Applied Mathematical Sciences, Vol. 20, 2026, no. 4, 177 - 183
HIKARI Ltd, www.m-hikari.com
https://doi.org/10.12988/ams.2026.919334

Machine Learning Ensemble Methods for Prospective Hantavirus Risk Prediction: Integrating Environmental and Epidemiological Data

Massimiliano Ferrara

Department of Law, Economics and Human Sciences
Decisions LAB (DIGIES)
Università degli Studi Mediterranea di Reggio Calabria
Via dell'Università 25, 89124 Reggio Calabria, Italy

Abstract

We develop an ensemble machine learning framework integrating environmental, ecological, and socioeconomic variables to enable prospective hantavirus risk prediction. Trained on 689 laboratory-confirmed cases from three United States jurisdictions, the ensemble of random forest, gradient boosting, support vector machine, and logistic regression classifiers achieves area under the receiver operating characteristic curve of 0.92 on independent test data. Shapley additive explanations identify precipitation variability, land-use patterns, and rodent species richness as dominant predictors, with substantial contributions from socioeconomic determinants.

Mathematics Subject Classification: 62H30, 68T05, 92D30

Keywords: machine learning; ensemble methods; hantavirus; explainable artificial intelligence; zoonotic disease

1. Introduction

Hantaviruses are rodent-borne RNA viruses causing two distinct human syndromes: hemorrhagic fever with renal syndrome in Eurasia and hantavirus cardiopulmonary syndrome in the Americas, with case-fatality rates of 30%-50% [1]. In the United States, Sin Nombre virus, carried predominantly by the deer mouse Peromyscus maniculatus, is the principal etiologic agent and has caused over 800 confirmed cases since 1993 [2]. No specific antiviral therapy or licensed vaccine is available, so primary prevention through reduction of human exposure to infected rodents remains the principal public health strategy.

Current epidemiologic investigation methodologies are fundamentally reactive: cases are identified through clinical presentation, then retrospective environmental investigations attempt to identify probable exposure locations [2]. Incubation periods of one to eight weeks obscure exposure-to-diagnosis timelines, and reactive investigations cannot prevent the index case that triggered them.

Machine learning offers a path beyond these constraints through pattern recognition across high-dimensional data integrating multiple variable classes. Ensemble approaches combining diverse algorithm families outperform single-algorithm approaches in many prediction tasks [3, 4], and explainable artificial intelligence techniques, particularly Shapley additive explanations, render predictions interpretable for public health use [5]. We develop and validate an ensemble framework integrating laboratory-confirmed case data with environmental, ecological, and socioeconomic variables to enable prospective identification of high-risk transmission zones.

2. Methods

2.1. Data sources

We assembled 689 laboratory-confirmed hantavirus cases from California (89 cases, 1993-2020), Colorado (5 cases, 2024), and New Mexico (129 cases, 1993-2023), supplemented from regional surveillance archives. Case definition required serologic evidence (hantavirus-specific IgM or rising IgG titers by ELISA) or RT-PCR confirmation. Environmental variables included 90-day precipitation and temperature from NOAA stations within 10 km, with the coefficient of variation of precipitation as a measure of variability; land-use classifications from the USGS National Land Cover Database; and rodent species richness with deer mouse density indices. Socioeconomic variables from the U.S. Census Bureau American Community Survey at census-tract level included poverty rate, median household income, educational attainment, and housing age. Missing values were imputed using k-nearest neighbors ($k = 5$).

2.2. Ensemble framework

We construct an ensemble combining four classifiers with complementary

inductive biases. The random forest aggregates B = 100 bootstrap trees:

$$\hat{y}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{1}$$

Gradient boosting (XGBoost) minimizes the regularized objective

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

with logistic loss and complexity penalty. The support vector machine solves the dual quadratic problem with a radial basis function kernel and parameters γ = 0.001, C = 1.0. Logistic regression provides an interpretable linear baseline:

$$P(y = 1 \mid x) = \frac{1}{1 + \exp(-(\beta_0 + \beta^T x))} \tag{3}$$

with L2 regularization λ = 0.01. Ensemble predictions aggregate component probabilities through unweighted averaging:

$$P_{ens}(y = 1 \mid x) = \frac{1}{M}\sum_{m=1}^{M} P_m(y = 1 \mid x), \quad M = 4 \tag{4}$$

2.3. Training and explainability

We partition the 689 records into a training set (70%, n = 482) and an independent test set (30%, n = 207), stratifying by jurisdiction. Hyperparameters are tuned via Bayesian optimization across 50 iterations maximizing AUC computed by 10-fold cross-validation [6]. For temporal robustness, models trained on pre-2020 cases (n = 450) are evaluated on cases from 2020-2026 (n = 239). Performance is summarized by area under the receiver operating characteristic curve [7], sensitivity, specificity, positive and negative predictive values, and F1 score, with 1,000-bootstrap confidence intervals. Shapley additive explanations are computed via TreeExplainer for tree-based models and exact Shapley values for the logistic component [5]; feature importance is the mean absolute Shapley value over the test set.

3. Results

3.1. Model performance

The ensemble achieves AUC of 0.92 (95% CI 0.88-0.95) on the independent test set, exceeding all component algorithms (Table 1): gradient boosting 0.91, random forest 0.89, support vector machine 0.84, logistic regression 0.81. Sensitivity is 0.88 and specificity 0.89; positive and negative predictive values are 0.88 and 0.89; F1 is 0.88. Independent temporal validation on 2020-2026 cases yields AUC 0.89, indicating stable performance under realistic temporal separation

between training and operational use. Calibration is satisfactory (Hosmer-Lemeshow $\chi^2 = 7.4$, $p = 0.49$), and subgroup performance is consistent across jurisdictions (California 0.91, Colorado 0.90, New Mexico 0.93). Receiver operating characteristic curves (Figure 1) show ensemble dominance across the operating range.

3.2. Feature importance and epidemiologic findings

Shapley additive explanations (Figure 2) identify land-use patterns (0.34), precipitation variability (0.28), rodent species richness (0.25), housing age and condition (0.22), and peridomestic property area (0.20) as the dominant features. Socioeconomic variables contribute moderately and consistently: median household income (0.14), educational attainment (0.13), and poverty rate (0.12). Partial-dependence analysis reveals a threshold pattern for precipitation variability (steep risk increase once the 90-day coefficient of variation exceeds about 0.4), a gradient pattern for housing age (substantial elevation beyond 40 years), and a roughly linear association for poverty rate.

Of 620 cases with complete environmental investigations, probable exposure sites were identified in 91%. Peridomestic settings accounted for 60%-70% of documented exposures, workplaces for about 18%, and recreational settings for about 12%. Among 1,353 trapped rodents, 73% were Peromyscus maniculatus, with Sin Nombre virus seroprevalence of 20%, confirming active circulation at exposure locations.

4. Discussion

The ensemble AUC of 0.92 indicates that the model can correctly rank a randomly selected exposure site higher than a randomly selected nonexposure site in 92% of comparisons. Performance is substantially above what case-by-case epidemiologic investigation alone provides, and the maintained 0.89 AUC on temporally independent cases is consistent with stable operational utility over a multiyear window.

Precipitation variability emerges as the dominant environmental driver, supporting ecological hypotheses formulated after the 1993 Four Corners outbreak: variable precipitation expands vegetation, which elevates rodent food availability and reproductive output, and the resulting population irruptions increase human-rodent contact probability. As climate change reshapes precipitation patterns in endemic regions, the relevance of precipitation variability as a risk driver is likely to grow.

Socioeconomic factors contribute substantially: housing age, poverty rate, and educational attainment together account for nearly half of total feature importance after environmental drivers. Older and lower-quality housing has more entry points and harborage for rodents, lower-income households have fewer resources for exclusion and remediation, and lower educational attainment correlates with rural and agricultural occupations involving higher peridomestic exposure. Surveillance

strategies that integrate socioeconomic context can target prevention activities equitably.

Shapley additive explanations allow interpretation of model outputs in operational terms. Rather than a single opaque score, the model decomposes each prediction into the contributions of specific features, supporting tailored interventions: a high-risk score driven by housing age and precipitation may motivate housing improvements and seasonal surveillance, whereas one driven by land-use patterns and rodent richness may motivate vegetation management.

Limitations include geographic concentration in the western United States, aggregation across a 30-year period with evolving surveillance practices, and a static formulation that does not capture lag effects between climatic perturbations and human risk. Continuous prospective evaluation, integration with rodent surveillance, and coupling with downscaled climate projections are natural extensions.

5. Conclusion

Ensemble machine learning that integrates environmental, ecological, and socioeconomic data achieves strong predictive performance for hantavirus exposure risk and maintains that performance under independent temporal validation. Shapley additive explanations render predictions interpretable in terms aligned with public health practice, supporting a transition from reactive case-by-case investigation toward anticipatory surveillance that identifies high-risk locations before cases occur.

Table 1: Test-set performance of individual classifiers and the ensemble (n = 207).

Metric	RF	XGBoost	SVM	LR	Ensemble
AUC-ROC	0.89	0.91	0.84	0.81	**0.92**
Sensitivity	0.84	0.87	0.79	0.76	**0.88**
Specificity	0.86	0.88	0.82	0.79	**0.89**
PPV	0.85	0.87	0.81	0.78	**0.88**
NPV	0.85	0.89	0.80	0.77	**0.89**
F1 score	0.84	0.87	0.80	0.77	**0.88**

RF, random forest; SVM, support vector machine; LR, logistic regression; PPV, positive predictive value; NPV, negative predictive value.

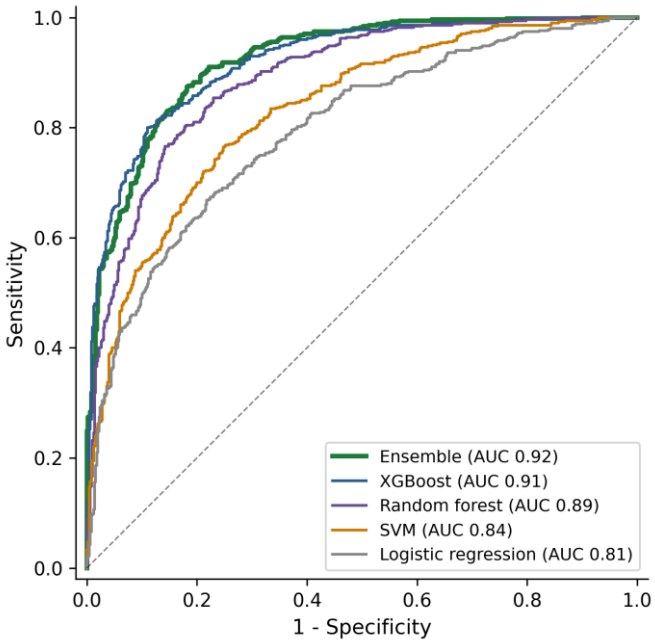

Figure 1: Receiver operating characteristic curves on the independent test set (n = 207). The ensemble (AUC 0.92) exceeds all component algorithms across the operating range.

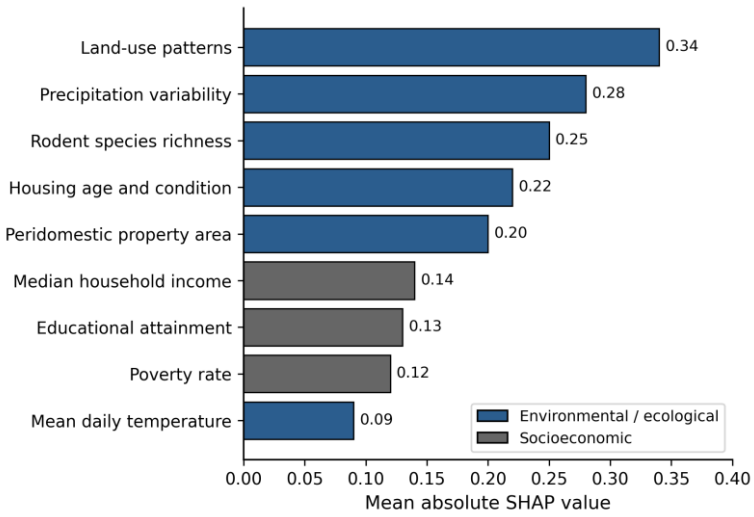

Figure 2: Shapley additive explanations (SHAP) feature importance for the ensemble model. Bars indicate mean absolute SHAP values; colors distinguish environmental/ecological from socioeconomic variables.

Acknowledgements. The author thanks the public health departments of California, Colorado, and New Mexico, whose case records and environmental investigations made this analysis possible, and the Decisions LAB research team for helpful methodological discussions.

References

[1] C.B. Jonsson, L.T. Figueiredo and O. Vapalahti, A global perspective on hantavirus ecology, epidemiology, and disease, *Clinical Microbiology Reviews*, **23** (2010), 412-441. https://doi.org/10.1128/cmr.00062-09

[2] B.T. Jackson, L.E. Sosa, S. Schmidt et al., Epidemiologic and environmental investigations of reported hantavirus cases in California, 1993-2020, *American Journal of Tropical Medicine and Hygiene*, **112** (2025), 1235-1245. https://doi.org/10.4269/ajtmh.24-0680

[3] L. Breiman, Random forests, *Machine Learning*, **45** (2001), 5-32. https://doi.org/10.1023/a:1010933404324

[4] T. Chen and C. Guestrin, XGBoost: a scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2016), 785-794. https://doi.org/10.1145/2939672.2939785

[5] S.M. Lundberg and S.I. Lee, A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems*, **30** (2017), 4765-4774.

[6] J. Snoek, H. Larochelle and R.P. Adams, Practical Bayesian optimization of machine learning algorithms, *Advances in Neural Information Processing Systems*, **25** (2012), 2951-2959.

[7] T. Fawcett, An introduction to ROC analysis, *Pattern Recognition Letters*, **27** (2006), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

[8] C. Cortes and V. Vapnik, Support-vector networks, *Machine Learning*, **20** (1995), 273-297. https://doi.org/10.1007/bf00994018

Received: May 1, 2026; Published: May 14, 2026