

Explainable Multi-View Machine Learning for Hantavirus Detection and Severity Classification Using Synthetic Data: A Computational Benchmark and Data-Integrity Assessment

Mohammad Yasir Mudasir^{1*}; Zaeem Akbar Khan²

^{1,2} Department of Computer Science, COMSATS University Islamabad, Islamabad Campus, Pakistan

ORCIDs: ¹<https://orcid.org/0009-0004-7275-2356>, ²<https://orcid.org/0009-0000-9465-343X>

Corresponding Author: Mohammad Yasir Mudasir^{1*}

Publication Date: 2026/09/29

Abstract: This study presents an explainable multi-view machine-learning framework for hantavirus detection and disease-severity classification using structured synthetic clinical and laboratory data. A public benchmark comprising 10,000 entirely synthetic records was evaluated under clinical-only, laboratory-only, and decision-level fusion settings. A genetic algorithm (GA) was used for task- and classifier-specific feature selection, and six machine-learning classifiers were evaluated using a stratified 80:20 development–test protocol. Clinical and laboratory probabilities were integrated through 10,000 Monte Carlo weight samples evaluated using out-of-fold development predictions. SHapley Additive exPlanations (SHAP) and a structured data-integrity audit were used to examine feature attribution, profile overlap, permutation performance, and dominant-feature dependence. For hantavirus detection, the fused GA–random forest (RF) model achieved a test F1-score of 0.9990, with secondary five-fold internal-validation F1 of 0.9953 plus or minus 0.0013 and area under the receiver operating characteristic curve (ROC–AUC) of 0.9998 plus or minus 0.0003. For disease-severity classification, the fused model achieved a test F1-score of 0.9981, with five-fold cross-validation F1 of 0.9989 plus or minus 0.0017 and ROC–AUC of 1.0000 plus or minus 0.0000. Laboratory variables provided the strongest predictive signal, particularly C-reactive protein (CRP), creatinine, and aspartate aminotransferase (AST) for detection, whereas severity classification was dominated by creatinine. The data-integrity audit detected no explicit identifier-based, duplicate-profile, cross-view, or target-variable leakage under the performed checks. However, the dataset creator confirmed that the records and target labels were generated synthetically and that biomarker differences were programmatically simulated. Accordingly, the near-perfect performance is interpreted as recovery of strong predictive structure within the synthetic benchmark rather than evidence of clinical validity. The framework therefore provides a transparent computational benchmark and data-integrity assessment, while evaluation on independently collected and clinically verified data would be required before any clinical interpretation or application.

Keywords: Hantavirus Detection, Disease-Severity Classification, Synthetic Data, Machine Learning, Explainable Artificial Intelligence, Decision-Level Fusion, Data-Integrity Audit.

How to Cite: Mohammad Yasir Mudasir; Zaeem Akbar Khan (2026), Explainable Multi-View Machine Learning for Hantavirus Detection and Severity Classification Using Synthetic Data: A Computational Benchmark and Data-Integrity Assessment. *International Journal of Innovative Science and Research Technology* 11(9), 1815-1837. <https://doi.org/10.38124/ijisrt/26sep765>

I. INTRODUCTION

Hantaviruses are negative-sense ribonucleic acid (RNA) viruses of the family Hantaviridae, commonly transmitted through aerosols contaminated by infected animal excreta [1–3]. Hemorrhagic fever with renal syndrome (HFRS) predominates in Asia and Europe, whereas hantavirus

cardiopulmonary syndrome (HCPS) is more common in the Americas [2, 3]. Early recognition can be difficult because symptoms overlap with those of other febrile illnesses, while progressive HFRS may involve thrombocytopenia, vascular leakage, hepatic and renal injury, hypotension, and shock [2, 3].

Laboratory indicators, including white blood cell count (WBC), platelet indices, blood urea nitrogen (BUN), creatinine, liver enzymes, coagulation variables, and inflammatory markers, have been associated with HFRS severity and prognosis [4–6]. Severity-related differences in platelet indices and other hematological, biochemical, and coagulation measures have also shown prognostic relevance [4–6], motivating computational investigation of complementary clinical and laboratory information.

Machine learning supports predictive modelling and decision-support research in health informatics [7, 8]. In hantavirus research, however, computational studies have more commonly addressed epidemiological forecasting, severity, complications, or prognosis than jointly examining record-level detection and severity classification [9, 10]. Clinical and laboratory information may provide complementary predictive structure, but unusually high performance requires careful validation and investigation of possible data-related artifacts [11, 12]. SHAP can characterize model-level predictor contributions [13], while data-integrity analyses can help determine whether high benchmark performance is associated with duplicate profiles, partition contamination, dominant predictors, or other dataset-specific dependencies.

Synthetic datasets can support reproducible computational benchmarking, algorithm development, and methodological evaluation, but performance on synthetic data must be distinguished from performance on independently collected clinical cohorts. The dataset used in this study was confirmed by its creator to be entirely synthetic: no records were collected from real patients, hospitals, laboratories, or published clinical sources, and both classification targets were assigned programmatically. Biomarker differences were also simulated to approximate plausible physiological patterns. Consequently, exceptionally high performance from individual laboratory variables may reflect predictor–target relationships introduced during synthetic data generation rather than clinically validated associations. Although the synthetic nature of the dataset is established, the complete generation algorithm and full predictor–target dependency structure are not formally documented. Evaluating predictive performance together with data integrity, feature dominance, and target dependence is therefore central to the present study.

This study presents an explainable multi-view machine-learning framework for hantavirus detection and disease-severity classification using a public synthetic structured-record benchmark. Clinical and laboratory information are treated as complementary data views. GA-based feature selection is performed separately for each task and classifier, six classifiers are evaluated, and the two views are integrated using training-only Monte Carlo decision-level fusion. SHAP provides model-level attribution, while a structured data-integrity and target-dependence audit evaluates duplicate profiles, partition overlap, feature–target associations, permutation performance, and dominant-feature sensitivity. The objective is therefore not only to compare predictive performance, but also to characterize the integrity and dependence structure underlying unusually high performance in the synthetic benchmark.

➤ *The Main Contributions are:*

- A unified explainable multi-view machine-learning framework is developed for hantavirus detection and disease-severity classification by modelling synthetic clinical and laboratory information as complementary data views and integrating their predictions at the decision level. Previous hantavirus studies have primarily examined diagnosis, severity, or prognosis as separate problems, with limited joint evaluation of clinical and laboratory information across both classification tasks [9, 10, 14–16].
- A controlled comparative evaluation is performed across clinical-only, laboratory-only, and decision-level fused settings using decision tree (DT), k-nearest neighbors (KNN), logistic regression (LR), support vector machine (SVM), RF, and Extreme Gradient Boosting (XGBoost) under the same experimental protocol. This design enables direct comparison of the predictive contribution of each data view and the incremental benefit provided by fusion. Previous hantavirus prediction studies have mainly evaluated individual statistical or machine-learning models within specific data settings, with limited systematic comparison of clinical, laboratory, and fused representations [9, 10, 16].
- A task-specific and classifier-specific GA feature-selection strategy is integrated with training-only Monte Carlo probability fusion and SHAP-based attribution. Clinical and laboratory fusion weights are selected from 10,000 candidate combinations using out-of-fold development probabilities before independent test evaluation. This design combines feature selection, decision-level integration, and model attribution within a reproducible benchmarking framework [9, 10, 13, 16].
- A structured data-integrity and target-dependence audit is incorporated to assess duplicate-profile and cross-partition overlap, cross-view alignment, single-feature dominance, label-permutation performance, preprocessing isolation, and dominant-feature sensitivity. In the context of an entirely synthetic dataset, this audit is used not only to check for conventional leakage but also to characterize strong predictor–target dependencies that may arise from the data-generation process. It therefore provides an explicit basis for distinguishing internally consistent benchmark performance from evidence of clinical validity [11, 12].

II. RELATED WORK

Hantavirus research spans epidemiology, diagnosis, prognosis, and computational surveillance. Infection commonly follows exposure to rodent-contaminated aerosols [1, 2], while HFRS predominates in Asia and Europe and may involve thrombocytopenia, vascular leakage, hepatic and renal injury, coagulation abnormalities, hypotension, and shock [1, 2, 17]. These manifestations have motivated research ranging from diagnostic testing and severity assessment to population-level forecasting and computational risk modelling.

➤ *Clinical, Diagnostic, and Prognostic Studies*

HFRS severity and prognosis have been associated with laboratory markers, including platelet indices, WBC, creatinine, BUN, liver markers, coagulation variables, and inflammatory biomarkers [4–6, 18]. Du et al. [4] examined

clinical characteristics and outcomes in critically affected patients with HFRS, while Fan et al. [5] investigated platelet distribution width as a potential indicator of disease severity and survival. Du et al. [18] evaluated pentraxin-3 in relation to severity and mortality risk, whereas Chen et al. [6] analyzed hematological, biochemical, and coagulation parameters in relation to prognosis. Collectively, these studies provide clinical evidence that laboratory measurements can contain information relevant to HFRS severity and outcomes.

Diagnostic research has examined several complementary approaches. Brockmann et al. [14] evaluated the Aachen Puumala Virus Score as a clinical screening tool intended to reduce unnecessary testing. Goeijenbier et al. [19] investigated underdiagnosis of Puumala hantavirus infection among patients suspected of leptospirosis, illustrating diagnostic overlap among febrile infectious diseases. Laboratory-based diagnostic studies have also evaluated rapid testing [20] and enzyme-immunoassay/immunoblot methods [21]. These studies address diagnosis through clinical scoring or laboratory assays rather than the multi-view machine-learning benchmark considered in the present work.

Severity modelling has included regression-based clinical tools, retrospective and pediatric nomograms [9, 15], least absolute shrinkage and selection operator (LASSO) with LR [10], and the multicenter EASTAR system [16]. These studies demonstrate the use of structured clinical and laboratory variables for severity or risk assessment in clinical cohorts. Their objectives and data sources, however, differ from the present study, which jointly benchmarks detection and severity classification across separate synthetic clinical and laboratory views.

➤ *Computational Hantavirus Studies*

Computational hantavirus research has also addressed population-level forecasting and regional risk. Autoregressive integrated moving average (ARIMA) modelling has been used to forecast HFRS incidence in China [22], while ensemble machine learning [23] and XGBoost-based forecasting [24] have been investigated for temporal incidence modelling. These studies demonstrate the use of statistical and machine-learning methods for epidemiological surveillance, but their targets concern population-level incidence rather than record-level infection-status or severity classification.

Regional and ecological modelling has similarly examined the geographic distribution and environmental risk of hantavirus infection. She et al. [25] investigated regional HFRS risk, Kazasidis and Jacob [26] used machine learning to identify early-warning rules for Puumala hantavirus outbreaks, and Tian et al. [27] analyzed distribution dynamics of Hantaan and Seoul virus infections using machine-learning modelling. These approaches address surveillance, outbreak prediction, and regional epidemiology and therefore represent a different computational setting from classification based on structured record-level variables.

For record-level classification, LR, SVM, KNN, and DT provide conventional machine-learning baselines, whereas RF and XGBoost can represent nonlinear relationships and interactions among predictors [28, 29]. Within the hantavirus

literature reviewed here, comparatively limited attention has been given to evaluating clinical and laboratory information as separate predictive views and then integrating their model probabilities at the decision level. This motivates a controlled comparison of clinical-only, laboratory-only, and fused representations within a common experimental protocol.

➤ *Research Gap and Study Positioning*

Existing screening, severity, and prognostic studies [9, 10, 14–16] provide evidence from their respective clinical settings, but their objectives differ from a unified comparison of infection-status and severity classification across separate clinical, laboratory, and probability-fused views. Likewise, forecasting and spatial studies [22–27] primarily address incidence, outbreak prediction, or regional risk rather than record-level classification.

A related methodological question concerns the relative predictive contribution of different data views. Clinical scores, individual biomarkers, nomograms, and task-specific modelling pipelines have been investigated in previous studies, but less attention has been given to modelling clinical and laboratory views separately under the same protocol and quantifying whether decision-level probability fusion provides incremental predictive information.

A further issue is particularly important for synthetic health-data benchmarks. Very high predictive performance may arise from strong predictor–target relationships deliberately or inadvertently introduced during data generation. Such relationships are conceptually distinct from conventional train–test leakage: a model may be evaluated under a properly isolated experimental protocol and still obtain near-deterministic performance because the synthetic target and predictor distributions were generated with strong dependencies. Consequently, evaluation of synthetic benchmarks should consider not only predictive metrics but also duplicate profiles, partition overlap, cross-view alignment, feature–target associations, permutation behavior, and dominant-feature sensitivity.

The present study evaluates Hantavirus_Positive and Disease_Severity as separate synthetic classification targets using the public dataset of [30]. The dataset creator confirmed through direct correspondence that all 10,000 records and both target labels were generated synthetically and were not derived from real patients, hospitals, laboratories, or published clinical sources. Biomarker differences were also programmatically simulated. Although this establishes the synthetic provenance of the benchmark, the complete generation algorithm and full predictor–target dependency structure are not formally documented. Accordingly, the dataset is treated strictly as a synthetic computational benchmark rather than a clinically validated cohort.

Within this setting, the proposed framework combines GA-based feature selection, six classifiers, training-only Monte Carlo decision-level fusion using out-of-fold development probabilities, SHAP attribution, and a structured data-integrity and target-dependence audit. The audit is used to distinguish conventional data leakage from strong benchmark dependencies that may arise from synthetic generation.

Transparent reporting of these issues is consistent with broader guidance emphasizing rigorous prediction-model reporting and risk-of-bias assessment [11, 12]. Table 1 summarizes

representative hantavirus studies and positions the present benchmark relative to their objectives and data settings.

Table 1 Representative Clinical and Computational Hantavirus Studies and Their Relation to the Present Synthetic Benchmark

Studies	Focus	Approach	Relation to Present Study
[1]; [2]	Clinical background	Epidemiology, pathogenesis, diagnosis	Provides clinical and epidemiological context rather than an ML benchmark.
[4]; [5]; [18]; [6]	Severity/prognosis	Laboratory and inflammatory biomarkers	Provides clinical evidence on severity-related biomarkers; no multi-view synthetic benchmark.
[14]; [19]; [20]; [21]	Diagnosis	Clinical scores, serology, diagnostic assays	Addresses clinical diagnosis rather than synthetic multi-view ML classification.
[9]; [15]; [10]; [16]	Severity/prognosis	Nomograms, LASSO, LR	Clinical severity/risk modelling rather than joint synthetic detection–severity benchmarking.
[22]; [23]; [24]	Forecasting	ARIMA and machine-learning forecasting	Population-level temporal prediction rather than record-level classification.
[25]; [26]; [27]	Regional risk	Geographic/ecological machine learning	Population and regional modelling rather than structured-record classification.
Proposed framework	Synthetic detection/severity benchmark	GA, six classifiers, fusion, SHAP, integrity audit	Compares clinical and laboratory synthetic views while explicitly auditing data integrity and generation-related target dependence.

III. DATASET DESCRIPTION AND PREPROCESSING

➤ Dataset Source and Scope

The public Hantavirus Detection Dataset [30] contains 10,000 entirely synthetic structured records. Direct clarification from the dataset creator established that the records were generated programmatically for machine-learning testing, coursework, and algorithmic evaluation and were not collected from real patients, hospitals, laboratories, or published clinical sources.

The dataset contains demographic, symptoms, exposure, virus-type, laboratory, and derived machine-learning fields. No genomic sequencing data are included; Hantavirus_Type is a categorical synthetic variable rather than a sequencing-derived measurement. The present study retained only demographic, symptom, exposure, virus-type, and laboratory variables relevant to the two classification tasks.

The dataset creator further clarified that both Hantavirus_Positive and Disease_Severity were assigned synthetically and that biomarker differences were programmatically simulated to approximate plausible physiological patterns. The creator specifically indicated that relationships involving laboratory variables such as creatinine

and BUN were introduced in the synthetic generation process. Severity was not defined using a disclosed rigid laboratory cutoff; however, the complete generation algorithm and full predictor–target dependency structure are not available in a formal technical publication or accompanying documentation. The dataset has not been reviewed or validated by healthcare professionals.

Accordingly, the dataset is treated strictly as a synthetic computational benchmark. Its class distributions, biomarker relationships, and predictive performance should not be interpreted as estimates of real-world disease prevalence, clinical diagnostic performance, or prognostic validity. Because the dataset contains no real human participants or identifiable patient data, human-subject ethical approval and informed consent are not applicable to the dataset used in this study. Table 2 summarizes its main characteristics.

The additional provenance and construction details reported above were obtained through direct correspondence with the dataset creator. Because no formal technical publication describing the complete synthetic generation process is available, this information is reported as a provenance clarification rather than as independently peer-reviewed evidence.

Table 2 Summary of the Synthetic Hantavirus Structured-Record Benchmark

Attribute	Value
Total synthetic records	10,000
Positive/negative labels	2,571/7,429
Positive-label proportion	25.71%
Male/female categories	5,068/4,932
Age, mean $\pm$ standard deviation	51.1 $\pm$ 19.36 years
Age, median/range	51; 18–84 years

➤ Data Views and Classification Tasks

The dataset was organized into two complementary data views. The clinical view comprised synthetic demographic, symptom, exposure, and task-specific virus-type variables,

whereas the laboratory view comprised WBC, platelet count, CRP, alanine aminotransferase (ALT), AST, BUN, and creatinine. Separating the views enabled their predictive

contributions within the synthetic benchmark to be evaluated independently before probability-level fusion.

The field Patient_ID was treated solely as a synthetic record identifier and was used only to align corresponding clinical and laboratory rows before model development. Region and Sample_Date were treated as metadata, while Symptom_Count was excluded because it was redundant with the individual symptom indicators.

Two binary classification tasks were defined:

▪ **Task 1: Hantavirus Detection:**

Hantavirus_Positive was the synthetic target label. Severity, virus type, identifiers, and outcome-related variables were excluded from the predictors. This task evaluated whether

the retained synthetic clinical and laboratory variables could distinguish records assigned positive and negative labels by the dataset-generation process.

▪ **Task 2: Severity Classification:**

Disease_Severity was the synthetic target among records assigned a positive Hantavirus_Positive label and having an available severity label. Hantavirus_Positive was excluded from the predictors, while virus-type indicators were retained. This task evaluated discrimination between the synthetically assigned severity categories within the positive-label subset.

Missing target labels were not imputed. Severity encoding and class distributions are reported in Section 5, while Table 3 summarizes the predictor configurations used for both tasks.

Table 3 Task-specific predictor configurations

Task	Setting	Predictors/configuration
Task 1	Clinical	Gender, age, symptoms, exposures
Task 1	Laboratory	WBC, platelet count, CRP, ALT, AST, BUN, creatinine
Task 1	Fusion	Training-only clinical/laboratory probability fusion
Task 2	Clinical	Gender, age, symptoms, exposures, encoded virus types
Task 2	Laboratory	WBC, platelet count, CRP, ALT, AST, BUN, creatinine
Task 2	Fusion	Training-only clinical/laboratory probability fusion

➤ **Preprocessing and Leakage Controls**

Symptoms were binary encoded, exposures were one-hot encoded, and virus types were encoded only for Task 2. Missing symptom indicators were treated as absent, unknown exposures were represented as Exposure_Unknown, laboratory values were median-imputed, and continuous variables were min–max scaled.

All preprocessing operations were fitted using training data only and were re-estimated within each validation n fold. This procedure prevented information from validation or test records from influencing imputation, scaling, or other preprocessing parameters during model development.

Task 1 excluded Disease_Severity, Hantavirus_Type, Sample_Date, Patient_ID, and other outcome-related variables. Task 2 excluded identifiers, unrelated metadata, and the constant Hantavirus_Positive field. GA-based feature selection and model development were confined to the development partition.

These controls were designed to reduce conventional information leakage through identifiers, target-related variables, preprocessing operations, or held-out records. They do not, however, remove predictor–target dependencies introduced during synthetic data generation. This distinction is important because the dataset creator confirmed that biomarker differences were programmatically simulated. Additional analyses of conventional leakage, feature dominance, and synthetic target dependence are therefore reported separately in the data-integrity audit.

➤ **Partition Alignment and Data Settings**

Clinical and laboratory records were aligned using the synthetic Patient_ID field before identifier removal. Identical development–test assignments and corresponding internal

folds were maintained across the two views. This ensured that clinical and laboratory probabilities being combined during decision-level fusion corresponded to the same synthetic records. Models were trained independently for each view and fused at the probability level rather than through direct feature concatenation.

The dataset was evaluated exclusively as an offline synthetic computational benchmark. No data were acquired through connected medical devices or real-time healthcare systems, and no device-level, network-level, or real-time communication experiments were performed. Accordingly, the study evaluates the analytical machine-learning pipeline independently of any healthcare data-acquisition or deployment infrastructure.

IV. PROPOSED METHODOLOGY

The proposed framework performs hantavirus detection and disease-severity classification on aligned synthetic clinical and laboratory records using task-specific preprocessing, GA-based feature selection, six classifiers, training-only decision-level fusion, and SHAP attribution. Clinical and laboratory views are first modelled independently, after which their predicted probabilities are combined to evaluate whether the two views provide complementary predictive information within the synthetic benchmark. Fig. 1 summarizes the complete workflow.

➤ **Framework and Data Processing**

Clinical predictors comprised synthetic demographic, symptom, exposure, and task-specific virus-type variables, whereas laboratory predictors represented synthetic hematological, inflammatory, hepatic, and renal measurements. Treating these variables as two independent data views enabled their predictive contributions within the

benchmark to be examined separately before decision-level fusion.

The two views were modelled independently and subsequently combined at the probability level. Records were aligned using the synthetic record identifier Patient_ID before identifier removal, with identical partitions maintained across views. This ensured that clinical and laboratory probabilities corresponded to the same synthetic records during fusion. Preprocessing followed the procedures described in Section 3.

➤ Genetic Algorithm-Based Feature Selection

GA-based feature selection was performed separately for each task, data view, and classifier. This design allowed the selected feature subset to adapt to the classification objective and the characteristics of the corresponding classifier rather than imposing a common subset across all models. Genetic algorithms provide a population-based evolutionary search strategy [31, 32]. Each candidate feature subset was encoded as a binary chromosome [33]:

$$z = (z_1, z_2, \dots, z_d), \quad z_j \in \{0, 1\} \quad (1)$$

Where $z_j = 1$ indicates inclusion of predictor x_j ;

$$S(z) = \{x_j \mid z_j = 1, j = 1, \dots, d\} \quad (2)$$

Candidate subsets were evaluated using five-fold stratified validation within the development partition according to

$$\mathcal{F}(z) = \frac{1}{5} \sum_{k=1}^5 F 1_k(S(z)) - \lambda \left(\frac{\|z\|_0}{d} \right) \quad (3)$$

Where $\|z\|_0$ denotes the number of selected features, d is the number of candidate features, and $\lambda = 0.01$. The first term rewards predictive performance, whereas the second applies a small penalty proportional to the selected-feature fraction to encourage more compact subsets.

The GA used a population of 40 chromosomes over 50 generations, tournament selection with size 3, single-point crossover with probability 0.80, bit-flip mutation with probability 0.05 per gene, elitism of two chromosomes, and random seed 42. The complete settings are summarized in Table 4. The GA is treated as the feature-selection procedure used within the present benchmark; the study does not claim that it outperforms simpler feature-selection strategies, which were not evaluated comparatively.

➤ Machine-Learning Models

Six classifiers were evaluated: LR [34], SVM [35], KNN [36], DT [37], RF [28], and XGBoost [29]. These models span linear, distance-based, kernel-based, tree-based, and ensemble approaches, allowing the GA-selected feature subsets to be evaluated across different classification mechanisms.

Each classifier was trained independently on its GA-selected clinical and laboratory feature subsets. Classifier configurations were specified during model development before final test evaluation and are reported in Table 5.

➤ Training-Only Decision-Level Fusion

Clinical and laboratory probabilities were combined through weighted decision-level fusion [38, 39]:

$$P_{\text{fused}}(x) = w_c P_c(x) + w_l P_l(x) \quad (4)$$

Subject to

$$w_c + w_l = 1, \quad 0 \leq w_c, w_l \leq 1 \quad (5)$$

Where $P_c(x)$ and $P_l(x)$ denote the clinical-view and laboratory-view positive-class probabilities, respectively. The weights w_c and w_l determine the relative contributions of the two views to the fused probability.

Following a Monte Carlo sampling strategy [40], w_c was sampled from $[0, 1]$, with $w_l = 1 - w_c$. A total of 10,000 candidate weight pairs were evaluated using out-of-fold development probabilities and F1-score. The weight pair yielding the highest development F1-score was fixed before evaluation on the independent test partition. Using out-of-fold probabilities reduced the risk of selecting fusion weights from predictions generated for the same observations used to fit the corresponding fold-specific models.

Because the constraint $w_l = 1 - w_c$ leaves only one independent scalar weight, this procedure is interpreted as stochastic weight selection over a one-dimensional search space rather than as a general-purpose optimization algorithm. The 10,000 samples provide dense stochastic coverage of the bounded weight interval; this sampling budget is not claimed to be necessary or optimal.

➤ Validation and Model Attribution

Preprocessing, GA-based feature selection, classifier development, and fusion-weight estimation were restricted to the development partition. Internal validation used stratified folds with corresponding record assignments maintained across the two views. After development-stage procedures were completed, the resulting pipelines were evaluated on the independent test partition, which was not used for feature or fusion-weight selection.

F1-score was used as the primary performance criterion, while accuracy, precision, recall, ROC-AUC, and confusion matrices were reported as additional evaluation measures. These metrics characterize complementary aspects of classification performance within the synthetic benchmark.

SHAP was applied to the final models to characterize model-level predictor attribution [13]. Mean absolute SHAP values were used to summarize global importance, while beeswarm plots illustrated the distribution and direction of feature contributions; TreeSHAP was used for tree-based models. SHAP values were interpreted strictly as model attributions rather than causal, biological, or clinical effects. Because the predictors and targets were generated synthetically and the complete generation rules are not formally documented, attribution patterns characterize dependencies learned within this benchmark only and must not be interpreted as evidence of real-world biomarker importance.

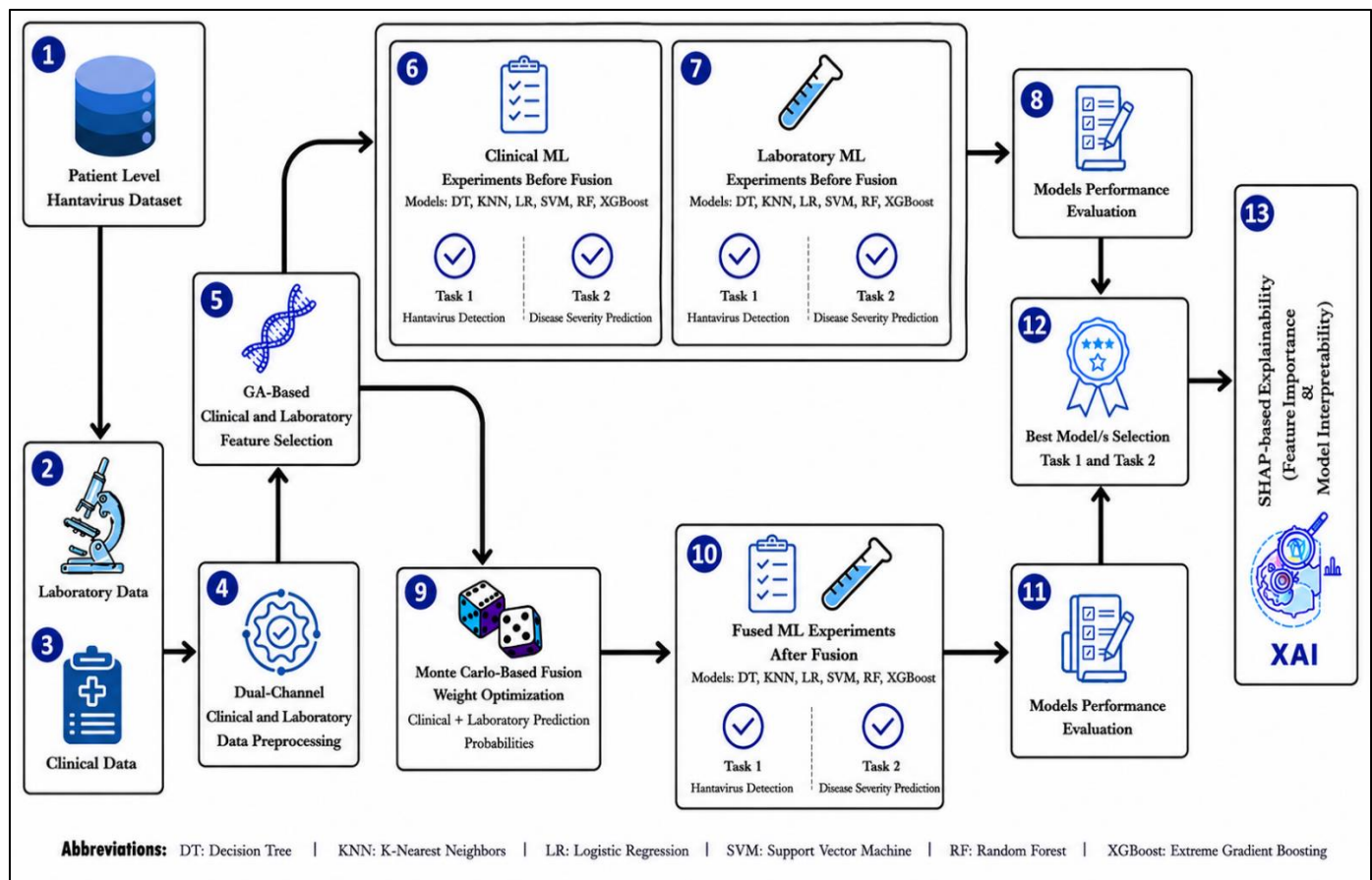


Fig 1 Workflow of the Proposed Explainable Machine-Learning Framework for Hantavirus Detection and Disease-Severity Classification on the Synthetic Clinical and Laboratory Benchmark

V. EXPERIMENTAL SETUP

➤ Implementation and Validation Protocol

Experiments were conducted using Python 3.10 with Pandas, NumPy, Scikit-learn, XGBoost, Matplotlib, and SHAP on an Intel Core i7-1265U system with 32 GB random-access memory (RAM). For each task, synthetic records were stratified into 80% development and 20% independent test partitions using seed 42, with identical record assignments maintained across the clinical and laboratory views. This alignment ensured that probabilities combined during decision-level fusion corresponded to the same synthetic records.

All preprocessing, GA-based feature selection, classifier development, and fusion-weight estimation were performed using the development partition. Five-fold stratified validation was used during development, and the resulting pipelines were subsequently evaluated on the independent test partition. The

test partition was therefore excluded from feature selection and fusion-weight estimation.

Task 1 used the synthetically assigned Hantavirus_Positive target across all 10,000 records, comprising 7,429 negative-labelled and 2,571 positive-labelled records. Task 2 retained the 2,571 records assigned a positive Hantavirus_Positive label and encoded the synthetic Disease_Severity target as

$$\text{Disease Severity} = \begin{cases} 0, & \text{Moderate,} \\ 1, & \text{Severe,} \end{cases} \quad (6)$$

Yielding 1,240 moderate-labelled and 1,331 severe-labelled records. Missing target labels were excluded rather than imputed. These class counts describe the synthetic benchmark and should not be interpreted as estimates of real-world disease prevalence or severity distributions.

Table 4 GA Configuration for Wrapper-Based Feature Selection

Component	Configuration
Representation	Binary feature-inclusion chromosome
Population / generations	40 / 50
Selection	Tournament, size 3
Crossover	Single point, 0.80
Mutation	Bit-flip, 0.05 per gene
Elitism	Two chromosomes
Validation	Five-fold stratified
Fitness	Mean F1 – 0.01 × selected-feature proportion
Seed	42

➤ Genetic Algorithm and Classifier Configuration

GA-based feature selection followed Equation 3 independently for each task, view, and classifier. Candidate subsets were evaluated using five-fold stratified validation within the development partition. Table 4 summarizes the settings.

LR, SVM, DT, and RF used balanced class weights, KNN used distance-weighted voting, and XGBoost used no explicit class weighting. These configurations were kept fixed after development-stage specification so that the independent test partition was not used for classifier tuning. The configurations in Table 5 were applied to both clinical and laboratory views.

Table 5 Final Classifier Configurations used in the Benchmark

Classifier	Configuration
LR	Limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) solver; 2,000 iterations; balanced weights; seed 42.
SVM	Radial basis function (RBF) kernel; $C = 1.0$; $\gamma = \text{scale}$; balanced weights; probability enabled; seed 42.
KNN	3 neighbors; distance weighting; Euclidean distance.
DT	Gini; depth 10; minimum split 10; minimum leaf 5; balanced weights; seed 42.
RF	200 trees; depth 20; $\text{max_features}=\text{sqrt}$; minimum split 10; minimum leaf 5; balanced weights; seed 42; $n_jobs=-1$.
XGBoost	300 estimators; depth 4; learning rate 0.05; subsample 0.80; column sampling 0.80; binary:logistic; logloss; seed 42; $n_jobs=-1$.

➤ Monte Carlo Fusion and Evaluation

Corresponding clinical and laboratory classifiers were combined according to Equation 4; GA-based feature selection was not repeated during fusion. Fusion therefore operated on probabilities generated by the already developed view-specific pipelines rather than introducing an additional feature-selection stage.

Using seed 42, 10,000 candidate weights were sampled according to

$$w_c \sim \mathcal{U}(0,1), \quad w_l = 1 - w_c \quad (7)$$

Candidate weight pairs were evaluated using out-of-fold development probabilities, F1-score, and a classification threshold of 0.50. When multiple sampled pairs produced the same maximum F1-score, the first encountered maximum was retained. The selected pair was then fixed before independent test evaluation. The same procedure was applied consistently to corresponding clinical and laboratory classifiers.

Preprocessing followed Section 3 and was estimated from training data within the relevant development folds. Stratification was used to preserve class proportions. F1-score was the primary selection metric; accuracy, precision, recall, ROC–AUC, and confusion matrices were additionally reported on the held-out test records.

The independent test partition was used only for final evaluation and was not used to select features or fusion weights. The five-fold results reported subsequently are secondary internal-validation summaries and should not be interpreted as external validation.

➤ Interpretation of Model Comparisons

All model comparisons are interpreted within the synthetic benchmark. Because the records, target labels, and biomarker differences were generated programmatically, high discrimination does not establish clinical diagnostic or prognostic validity. In particular, separation of development and test partitions can reduce conventional train–test contamination but cannot eliminate predictor–target dependencies introduced during synthetic data generation.

These dependencies are examined separately through the data-integrity and target-dependence audit.

Performance differences between classifiers and data views are reported descriptively. No paired statistical significance testing or confidence-interval analysis was performed; consequently, small numerical differences are not interpreted as evidence of statistically significant superiority. Similarly, the GA is evaluated as the feature-selection procedure used in the present pipeline rather than as a method demonstrated to outperform simpler feature-selection alternatives.

The decision-level fusion procedure used the classifiers predicted probabilities directly. A separate probability-calibration analysis was not performed; consequently, the fused probabilities are used for comparative classification rather than interpreted as calibrated estimates of clinical risk. Under the constraint $w_l = 1 - w_c$, fusion involves a one-dimensional weight search. The 10,000 Monte Carlo samples provide dense stochastic coverage of this bounded interval but are not claimed to represent a necessary or optimal sampling budget.

Accordingly, the experimental protocol is intended to evaluate computational behavior, feature dependence, and complementary information across synthetic data views rather than to establish clinical utility, statistical superiority, or deployment readiness.

➤ Computational Reproducibility and Control

All stochastic procedures used fixed random seed 42, and identical development–test assignments were maintained across the clinical and laboratory views. The GA settings, classifier configurations, preprocessing protocol, and Monte Carlo fusion procedure were documented and applied consistently across the two classification tasks. These controls support computational repeatability of the reported experimental pipeline and provide a common basis for comparing clinical, laboratory, and fused settings.

The use of a single fixed random seed should not be interpreted as evidence of robustness to stochastic variation. Repeated experiments across multiple random seeds were not

performed and therefore remain a limitation of the present benchmark.

VI. DATA-INTEGRITY AND TARGET-DEPENDENCE AUDIT

A structured audit was conducted to examine conventional information leakage, cross-view alignment, duplicate profiles, feature–target associations, single-feature performance, label-permutation behavior, and dominant-feature dependence across the 10,000 Task 1 records and 2,571 Task 2 records. The audit was motivated by the unusually high predictive performance observed in the synthetic benchmark and was designed to distinguish two conceptually different sources of high performance: (i) conventional data leakage or experimental contamination and (ii) strong predictor–target dependencies arising from the synthetic data-generation process.

The first category includes prohibited target or identifier inputs, train–test profile overlap, preprocessing contamination, and misalignment of the clinical and laboratory views. The second category does not necessarily constitute conventional leakage: a synthetically generated dataset may contain strong or near-deterministic predictor–target relationships even when development and test partitions are correctly isolated. This distinction is particularly important here because the dataset creator confirmed that target labels were assigned synthetically and that biomarker differences were programmatically simulated. The audit therefore evaluates both conventional data-integrity risks and the observable dependence of benchmark performance on individual synthetic predictors.

➤ *Input, Alignment, and Profile Checks*

The synthetic record identifier `Patient_ID` was used only for cross-view alignment and was removed before modelling. For Task 1, `Hantavirus_Positive` was used only as the target, while `Disease_Severity`, `Hantavirus_Type`, `Sample_Date`, identifiers, and other outcome-related variables were excluded from the predictor set. For Task 2, `Disease_Severity` was used only as the target, while the constant `Hantavirus_Positive` field, identifiers, and unrelated metadata were excluded. Final model inputs were programmatically inspected before feature selection to verify that prohibited identifiers, target variables, and excluded outcome-related fields were not supplied to the classifiers.

Clinical and laboratory records were matched one-to-one using the synthetic record identifier before its removal. Corresponding records were required to have identical target labels and development–test assignments across views. This check was necessary because decision-level fusion assumes that clinical and laboratory probabilities correspond to the same synthetic record.

Duplicate identifiers and exact predictor profiles were examined using row-level hashes within and across partitions. Exact cross-partition profile matches were treated as potential evidence of contamination, while repeated predictor profiles were additionally compared for conflicting target labels. These checks were designed to detect direct record duplication or partition contamination; they do not assess whether the

synthetic generator produced different records according to similar underlying rules.

➤ *Association, Single-Feature, and Sensitivity Analyses*

Mutual information and Spearman correlation were used to characterize feature–target associations. Mutual information was used to detect general statistical dependence, whereas Spearman correlation characterized monotonic relationships. Strong associations were treated as signals requiring further investigation rather than, by themselves, as evidence of data leakage.

Each predictor was additionally evaluated in isolation using five-fold univariate LR and depth-one DT models, with F1-score and ROC–AUC used to quantify predictive performance. Single predictors approaching the performance of the corresponding multivariable model were considered evidence of strong feature dominance within the synthetic benchmark. Such dominance does not by itself demonstrate conventional leakage, particularly because the dataset creator confirmed that biomarker differences were deliberately introduced during synthetic generation.

Dominant predictors were identified using evidence from GA selection, association analysis, single-feature performance, and SHAP attribution. The development pipeline was then repeated after removal of the dominant predictor, and changes in F1-score and ROC–AUC were used to quantify model dependence on that variable. A large performance decline indicates that much of the benchmark’s discriminative information is concentrated in the removed predictor, whereas a smaller decline indicates that predictive information is more distributed across the retained variables.

➤ *Label-Permutation Test*

Label-permutation testing was used to determine whether the observed predictive performance exceeded that expected after disrupting the feature–label relationship. Development labels were permuted 1,000 times using seed 42, with performance evaluated through five-fold stratified validation. The resulting scores formed an empirical null distribution.

The empirical significance level was calculated as

$$p_{\text{perm}} = \frac{1 + \sum_{b=1}^B \mathbb{I}(F1_b \geq F1_{\text{obs}})}{B + 1}, \quad B = 1,000, \quad (8)$$

Where $F1_{\text{obs}}$ denotes the observed F1-score and $F1_b$ denotes the score obtained for permutation b . Values of $p_{\text{perm}} < 0.05$ indicate that the observed performance exceeds the empirical null distribution under the specified permutation procedure.

Importantly, this test evaluates whether predictive structure remains when the original feature–label association is preserved relative to randomized labels; it does not establish that the observed association is clinically valid or independently derived. In a synthetic dataset, a significant permutation result is also expected when strong predictor–target relationships were deliberately introduced by the generator. Accordingly, the permutation analysis is interpreted here as evidence of nonrandom predictive structure within the

benchmark rather than evidence against synthetic target-construction dependence.

➤ *Audit Protocol and Interpretation*

Table 6 summarizes the audit components, procedures, and interpretation criteria.

The audit is intended as a diagnostic assessment of both conventional data-integrity problems and unusually strong predictive dependencies in a synthetic health-data benchmark. The input, duplicate, alignment, and partition checks address conventional sources of information leakage and experimental contamination. Association analysis, single-feature baselines, permutation testing, and dominant-feature removal address a different question: whether high benchmark performance is concentrated in strong predictor–target relationships.

This distinction is essential for interpreting the present dataset. The dataset creator confirmed that biomarker

differences were programmatically simulated, while the complete generation algorithm and full predictor–target dependency structure are not formally documented. Consequently, absence of conventional leakage under the performed checks cannot establish independence between predictors and synthetic target construction. Likewise, significant permutation performance and strong single-feature discrimination demonstrate recoverable structure within the synthetic benchmark but do not establish clinical validity.

The audit therefore does not attempt to validate the dataset as a clinical cohort. Instead, it provides a structured procedure for identifying direct experimental contamination and characterizing feature dominance and generation-related target dependence. Its findings are used alongside the predictive results to prevent near-perfect synthetic-benchmark performance from being interpreted as evidence of real-world diagnostic or prognostic performance.

Table 6 Data-Integrity and Target-Dependence Audit Protocol for the Synthetic Benchmark

Component	Procedure	Interpretation
Input inspection	Identifiers, targets, metadata, and outcome-related fields checked	Detects prohibited model inputs.
Duplicate analysis	Identifier counts and exact-profile hashes within/across partitions	Cross-partition matches indicate potential contamination.
Conflicting labels	Repeated predictor profiles compared across target labels	Identifies identical profiles assigned different labels.
Cross-view alignment	One-to-one synthetic-record matching across views	Targets and partition assignments must correspond.
Association analysis	Mutual information and Spearman correlation	Strong dependence triggers further investigation; not leakage by itself.
Single-feature baselines	Five-fold univariate LR and depth-one DT	Near-full-model performance indicates feature dominance.
Label permutation	1,000 permutations; seed 42	Tests observed performance against a randomized-label empirical null.
Dominant-feature removal	Development pipeline repeated without dominant predictor	Performance loss quantifies dependence on the removed feature.

VII. RESULTS AND ANALYSIS

Clinical, laboratory, and fused models were evaluated on identical held-out test records. Accuracy, precision, recall, and F1-score were reported, with ROC–AUC additionally summarized in the secondary five-fold analysis. This common evaluation setting enables direct comparison of the predictive contribution of the clinical and laboratory views and the incremental effect of decision-level fusion. Because the records, labels, and biomarker differences were generated synthetically, all findings are interpreted as performance within the synthetic computational benchmark rather than as estimates of clinical diagnostic or prognostic performance.

➤ *Data-Integrity and Target-Dependence Audit Results*

Table 7 summarizes the audit findings. No duplicate identifiers, exact cross-partition predictor profiles, conflicting labels among repeated profiles, or cross-view alignment mismatches were detected. All 10,000 Task 1 records and 2,571

Task 2 records were consistently aligned between the clinical and laboratory views. These findings provide evidence against the specific forms of identifier-based, exact-profile, cross-partition, and cross-view contamination examined by the audit. They do not exclude dependencies introduced during synthetic data generation.

The single-feature analyses revealed unusually strong benchmark-level predictive structure. For Task 1, CRP_Scaled alone achieved $F1 = 0.9824$ and $ROC - AUC = 0.9988$. For Task 2, Creatinine_Scaled alone achieved $F1 = 0.9981$ and $ROC - AUC = 0.9999$. Thus, a single laboratory variable approached the multivariable performance in each task, with the effect particularly pronounced for Task 2.

The permutation analyses further confirmed that the observed feature–label structure was substantially stronger than that obtained after randomization. For Task 1, the observed F1-score was 0.9990 compared with a permutation-null mean of

0.206 ± 0.018 . For Task 2, the observed F1-score was 0.9981 compared with a null mean of 0.501 ± 0.024 . Both analyses yielded $p_{perm} = 0.001$. These results demonstrate recoverable non-random predictive structure within the synthetic benchmark. They do not establish that the structure represents independently observed clinical relationships, because both the labels and biomarker distributions were generated programmatically.

Dominant-feature removal provided additional evidence about the concentration of predictive information. Removing

CRP reduced Task 1 F1 from 0.9990 to 0.9912, indicating that substantial predictive information remained in the other variables. In contrast, removing creatinine reduced Task 2 F1 from 0.9981 to 0.6897, demonstrating strong dependence of severity classification on this variable. This finding is consistent with the dataset creator's clarification that biomarker differences between synthetic severity classes were programmatically simulated. However, because the complete generation algorithm and full predictor–target dependency structure are not formally documented, the analysis cannot determine the exact role of creatinine in constructing the synthetic severity labels.

Table 7 Summary of the Data-Integrity and Target-Dependence Audit

Audit component	Result	Interpretation
Identifier/profile checks	No duplicate identifiers, exact cross-partition profiles, or conflicting labels	No conventional contamination detected under these checks
Cross-view alignment	Task 1: 10,000; Task 2: 2,571; 0 mismatches	Clinical and laboratory views consistently aligned
Task 1 single feature	CRP_Scaled: F1 = 0.9824; ROC–AUC = 0.9988	Strong single-feature benchmark discrimination
Task 2 single feature	Creatinine_Scaled: F1 = 0.9981; ROC–AUC = 0.9999	Near-complete single-feature benchmark discrimination
Task 1 permutation	F1 = 0.9990; null = 0.206 ± 0.018 ; $p_{perm} = 0.001$	Observed structure substantially exceeds randomized-label null
Task 2 permutation	F1 = 0.9981; null = 0.501 ± 0.024 ; $p_{perm} = 0.001$	Observed structure substantially exceeds randomized-label null
Task 1 feature removal	F1: 0.9990 $\rightarrow$ 0.9912	Predictive information remains after CRP removal
Task 2 feature removal	F1: 0.9981 $\rightarrow$ 0.6897	Strong dependence on creatinine

Taken together, the audit detected no conventional identifier-, exact-profile, cross-view-, or direct target-variable leakage under the performed checks. This result must be distinguished from generation-induced predictor–target dependence. The dataset creator confirmed that the targets and biomarker differences were generated programmatically, while the complete generation rules are not formally documented. Accordingly, the audit supports the internal integrity of the evaluated experimental pipeline under the specified checks but does not validate the synthetic target-construction process or establish clinical ground truth.

➤ Task 1: Hantavirus Detection

Table 8 presents the independent test results for Task 1. Among the clinical models, GA-LR achieved the highest observed F1-score of 0.8691. Laboratory models produced

substantially higher benchmark performance, with GA-XGBoost achieving F1 = 0.9981. The difference between the two views indicates that the synthetic laboratory variables contained substantially stronger information for discriminating the programmatically assigned Task 1 labels than the clinical variables.

Fused GA-LR and GA-RF each achieved F1 = 0.9990. This represents a numerical increase of only 0.0009 over the strongest laboratory-only model and corresponds to one additional correctly classified test record. Because no paired statistical significance analysis or confidence-interval comparison was performed, this small numerical difference is not interpreted as evidence that fusion is statistically superior to the laboratory-only model.

Table 8 Independent Test Performance for Task 1 Synthetic Hantavirus Detection Classification

Data view	Classifier	Accuracy	Precision	Recall	F1-score
Clinical	GA-DT	0.9225	0.9432	0.7432	0.8313
Clinical	GA-KNN	0.9275	0.9868	0.7276	0.8376
Clinical	GA-LR	0.9405	1.0000	0.7685	0.8691
Clinical	GA-SVM	0.9310	0.9455	0.7763	0.8526
Clinical	GA-RF	0.9305	0.9454	0.7743	0.8513
Clinical	GA-XGBoost	0.9365	1.0000	0.7529	0.8590
Laboratory	GA-DT	0.9950	0.9903	0.9903	0.9903
Laboratory	GA-KNN	0.9980	1.0000	0.9922	0.9961
Laboratory	GA-LR	0.9980	1.0000	0.9922	0.9961
Laboratory	GA-SVM	0.9985	1.0000	0.9942	0.9971
Laboratory	GA-RF	0.9985	0.9981	0.9961	0.9971
Laboratory	GA-XGBoost	0.9990	1.0000	0.9961	0.9981
Fused	GA-DT	0.9970	0.9942	0.9942	0.9942

Fused	GA-KNN	0.9980	1.0000	0.9922	0.9961
Fused	GA-LR	0.9995	1.0000	0.9981	0.9990
Fused	GA-SVM	0.9985	1.0000	0.9942	0.9971
Fused	GA-RF	0.9995	1.0000	0.9981	0.9990
Fused	GA-XGBoost	0.9990	1.0000	0.9961	0.9981

Laboratory models frequently retained CRP, AST, ALT, BUN, platelet count, and creatinine, suggesting that Task 1 predictive structure was distributed across several synthetic laboratory variables rather than being attributable to CRP alone. This interpretation is also supported by the comparatively small reduction in F1 after CRP removal. GA-RF used clinical/laboratory fusion weights of 0.40/0.60 and was retained for TreeSHAP analysis because it tied GA-LR for the highest observed fused F1-score.

Fig. 2 shows the independent test confusion matrix for the selected fused GA-RF model, comprising 1,486 true negatives, 513 true positives, one false negative, and no false positives. The model therefore misclassified one of the 2,000 held-out synthetic records. These counts describe classification of synthetic benchmark labels and should not be interpreted as clinical diagnostic errors.

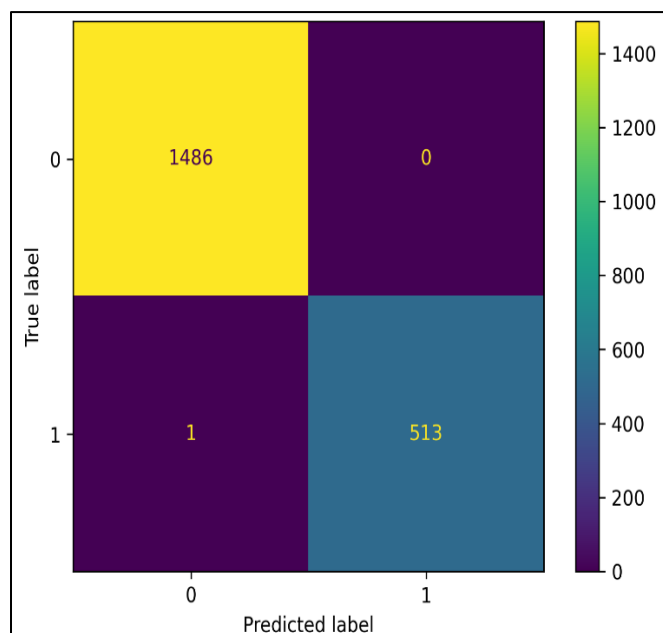


Fig 2 Confusion Matrix of Fused GA-RF on the Independent Task 1 Synthetic Benchmark Test Partition

The secondary five-fold analysis yielded accuracy 0.9976 ± 0.0007 , $F1 = 0.9953 \pm 0.0013$, and $ROC-AUC = 0.9998 \pm 0.0003$, with laboratory fusion weights ranging from 0.61 to 0.64. These internal-validation results consistently with a strong laboratory contribution within the synthetic benchmark. They should not be interpreted as external validation or as evidence of robustness across independently collected clinical populations or multiple random seeds.

➤ Task 2: Disease-Severity Classification

Table 9 presents the independent test results for Task 2. The clinical models provided comparatively limited discrimination, with GA-XGBoost producing the highest observed clinical F1-score of 0.6849. In contrast, all laboratory models selected only Creatinine_Scaled and achieved $F1 = 0.9981$. The resulting difference between the clinical and laboratory views demonstrates that Task 2 classification within this synthetic benchmark was overwhelmingly dependent on creatinine.

Decision-level fusion did not improve upon the strongest laboratory-only F1-score. For most fused models, the optimized weighting therefore provided little or no contribution from the clinical view once the laboratory probability based on creatinine was available. This finding should be interpreted as a property of the synthetic benchmark rather than evidence that clinical information is unimportant for severity assessment in real-world HFRS.

Table 9 Independent test performance for Task 2 synthetic disease-severity classification

Data view	Classifier	Accuracy	Precision	Recall	F1-score
Clinical	GA-DT	0.5437	0.5552	0.6030	0.5781
Clinical	GA-KNN	0.5068	0.5265	0.4831	0.5039
Clinical	GA-LR	0.5087	0.5164	0.8240	0.6349
Clinical	GA-SVM	0.5068	0.5152	0.8240	0.6340
Clinical	GA-RF	0.5243	0.5225	0.9551	0.6755
Clinical	GA-XGBoost	0.5301	0.5250	0.9850	0.6849
Laboratory	GA-DT	0.9981	0.9963	1.0000	0.9981
Laboratory	GA-KNN	0.9981	1.0000	0.9963	0.9981
Laboratory	GA-LR	0.9981	1.0000	0.9963	0.9981
Laboratory	GA-SVM	0.9981	0.9963	1.0000	0.9981
Laboratory	GA-RF	0.9981	0.9963	1.0000	0.9981
Laboratory	GA-XGBoost	0.9981	0.9963	1.0000	0.9981

Fused	GA-DT	0.9981	0.9963	1.0000	0.9981
Fused	GA-KNN	0.9961	0.9963	0.9963	0.9963
Fused	GA-LR	0.9981	0.9963	1.0000	0.9981
Fused	GA-SVM	0.9981	0.9963	1.0000	0.9981
Fused	GA-RF	0.9981	0.9963	1.0000	0.9981
Fused	GA-XGBoost	0.9981	0.9963	1.0000	0.9981

GA-DT, GA-SVM, GA-RF, and GA-XGBoost assigned approximately full laboratory weight, consistent with the limited incremental contribution of the clinical view. Fused GA-RF was retained for TreeSHAP analysis as a representative high-performing tree-based fused model.

Fig. 3 shows 247 true negatives, 267 true positives, one false positive, and no false negatives, corresponding to one misclassified record among 515 held-out synthetic records. As in Task 1, these counts refer to agreement with synthetically assigned benchmark labels rather than clinically adjudicated outcomes.

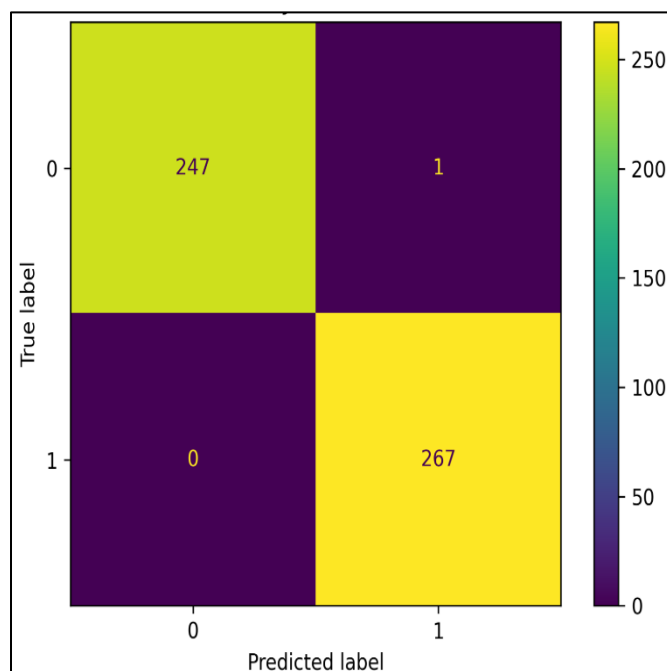


Fig 3 Confusion Matrix of Fused GA-RF on the Independent Task 2 Synthetic Benchmark Test Partition

The secondary five-fold analysis yielded accuracy 0.9988 ± 0.0017 , $F1 = 0.9989 \pm 0.0017$, and $ROC-AUC = 1.0000 \pm 0.0000$, with approximately full laboratory weight. These values demonstrate consistently high discrimination across the evaluated folds of the synthetic benchmark, but they do not constitute external validation or demonstrate robustness across alternative random seeds. The near-perfect performance, together with the single-feature and feature-removal results, indicates unusually strong creatinine-severity dependence within the generated data. This observation is consistent with the creator's confirmation that biomarker differences between synthetic severity classes were programmatically simulated; the exact role of creatinine in the target-generation process cannot be established because the complete generation rules are unavailable.

➤ Overall Comparative Analysis

The highest observed clinical, laboratory, and fused F1-scores were 0.8691, 0.9981, and 0.9990, respectively, for Task 1 and 0.6849, 0.9981, and 0.9981 for Task 2. Laboratory variables therefore provided substantially stronger discrimination of the synthetic target labels than the clinical variables in both tasks. Fusion produced a very small numerical increase over the strongest laboratory model for Task 1 and no improvement for Task 2. Because these differences were not subjected to paired significance testing or confidence-interval comparison, they are reported descriptively rather than as evidence of model superiority.

The two tasks exhibited different dependence patterns. Task 1 retained some complementary information across the clinical and laboratory views and continued to perform strongly after CRP removal, suggesting that its predictive structure was distributed across multiple variables. Task 2, by contrast, was dominated by the laboratory view and specifically by Creatinine_Scaled, as demonstrated by single-feature performance, feature selection, fusion weighting, and the large decline following creatinine removal.

Taken together, the results demonstrate that the evaluated pipelines can recover strong predictive structure embedded in the synthetic benchmark. The audit found no conventional leakage under the specific checks performed, but the creator-confirmed synthetic generation process provides an important explanation for the unusually strong biomarker-target relationships. Accordingly, the near-perfect laboratory and fused results should not be interpreted as evidence of real-world diagnostic accuracy, prognostic performance, clinical utility, or generalizability. Evaluation on independently collected data with clinically established outcomes would be required before drawing such conclusions.

VIII. ABLATION STUDY

Ablation analysis examined the relative contribution of the clinical and laboratory data views and the sensitivity of the predictive pipeline to dominant features. GA-RF was evaluated across clinical, laboratory, and fused settings to maintain the same classifier configuration, while GA-based feature selection was performed independently within each predictor space. All ablation findings are interpreted as properties of the synthetic benchmark rather than as evidence of clinical variable importance.

➤ Data-View Ablation

Table 10 shows that Task 1 F1 increased from 0.8513 for the clinical-only GA-RF model to 0.9971 for the laboratory-only model and 0.9990 after decision-level fusion. Thus, most of the discriminative information used by GA-RF for Task 1 was contained in the synthetic laboratory view.

The laboratory-to-fusion accuracy increase from 0.9985 to 0.9995 corresponds to two additional correctly classified records among the 2,000 held-out test records. The corresponding F1 increase was 0.0019. Therefore, within the GA-RF ablation, fusion produced only a small numerical improvement over the laboratory-only model. Because this difference was not subjected to paired statistical testing or confidence-interval analysis, it is not interpreted as evidence of statistically significant superiority.

For Task 2, the clinical-only GA-RF model achieved F1 = 0.6755, whereas the laboratory-only and fused models both achieved F1 = 0.9981. Fusion therefore provided no observed F1 improvement over the laboratory-only model. Within this synthetic benchmark, the clinical view contributed little additional predictive information once the selected laboratory representation was available.

Table 10 Data-View Ablation Using GA-RF on the Independent Synthetic Benchmark Test Partitions

Task	Data view	Accuracy	Precision	Recall	F1-score
Task 1	Clinical	0.9305	0.9454	0.7743	0.8513
Task 1	Laboratory	0.9985	0.9981	0.9961	0.9971
Task 1	Monte Carlo fused	0.9995	1.0000	0.9981	0.9990
Task 2	Clinical	0.5243	0.5225	0.9551	0.6755
Task 2	Laboratory	0.9981	0.9963	1.0000	0.9981
Task 2	Monte Carlo fused	0.9981	0.9963	1.0000	0.9981

Note: The RF configuration was fixed across data views, whereas GA-based feature selection was performed independently within each predictor space. Numerical differences are reported descriptively and were not subjected to paired significance testing.

➤ Feature-Level Sensitivity Analysis

Task 1 laboratory models frequently retained CRP, creatinine, AST, ALT, BUN, and platelet count, indicating that discriminative information was distributed across several synthetic laboratory variables. Removing CRP_Scaled reduced the best F1 from 0.9990 to 0.9912. Although this decline demonstrates dependence on CRP within the benchmark, the remaining F1 of 0.9912 indicates that substantial predictive information was retained in other variables.

For Task 2, all six laboratory models selected Creatinine_Scaled, and removal of this variable reduced F1 from 0.9981 to 0.6897. This substantially larger decline demonstrates that the predictive structure for the synthetic severity target was highly concentrated in creatinine. Table 11 summarizes the dominant predictors and feature-removal sensitivity observed for both tasks.

Table 11 Dominant Predictors and Feature-Removal Sensitivity within the Synthetic Benchmark

Task	View/pipeline	Dominant predictors	Sensitivity
Task 1	Clinical	Exposure indicators, headache, cough, fatigue, dyspnea	Lower benchmark performance than laboratory view
Task 1	Laboratory	CRP, creatinine, AST, ALT, BUN, platelet count	CRP showed strong individual discrimination
Task 1	Best pipeline	CRP_Scaled	F1: 0.9990 → 0.9912 after removal
Task 2	Clinical	Exposure and virus-type indicators	Lower benchmark performance than laboratory view
Task 2	Laboratory/fused	Creatinine_Scaled	F1: 0.9981 → 0.6897 after removal

Overall, Task 1 retained strong discrimination after CRP removal, whereas Task 2 showed pronounced dependence on creatinine. These findings are consistent with strong predictor–target relationships within the synthetic benchmark and with the dataset creator's clarification that biomarker differences were programmatically simulated. Because the complete generation algorithm and full predictor–target dependency structure are not formally documented, the precise mechanisms producing these dependencies cannot be reconstructed. The ablation findings therefore characterize the synthetic benchmark and should not be interpreted as evidence of real-world biomarker importance or clinical validity.

IX. EXPLAINABILITY ANALYSIS

SHAP was used to characterize the behavior of the final GA-RF models [13]. Mean absolute SHAP values summarized global feature attribution, whereas beeswarm plots showed both attribution magnitude and feature-value directionality.

Because the clinical and laboratory models were trained independently, fusion-weighted summaries are descriptive combinations of within-view attributions rather than SHAP values computed for a single jointly trained fused model. They are therefore used only to summarize the relative attribution patterns of the two independently modelled views under the selected fusion weights. They should not be interpreted as a formal SHAP decomposition of the fused prediction function.

Furthermore, because both predictors and targets were generated synthetically, all attribution patterns describe relationships learned from the benchmark's generated structure. They do not establish causal, biological, or clinically validated effects.

➤ *Task 1: Hantavirus Detection*

The final GA-RF model used clinical/laboratory fusion weights of 0.40/0.60 (Fig. 4). Within the selected fusion procedure, the laboratory probability therefore received the larger weight while the clinical probability retained a smaller contribution.

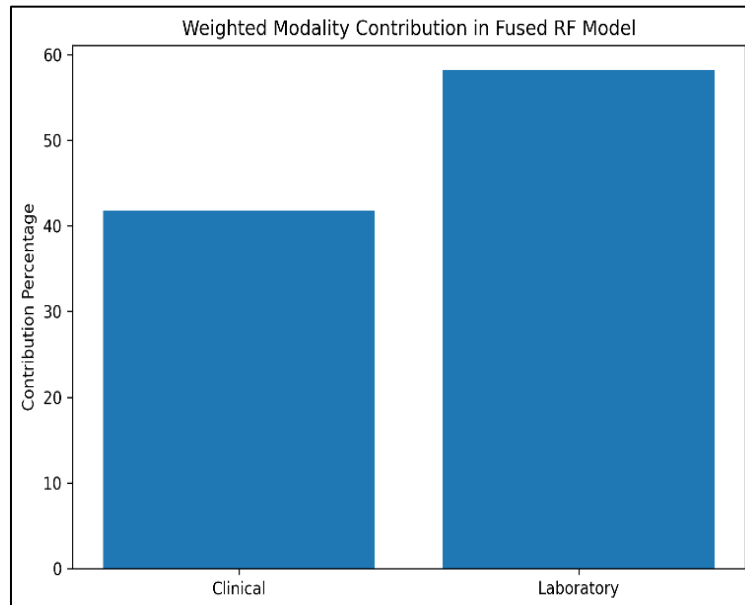


Fig 4 Clinical and laboratory fusion weights for the Task 1 GA-RF model

The descriptive fusion-weighted attribution summary in Fig. 5 was dominated by CRP, creatinine, and AST, with smaller contributions from clinical variables. This pattern is consistent with the substantially stronger laboratory-only benchmark performance reported in Table 8.

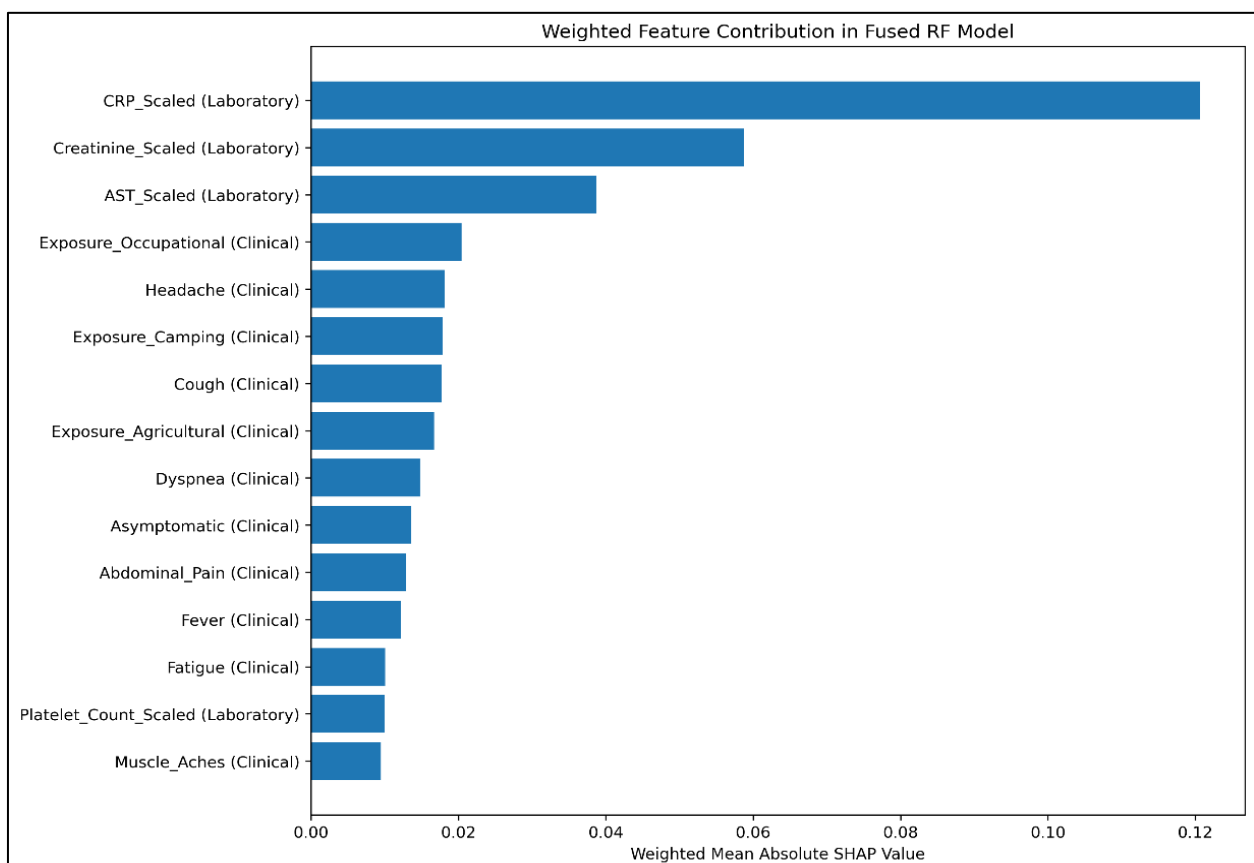


Fig 5 Descriptive Fusion-Weighted Global Attribution Summary for Task 1

The laboratory SHAP analysis (Fig. 6) showed that, within the fitted synthetic-benchmark model, higher values of CRP, creatinine, and AST generally shifted model output toward the positive synthetic label. These directional patterns characterize model behaviour within the generated dataset and should not be interpreted as causal or clinically validated biomarker effects.

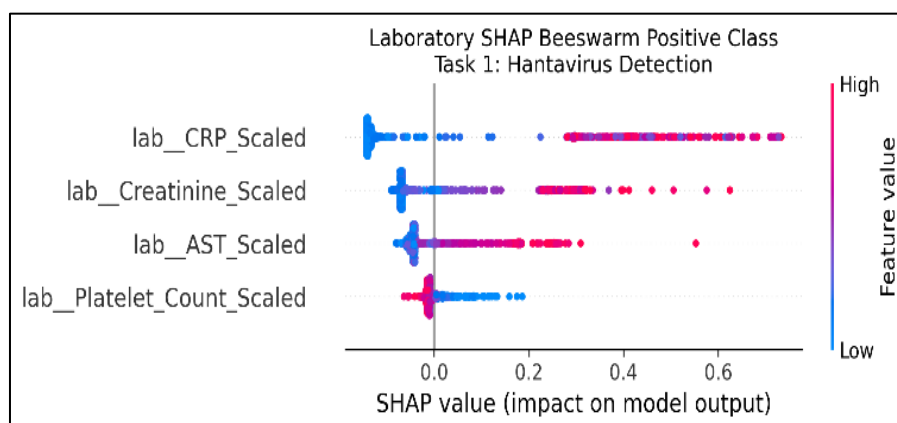


Fig 6 Laboratory SHAP Beeswarm Plot for the Task 1 GA-RF Model

For the standalone clinical model, occupational, camping, and agricultural exposures showed the strongest attributions, followed by headache, cough, dyspnea, fatigue, and fever (Fig. 7). These variables contributed to the predictions of the independently fitted clinical model, but their contribution to the descriptive fusion-weighted summary was smaller because the clinical view received the lower fusion weight.

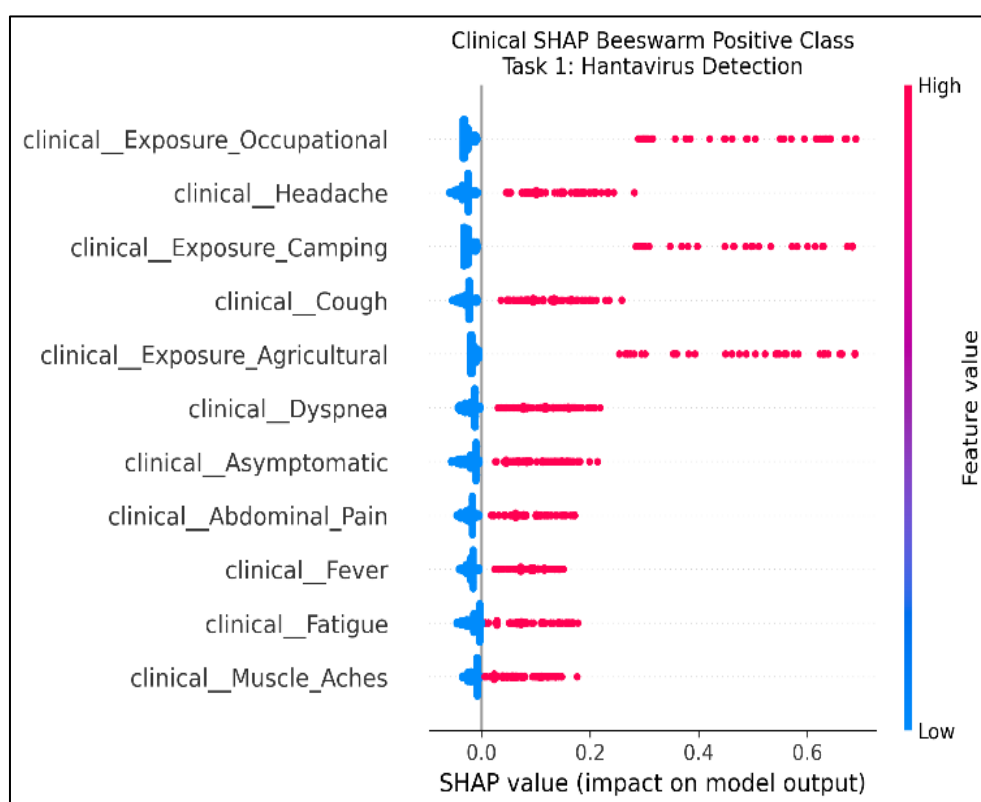


Fig 7 Clinical SHAP Beeswarm Plot for the Standalone Task 1 GA-RF Model

Together, Figs. 4–7 show that Task 1 predictions were predominantly associated with the laboratory view within the synthetic benchmark, while the clinical model provided a smaller complementary contribution under the selected fusion weights.

➤ Task 2: Disease-Severity Classification

For Task 2, the selected fusion solution assigned approximately zero weight to the clinical probability and approximately full weight to the laboratory probability (Fig. 8). The fused prediction was therefore effectively determined by the laboratory model. This weighting pattern is consistent with the absence of an observed performance improvement from fusion over the laboratory-only model.

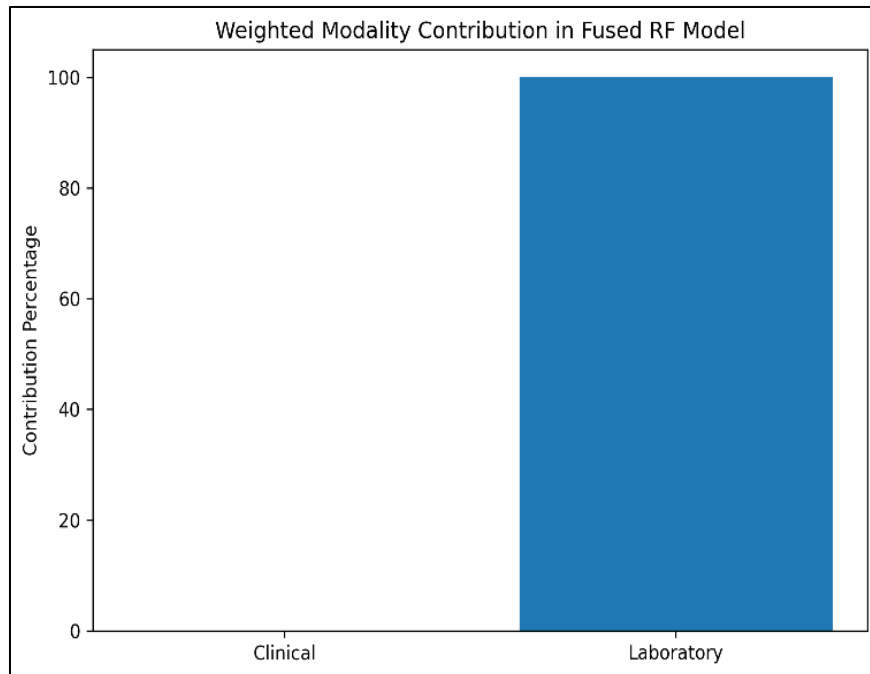


Fig 8 Clinical and Laboratory Fusion Weights for the Task 2 GA-RF Model

Accordingly, Creatinine_Scaled dominated the descriptive global attribution summary in Fig. 9. This finding is consistent with the single-feature and feature-removal analyses reported in Sections VII-A and VIII-B.

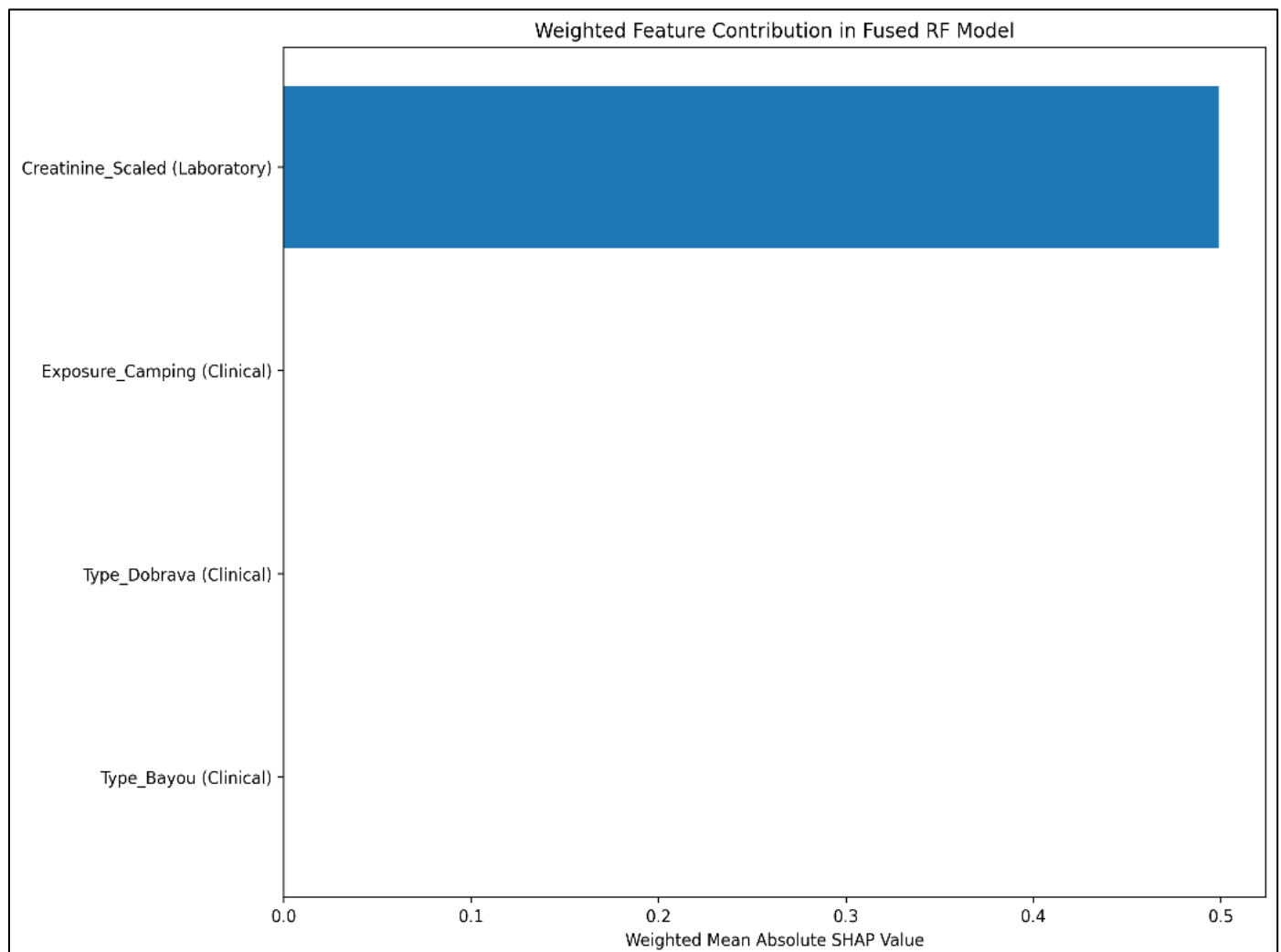


Fig 9 Descriptive Fusion-Weighted Global Attribution Summary for Task 2

The laboratory SHAP plot (Fig. 10) showed that higher creatinine values generally shifted the fitted model toward the synthetic severe label, whereas lower values shifted it toward the synthetic moderate label. This pattern is consistent with the strong creatinine dependence identified by the audit and ablation analyses. Because the severity labels and biomarker distributions were generated programmatically, this directional pattern must not be interpreted as evidence that creatinine alone determines real-world HFRS severity.

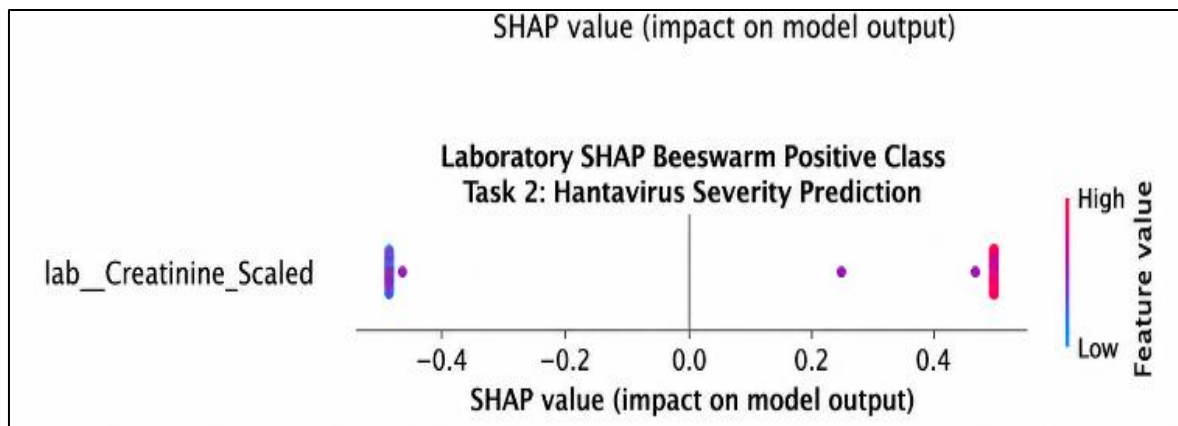


Fig 10 Laboratory SHAP Beeswarm Plot for the Task 2 GA-RF Model

Although the clinical model received negligible weight in the selected fused solution, its standalone SHAP analysis identified Exposure_Camping, Type_Dobrava, and Type_Bayou among its strongest attributed variables (Fig. 11). These findings describe only the behaviour of the independently fitted synthetic clinical model and should not be interpreted as meaningful contributions to the final fused classification or as clinically established severity factors.

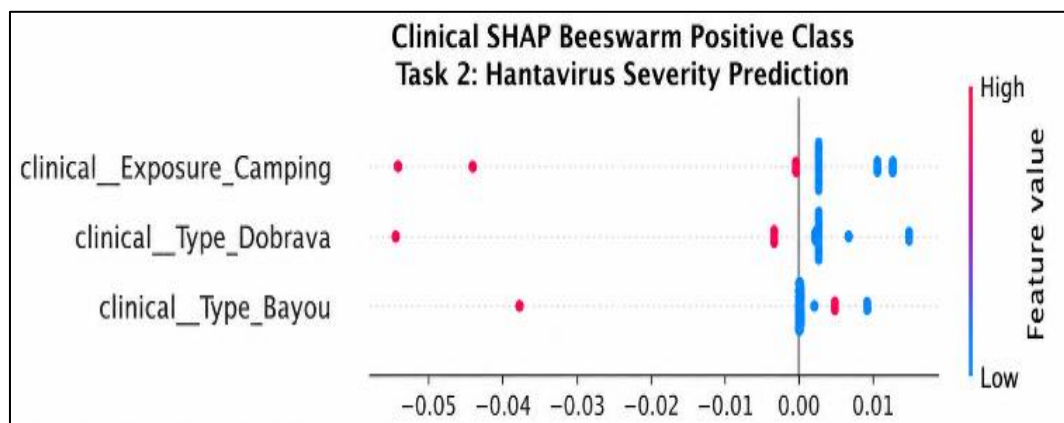


Fig 11 Clinical SHAP Beeswarm Plot for the Standalone Task 2 GA-RF Model

➤ Overall Explainability Findings

Task 1 attribution was concentrated mainly in CRP, creatinine, and AST, with smaller exposure- and symptom-related contributions from the clinical model, whereas Task 2 attribution was overwhelmingly concentrated in creatinine. The SHAP analyses therefore provide a model-level explanation consistent with the comparative performance, single-feature, ablation, and dominant-feature-removal findings.

These attribution patterns describe fitted-model behaviour rather than biological causality. In particular, the strong Task 2 dependence on creatinine must be interpreted in the context of the confirmed synthetic data-generation process. Although the creator confirmed that biomarker differences between severity classes were programmatically simulated, the complete severity-generation algorithm and full dependency structure are not formally documented.

Accordingly, the explainability analysis characterizes how the evaluated models use variables within the synthetic benchmark. It does not establish clinical biomarker significance, causal relationships, or operational decision-making value.

X. DISCUSSION

This study evaluated an explainable multi-view machine-learning framework for hantavirus detection and disease-severity classification using entirely synthetic structured clinical and laboratory records. The framework combined GA-based feature selection, six classifiers, training-only Monte Carlo decision-level fusion, SHAP attribution, and a structured data-integrity and target-dependence audit. The principal finding is not simply the near-perfect performance of several laboratory and fused models, but the strong concentration of predictive information in synthetic laboratory variables and the different dependence structures observed across the two tasks.

For Task 1, the highest observed clinical and laboratory F1-scores were 0.8691 and 0.9981, respectively. CRP, creatinine, AST, ALT, BUN, and platelet count were frequently retained by the laboratory models, while SHAP attribution emphasized CRP, creatinine, and AST. Similar classes of inflammatory, renal, hepatic, and hematological variables have been associated with HFRS severity and outcomes in clinical studies [4–6, 18]. However, this similarity should not be interpreted as validation of the present synthetic benchmark. The dataset creator confirmed that biomarker differences were programmatically simulated to approximate plausible physiological patterns. Agreement between generated patterns and published clinical literature may therefore reflect aspects of the synthetic design rather than independently reproduced clinical evidence.

Decision-level fusion increased the highest Task 1 F1-score from 0.9981 for the strongest laboratory model to 0.9990 for the best fused models, a difference of 0.0009 corresponding to one additional correctly classified record when comparing the overall best laboratory and fused models. Within the GA-RF data-view ablation specifically, laboratory-to-fusion accuracy increased from 0.9985 to 0.9995, corresponding to two additional correctly classified records. These are small numerical differences, and no paired statistical test or confidence-interval comparison was performed. The results therefore do not establish statistically significant superiority of fusion over laboratory-only modelling.

Task 2 exhibited a substantially different pattern. The highest clinical F1-score was 0.6849, whereas all laboratory models achieved $F1 = 0.9981$ using *Creatinine_Scaled* alone. Removing creatinine reduced F1 from 0.9981 to 0.6897, compared with the smaller Task 1 reduction from 0.9990 to 0.9912 after CRP removal. The convergence of feature selection, single-feature performance, fusion weighting, SHAP attribution, and feature-removal sensitivity therefore indicates that Task 2 benchmark performance was overwhelmingly concentrated in creatinine.

Creatinine is clinically relevant to renal dysfunction in HFRS [4, 6]; however, its near-complete discrimination in the present dataset must be interpreted primarily as a property of the synthetic generation process. The dataset creator confirmed that biomarker differences between synthetic severity classes were programmatically simulated and that severity was not defined using a disclosed rigid laboratory cutoff. Because the complete generation algorithm and full predictor–target dependency structure are unavailable, the analysis cannot determine whether creatinine participated directly in label assignment, indirectly through simulated class distributions, or through another generation rule. Consequently, the Task 2 result demonstrates recovery of a strong synthetic creatinine–severity relationship and does not establish that creatinine alone is sufficient for clinical HFRS severity assessment.

The data-integrity and target-dependence audit provides an important context for these results. No duplicate identifiers, exact cross-partition predictor profiles, conflicting labels among repeated profiles, cross-view mismatches, or prohibited target variables were detected under the performed checks. These findings reduce concern about the specific forms of

conventional information leakage examined by the audit. They do not, however, exclude generation-induced predictor–target dependence.

Permutation testing produced $p_{\text{perm}} = 0.001$ for both tasks, demonstrating that the observed predictive structure substantially exceeded the corresponding randomized-label empirical null distributions. This result should not be interpreted as validation of target construction. In a synthetic dataset with deliberately generated predictor–target relationships, strong performance relative to permuted labels is expected. The permutation analysis therefore confirms recoverable non-random structure within the benchmark rather than independent clinical validity.

The feature-selection results should likewise be interpreted within the scope of the implemented experiments. GA produced task- and classifier-specific feature subsets, but its performance was not compared with all-feature modelling, filter-based selection, embedded selection, or other feature-selection strategies. The present findings therefore demonstrate the operation of GA-based selection within the evaluated framework rather than superiority of GA over alternative approaches. Furthermore, because the experiments used a single fixed random seed, robustness of GA-selected subsets to stochastic variation was not established.

The decision-level fusion results also warrant a limited interpretation. Under the constraint $w_l = 1 - w_c$, only one independent scalar weight was searched. Monte Carlo sampling is therefore best regarded as stochastic weight selection over a bounded one-dimensional interval rather than as a general-purpose optimization method. The 10,000 samples provided dense coverage of this interval but were not demonstrated to be necessary or optimal. Comparisons with deterministic grid search, equal weighting, and repeated random seeds were not performed. In addition, a separate probability-calibration analysis was not conducted; consequently, the fused probabilities should be interpreted as inputs to comparative classification rather than as calibrated estimates of clinical risk.

SHAP analysis further characterized the different dependence structures of the two tasks. Task 1 showed laboratory-dominant but comparatively distributed attribution, whereas Task 2 attribution was concentrated almost entirely in creatinine. The fusion-weighted attribution summaries are descriptive combinations of independently calculated within-view SHAP values rather than formal SHAP decompositions of a jointly trained fused model. More generally, SHAP describes the behaviour of the fitted models and does not establish biological causality or clinical biomarker importance.

The secondary five-fold analyses produced consistently high discrimination within the evaluated synthetic folds. However, these analyses constitute internal validation rather than external validation, and the use of a single random seed does not establish robustness to alternative partitions or stochastic variation. Similarly, small numerical differences among high-performing models were not evaluated with paired statistical tests or confidence intervals and should therefore not

be interpreted as evidence of statistically significant model superiority.

Overall, the results show that the evaluated pipelines recover strong predictive relationships embedded in the synthetic benchmark. Task 1 exhibited laboratory-dominant but more distributed predictive structure and only a small numerical fusion gain, whereas Task 2 showed near-total dependence on a single synthetic laboratory variable. The combination of comparative modelling, single-feature analysis, ablation, SHAP attribution, permutation testing, and data-integrity auditing makes these dependencies explicit rather than treating near-perfect performance as evidence of clinical success.

Accordingly, the primary interpretation of this study is computational and diagnostic with respect to the dataset itself: it demonstrates how apparently excellent machine-learning performance on a synthetic health-data benchmark can coexist with strong generation-related predictor–target dependence. The results do not establish real-world hantavirus detection accuracy, severity prediction, clinical utility, or generalizability. Such conclusions would require evaluation using independently collected clinical cohorts with documented outcome definitions and clinically established reference standards.

XI. LIMITATIONS

Several limitations constrain the interpretation of this study. First, the dataset is entirely synthetic and contains no observations collected from patients, hospitals, clinical laboratories, or published clinical cohorts. Both classification targets were assigned synthetically, and biomarker differences were introduced programmatically. The dataset creator clarified the general synthetic nature and intended use of the data, but no formal technical document describes the complete generation algorithm, target-assignment procedure, or full predictor–target dependency structure. Furthermore, the dataset has not been independently reviewed or clinically validated by healthcare professionals. Consequently, its class distributions, feature relationships, and model performance cannot be interpreted as estimates of disease prevalence, diagnostic validity, severity validity, or generalizability to real healthcare populations.

Second, Task 2 exhibited extremely strong dependence on `Creatinine_Scaled`. The audit detected no conventional identifier-based, exact-profile, cross-view, or direct target-variable leakage under the checks performed; however, absence of these forms of contamination does not exclude predictor–target dependencies introduced by the synthetic generation process. The creator confirmed that biomarker differences between synthetic severity classes were programmatically simulated, while the exact generation rules remain unavailable. The near-deterministic Task 2 performance must therefore be interpreted as a property of the synthetic benchmark rather than as evidence that creatinine alone is sufficient for real-world severity classification.

Third, the primary evaluation used a single stratified 80:20 development–test split with a fixed random seed. The secondary five-fold analysis provided additional internal estimates across the evaluated folds but did not constitute external validation and did not assess robustness across alternative random seeds or independently generated partitions. Repeated nested cross-validation, confidence intervals, and formal paired statistical comparisons were not performed. Consequently, small numerical differences among closely performing models should not be interpreted as statistically established superiority.

Fourth, GA-based feature selection was not compared with all-feature modelling, filter-based selection, embedded methods, or other wrapper strategies. Similarly, Monte Carlo fusion was not benchmarked against equal weighting, deterministic grid search, or repeated multi-seed sampling. The present experiments therefore characterize the implemented procedures but do not establish the superiority of GA-based selection or stochastic fusion over alternative strategies.

Fifth, only six classifiers and the structured variables available in the synthetic dataset were evaluated. Potentially informative real-world modalities such as clinically verified serology, viral load, imaging, longitudinal laboratory measurements, treatment information, environmental measurements, and genomic sequencing were unavailable. In particular, `Hantavirus_Type` is a synthetic categorical field and does not represent genomic sequencing information.

Sixth, the classifiers' predicted probabilities were used directly for decision-level fusion without a separate calibration analysis. Therefore, the probability outputs and their weighted combinations should not be interpreted as calibrated estimates of clinical risk. Decision-curve analysis, uncertainty quantification, subgroup robustness, fairness, usability, cost-effectiveness, and prospective clinical utility were also not evaluated.

Finally, SHAP provides model attribution rather than causal explanation. The reported attribution patterns describe how the fitted models use variables within the synthetic benchmark and cannot establish biological causality or clinical biomarker importance. No independently collected hospital cohort, prospective evaluation, or operational healthcare deployment was available. The present work should therefore be regarded as a synthetic-data computational benchmark and data-integrity assessment rather than a clinically validated prediction system.

XII. FUTURE WORK

An immediate priority for future work is to obtain or formally document the complete synthetic data-generation procedure, including the mechanisms used for target assignment and the dependencies introduced between predictors and labels. This is particularly important for clarifying the strong relationship between creatinine and the synthetic Task 2 severity target. Such documentation would allow future benchmark studies to distinguish intended simulation rules from unintended target dependence more precisely.

More importantly, the modelling framework should be evaluated using independently collected clinical data with documented provenance, clinically established reference outcomes, and clearly defined severity criteria. Multicentre and temporally separated cohorts would be valuable for assessing whether any patterns observed in the synthetic benchmark are reproducible across real healthcare settings and populations.

Methodological evaluation should include repeated nested cross-validation, multiple random seeds, confidence intervals, paired statistical testing, probability calibration, precision–recall analysis, decision-curve analysis, and uncertainty estimation. Subgroup performance should also be examined where clinically and ethically appropriate. GA-based feature selection should be compared with all-feature modelling and alternative filter, wrapper, and embedded approaches. Likewise, Monte Carlo fusion should be benchmarked against equal weighting, deterministic grid search, and other transparent weight-selection strategies.

Future clinical datasets should, where available and appropriate, include clinically verified serology, viral load, longitudinal laboratory measurements, imaging, treatment history, environmental exposure variables, and validated genomic information. These additional modalities could enable a more comprehensive investigation of infection-status and disease-severity prediction than is possible with the present synthetic structured dataset.

Explainability research should also extend beyond global feature importance. Future studies could evaluate attribution stability across resamples, record-level explanations, counterfactual analyses, and uncertainty in model explanations. Where models are evaluated on real clinical data, healthcare professionals should additionally assess whether the resulting explanations are understandable, clinically plausible, and useful for decision support.

Only after satisfactory evaluation on independent real-world clinical data should implementation-oriented studies be considered. Such work could examine integration with interoperable hospital and laboratory information systems and, where appropriate, connected healthcare infrastructure. Deployment-oriented evaluation would need to address secure data acquisition and transmission, authentication, privacy, cybersecurity, latency, reliability, workflow integration, usability, fairness, clinical utility, and cost-effectiveness before any operational healthcare use.

XIII. CONCLUSION

This study presented an explainable multi-view machine-learning framework for hantavirus detection and disease-severity classification on an entirely synthetic structured-data benchmark. The framework combined GA-based feature selection, six classifiers, training-only Monte Carlo decision-level fusion, SHAP attribution, and a structured data-integrity and target-dependence audit. Clinical and laboratory views were evaluated independently and then combined at the probability level to characterize their individual and complementary predictive contributions.

For Task 1, the highest fused F1-score was 0.9990. Laboratory variables provided substantially stronger discrimination of the synthetic target than the clinical variables, with CRP, creatinine, and AST among the most influential laboratory predictors. Fusion produced only a very small numerical improvement over the strongest laboratory-only model, corresponding to one additional correctly classified record in the overall best-model comparison. This difference was not statistically tested and therefore does not establish superiority of fusion.

For Task 2, the highest fused F1-score was 0.9981, but performance was overwhelmingly dependent on `Creatinine_Scaled`. The laboratory and fused models achieved the same F1-score, while removal of creatinine reduced F1 from 0.9981 to 0.6897. The agreement among single-feature analysis, feature selection, fusion weighting, SHAP attribution, and feature-removal sensitivity demonstrates pronounced creatinine dependence within the synthetic benchmark.

The data-integrity audit detected no conventional identifier-based, exact-profile, cross-view, or direct target-variable leakage under the performed checks. However, these findings do not exclude dependencies created during synthetic data generation. The dataset creator confirmed that the targets were synthetically assigned and that biomarker differences were programmatically simulated, while the complete generation algorithm and full predictor–target dependency structure are not formally documented. Accordingly, the near-perfect predictive results are best interpreted as recovery of strong programmed structure within the synthetic benchmark rather than as evidence of clinical diagnostic or prognostic performance.

The main contribution of the study is therefore computational and diagnostic with respect to the benchmark itself: it combines comparative modelling, feature selection, decision-level fusion, explainability, single-feature analysis, permutation testing, feature-removal analysis, and data-integrity auditing to make unusually strong predictor–target dependencies explicit. This distinction is important when evaluating apparently exceptional machine-learning performance on synthetic health datasets.

The present results do not establish real-world hantavirus detection accuracy, severity prediction, clinical utility, or generalizability. Evaluation using independently collected clinical cohorts with documented provenance, clinically established outcomes, and appropriate methodological validation is required before any clinical interpretation or operational use of the framework can be considered.

ACKNOWLEDGMENT

The authors acknowledge the creator of the public Hantavirus Detection Dataset [30] for providing additional clarification regarding the synthetic origin, intended use, target-generation process, and programmatically simulated biomarker differences in the dataset. The authors also acknowledge the use of publicly available scientific literature and open-source computational libraries that supported the implementation and evaluation of the proposed framework.

DATA AVAILABILITY STATEMENT

The dataset used in this study is the publicly available Hantavirus Detection Dataset cited in [30]. The dataset contains 10,000 entirely synthetic structured records and does not contain data collected from real patients, hospitals, clinical laboratories, or published clinical cohorts. Additional clarification regarding the synthetic provenance and construction of the dataset was obtained through direct correspondence with the dataset creator.

The complete synthetic data-generation algorithm and full predictor–target dependency structure were not available as formal technical documentation at the time of this study.

ETHICS STATEMENT

This study used an entirely synthetic, publicly available dataset and did not involve real human participants, identifiable patient information, biological specimens, or clinical interventions. Therefore, institutional human-subject ethical approval and informed consent were not applicable to the computational experiments reported in this study.

INFORMED CONSENT STATEMENT

Not applicable. The study did not involve human participants or identifiable patient data.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest related to this work.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS

Mohammad Yasir Mudasir performed the conceptualization, methodology development, implementation, data processing, experimentation, analysis, visualization, and preparation of the original manuscript. Zaeem Akbar Khan assisted in the research, provided methodological and academic guidance, contributed to interpretation of the findings, and reviewed and edited the manuscript. Both authors reviewed and approved the final version of the manuscript.

REFERENCES

- [1]. C. B. Jonsson, L. T. M. Figueiredo, and O. Vapalahti, “A global perspective on hantavirus ecology, epidemiology, and disease,” *Clinical Microbiology Reviews*, vol. 23, no. 2, pp. 412–441, 2010.
- [2]. M. Tariq and D.-M. Kim, “Hemorrhagic fever with renal syndrome: Literature review, epidemiology, clinical picture and pathogenesis,” *Infection & Chemotherapy*, vol. 54, no. 1, pp. 1–19, 2022.

- [3]. P. A. Vial, M. Ferrés, C. Vial, J. Klingström, C. Ahlm, R. López, N. Le Corre, and G. J. Mertz, “Hantavirus in humans: A review of clinical aspects and management,” *The Lancet Infectious Diseases*, vol. 23, no. 9, pp. e371–e382, 2023.
- [4]. H. Du, P.-Z. Wang, J. Li, L. Bai, H. Li, H.-T. Yu, W. Jiang, Y. Zhang, J.-N. Wang, and X.-F. Bai, “Clinical characteristics and outcomes in critical patients with hemorrhagic fever with renal syndrome,” *BMC Infectious Diseases*, vol. 14, p. 191, 2014.
- [5]. X. Fan, Z. Liu, S. Fu, J. Sang, H. Deng, F. Li, X. Zhang, N. Li, Q. Han, and Z. Liu, “Platelet distribution width at first day of hospital admission in patients with hemorrhagic fever with renal syndrome caused by Hantaan virus may predict disease severity and critical patients’ survival,” *Disease Markers*, vol. 2018, p. 9701619, 2018.
- [6]. W.-j. Chen, H. Du, H.-f. Hu, J.-q. Lian, H. Jiang, J. Li, Y. Zhang, and P.-z. Wang, “Levels of peripheral blood routine, biochemical and coagulation parameters in patients with hemorrhagic fever with renal syndrome and their relationship with prognosis: An observational cohort study,” *BMC Infectious Diseases*, vol. 24, p. 75, 2024.
- [7]. E. J. Topol, “High-performance medicine: The convergence of human and artificial intelligence,” *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.
- [8]. A. Rajkomar, J. Dean, and I. Kohane, “Machine learning in medicine,” *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.
- [9]. Z. Yang, Q. Hu, Z. Feng, and Y. Sun, “Development and validation of a nomogram for predicting severity in patients with hemorrhagic fever with renal syndrome: A retrospective study,” *Open Medicine*, vol. 16, no. 1, pp. 1582–1591, 2021.
- [10]. L. Huang, J. Wu, J. Luo, and W. Gu, “Predictors of severity in hemorrhagic fever with renal syndrome,” *International Journal of General Medicine*, vol. 18, pp. 2033–2045, 2025.
- [11]. G. S. Collins, K. G. M. Moons, P. Dhiman, R. D. Riley, A. L. Beam, et al., “TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods,” *BMJ*, vol. 385, p. e078378, 2024.
- [12]. K. G. M. Moons, J. A. A. Damen, T. Kaul, L. Hooft, C. Andaur Navarro, et al., “PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods,” *BMJ*, vol. 388, p. e082505, 2025.
- [13]. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, Curran Associates, Inc., 2017.
- [14]. J. Brockmann, M. Kleines, N. Ghaffari Laleh, J. N. Kather, S. Wied, J. Floege, and G. S. Braun, “A simple clinical score to reduce unnecessary testing for Puumala hantavirus,” *PLOS ONE*, vol. 19, no. 6, p. e0304500, 2024.
- [15]. R. Li, J. Sun, Y. Chen, X. Fan, X. Wang, X. Zhang, K. Zhang, Q. Han, and Z. Liu, “Clinical and laboratory features and factors predicting disease severity in pediatric patients with hemorrhagic fever with renal

- syndrome caused by Hantaan virus,” *Journal of Medical Virology*, vol. 95, no. 1, p. e28339, 2023.
- [16]. K. Ma, T. Wu, W. Guo, J. Wang, Q. Ming, and J. Zhu, “Clinical characteristics and a novel prediction nomogram (EASTAR) for patients with hemorrhagic fever with renal syndrome: A multicenter retrospective study,” *Tropical Medicine and Infectious Disease*, vol. 10, no. 2, p. 51, 2025.
- [17]. A. Sehgal, S. Mehta, K. Sahay, E. Martynova, A. Rizvanov, M. Baranwal, and S. Chandy, “Hemorrhagic fever with renal syndrome in asia: History, pathogenesis, diagnosis, treatment, and prevention,” *Viruses*, vol. 15, no. 2, p. 561, 2023.
- [18]. H. Du, H. Hu, P. Wang, X. Wang, Y. Zhang, H. Jiang, J. Li, X. Bai, and J. Lian, “Predictive value of pentraxin-3 on disease severity and mortality risk in patients with hemorrhagic fever with renal syndrome,” *BMC Infectious Diseases*, vol. 21, p. 445, 2021.
- [19]. M. Goeijenbier, R. A. Hartskeerl, J. Reimerink, J. Verner-Carlsson, J. F. Wage-naar, M. G. Goris, B. E. Martina, Å. Lundkvist, M. Koopmans, A. D. Osterhaus, E. C. van Gorp, and C. B. Reusken, “The hanta hunting study: Underdiagnosis of Puumala hantavirus infections in symptomatic non-travelling leptospirosis-suspected patients in the netherlands, in 2010 and april to november 2011,” *Euro-surveillance*, vol. 19, no. 32, p. 20878, 2014.
- [20]. H. Hujakka, V. Koistinen, I. Kuronen, P. Eerikainen, M. Parviainen, A. Lundkvist, A. Vaheri, A. Narvanen, and O. Vapalahti, “Diagnostic rapid tests for acute hantavirus infections: Specific tests for Hantaan, Dobrava and Puumala viruses versus a hantavirus combination test,” *Journal of Virological Methods*, vol. 108, no. 2, pp. 117–122, 2003.
- [21]. J. Schubert, F. Tollmann, and B. Weissbrich, “Evaluation of a pan-reactive hanta-virus enzyme immunoassay and of a hantavirus immunoblot for the diagnosis of nephropathia epidemica,” *Journal of Clinical Virology*, vol. 21, no. 1, pp. 63–74, 2001.
- [22]. Q. Liu, X. Liu, B. Jiang, and W. Yang, “Forecasting incidence of hemorrhagic fever with renal syndrome in china using ARIMA model,” *BMC Infectious Diseases*, vol. 11, p. 218, 2011.
- [23]. G. Ye, M. Alim, P. Guan, D. Huang, B. Zhou, and W. Wu, “Improving the precision of modeling the incidence of hemorrhagic fever with renal syndrome in mainland china with an ensemble machine learning approach,” *PLOS ONE*, vol. 16, no. 3, p. e0248597, 2021.
- [24]. C. X. Lv, S. Y. An, B. J. Qiao, and W. Wu, “Time series analysis of hemorrhagic fever with renal syndrome in mainland china by using an XGBoost forecasting model,” *BMC Infectious Diseases*, vol. 21, p. 839, 2021.
- [25]. K. She, C. Li, C. Qi, T. Liu, Y. Jia, Y. Zhu, L. Liu, Z. Wang, Y. Zhang, and X. Li, “Epidemiological characteristics and regional risk prediction of hemorrhagic fever with renal syndrome in shandong province, china,” *International Journal of Environmental Research and Public Health*, vol. 18, no. 16, p. 8495, 2021.
- [26]. O. Kazasidis and J. Jacob, “Machine learning identifies straightforward early warning rules for human Puumala hantavirus outbreaks,” *Scientific Reports*, vol. 13, p. 3585, 2023.
- [27]. Y. Tian, T. Wang, J.-J. Chen, Q. Xu, G.-L. Wang, B.-G. Jiang, L.-P. Wang, C.-L. Lv, T. Jiang, and L.-Q. Fang, “Distribution dynamics and urbanization-related factors of Hantaan and Seoul virus infections in china between 2001 and 2020: A machine learning modelling analysis,” *Heliyon*, vol. 10, no. 21, p. e39852, 2024.
- [28]. L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [29]. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, Association for Computing Machinery, 2016.
- [30]. T. A. Rahat, “Hantavirus detection dataset.” Kaggle, 2026. Accessed May 30, 2026.
- [31]. J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI, USA: University of Michigan Press, 1975.
- [32]. D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA, USA: Addison-Wesley, 1989.
- [33]. J. Yang and V. Honavar, “Feature subset selection using a genetic algorithm,” *IEEE Intelligent Systems and Their Applications*, vol. 13, no. 2, pp. 44–49, 1998.
- [34]. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. John Wiley & Sons, 3 ed., 2013.
- [35]. C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [36]. T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [37]. J. R. Quinlan, “Induction of decision trees,” *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.
- [38]. J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, “On combining classifiers,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.
- [39]. L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, NJ, USA: John Wiley & Sons, 2004.
- [40]. N. Metropolis and S. Ulam, “The monte carlo method,” *Journal of the American Statistical Association*, vol. 44, no. 247, pp. 335–341, 1949.