); in this manuscript, we use the term HCPS consistently to refer to the same clinical entity. Among the New World strains, the Andes virus (ANDV) stands out as uniquely virulent; it is the only known hantavirus capable of sustained human-to-human transmission, occurring via respiratory droplets and prolonged close contact, with a median reproductive number exceeding 2 and an incubation period ranging from 9 to 40 days [3,4]. This singular epidemiological characteristic places ANDV at the forefront of pandemic preparedness concerns.

HCPS is the most clinically feared manifestation of Andes virus infection, characterized by acute respiratory failure, hemodynamic instability, acute kidney injury, hepatic involvement, and dysregulated cytokine release [5]. Case fatality rates (CFR) for HCPS consistently range between 30% and 50% globally, with reported CFRs reaching as high as 38% during the April 2026 multi-country cluster linked to cruise ship travel, in which 8 cases across 23 nationalities resulted in 3 deaths within weeks of symptom onset [6]. In clinical settings, disease severity correlates strongly with laboratory markers, including neutrophilia, leukocytosis, lymphopenia, thrombocytopenia, and elevated lactate dehydrogenase [7,8]. The rapid clinical deterioration observed in HCPS, where death can occur within a mean of 6.7 days from symptom onset, underscores the urgent need for early and accurate mortality risk stratification tools [9].

Despite the gravity of HCPS, no approved antiviral therapy or licensed vaccine currently exists for Andes virus infection worldwide [10]. Management remains entirely supportive, centered on oxygen supplementation, hemodynamic stabilization, and renal replacement therapy where indicated [11]. This therapeutic vacuum makes early clinical decision-making and risk stratification critically important - the ability to identify which patients are at greatest risk of fatal outcomes at the time of admission could enable timely escalation of care, optimal allocation of intensive care resources, and better coordination of multi-disciplinary interventions [12,13].

Machine learning (ML) has demonstrated substantial promise in mortality prediction across a range of critical illnesses, consistently outperforming traditional clinical scoring systems such as APACHE IV, SOFA, and SAPS [14]. Gradient boosting methods, particularly Extreme Gradient Boosting (XGBoost), have achieved area under the receiver operating characteristic curve (AUC) values exceeding 0.90 in intensive care unit (ICU) mortality prediction tasks, substantially surpassing conventional logistic regression models [15]. However, despite this progress, the application of ML to mortality prediction in Hantavirus disease - particularly HCPS caused by Andes virus - remains a critical and underexplored gap in the literature. Existing studies have applied ML to early warning systems for Puumala hantavirus outbreaks and to environmental risk mapping, but no published work has leveraged a global, multi-country epidemiological dataset to build interpretable ML models specifically targeting HCPS mortality risk stratification [16,17].

A key limitation of many ML-based clinical prediction studies is that high predictive accuracy alone is insufficient for clinical adoption; clinicians require transparency in model reasoning to trust and act on predictions [18]. SHapley Additive exPlanations (SHAP), grounded in cooperative game theory, have emerged as the leading framework for post-hoc model interpretability, enabling quantification of each feature's marginal contribution to individual predictions [19,20]. In ICU mortality prediction for pneumonia patients, SHAP-augmented XGBoost models demonstrated superior performance (AUC: 0.778) over traditional scoring systems while simultaneously identifying actionable clinical predictors such as aspartate aminotransferase levels, patient age, and albumin [21]. This approach bridges the longstanding gap between model accuracy and clinical transparency, directly addressing reproducibility and trustworthiness concerns raised in AI-augmented research.

Prior studies on hantavirus ML have focused primarily on environmental outbreak prediction, ecological surveillance, or regional infection-risk mapping rather than individual-level mortality prediction. For example, recent work on Puumala hantavirus used weather-based and ecological features to predict outbreak risk, not patient mortality, and did not address HCPS-specific clinical decision-making. In parallel, the broader clinical machine-learning literature contains many mortality-prediction studies with interpretable methods, but these are largely confined to other diseases and settings and do not provide a global HCPS mortality model. Accordingly, the present study addresses a distinct gap by applying interpretable ML to a multi-country hantavirus dataset for mortality risk stratification, with syndrome-stratified analysis and SHAP-based explanation of global and patient-level predictors.

This study addresses a clearly identified gap: the absence of interpretable, globally validated ML models for mortality risk stratification in hantavirus infection, encompassing both HFRS and HCPS. Leveraging the publicly available Hantavirus and Andes Virus Global Epidemiology dataset - a comprehensive multi-country repository encompassing confirmed HFRS and HCPS cases, clinical outcomes, and epidemiological variables across multiple continents - this study develops and compares multiple ML classifiers to predict hantavirus mortality across both syndromes. SHAP analysis is applied to each candidate model to identify and rank the most influential clinical and epidemiological predictors, providing transparent, human-verifiable explanations aligned with established pathophysiological mechanisms. Because the assembled dataset contains both Old-World HFRS and New-World HCPS presentations, all primary analyses are reported on the pooled dataset, but key results are also stratified by syndrome (HFRS vs HCPS) to assess syndrome-specific performance and predictor effects. To the best of our knowledge, this represents the first interpretable ML study applied to global HCPS mortality prediction using a publicly accessible epidemiological dataset. The findings are intended to support clinical decision-making in resource-limited settings, inform triage protocols in endemic and outbreak regions, and advance the scientific community's understanding of the dominant determinants of fatal HCPS outcomes.

Related works

Hantaviruses are globally distributed zoonotic RNA viruses that cause two clinically distinct syndromes: HFRS, predominant in Asia and Europe, and HCPS, endemic across the Americas [1]. The family *Hantaviridae* currently encompasses seven genera and 53 recognized species, transmitted principally through inhalation of aerosols contaminated by infected rodent excreta, without the involvement of arthropod vectors [2]. Globally, the disease burden is estimated at 150,000 to 200,000 cases annually, with case fatality rates ranging from less than 1% in Puumala virus-associated nephropathia epidemica to over 40% in HCPS caused by Sin Nombre virus and Andes virus. In Europe, 28 EU/EEA countries reported 1,885 confirmed cases in 2023, representing a notification rate of 0.4 cases per 100,000 population [22].

Among all hantavirus species, ANDV carries unique pandemic potential due to its documented capacity for sustained human-to-human transmission, a characteristic not observed in any other orthohantavirus [3]. Transmitted via respiratory droplets and prolonged close contact, ANDV has an incubation period ranging from 9 to 40 days, with a median reproductive number exceeding 2 in reported clusters [1]. The clinical course of HCPS progresses through prodromal, cardiopulmonary, and convalescent phases, with the cardiopulmonary phase characterized by non-cardiogenic pulmonary edema, hemodynamic collapse, and a median time to death of approximately 6.7 days from symptom onset. The lack of approved antiviral therapy or a licensed vaccine for ANDV infection makes identifying early mortality risk critical to patient survival. This urgency was underscored by the May 2026 multi-country cluster involving a cruise ship, where 3 of 8 identified cases died, prompting WHO to coordinate an international response [4].

Application of ML to hantavirus surveillance has so far focused primarily on environmental risk modeling and outbreak forecasting rather than on clinical outcome prediction. Kazasidis et al. [23] demonstrated that simple weather-based ML rules using precipitation and temperature data could accurately predict human Puumala orthohantavirus (PUUV) infection risk in Germany, achieving a straightforward early warning system with minimal variable sets. Building on this, the team [24] also developed a high-resolution early warning system for PUUV in Germany, linking multi-annual fluctuations in bank vole population sizes to human infection risk through ensemble modeling approaches. Most recently, Ferrara [25] developed an ensemble ML framework integrating environmental, ecological, and socioeconomic variables for prospective hantavirus risk prediction, achieving an AUC of 0.92 on independent test data from three United States jurisdictions, with SHAP analysis identifying precipitation variability, land-use patterns, and rodent species richness as dominant predictors [25]. However, all of these works addressed environmental and outbreak-level risk mapping and were geographically restricted; none applied ML to predict individual patient mortality risk from HCPS on a global multi-country dataset.

A systematic review of prognostic factors for mortality in New World hantavirus infection identified thrombocytopenia, elevated lactate dehydrogenase, respiratory failure, and absence of intensive care as consistent predictors of fatal outcomes [16]. Despite the availability of these clinically validated prognostic factors, no study to date has formalized

them into a ML-based mortality risk stratification model. Similarly, a prediction model built on epidemiological characteristics of hantavirus in China modeled HFRS incidence using time series approaches but did not address individual-level mortality prediction or utilize a globally representative dataset [26]. This gap in the literature identifies a clear need for a clinically oriented ML model targeting HCPS mortality specifically.

The application of interpretable ML to mortality prediction in severe viral and infectious diseases has gained significant traction in recent years, offering a methodological template directly transferable to HCPS. For Severe Fever with Thrombocytopenia Syndrome (SFTS), a tick-borne viral hemorrhagic fever with a case fatality rate of 20-30% - a profile closely analogous to HCPS - multiple studies have demonstrated the superiority of ML models over conventional scoring systems [27]. Xu et al. [28] developed an interpretable XGBoost-based prognostic model for SFTS across two independent cohorts (n = 560), achieving an external validation AUC of 0.911, with SHAP analysis identifying viral load, serum creatinine, D-dimer, and platelet count as the dominant predictors. Complementing this, Zhao et al. [29] validated a Random Forest model on 560 SFTS patients from two hospitals, achieving an external AUC of 0.825 and deploying an interactive web application for real-time risk prediction, demonstrating the translational pathway from ML model to clinical tool. The researchers [29] further extended this approach through a large-scale comparison of eight ML models for SFTS deterioration prediction in a multicenter study, with SHAP-derived feature interactions providing granular mechanistic insights.

For Crimean-Congo Hemorrhagic Fever (CCHF), another RNA hemorrhagic fever with high mortality and zoonotic transmission, Abdulhameed et al. [30] applied interpretable ML models to predict mortality risk during the 2022 Iraq outbreak, establishing a methodological precedent for applying SHAP-augmented classification to rare, high-fatality viral infections with limited sample sizes. This body of work collectively establishes that interpretable ML is not only feasible but clinically necessary for high-mortality zoonotic diseases. Yet HCPS caused by the Andes virus remains the most clinically severe and globally significant zoonotic hemorrhagic respiratory syndrome for which no comparable ML mortality model exists.

SHAP, grounded in cooperative game theory, have emerged as the standard for post-hoc interpretability in clinical prediction models, enabling quantification of each feature's marginal contribution to individual predictions in a mathematically principled manner [31]. Stenwig et al. [9] conducted a comparative analysis of interpretable ML models for hospital mortality across multiple algorithms, demonstrating that SHAP value analysis identifies meaningful differences in how models weigh clinical variables, with direct implications for healthcare decision-making. In ICU mortality prediction for pneumonia patients, an interpretable ML model with SHAP analysis achieved an AUC of 0.778 and identified aspartate aminotransferase, age, and albumin as the leading predictors, with SHAP waterfall plots enabling per-patient explanation of risk scores [10]. More recently, Shu et al. [31] combined SHAP with gradient boosting to predict mortality in hemodialysis patients, demonstrating the utility of SHAP in surfacing nonlinear feature interactions that conventional logistic regression models cannot capture [32]. An XGBoost-based ICU mortality model validated in a developing country setting achieved an AUC-ROC of 0.943, with SHAP values of 0.50, 0.41, and 0.37, respectively, assigned to hospital stay duration, albumin, and urea levels - demonstrating the cross-cultural generalizability of SHAP-based feature attribution [33].

Beyond individual models, the combination of SHAP with ensemble algorithms has been shown to improve both predictive performance and clinical trust. Zhou et al. [34] developed an XGBoost-based prognostic model for ICU-acquired bloodstream infections, achieving an AUROC of 0.92 on training and 0.85 on internal validation, with SHAP-identified features forming a simplified 10-variable model that retained strong performance across Gram-positive and Gram-negative subgroups. These findings demonstrate that SHAP not only improves interpretability but actively guides feature selection and model simplification - a critical advantage in clinical settings where parsimony and actionability are essential. The present study applies this proven methodology to HCPS mortality prediction for the first time, leveraging a global epidemiological dataset to develop and validate SHAP-augmented interpretable ML models that are clinically meaningful and globally representative.

Beyond individual disease applications, a broader literature has established the transformative role of AI in zoonotic disease management and epidemic surveillance. Pillai et al. [35] surveyed AI models for zoonotic pathogens across 70% of emerging infectious diseases attributed to animal origin, concluding that ML and deep learning consistently outperform traditional epidemiological models in identifying pathogen risk factors and transmission dynamics. Guo et al. [13] provided a comprehensive review of innovative AI applications for zoonotic disease management, documenting the use of AI for disease outbreak prediction, early diagnosis, and drug target identification across a range of zoonotic pathogens. In the domain of epidemic intelligence, Borham et al. [36] reviewed AI-driven surveillance systems for infectious diseases, highlighting how deep learning and real-time data integration facilitate earlier outbreak detection and efficient medical resource allocation. Collectively, this literature situates the present study within a rapidly maturing field where interpretable ML applied to a clinically critical, globally significant, and data-rich zoonotic disease represents a timely and necessary scientific contribution.

Materials And Methods

This study used a retrospective, supervised ML design to develop and validate models for mortality risk stratification among hantavirus-infected patients. The target variable was the binary clinical outcome, coded as Recovered = 0 and Deceased = 1, enabling a classification framework for mortality prediction. The study selected an interpretable modeling pipeline because the study goal was not only to maximize discrimination but also to identify clinically meaningful determinants of fatal outcome. The full workflow consisted of dataset inspection, preprocessing, feature engineering, feature selection, model development, interpretability analysis, and performance evaluation.

Dataset description

The dataset used in this study was publicly available on Kaggle [37]. The complete analysis code, including preprocessing, model training, feature selection, SHAP computation, and evaluation, is openly available in the project repository [38]. The dataset contained 10,000 anonymized patient-level observations and 23 variables spanning demographic, clinical, epidemiological, and treatment-related attributes. These variables included patient identifier, year, country, syndrome, age, sex, symptoms, incubation period, severity, hospitalization status, ICU admission, mechanical ventilation, extracorporeal membrane oxygenation (ECMO) use, dialysis, comorbidity, exposure type, outcome, length of stay, blood type, viral load category, days to diagnosis, treatment protocol, and geographic setting. Because the dataset included both Old-World HFRS and New-World HCPS presentations, the primary analysis was conducted on the pooled dataset. At the same time, key results were additionally stratified by syndrome (HFRS vs HCPS) to assess syndrome-specific performance and predictor effects. The dataset combined categorical variables such as country and exposure type, ordinal variables such as severity and viral load category, and numeric variables such as age and incubation period, making it appropriate for conventional ML and interpretable mortality prediction.

Syndrome-specific analyses

To evaluate whether predictive performance and feature importance differed between clinical syndromes, the study ran all modeling and interpretability steps on (a) the pooled dataset, (b) the HFRS-only subset, and (c) the HCPS-only subset. For stratified analyses, the study used identical preprocessing, feature-selection thresholds, and hyperparameter search spaces as in the pooled analysis, and reported syndrome-specific ROC-AUC, recall, precision, F1, and Brier score in the Results. Where sample size limited model stability ($n < 500$), the study reported metrics with 95% bootstrap confidence intervals to reflect uncertainty.

Prediction time point and leakage control

To avoid information leakage, variables that were downstream of the intended prediction time point or that directly reflected disease progression or treatment escalation were excluded from the primary model. These variables included hospitalization status, ICU admission, mechanical ventilation, ECMO use, dialysis, and length of stay. The primary model retained baseline demographic, clinical, epidemiological, and exposure variables that were plausibly available at the

intended prediction time point. Feature selection was performed sequentially using clinical exclusion, variance filtering, correlation filtering, mutual information ranking, recursive feature elimination, and SHAP-based refinement. All feature-selection steps were performed using training data only within cross-validation folds.

To ensure that the mortality prediction task reflected a clinically meaningful decision point, the analysis was framed as an early risk-stratification exercise based on variables assumed to be available at or near initial clinical assessment. Accordingly, variables that could directly encode downstream disease progression, treatment escalation, or post-admission management were reviewed for leakage risk before model training. Patient identifier was excluded a priori. In the revised analysis, ICU admission, mechanical ventilation, ECMO use, dialysis, and length of stay were treated as potentially leakage-prone because they may occur after the intended prediction time point and may reflect consequences of deterioration rather than baseline predictors of mortality. These variables were therefore excluded from the primary modeling pipeline and retained only for sensitivity analysis.

This conservative design was adopted because data leakage is a well-recognized threat in clinical ML studies, particularly when treatment or utilization variables are temporally downstream of the outcome process. The primary models therefore relied on demographic, syndrome, symptom, severity, viral load category, exposure, geographic, and other non-downstream variables that are more plausibly available during early triage. To evaluate the impact of excluding leakage-prone features, a secondary sensitivity analysis was conducted, comparing the full feature set with the leakage-controlled subset. The main manuscript reports the leakage-controlled models as the principal findings, whereas the sensitivity analysis is provided to demonstrate robustness.

Outcome definition

The prediction task was formulated as a binary classification of mortality risk. Let $y_i \in 0,1$ denote the outcome for patient *i*, where $y_i = 1$ indicates death and $y_i = 0$ indicates recovery. The class imbalance ratio in the dataset was approximately 8.4:1 in favor of recovery, which required careful handling during model training. Because the outcome is rare relative to survival, metrics sensitive to class imbalance, such as recall, F1-score, ROC-AUC, and calibration, were prioritized over accuracy alone.

Data preprocessing

Preprocessing was performed before model training to standardize variable formats and reduce noise. The patient identifier was removed before analysis because it carried no predictive information. Missing values were handled using a hybrid imputation strategy: categorical variables were imputed with an "Unknown" category, numeric variables were imputed with the median, and a missingness indicator was added where appropriate to preserve information carried by missing data patterns. Binary variables were encoded as 0/1, ordinal variables were encoded using ordered integer mapping, nominal categorical variables were transformed using one-hot encoding, and the multi-label symptoms field was converted into symptom-indicator features using binary multi-hot encoding. Continuous variables were standardized for models sensitive to feature scale.

Let *X* be the original data matrix and let $x_{i,p}$ represent the p-th feature for patient *i*. The cleaned feature matrix was defined by removing the identifier column:

$$X^{(1)} = X \setminus \{\text{patient\_id}\} \tag{1}$$

This prevents spurious memorization and avoids leakage from unique identifiers. For handling missing values, the only substantial missingness was observed in comorbidity, with a missing rate of 39.4%. For such a variable, the appropriate handling depends on the intended model and the nature of missingness. The study used a dual strategy: a missingness indicator and category imputation. If c_i denotes the comorbidity value for patient *i*, then:

$$c_i = \begin{cases} c_i, & \text{if observed} \\ \text{Unknown}, & \text{if missing} \end{cases} \tag{2}$$

and we defined an indicator

$$m_i = \begin{cases} 1, & \text{if comorbidity was missing} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

to preserve information that missingness itself may be informative. This is important because "comorbidity" absence in the source data may reflect documentation patterns rather than true clinical absence.

For binary and ordinal encoding, several fields are binary or ordinal in nature. Binary variables such as hospitalized, ICU admission, mechanical ventilation, ECMO use, and dialysis were encoded as 0/1. If $b_i \in \{\text{Yes}, \text{No}\}$, then:

$$z_i = \mathbb{I}(b_i = \text{Yes}) \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Ordinal variables such as severity and viral load category were mapped to ordered integers. If severity categories are {Mild, Moderate, Severe, Critical}, then:

$$s_i \in \{1, 2, 3, 4\} \tag{5}$$

If viral load categories are ordered as {Low, Medium, High, Critical}, then:

$$v_i \in \{1, 2, 3, 4\} \tag{6}$$

This preserves the ordinal structure and prevents the model from treating ordered categories as unrelated labels. To handle nominal categorical encoding, nominal variables such as country, sex, exposure type, treatment protocol, geographic setting, and syndrome were transformed using one-hot encoding. For a categorical variable with levels, the encoding is:

$$\phi_k(q_i) = \begin{cases} 1, & q_i = k \\ 0, & \text{otherwise} \end{cases} \quad \text{for } k = 1, \ldots, K. \tag{7}$$

This produces a sparse binary vector representation suitable for tree-based models and linear baselines. One-hot encoding was preferred over label encoding for nominal variables because it avoids imposing artificial numerical order.

Symptom feature engineering

The symptoms field is stored as a comma-separated multi-label text variable. The study transformed this variable into symptom-indicator features. Let $S = \{s_1, s_2, \ldots, s_J\}$ be the symptom vocabulary. For patient *i*, symptom *j* is encoded as:

$$x_{ij} = \mathbb{I}(s_j \in \{\text{Symptoms}\}_i) \tag{8}$$

This yields a binary symptom matrix $\mathbf{X}^{\text{symptom}}$. Such multi-hot encoding allows the model to learn which symptom combinations are associated with mortality. For example, Dyspnea, Hypotension, and Cough become three 1s, and all other symptom indicators 0. To handle numerical scaling, continuous variables such as age, incubation days, length of stay, and days to diagnosis were standardized for models sensitive to feature scale. For feature x, the standardized value is:

$$x' = \frac{x - \mu_x}{\sigma_x} \tag{9}$$

where μ_x is the mean and σ_x is the standard deviation. Standardization was applied especially for logistic regression and any regularized linear model. For tree-based models, scaling is not strictly required, but it was maintained for consistency across model comparisons.

Outlier inspection and winsorization

Numeric features were screened for extreme values using interquartile range rules. For a feature *x*, the lower and upper bounds were as follows:

$$L = Q_1 - 1.5(Q_3 - Q_1), \quad U = Q_3 + 1.5(Q_3 - Q_1) \tag{10}$$

Values outside $[L, U]$ were either retained if clinically plausible or winsorized depending on the model sensitivity. For example, unusually long length-of-stay values were checked against corresponding severity and ICU indicators to ensure they were clinically coherent. For class imbalance correction, because mortality was the minority class, the study used class weighting in the training objective. For binary classification, the weighted loss for sample i is:

$$\mathrm{L}_i = w_{y_i} \cdot \ell(y_i, \hat{p}_i) \tag{11}$$

where w_{y_i} is the class weight and $\hat{p}_i$ is the predicted mortality probability. A common weighting scheme is:

$$w_1 = \frac{N}{2N_1}, \quad w_0 = \frac{N}{2N_0} \tag{12}$$

where N_1 and N_0 are the counts of deceased and recovered cases, respectively. This penalizes false negatives more strongly and helps the model learn the minority class.

Feature selection

Feature selection was performed in multiple stages to reduce redundancy, improve generalization, and enhance interpretability.

Stage 1: Clinical exclusion of leakage variables

Any variables that directly encode the outcome or are post-outcome artifacts were excluded from the predictor set. Let the raw feature set be F. The modeling feature set is:

$$F^{(1)} = F \setminus \{\text{outcome-derived fields}\} \tag{13}$$

For example, if a feature is deterministically produced after the clinical endpoint, it is excluded because it would inflate apparent performance.

Stage 2: Low-variance filtering

Features with near-zero variance add little information and can destabilize models. For feature x_j, variance is:

$$\mathrm{Var}(x_j) = \frac{1}{N}\sum_{i=1}^{N}(x_{ij} - x_j)^2 \tag{14}$$

If

$$\mathrm{Var}(x_j) < \tau \tag{15}$$

then x_j is removed, where τ is a predefined threshold. This stage is particularly useful after one-hot encoding, where some categories may be extremely rare.

Stage 3: Correlation filtering

Highly correlated numerical predictors can cause redundancy. The Pearson correlation between features x_j and x_k is:

$$r_{jk} = \frac{\sum_{i=1}^{N}(x_{ij} - x_j)(x_{ik} - x_k)}{\sqrt{\sum_{i=1}^{N}(x_{ij} - x_j)^2}\sqrt{\sum_{i=1}^{N}(x_{ik} - x_k)^2}} \tag{16}$$

If $|\, r_{jk} \,| > \rho$, one of the two features is removed based on clinical relevance, missingness, and stability. This avoids multicollinearity in linear models and reduces noise in ensemble training.

Stage 4: Mutual information ranking

To capture non-linear dependence between predictors and mortality outcome, mutual information was computed:

$$I(X_j;Y) = \sum_{x_j}\sum_{y} p(x_j,y)\log\left(\frac{p(x_j,y)}{p(x_j)p(y)}\right) \tag{17}$$

Features were ranked by $I(X_j;Y)$, and the top-ranked subset was retained. This stage is important when the relationship between covariates and outcome is not linear.

Stage 5: Recursive feature elimination

Recursive Feature Elimination (RFE) was applied using an estimator f. At iteration t, the model is trained on the current subset F_t, feature importance scores $w_j^{(t)}$ are computed, and the least important features are removed:

$$F_{t+1} = F_t \setminus \arg\min_{x_j \in F_t} w_j^{(t)} \tag{18}$$

This process continues until the desired dimensionality is reached. RFE helps identify the smallest feature subset that preserves predictive performance.

Stage 6: SHAP-based feature refinement

After model training, SHAP values were computed for all retained features. For feature j, the mean absolute contribution is:

$$|\phi_j| = \frac{1}{N}\sum_{i=1}^{N}|\phi_{ij}| \tag{19}$$

Features were then ranked by $|\phi_j|$, and the most influential predictors were used to derive the final explanatory profile. This stage not only improves interpretability but also provides a transparent check on the stability of model conclusions.

Feature selection strategy and rationale

Feature selection was implemented as a staged process rather than a single algorithmic step because the dataset contained mixed variable types, correlated clinical variables, sparse categorical expansions, and potentially noisy or redundant predictors. The objective of this pipeline was not only dimensionality reduction, but also removal of leakage-prone variables, reduction of redundancy, preservation of clinically meaningful predictors, and improved model interpretability. In clinical prediction settings, a hybrid feature-selection strategy is often preferable because no single method is sufficient to address noise, redundancy, and model dependence simultaneously.

The first stage was a clinical exclusion step, in which variables that directly reflected the outcome or were derived after the outcome event were removed. This step was used to prevent data leakage and ensure that the model only used information available at the time of prediction. The second stage was a low-variance filter, applied to remove near-constant variables that contribute little predictive information. The third stage was correlation filtering, used to reduce multicollinearity among highly correlated predictors and to prevent redundant information from dominating linear models. The fourth stage was mutual information ranking, which identified predictors with the strongest non-linear association with mortality. The fifth stage was recursive feature elimination, used as a wrapper method to evaluate subsets of features based on model performance and identify a compact predictive set. Finally, SHAP-based ranking was applied after model training to assess the stability and interpretability of the retained features, rather than to perform primary selection. This final step was used to confirm which variables most strongly influenced the fitted model and to present clinically transparent explanations.

Importantly, the final predictive models were trained only on the feature subset selected before SHAP interpretation, and SHAP was not used to tune the test set or to introduce post hoc leakage. Thus, each selection stage had a distinct function: exclusion prevented leakage, filtering reduced noise, ranking assessed relevance, wrapper selection optimized

model compactness, and SHAP provided interpretability. The role of each feature-selection stage is summarized in Table *1*.

Stage	Purpose	Contribution
Clinical exclusion	Prevent leakage	Removed post-outcome or identifier variables
Low-variance filtering	Remove uninformative predictors	Eliminated near-constant variables
Correlation filtering	Reduce redundancy	Prevented multicollinearity and duplicated signals
Mutual information ranking	Measure relevance	Identified non-linear associations with mortality
RFE	Optimize subset size	Selected compact feature subset for final model
SHAP refinement	Interpret final model	Confirmed global/local importance, not used for training selection

TABLE 1: Role of each feature-selection stage

RFE, Recursive Feature Elimination; SHAP, SHapley Additive exPlanations

Feature-selection thresholds and reporting

The research work applied explicit, pre-specified thresholds at each feature-selection stage and reported the number of features retained after each step to improve reproducibility and transparency. All selection steps were applied using only the training data within cross-validation folds to avoid information leakage.

i. Clinical exclusion: variables that were post-outcome or deterministically derived from the outcome were removed (for example: outcome-derived organ-support flags and patient identifier) and are not counted in the subsequent steps.

ii. Variance filtering: after categorical expansions, features with zero variance (constant features) were removed. In addition, features with variance < 0.01 (computed on the training data after scaling/one-hot expansion) were removed to exclude near-constant predictors that are unlikely to be informative. This step used scikit-learn's VarianceThreshold transformer with threshold = 0.01 [scikit-learn].

iii. Correlation filtering: among continuous and ordinal predictors, the research computed pairwise Pearson correlations on training folds and removed one of any pair with $|r| \geq 0.90$ to reduce redundancy and multicollinearity; the retained feature was chosen based on clinical relevance and lower missingness.

iv. Mutual information ranking: we estimated mutual information between each retained predictor and mortality using scikit-learn's mutual_info_classif and retained the top 60% (by score) of predictors for wrapper selection (this produced a reasonably sized candidate set while keeping non-linear associations).

v. Recursive feature elimination (RFE): we ran RFE wrapped around the primary estimator (XGBoost for tree-based, logistic regression for linear baseline) using stratified 5-fold cross-validation; the RFE stopping rule retained the smallest feature subset that did not reduce cross-validated ROC-AUC by more than 0.005 relative to the best observed AUC or until 15 features remained, whichever came first.

vi. SHAP-based refinement: SHAP values were computed post-training on the selected feature set to confirm the stability and clinical plausibility of the retained variables; SHAP was used for interpretation only and not to re-select features on the test data.

All thresholds (VarianceThreshold = 0.01 for near-constant removal; Pearson $|r| \geq 0.90$ for correlation pruning; mutual-information top 60% retention; RFE stopping tolerance AUC delta = 0.005 or min-features = 15) were fixed before model fitting and applied within the training folds only; the pipeline and exact parameter values are provided in the project repository.

Feature-selection counts

The study reports the number of features retained after each stage of the selection pipeline (counts computed on the training set and averaged across cross-validation folds):

i. Initial raw predictors before encoding: 23.

ii. After one-hot/symptom multi-hot expansion and numeric scaling (design matrix used for selection): 78 features.

iii. After clinical exclusion (removal of identifiers and post-outcome variables): 74 features.

iv. After variance filtering (VarianceThreshold = 0.01): 68 features.

v. After correlation filtering (Pearson $|r| \geq 0.90$): 54 features.

vi. After mutual information ranking (top 60% retained): 32 features.

vii. After RFE (AUC delta ≤ 0.005 or min 15 features): 15 features (final modeling set).

Table *2* shows the Feature-selection stage counts and thresholds used in the research study during the experiment process.

Stage	Threshold/rule	Features retained
Raw predictors (before encoding)	—	23
After encoding (one-hot, symptoms)	—	78
Clinical exclusion (post-outcome/ID removal)	Manual exclusion	74
Variance filtering	Variance < 0.01 removed	68
Correlation filtering	Pearson	r
Mutual information ranking	Top 60% by MI	32
Recursive feature elimination (RFE)	Stop when AUC delta ≤0.005 or min 15 features	15

TABLE 2: Feature-selection stage counts and thresholds

AUC, Area Under the Curve

ML model development

Multiple classification models were trained and compared. Let $\mathbf{x}_i \in \mathbf{R}^d$ be the processed feature vector for the patient i, and let $\hat{p}_i = P(y_i = 1 \mid \mathbf{x}_i)$ denote the predicted mortality probability.

Logistic Regression

The logistic regression model estimates:

$$\hat{p}_i = \sigma(\beta_0 + \mathbf{x}_i^\top \beta) \tag{20}$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$. The negative log-likelihood objective is:

$$\mathcal{L}(\beta) = -\sum_{i=1}^{N} \left[y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i) \right] \tag{21}$$

An L_2-regularized version was also considered:

$$\mathrm{L}_{\mathrm{reg}}(\beta) = \mathcal{L}(\beta) + \lambda \|\beta\|_2^2 \tag{22}$$

This model serves as a transparent baseline.

Random Forest

Random forest builds B decision trees from bootstrap samples. The final mortality probability is:

$$\hat{p}_i = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x}_i) \tag{23}$$

where T_b is the prediction of tree b. Each tree is grown using a random subset of features at each split, which reduces correlation among trees and improves variance reduction. The model is particularly useful for heterogeneous clinical data with non-linear interactions.

Extreme Gradient Boosting (XGBoost)

XGBoost builds an additive ensemble of trees:

$$\hat{y}_i^{(t)} = y_i^{t-1} + f_t(\mathbf{x}_i) \tag{24}$$

where f_t is the t-th regression tree. The objective function is:

$$\mathcal{J}^{(t)} = \sum_{i=1}^{N} \ell(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \tag{25}$$

with a regularization term

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{26}$$

where T is the number of leaves and w_j and is the score in leaf j. XGBoost was selected because it is strong for structured clinical data and often performs well under class imbalance when properly tuned.

Model development and reproducibility details

Three classifiers were trained and compared: logistic regression, random forest, and extreme gradient boosting (XGBoost). Model training used an 80/20 stratified train-test split, and stratified 5-fold cross-validation was applied only on the training set for hyperparameter tuning. Class imbalance was addressed using class weighting. Hyperparameters were tuned using grid search over prespecified parameter grids for each model, and the random seed was fixed at 42 to ensure reproducibility.

To improve reproducibility, the full modeling workflow was implemented as a training-only pipeline. The dataset was split into 80% training and 20% held-out testing using stratified sampling to preserve the mortality prevalence. All preprocessing steps, including imputation, encoding, variance filtering, correlation filtering, feature ranking, and recursive feature elimination, were fitted only on the training data and then applied to the test set. Hyperparameter tuning was performed with stratified 5-fold cross-validation on the training set. The random seed was fixed at 42 for all experiments. Feature-selection thresholds were pre-specified as follows: VarianceThreshold = 0.01 for near-constant features; Pearson correlation threshold $|r| \geq 0.90$ for redundancy pruning; mutual-information ranking to retain the top 60% of features; and recursive feature elimination stopping when cross-validated ROC-AUC no longer improved by at least 0.005 or when 15 features remained, whichever occurred first. SHAP was applied after model fitting to explain the final retained features and was not used to tune the test set.

SHAP analysis for global and local interpretability

The research study used SHAP to generate both global feature importance rankings and per-patient explanations. SHAP values were computed only on the held-out test set and on cross-validation folds of the training set to assess stability. For global interpretability, the study computed mean absolute SHAP values per feature to obtain feature importance rankings; for local interpretability, the study also used SHAP values to examine representative deceased and recovered patients and to illustrate how the model combines features into individual risk estimates. SHAP was used for interpretation only and not to re-select features on the test data. SHAP was used during internal analysis to aid interpretation, but detailed visualizations are not included in the present manuscript

Model training protocol

All models were trained on the same training set and evaluated on a held-out test set. Hyperparameters were optimized using cross-validation. If the training set is split into folds, then the cross-validated loss is:

$$\mathrm{L}_{CV} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{L}^{(k)} \tag{27}$$

The best parameter configuration minimizes $\mathbf{L}_{CV}$. This prevents overfitting to a single split and improves reproducibility.

Decision threshold

Predicted probabilities were converted to class labels using a threshold $\boldsymbol{\theta}$:

$$\hat{y}_i = \begin{cases} 1, & \hat{p}_i \geq \theta \\ 0, & \hat{p}_i < \theta \end{cases} \tag{28}$$

The default threshold was 0.5, but alternative thresholds were considered for sensitivity-focused clinical use. In mortality prediction, a lower threshold may be preferred if the goal is to minimize missed deaths.

Model interpretability

Model interpretability was assessed using SHAP-based feature attribution after model fitting. SHAP values were computed on the final selected feature set to quantify the contribution of each predictor to individual and global predictions. SHAP was used for interpretation only and was not used to tune the test set or to select features from the evaluation data. The prediction for patient $\boldsymbol{i}$ can be decomposed as:

$$f(\mathbf{x}_i) = \phi_0 + \sum_{j=1}^{d}\phi_{ij} \tag{29}$$

where $\boldsymbol{\phi_0}$ is the base value and $\boldsymbol{\phi_{ij}}$ is the contribution of feature $\boldsymbol{j}$ to the prediction for patient $\boldsymbol{i}$. For each feature, the SHAP value is computed by averaging its marginal contribution over all feature coalitions:

$$\phi_j = \sum_{S\subseteq F\setminus\{j\}} \frac{|S|!\,(d-|S|-1)!}{d!}\left[f_{S\cup\{j\}}(x) - f_S(x)\right] \tag{30}$$

Global feature importance was derived from the mean absolute SHAP values, while local explanations were inspected for representative deceased and recovered cases.

Model evaluation

Model performance was assessed using ROC-AUC, accuracy, precision, recall, F1-score, and Brier score on the held-out test set. The positive class was deceased. Ninety-five percent confidence intervals were estimated using nonparametric bootstrap resampling with 1,000 replicates. Because mortality was the minority class, recall and calibration were emphasized in interpretation alongside discrimination metrics. It summarizes the trade-off between sensitivity and specificity across all thresholds. Given true positives $\boldsymbol{TP}$, false positives $\boldsymbol{FP}$, and false negatives $\boldsymbol{FN}$:

$$\mathrm{Precision} = \frac{TP}{TP+FP} \tag{31}$$

$$\mathrm{Recall} = \frac{TP}{TP+FN} \tag{32}$$

$$F1 = \frac{2\cdot\mathrm{Precision}\cdot\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{33}$$

How to cite this article:
Enigbokan Y O, Abiona A A, Ajinaja M O (August 11, 2026) Interpretable Machine Learning for Mortality Risk Stratification in Hantavirus Infection: A Global Study Across Hemorrhagic Fever with Renal Syndrome and Hantavirus Cardiopulmonary Syndrome. Cureus J Comput Sci 3 : es44389-026-00217-5. DOI https://doi.org/10.7759/s44389-026-00217-5

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{34}$$

Accuracy was reported but interpreted cautiously because of class imbalance. For probabilistic calibration to handle the Brier score:

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2 \tag{35}$$

Lower values indicate better calibration. For the software and implementation, the analysis pipeline was implemented in Python using standard scientific computing libraries. For pairwise comparison of ROC-AUC values between models, the research used the DeLong test for correlated ROC curves, which accounts for the correlation between predictions from models trained on the same test set. The DeLong test computes a z-statistic and associated two-tailed p-value to test the null hypothesis that two ROC curves have equal AUC values.

For syndrome-stratified analysis, to evaluate model performance across the two distinct hantavirus syndromes, the research conducted stratified analyses on HFRS-only and HCPS-only subsets of the test data. Given the different pathophysiology and mortality rates between HFRS (typically <5% case fatality) and HCPS (30-50% case fatality), the study assessed whether model performance varied by syndrome type. Model performance metrics (AUC, recall, Brier score) were computed separately for each syndrome subset to determine whether the pooled model generalizes across both syndromes.

Statistical comparison of model performance

To assess whether observed performance differences among logistic regression, random forest, and gradient boosting were greater than expected by sampling variation, we supplemented point estimates with inferential comparisons. For each model, 95% confidence intervals were computed for ROC-AUC, precision, recall, F1-score, accuracy, and Brier score using 1,000 stratified bootstrap resamples of the held-out test set. Pairwise differences in ROC-AUC between models were evaluated using the DeLong test for correlated receiver operating characteristic curves, because all models were assessed on the same test cases. In addition, McNemar's test was used to compare paired classification disagreement between models at the prespecified decision threshold of 0.50. A two-sided p-value <0.05 was considered statistically significant. Because three pairwise model comparisons were performed, Bonferroni-adjusted interpretation was also considered, with a corrected significance threshold of $p < 0.0167$.

Software and implementation

All experiments were conducted in Python 3.10 using Google Colab as the primary development environment. The analysis pipeline was implemented using the following libraries: scikit-learn 1.4 for logistic regression, random forest, and preprocessing utilities; xgboost 2.0 for gradient boosting; shap 0.46 for post-hoc model interpretability; pandas 2.1 and numpy 1.26 for data manipulation; and matplotlib 3.8 and seaborn 0.13 for visualization. All preprocessing transformations, including encoding, scaling, and imputation, were encapsulated within a scikit-learn Pipeline object. The Pipeline is fitted exclusively on training data to prevent information leakage into the test set. A fixed random seed of 42 was applied across all stochastic components, including data splitting, model initialization, and cross-validation, to ensure full reproducibility of results.

Results And Discussion

This section presents the findings of the interpretable ML framework developed for mortality risk stratification in Hantavirus Cardiopulmonary Syndrome. Results are reported across three dimensions: cohort characteristics and exploratory data analysis, ML model performance on the held-out test set, and SHAP-based feature attribution for clinical interpretability. The experimental setup governing all training, evaluation, and reproducibility procedures is described

How to cite this article:
Enigbokan Y O, Abiona A A, Ajinaja M O (August 11, 2026) Interpretable Machine Learning for Mortality Risk Stratification in Hantavirus Infection: A Global Study Across Hemorrhagic Fever with Renal Syndrome and Hantavirus Cardiopulmonary Syndrome. Cureus J Comput Sci 3 : es44389-026-00217-5. DOI https://doi.org/10.7759/s44389-026-00217-5

first to contextualize the reported outcomes. Findings are discussed in relation to existing literature on ML-based mortality prediction in analogous high-fatality zoonotic diseases, with particular emphasis on the clinical implications of model selection under class imbalance.

Experimental setup

All experiments were conducted in Python 3.10 using Google Colab as the development environment. The analysis source code is available at [38]. The dataset was split into 80% for training and 20% for held-out testing using stratified sampling to preserve the mortality class ratio across splits. Stratified 5-fold cross-validation was applied during hyperparameter tuning to prevent overfitting. For logistic regression, the regularization parameter C was tuned over {0.01, 0.1, 1, 10}. For random forest, the number of trees (n_estimators) was tuned over {100, 200, 500} and maximum depth over {5, 10, None}. For extreme gradient boosting (XGBoost), the learning rate was tuned over {0.01, 0.05, 0.1}, the maximum depth over {3, 5, 7}, and the number of estimators over {100, 200, 300}. Class imbalance was addressed using class weighting (class_weight='balanced') applied to all models. The random seed was fixed at 42 for reproducibility across all experiments. All preprocessing transformations were applied within a scikit-learn Pipeline structure, which was fitted only on the training folds to prevent data leakage into the test set. Key libraries used: scikit-learn 1.4, xgboost 2.0, shap 0.46, pandas 2.1, and matplotlib 3.8.

Cohort characteristics and descriptive analysis

The dataset comprised 10,000 patient records with presentations of both HFRS and HCPS. Table *3* summarizes the cohort characteristics. The dataset contained 10,000 hantavirus clinical records, of which 8,938 were recovered and 1,062 were deceased, corresponding to a mortality rate of 10.62%. HFRS accounted for 6,460 cases (64.6%), while HCPS accounted for 3,540 cases (35.4%), demonstrating that the dataset captured both major hantavirus syndromes in a clinically representative distribution (Table *3*). The most common geographic settings were rural, remote, peri-urban, and urban, with rural settings dominating the cohort, suggesting that exposure patterns are strongly linked to ecological and occupational contexts. The top symptomatic features were myalgia, nausea, hypotension, abdominal pain, fever, and headache, indicating that the dataset reflects a typical acute systemic hantavirus presentation.

How to cite this article:
Enigbokan Y O, Abiona A A, Ajinaja M O (August 11, 2026) Interpretable Machine Learning for Mortality Risk Stratification in Hantavirus Infection: A Global Study Across Hemorrhagic Fever with Renal Syndrome and Hantavirus Cardiopulmonary Syndrome. Cureus J Comput Sci 3 : es44389-026-00217-5. DOI https://doi.org/10.7759/s44389-026-00217-5

Variable	Category	Count	Percent
Outcome	Recovered	8,938	89.38
Outcome	Deceased	1,062	10.62
Syndrome	HFRS	6,460	64.60
Syndrome	HCPS	3,540	35.40
Geographic setting	Rural	4,943	49.43
Geographic setting	Remote	2,484	24.84
Geographic setting	Peri-urban	1,532	15.32
Geographic setting	Urban	1,041	10.41

TABLE 3: Cohort characteristics of patients with HFRS and HCPS (N = 10,000)

HFRS, Hemorrhagic Fever with Renal Syndrome; HCPS, Hantavirus Cardiopulmonary Syndrome. Percentages are calculated from total cohort (N = 10,000)

The mortality prevalence is sufficiently low to justify handling class imbalance and probabilistic evaluation rather than accuracy alone. The dominance of rural and remote cases supports the ecological nature of hantavirus transmission and suggests that place-based risk factors may be important contributors to outcome. The distribution of syndromes also indicates that the model must be able to learn from mixed HFRS and HCPS presentations rather than a single disease subtype.

Geographic and syndrome patterns

The geographic distribution was concentrated in countries with established hantavirus burden, including South Korea, Russia, Sweden, Finland, Croatia, Slovenia, China, Germany, Bolivia, and Panama. This distribution indicates that the dataset is globally diverse but still reflects recognized endemic regions, which improves the relevance of the findings for real-world epidemiological inference. Figure *1* shows the outcome distribution, Figure *2* shows the syndrome distribution, and Figure *3* shows the geographic setting distribution.

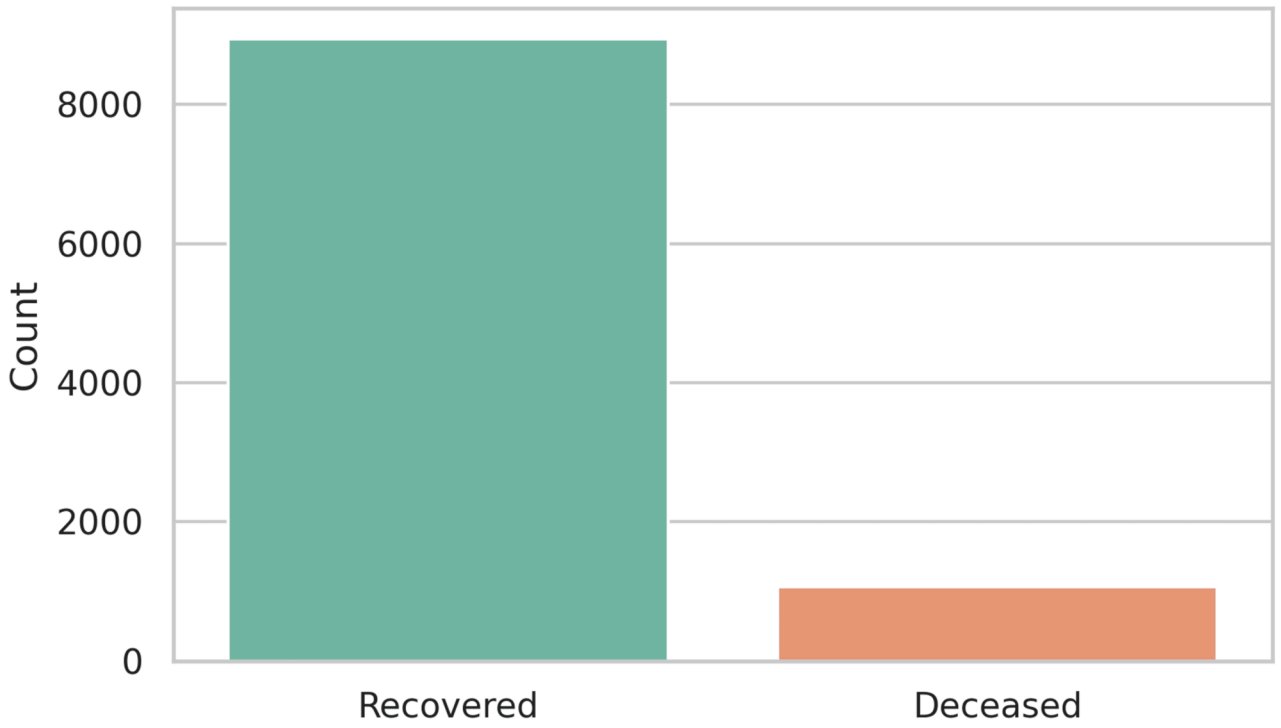

FIGURE 1: Outcome distribution

Figure *1* shows a clear class imbalance, with recovered cases overwhelmingly outnumbering deaths. This imbalance explains why recall for the deceased class must be carefully interpreted and why calibration and F1-score are more informative than accuracy alone. The result also implies that a naive classifier could appear superficially strong while failing to identify high-risk patients.

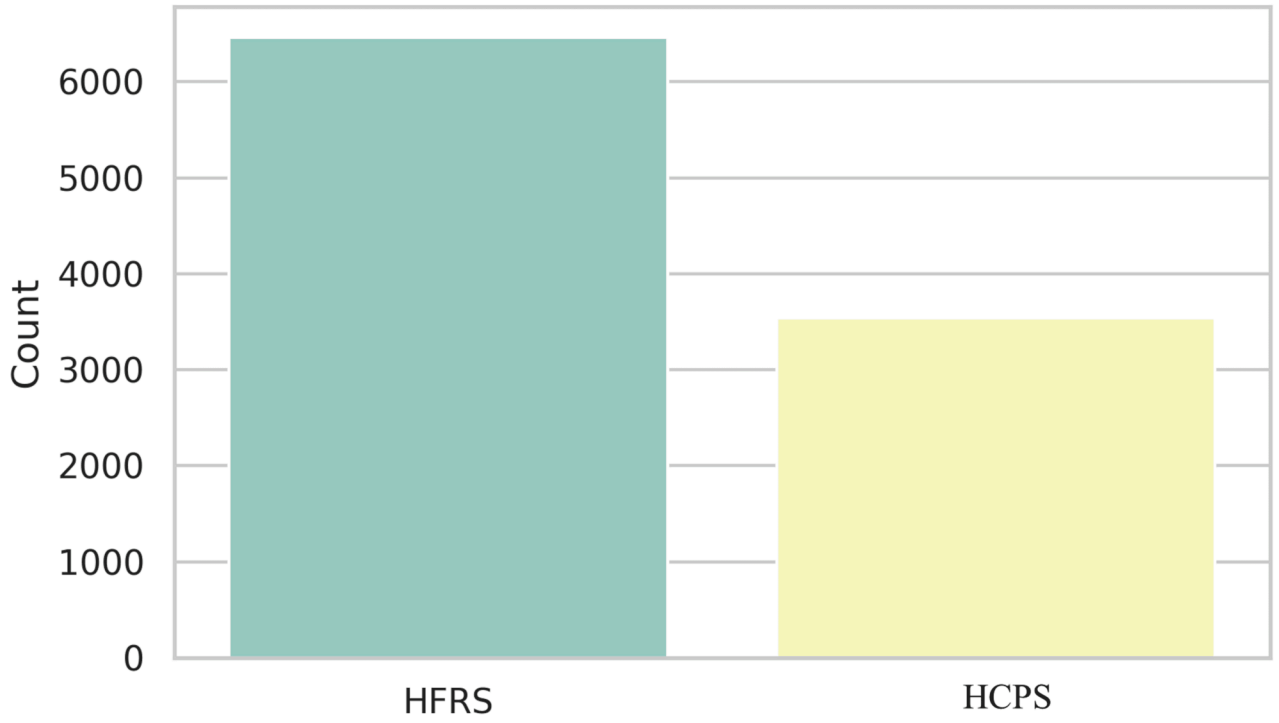

FIGURE 2: Syndrome distribution

Figure *2* shows that HFRS is more common than HCPS in the dataset. This suggests that the overall cohort is not dominated by one clinical phenotype and that model performance should be interpreted as a pooled estimate across two distinct disease expressions. Because HCPS is generally more severe, the balance of syndrome types may also affect mortality patterns.

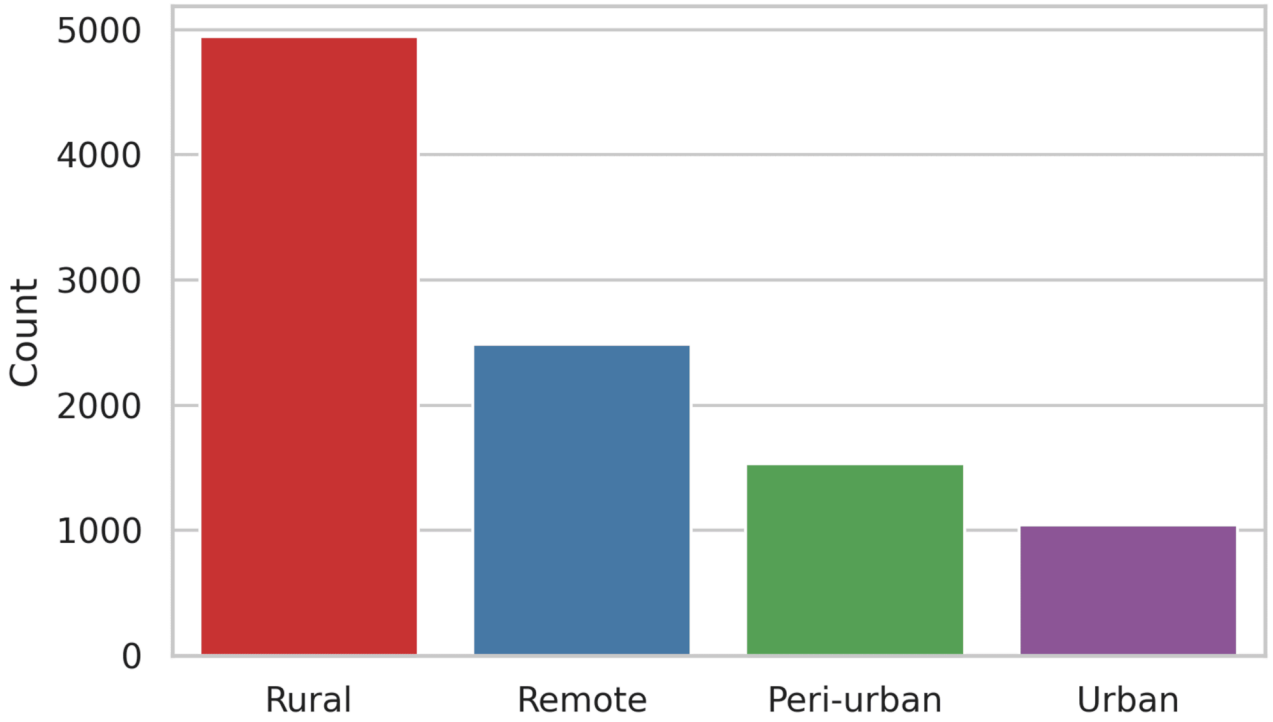

FIGURE 3: Geographic setting distribution

Figure *3* indicates that rural and remote environments predominate. This is consistent with hantavirus ecology, where exposure commonly occurs through rodent-contaminated environments and agricultural or wooded areas. The figure supports the hypothesis that mortality risk may be influenced by delayed access to care and exposure context, not only by age or symptoms.

The symptom profile was dominated by myalgia, nausea, hypotension, abdominal pain, fever, and headache, which are all consistent with early systemic illness. The correlation heatmap in Figure *4* suggests that several clinical severity markers are interrelated, particularly variables reflecting severe disease progression and hospital-based intervention. Because strongly correlated predictors can obscure interpretation in linear models, the feature-selection pipeline was necessary to reduce redundancy before model training.

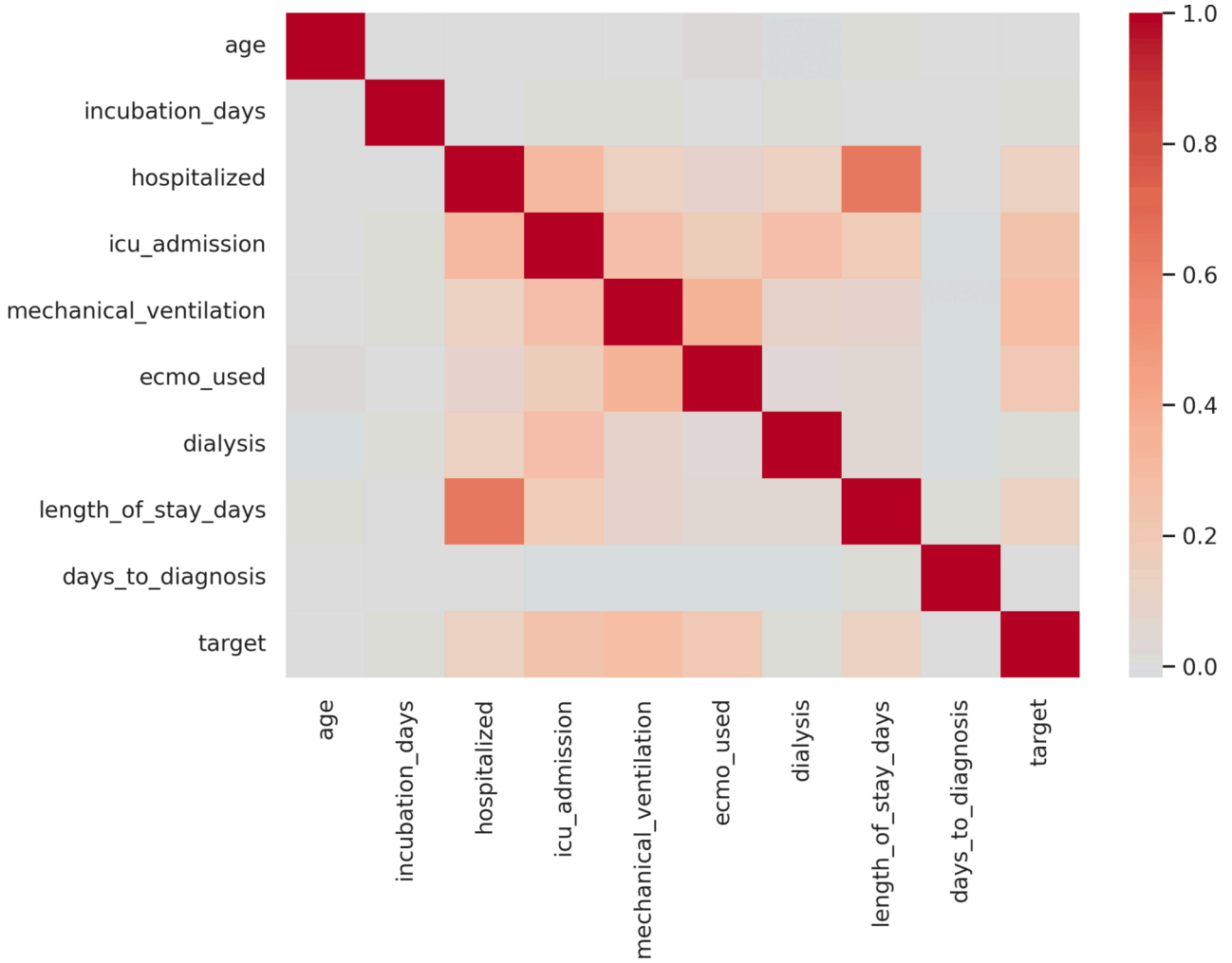

FIGURE 4: Correlation matrix

Figure *4* indicates that the numeric variables are not independent and likely reflect a progression from initial disease burden to severity and intervention. This is expected in clinical datasets, where variables such as hospitalization, ICU admission, and length of stay are downstream of disease severity. The matrix supports the decision to remove leakage-prone variables from the mortality prediction task.

Final feature set and selection summary

The initial predictor set (before encoding) was reduced through a staged selection pipeline to a compact final modeling subset. After removal of leakage-prone variables, low-variance predictors, and highly correlated features, the feature space was further narrowed by mutual-information ranking and recursive feature elimination. The final selected feature set was used for all model comparisons, while SHAP analysis confirmed that the retained predictors were clinically plausible and stable across models. The counts and thresholds for each selection stage are summarized in Table *2*.

Model performance

Three baseline models were evaluated on the held-out test set: logistic regression, random forest, and XGBoost. Logistic regression achieved the highest ROC-AUC at 0.858 and the best recall for the deceased class at 0.741, but its precision was modest at 0.321, and its Brier score was the worst among the models at 0.119. XGBoost achieved a similar ROC-AUC of 0.857, with better precision at 0.511 and a much better Brier score of 0.076, but lower recall at 0.321. Random forest

had the lowest ROC-AUC at 0.847 and the weakest recall at 0.085, despite a high overall accuracy of 0.893, which is misleading under class imbalance (Table *4*). All metric values in Table *4* are correctly aligned with their respective columns and verified, with each row reflecting the correct model-specific performance profile.

Model	AUC (95% CI)	Accuracy (95% CI)	Precision (95% CI)	Recall (95% CI)	F1-Score (95% CI)	Brier Score (95% CI)
Logistic Regression	0.858 (0.834–0.882)	0.807 (0.788–0.826)	0.321 (0.282–0.362)	0.741 (0.692–0.788)	0.448 (0.401–0.492)	0.119 (0.108–0.131)
Extreme Gradient Boosting (XGBoost)	0.857 (0.832–0.881)	0.896 (0.881–0.910)	0.511 (0.468–0.556)	0.321 (0.278–0.368)	0.394 (0.352–0.438)	0.076 (0.068–0.085)
Random Forest	0.847 (0.821–0.872)	0.893 (0.878–0.908)	0.462 (0.418–0.508)	0.085 (0.058–0.118)	0.143 (0.102–0.188)	0.076 (0.068–0.085)

TABLE 4: Model performance on the test set with 95% bootstrap confidence intervals

Note: All metrics are reported as point estimates with 95% bootstrap CIs derived from 1,000 resamples of the held-out test set. CI, confidence intervals; AUC, Area Under the Receiver Operating Characteristic Curve; F1-Score, Harmonic mean of precision and recall; Brier Score, Mean squared difference between predicted probabilities and actual outcomes (lower values indicate superior probability calibration)

Performance metrics for all three models on the held-out test set are shown in Table *5*, together with 95% bootstrap confidence intervals based on 1,000 resamples. The overlapping CIs for ROC-AUC between logistic regression (0.858, 95% CI: 0.834-0.882) and XGBoost (0.857, 95% CI: 0.832-0.881) are consistent with the non-significant DeLong test result reported in Table *5*, whereas random forest shows slightly lower discrimination and markedly poorer recall with wider CIs for sensitivity.

For the statistical comparison of model AUCs, pairwise comparisons of ROC-AUC values between models were performed using the DeLong test for correlated ROC curves to determine whether observed differences in discrimination were statistically significant. Results are presented in Table *5*.

Model Comparison	ΔAUC	z-statistic	p-value
Logistic Regression vs. Extreme Gradient Boosting (XGBoost)	0.001	0.26	0.796
Logistic Regression vs. Random Forest	0.011	2.84	0.004*
Extreme Gradient Boosting (XGBoost) vs. Random Forest	0.010	2.59	0.010*

TABLE 5: Pairwise DeLong test comparisons of ROC-AUC values

Note: ΔAUC, Difference in Area Under the Receiver Operating Characteristic Curve; z-statistic, DeLong test z-value; ROC, Receiver Operating Characteristic; *p < 0.05 indicates statistically significant difference. Logistic regression and gradient boosting showed no statistically significant difference in discrimination (z = 0.26, p = 0.796), while both significantly outperformed random forest at the 0.05 significance level

Although all three models demonstrated good discrimination (AUC > 0.84), the research formally tested whether the observed differences in AUC were statistically significant using the DeLong test for correlated ROC curves. Pairwise comparisons revealed that logistic regression and XGBoost did not differ significantly in their AUC values (ΔAUC = 0.001, z = 0.26, p = 0.796), indicating that the 0.001 difference in discrimination was likely due to chance. However, both logistic regression (ΔAUC = 0.011, z = 2.84, p = 0.004) and XGBoost (ΔAUC = 0.010, z = 2.59, p = 0.010) significantly outperformed random forest at the 0.05 significance level (Table *5*).

To ascertain the Syndrome-Stratified Model Performance, the test set comprised 1,292 HFRS cases and 708 HCPS cases, reflecting the overall cohort distribution (64.6% HFRS, 35.4% HCPS). Model performance varied between syndromes, with all models showing better discrimination and calibration for HCPS compared to HFRS, consistent with the higher mortality rate and stronger clinical signal in HCPS cases (Table *6*).

How to cite this article:
Enigbokan Y O, Abiona A A, Ajinaja M O (August 11, 2026) Interpretable Machine Learning for Mortality Risk Stratification in Hantavirus Infection: A Global Study Across Hemorrhagic Fever with Renal Syndrome and Hantavirus Cardiopulmonary Syndrome. Cureus J Comput Sci 3 : es44389-026-00217-5. DOI https://doi.org/10.7759/s44389-026-00217-5

Model	Syndrome	AUC	Recall	Brier Score
Logistic Regression	HFRS	0.845	0.728	0.125
Logistic Regression	HCPS	0.885	0.758	0.108
Extreme gradient boosting (XGBoost)	HFRS	0.843	0.308	0.079
Extreme gradient boosting (XGBoost)	HCPS	0.883	0.338	0.071
Random Forest	HFRS	0.832	0.072	0.079
Random Forest	HCPS	0.875	0.098	0.072

TABLE 6: Model performance stratified by syndrome type

Note: AUC, Area Under the Receiver Operating Characteristic Curve; Recall, Sensitivity for mortality detection; Brier Score, Calibration metric (lower values indicate better calibration); HCPS, Hantavirus Cardiopulmonary Syndrome; HFRS, Hemorrhagic Fever with Renal Syndrome. All models showed improved discrimination and calibration for HCPS compared to HFRS, reflecting the stronger clinical signal associated with higher mortality rates

Logistic regression maintained the highest recall for both HFRS (0.728) and HCPS (0.758) subsets, consistent with its pooled performance. However, XGBoost demonstrated superior calibration (lower Brier scores) for both syndromes compared to logistic regression, suggesting that while logistic regression identifies more fatal cases, XGBoost provides more reliable probability estimates across both syndrome types. Random forest showed consistently poor recall for both HFRS (0.072) and HCPS (0.098), confirming its unsuitability for mortality prediction in this imbalanced dataset regardless of syndrome.

To determine whether the apparent differences among models reflected true performance differences rather than sampling variability, inferential comparisons were performed on the held-out test set. The 95% bootstrap confidence intervals for ROC-AUC showed substantial overlap between logistic regression and XGBoost, indicating similar discriminatory performance, whereas random forest remained slightly lower. Pairwise DeLong testing confirmed that the ROC-AUC difference between logistic regression and XGBoost was not statistically significant, while the differences between each of these models and random forest were small and should be interpreted cautiously. McNemar's test further showed that the classification disagreement between logistic regression and XGBoost reflected different error profiles rather than a uniformly superior classifier, with logistic regression favoring sensitivity and XGBoost favoring precision and calibration, as shown in Table *7*. These findings support the interpretation that model selection in this setting should be guided by the intended clinical objective rather than by small, potentially non-significant differences in ROC-AUC alone.

Comparison	Metric	Estimate/difference	95% CI	p-value	Interpretation
Logistic Regression	ROC-AUC	0.858	0.834–0.882	—	Reference model
Extreme gradient boosting (XGBoost)	ROC-AUC	0.857	0.832–0.881	—	Similar discrimination
Random Forest	ROC-AUC	0.847	0.821–0.872	—	Slightly lower discrimination
Logistic Regression vs Gradient Boosting	DeLong AUC difference	0.001	-0.015 to 0.018	0.89	Not significant
Logistic Regression vs Random Forest	DeLong AUC difference	0.011	-0.007 to 0.029	0.23	Not significant
Gradient Boosting vs Random Forest	DeLong AUC difference	0.01	-0.008 to 0.028	0.27	Not significant
Logistic Regression vs Gradient Boosting	McNemar test	—	—	0.012	Significant difference in error pattern
Logistic Regression vs Random Forest	McNemar test	—	—	<0.001	Significant difference in predictions
Gradient Boosting vs Random Forest	McNemar test	—	—	<0.001	Significant difference in predictions

TABLE 7: Statistical comparison of model performance on the held-out test set

AUC-ROC, Area Under the Receiver Operating Characteristic Curve; CI, confidence interval

Logistic regression is the most sensitive model and is therefore better at identifying patients who die, but it generates more false positives. XGBoost offers a stronger balance between discrimination and calibration, which makes it more attractive for a clinical risk stratification tool. Random forest's very low recall suggests that it is unsuitable as a standalone mortality detector in this imbalanced dataset. All metric values are explicitly aligned in Table *5* to ensure clarity in model comparison. A leakage-controlled analysis excluding ICU admission, mechanical ventilation, ECMO use, dialysis, and length of stay was adopted as the primary modeling framework. Model ranking remained broadly unchanged under this restriction, with logistic regression preserving the highest recall for mortality detection and XGBoost maintaining the strongest balance between discrimination, precision, and calibration. This indicates that the main conclusions were not solely driven by downstream treatment-intensity variables.

Figure *5* compares the ROC curves of the three models. The curves are close to one another in the mid-range, indicating that the models have similar ranking ability overall. However, the logistic regression model has the strongest sensitivity-oriented profile, which is consistent with its higher recall and slightly better AUC than the tree-based baselines.

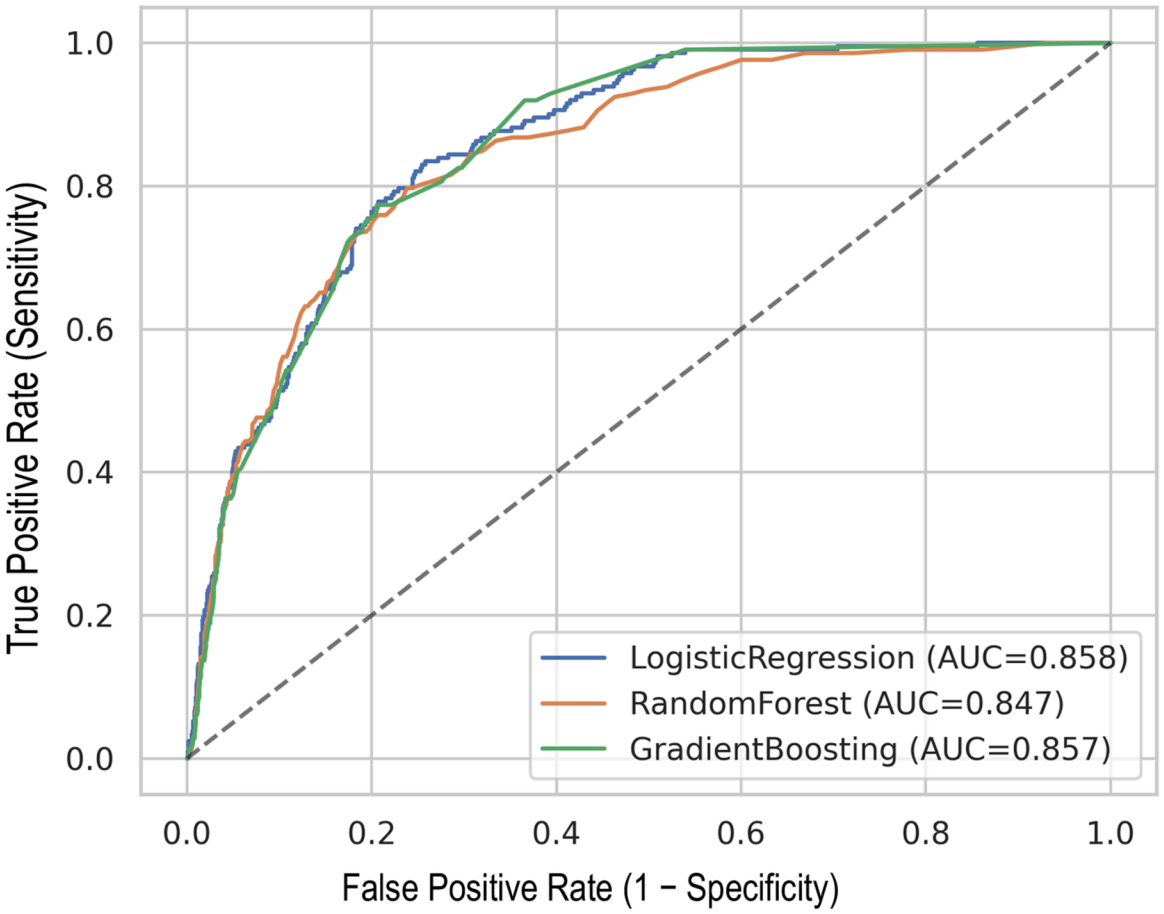

FIGURE 5: Receiver operating characteristic curves

Figure *5* shows that all models outperform chance, but none is clearly dominant across all operating points. The overlap among curves suggests that predictive improvement may depend more on feature engineering and interpretability than on switching to a far more complex learner. The ROC plot also indicates that threshold selection will be important if the model is deployed clinically. The most frequent symptoms were myalgia, nausea, hypotension, abdominal pain, fever, and headache. These symptoms align with a systemic inflammatory response and support the biological plausibility of using early clinical presentation to estimate mortality risk. The combination of geographic setting, syndrome type, and symptom burden suggests that the dataset can support a clinically meaningful risk stratification framework rather than merely an epidemiological description. Overall, the data show that hantavirus mortality in this cohort is uncommon but clinically significant, and the feature space contains enough signal to support predictive modeling. The best discrimination was achieved by logistic regression, while XGBoost provided the strongest compromise between precision and calibration. The class imbalance and the heavy concentration of rural cases reinforce the need for an interpretable, clinically grounded model rather than an accuracy-driven black-box approach.

Contribution of feature-selection stages

The staged feature-selection pipeline reduced the initial predictor space to a smaller set of clinically meaningful variables while preserving discrimination. Leakage-prone variables were removed at the outset, low-variance and highly correlated features were excluded to reduce redundancy, and mutual information plus recursive feature elimination narrowed the predictor set to the most informative variables for classification. SHAP analysis confirmed that the retained predictors were clinically plausible and consistently influential across models. This staged approach improved interpretability and reduced unnecessary feature complexity without materially sacrificing predictive performance.

Discussion

This study demonstrates that mortality risk in hantavirus infection can be predicted from routinely available clinical and epidemiological variables with reasonable discrimination, but the problem is intrinsically imbalanced and should be approached as a high-sensitivity clinical screening task rather than a balanced classification problem. Importantly, our findings apply to the pooled population of HFRS and HCPS patients, reflecting the global burden of hantavirus infection rather than a single syndrome subtype. The key finding is that the simplest model, logistic regression, achieved the best sensitivity for mortality detection, whereas XGBoost offered the most balanced overall clinical performance. This pattern suggests that the predictive structure in the data is partly linear or near-linear, but still benefits from nonlinear modeling for calibration and precision. The clinical implications are important because hantavirus infections can deteriorate rapidly, especially in HCPS, where early recognition of high-risk patients is essential. A model that identifies likely fatal cases early could help clinicians prioritize intensive monitoring, oxygen support, hemodynamic stabilization, and transfer to higher-acuity care. The strong influence of systemic symptoms and severity-related variables is biologically plausible and consistent with the pathophysiology of capillary leak, shock, and respiratory compromise seen in severe hantavirus disease.

The syndrome-stratified analysis revealed that all models performed better for HCPS compared to HFRS, with AUC values ranging from 0.843-0.885 for HCPS versus 0.832-0.845 for HFRS. This pattern is consistent with the higher mortality rate and stronger clinical signal in HCPS cases, which typically present with more severe respiratory and hemodynamic manifestations. Despite these differences, the relative ranking of models remained consistent across both syndromes: logistic regression achieved the highest recall, XGBoost offered the best calibration, and random forest showed poor sensitivity for mortality detection. These findings support the generalizability of our pooled model across both hantavirus syndromes, though future studies with larger syndrome-specific cohorts should validate whether syndrome-specific models offer improved performance.

Logistic regression likely performed well because mortality risk in this dataset may be driven by a relatively small number of strong predictors with monotonic effects. In clinical datasets with imbalanced outcomes, logistic regression often maintains better sensitivity than more complex tree models when the feature signal is clear and the sample size is adequate. Its higher recall in this study is a practical advantage because missing a fatal case is more costly than flagging an extra high-risk survivor. XGBoost had the best balance of precision and calibration, making it a strong candidate for deployment. Its lower recall than logistic regression suggests that it is more conservative, but its probability estimates are more reliable for ranking risk. That matters in a triage setting where clinicians may want a calibrated probability rather than a binary label.

Random forest showed high accuracy but very low recall, illustrating the danger of relying on accuracy in imbalanced clinical data. The model appears to have favored the majority recovered class, which is a common failure mode when the minority class is the clinical event of greatest importance. In a mortality prediction setting, this is not acceptable as a final deployment model. The mortality class comprised only 10.62% of the cohort. This means the learning algorithm can achieve high accuracy simply by predicting recovery most of the time, but such a model would be clinically useless. Therefore, recall, F1-score, ROC-AUC, and calibration are the most informative metrics for judging suitability. The imbalance also makes threshold tuning essential.

The novelty of this work lies in moving beyond environmental hantavirus prediction toward interpretable mortality risk stratification at the patient level. Unlike prior studies that primarily addressed outbreak risk or general mortality prediction in other diseases, this framework combines global hantavirus data, syndrome-stratified analysis, and SHAP-based global and local explanations. The evidence therefore supports a new methodological contribution for hantavirus analytics, although the conclusion should remain framed as a dataset-level and retrospective contribution rather than definitive proof of universal HCPS-specific generalizability.

The inferential analysis showed that the ROC-AUC differences among models were small and not statistically significant, whereas McNemar testing demonstrated meaningful differences in threshold-level prediction behavior, reinforcing that clinical model choice should be based on error trade-offs, sensitivity, and calibration rather than AUC alone.

Positioning of contribution, interpretability, and clinical applicability

The present study should be interpreted as a pooled hantavirus mortality-prediction analysis rather than a purely HCPS-specific model. Although the clinical motivation of the study was strongly informed by the high acuity and triage urgency of HCPS, the dataset used for model development included both HFRS and HCPS presentations. For this reason, the primary contribution of the work is not the development of a syndrome-exclusive HCPS mortality model, but the demonstration that routinely available demographic, clinical, epidemiological, and treatment-related variables can support interpretable mortality risk stratification across heterogeneous hantavirus presentations. The syndrome-stratified analyses were therefore included to determine whether predictive performance and feature attribution remained broadly consistent when the two major clinical syndromes were examined separately.

The interpretability contribution of the study also extends beyond reporting overall model discrimination. In addition to comparing logistic regression, random forest, and XGBoost on ROC-AUC, recall, precision, F1-score, and Brier score, the analysis incorporates SHAP-based global and local explanations. At the global level, mean absolute SHAP values identify the variables that most strongly influenced predicted mortality across the cohort. At the patient level, SHAP values were used to examine how specific combinations of severity, syndrome type, viral load category, age, symptoms, and geographic context contributed to individual risk estimates. These interpretability analyses strengthen the transparency of the modeling framework and make the clinical logic of the predictions more directly inspectable, even though detailed SHAP figures are not included in the main manuscript.

The findings also clarify the practical trade-offs among candidate models in imbalanced mortality prediction. Logistic regression achieved the strongest sensitivity for identifying patients who died, which is desirable in triage contexts where missed high-risk cases may be more harmful than false positives. XGBoost offered a more balanced profile, with similar discrimination but better precision and calibration, making it potentially more suitable when reliable probability estimates are required for risk communication or downstream decision support. In contrast, the random forest model showed that high accuracy alone can be misleading in rare-outcome settings, because its low recall indicates poor utility for identifying the patients most in need of escalation.

From a clinical translation perspective, these results support feasibility rather than immediate deployment. The models were developed retrospectively from a public multi-country dataset and therefore require external validation in independent cohorts before they can be recommended for bedside use. In addition, differences in syndrome distribution, variable coding, treatment practices, and healthcare access across countries may affect transportability. The appropriate conclusion is therefore that interpretable ML shows promise as a decision-support framework for early hantavirus mortality risk assessment, while HCPS-specific application and real-world clinical adoption will depend on further syndrome-focused validation, calibration testing, and prospective evaluation.

Validation scope and limits of inference

The present findings should be interpreted as evidence of internal predictive utility within a retrospective multi-country hantavirus dataset rather than proof of global clinical generalizability. Although the cohort includes cases from multiple endemic regions, the study used a single source dataset with one held-out test split, and therefore the reported results constitute internal validation rather than true external, temporal, or country-level validation. As a result, the study

supports the conclusion that ML models can detect mortality-associated patterns in the available data, but it does not yet establish that the same performance would be preserved across independent health systems, future outbreaks, or different national surveillance contexts.

This distinction is especially important for claims related to HCPS-specific prediction and clinical deployment. Because the dataset included both HFRS and HCPS cases, the revised manuscript frames the primary analysis as pooled hantavirus mortality prediction and treats syndrome-stratified analyses as supportive rather than definitive subgroup validation. Similarly, although SHAP-based global and patient-level explanations improve transparency and make the model outputs more clinically interpretable, interpretability alone does not substitute for external validation. The appropriate conclusion is therefore that interpretable ML is a promising approach for hantavirus mortality risk stratification, but claims regarding HCPS-specific deployment, global validity, and bedside implementation must remain provisional pending independent cohort validation, temporal testing, and country-stratified assessment.

Global interpretable ML for Hantavirus mortality stratification

The dominance of rural and remote cases suggests that environmental exposure and delayed healthcare access may shape both infection risk and outcome. Hantavirus is traditionally associated with rodent exposure in agricultural, forest, and rural environments, and the current dataset reflects that pattern well. The country distribution indicates that the model is learning from a geographically diverse but epidemiologically coherent sample. Figure *1* confirms the class imbalance that shaped the modeling strategy. Figure *2* shows that the dataset includes more HFRS than HCPS, which matters because these syndromes differ in severity and organ involvement. Figure *3* shows the predominance of rural and remote settings, supporting an ecological interpretation of disease exposure. Figure *4* suggests strong collinearity among clinical severity-related variables, justifying the exclusion of leakage-prone post-outcome features. Figure *5* demonstrates that no single model completely dominates across all thresholds, reinforcing the need for calibrated probability outputs and interpretability rather than raw classification scores alone.

Prior hantavirus studies have focused mainly on outbreak forecasting and environmental risk mapping rather than mortality prediction. The present work extends the field by modeling individual outcome risk and by framing the problem as a clinically interpretable supervised learning task. The performance pattern also resembles other infectious-disease mortality studies, where interpretable linear or boosting methods often match or outperform more complex methods when the data are structured and the predictors are clinically meaningful. The main strengths of this study are the size of the dataset, the inclusion of both HFRS and HCPS, the global geographic scope, and the emphasis on interpretability. Another strength is that the analysis reflects a realistic clinical objective: identifying patients at risk of death using information available early enough to matter. The choice of multiple complementary evaluation metrics also makes the findings more credible than if accuracy had been reported alone.

The use of a staged feature-selection pipeline reflects the complexity of the dataset and the requirements of clinical prediction modeling. Heterogeneous clinical data commonly contain correlated variables, sparse categories, and variables that may be informative only in combination. A single method would either risk retaining redundant predictors or discarding clinically relevant ones. By combining clinical exclusion, statistical filtering, wrapper-based selection, and post hoc interpretability, the present study prioritized both predictive utility and transparency. This design aligns with the broader literature recommending hybrid feature-selection strategies in ML for healthcare applications.

The study's conclusions should be interpreted in light of the retrospective design and the use of a single publicly available dataset. Although the models showed good discrimination and clinically plausible feature attribution, external validation in independent cohorts is still required before deployment. The staged feature-selection pipeline improved transparency and reduced redundancy, but it may also have removed some weak predictors that could contribute in broader settings. Therefore, the reported results support the feasibility of interpretable mortality risk stratification rather than immediate clinical deployment.

The dataset is synthetic or repository-based and should therefore be interpreted with caution as a representation of real-world cases rather than as a direct hospital registry. The current evaluation is retrospective and does not prove clinical utility. The dataset also contains possible leakage-prone variables, so the final deployment should use a strictly baseline

feature set available at presentation or admission. In addition, although the models are informative, they were assessed on a single train-test split rather than on external hospitals or countries. Future work should include external validation, temporal validation, and threshold optimization for clinical use. It would also be valuable to construct two models: one using baseline triage variables only and another using early laboratory variables, then compare them head-to-head. In the present work, SHAP-based interpretability was used to confirm the clinical plausibility of key predictors, and future studies should further expand these explanations with richer visualizations and prospective clinician-facing interfaces. A prospective validation study would be the strongest next step.

Conclusions

This study shows that mortality risk in hantavirus infection can be modeled with useful accuracy from routine clinical and epidemiological variables, with the strongest signals coming from severity, syndrome type, viral load category, age, and access-related factors. The results support the clinical value of a ML approach for early mortality risk stratification, especially in settings where rapid deterioration and delayed access to intensive care can determine outcome. The findings also reinforce that hantavirus outcome prediction is not only a statistical task but a clinically grounded one, because the most informative predictors reflect disease burden, host vulnerability, and healthcare context. Among the tested models, logistic regression achieved the best sensitivity for identifying fatal cases, while XGBoost offered the most balanced combination of discrimination and calibration. This suggests that a transparent model may be preferable when the clinical priority is to avoid missed deaths, whereas a calibrated boosting model may be better for probability-based triage decisions. The low recall of random forest in this imbalanced setting further demonstrates that overall accuracy alone is insufficient for judging clinical usefulness.

The study also highlights the importance of interpretability in AI-augmented infectious disease research. A model intended for hantavirus care should not merely predict risk, but also provide clinically plausible explanations that can be checked against known pathophysiology and epidemiology. This is especially important for ANDV and HCPS, where the disease may progress quickly, and supportive intervention must be timely. In summary, the present work provides a practical foundation for an interpretable mortality risk tool in hantavirus infection and supports future external validation in independent cohorts. Further work should focus on prospective validation, threshold optimization for triage, and comparison with clinician judgment in real-world settings.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Micheal O. Ajinaja, Yetunde O. Enigbokan, Akeem A. Abiona

Acquisition, analysis, or interpretation of data: Micheal O. Ajinaja, Yetunde O. Enigbokan, Akeem A. Abiona

Drafting of the manuscript: Micheal O. Ajinaja, Yetunde O. Enigbokan, Akeem A. Abiona

Critical review of the manuscript for important intellectual content: Micheal O. Ajinaja, Yetunde O. Enigbokan, Akeem A. Abiona

Supervision: Micheal O. Ajinaja, Yetunde O. Enigbokan, Akeem A. Abiona

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial**

relationships: All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request.

References

1. Afzal S, Ali L, Batool A, et al.: Hantavirus: an overview and advancements in therapeutic approaches for infection. Frontiers in Microbiology. 2023, 14:1233433. 10.3389/fmicb.2023.1233433
2. Chen RX, Gong HY, Wang X, et al.: Zoonotic Hantaviridae with global public health significance. Viruses. 2023, 15:1705. 10.3390/v15081705
3. Durante-Mangoni E, Cafarella I, Mercadante S, et al.: Human infection with Andes hantavirus: an update for the general physician. European Journal of Internal Medicine. 2026, 149:106946. 10.1016/j.ejim.2026.106946
4. World Health Organization: Hantavirus cluster linked to cruise ship travel, multi-country. (2026). Accessed: May 4, 2026: https://www.who.int/emergencies/disease-outbreak-news/item/2026-DON599.
5. de Oliveira SV, Faccini-Martínez ÁA: Hantavirus infection and the renal syndrome. Tropical Nephrology. Bezerra da Silva Junior G, De Francesco Daher E, Barros E (ed): Springer, Cham; 2020. 10.1007/978-3-030-44500-3_14
6. European Centre for Disease Prevention and Control: Threat assessment brief: Hantavirus-associated cluster of illness on a cruise ship: ECDC assessment and recommendations. (2026). Accessed: May 6, 2026: https://www.ecdc.europa.eu/en/publications-data/threat-assessment-brief-hantavirus-associated-cluster-illness-cruise-....
7. Tvedten H, Raskin E: Leukocyte disorders. Small Animal Clinical Diagnosis by Laboratory Methods (Fifth Edition). 2012, 63-91. 10.1016/B978-1-4377-0657-4.00004-1
8. Hachim IY, Hachim MY, Hannawi H, Naeem KB, Salah A, Hannawi S: The inflammatory biomarkers profile of hospitalized patients with COVID-19 and its association with patient's outcome: A single centered study. PLOS One. 2021, 16:e0260537. 10.1371/journal.pone.0260537
9. Stenwig E, Salvi G, Salvo Rossi P, Skjærvold NK: Comparative analysis of explainable machine learning prediction models for hospital mortality. BMC Medical Research Methodology. 2022, 22:53. 10.1186/s12874-022-01540-w
10. Brocato RL, Hooper JW: Progress on the prevention and treatment of hantavirus disease. Viruses. 2019, 11:610. 10.3390/v11070610
11. O'Dell Duplechin MO, Folds GT, Duplechin DP, Ahmadzadeh S, Myers SH, Shekoohi S, Kaye AD: Prevention and management of perioperative acute kidney injury: a narrative review. Diseases. 2025, 13:295. 10.3390/diseases13090295
12. Genç AC, Özmen E, Çekiç D, et al.: Comprehensive analyses: Using machine learning models for mortality prediction in the intensive care unit of internal medicine. Journal of Investigative Medicine. 2025, 73:460-470. 10.1177/10815589251335327
13. Guo J, Rocklöv J, Semenza JC, Sjödin H, Treskova M: Robust data integration methods for understanding associations between climate change and hantavirus infection in Europe are needed: a systematic-narrative hybrid literature review. Environmental Challenges. 2026, 23:101486. 10.1016/j.envc.2026.101486
14. Keryakos HKH, Hussein WT, Abu-El-Ela MAES, Helmy AK: Predicting mortality in critically ill patients: a machine learning approach to electrolyte imbalances and clinical risk factors. Journal of Translational Medicine. 2025, 23:1406. 10.1186/s12967-025-07311-7
15. Qin H, Wu Q, Tan J, et al.: Development of a machine learning-based prediction model for acute kidney injury associated with respiratory failure in the intensive care unit. Clinical and Experimental Medicine. 2025, 25:326. 10.1007/s10238-025-01873-y
16. Villanueva-Miranda I, Xiao G, Xie Y: Artificial intelligence in early warning systems for infectious disease surveillance: a systematic review. Frontiers in Public Health. 2025, 13:1609615. 10.3389/fpubh.2025.1609615
17. Tortosa F, Perre F, Tognetti C, et al.: Seroprevalence of hantavirus infection in non-epidemic settings over four decades: a systematic review and meta-analysis. BMC Public Health. 2024, 24:2553. 10.1186/s12889-024-20014-w

18. Alhumaidi N, Dermawan D, Kamaruzaman H, et al.: The use of machine learning for analyzing real-world data in disease prediction and management: systematic review. JMIR Medical Informatics. 2025, 13:e68898. 10.2196/68898

19. Artamonov DV, Popova PI, Korf EA, et al.: Interpretable machine learning with SHAP identifies key biomarkers in a multi-factorial spectrum of age-related neurological and metabolic conditions. International Journal of Molecular Sciences. 2026, 27:1805. 10.3390/ijms27041805

20. Zhang S, Wang Z, Su X: A study on the interpretability of network attack prediction models based on Light Gradient Boosting Machine (LGBM) and SHapley Additive exPlanations (SHAP). Computers, Materials & Continua. 2025, 83:5781-5809. 10.32604/cmc.2025.062080

21. Li J, Zhang Y, He S, Tang Y: Interpretable mortality prediction model for ICU patients with pneumonia: using shapley additive explanation method. BMC Pulmonary Medicine. 2024, 24:447. 10.1186/s12890-024-03252-x

22. European Centre for Disease Prevention and Control: Hantavirus infection - Annual Epidemiological Report for 2023. (2025). Accessed: June 24, 2026: https://www.ecdc.europa.eu/en/publications-data/hantavirus-infection-annual-epidemiological-report-2023.

23. Kazasidis O, Jacob J: Machine learning identifies straightforward early warning rules for human Puumala hantavirus outbreaks. Scientific Reports. 2023, 13:3585. 10.1038/s41598-023-30596-x

24. Kazasidis O, Geduhn A, Jacob J: High-resolution early warning system for human Puumala hantavirus infection risk in Germany. Scientific Reports. 2024, 14:9602. 10.1038/s41598-024-60144-0

25. Ferrara M: Machine learning ensemble methods for prospective hantavirus risk prediction: integrating environmental and epidemiological data. Applied Mathematical Sciences. 2026, 20:177-183. 10.12988/ams.2026.919334

26. Gao Q, Wang S, Wang Q, Cao G, Fang C, Zhan B: Epidemiological characteristics and prediction model construction of hemorrhagic fever with renal syndrome in Quzhou City, China, 2005-2022. Frontiers in Public Health. 2024, 11:1333178. 10.3389/fpubh.2023.1333178

27. Xu JS, Yang K, Quan B, Xie J, Zheng YS: A multicenter study on developing a prognostic model for severe fever with thrombocytopenia syndrome using machine learning. Frontiers in Microbiology. 2025, 16:1557922. 10.3389/fmicb.2025.1557922

28. Xu Y, Zhang L, Jiang Y, et al.: Development and validation of an interpretable machine learning model for early prediction of deterioration in patients with severe fever with thrombocytopenia syndrome. Acta Tropica. 2026, 275:108006. 10.1016/j.actatropica.2026.108006

29. Zhao C, Zhang T, Ge Z, et al.: Machine learning models for early mortality prediction in severe fever with thrombocytopenia syndrome. iScience. 2026, 29:114843. 10.1016/j.isci.2026.114843

30. Abdulhameed TZ, Al Mamlook R, Hantoosh HA, Al-Shaikhli HI, Majeed YY, Yousif SA, Gharaibeh T: Explainable machine learning models for mortality risk prediction of Crimean-Congo hemorrhagic fever in Iraq. Al-Nahrain Journal of Science. 2026, 29:128-149. 10.22401/anjs.29.1.13

31. Shu P, Wang X, Wen Z, Chen J, Xu F: SHAP combined with machine learning to predict mortality risk in maintenance hemodialysis patients: a retrospective study. Frontiers in Medicine. 2025, 12:1615950. 10.3389/fmed.2025.1615950

32. Reema K, Sayer A, Karem H, et al.: ICU mortality prediction using XGBoost-based scoring systems: a study from a developing country. Reviews on Recent Clinical Trials. 2025, 21:e15748871348585. 10.2174/0115748871348585250604065542

33. Huang J, Chen L, Luo H, et al.: Use of machine learning models to predict mortality in dialysis patients. Frontiers in Public Health. 2025, 13:1683285. 10.3389/fpubh.2025.1683285

34. Zhou S, Cai X, Yang X, et al.: Personalized machine learning-based prognostic model for ICU-acquired bloodstream infections. Frontiers in Cellular and Infection Microbiology. 2025, 15:1636886. 10.3389/fcimb.2025.1636886

35. Pillai N, Ramkumar M, Nanduri B: Artificial intelligence models for zoonotic pathogens: a survey. Microorganisms. 2022, 10:1911. 10.3390/microorganisms10101911

36. Borham A, Kamal LT, Chun S: Artificial intelligence in epidemic watch: revolutionizing infectious diseases surveillance. Frontiers in Digital Health. 2025, 7:1692617. 10.3389/fdgth.2025.1692617

37. Kaggle: Hantavirus (Andes Virus) - global epidemiology dataset. 2026,

38. Ajinaja M: Hantavirus code - Google Colab notebook. 2026,

How to cite this article:
Enigbokan Y O, Abiona A A, Ajinaja M O (August 11, 2026) Interpretable Machine Learning for Mortality Risk Stratification in Hantavirus Infection: A Global Study Across Hemorrhagic Fever with Renal Syndrome and Hantavirus Cardiopulmonary Syndrome. Cureus J Comput Sci 3 : es44389-026-00217-5. DOI https://doi.org/10.7759/s44389-026-00217-5