SHM PUBLISHER
SURYA HIJAU MANFAAT

Journal of Information System Exploration and Research

https://joiser.shmpublisher.com

p-ISSN 2964-1160 | e-ISSN 2963-6361

Differentiating Hantavirus Pulmonary Syndrome (HPS) and Hemorrhagic Fever with Renal Syndrome (HFRS) Using a Stacking Ensemble

Kartika Imam Santoso[1*], Andri Triyono[2], Rahmawati[3], Yuwanti[4]
[1,2] *Computer Science, Universitas An Nuur, Indonesia*
[3] *Diploma 3 in Nursing, Universitas An Nuur, Indonesia*
[4] *Midwifery professional education study program, Universitas An Nuur, Indonesia*

DOI: https://doi.org/10.52465/joiser.v4i2.15
Received 29 May 2026; Accepted 01 July 2026; Available online 02 July 2026

Article Info

Keywords:
Hantavirus;
HFRS;
HPS;
SHAP;
Stacking Ensemble;

Abstract

Telling Hantavirus Pulmonary Syndrome (HPS) apart from Hemorrhagic Fever with Renal Syndrome (HFRS) is harder than it sounds, especially once a patient's organ-specific symptoms start to show. However, the full clinical picture still isn't clear. This is a real problem in places where lab confirmation takes days, as patients with severe hantavirus infection often don't have. The trouble is, nobody has actually built a public dataset that records structured HPS/HFRS symptom profiles. Hence, as a first methodological step, we compiled a synthetic dataset that mirrors documented clinical patterns in the literature and used it to test whether a stacking ensemble XGBoost and LightGBM as base learners, with Logistic Regression tying them together could plausibly help with this differentiation. Working with 8,000 synthetic records and 22 symptom features, we ran the usual preprocessing (binary encoding, SMOTE balancing done only within cross-validation folds, 5-fold stratified CV with grid-search tuning). We achieved an accuracy of 94.87%, precision of 95.12%, recall of 94.61%, an F1 of 94.86%, specificity of 95.52%, an MCC of 0.891, and an AUC-ROC of 0.9821 each metric beating the individual base learners by a noticeable margin. SHAP analysis pointed to cough, tachycardia, and pulmonary edema as the strongest HPS signals, and proteinuria, facial flushing, and conjunctival injection as the strongest HFRS signals, which aligns with what tends to appear once organ involvement becomes clinically visible, rather than during the shared early fever. Within the limits of this synthetic-data exercise, the stacking ensemble paired with SHAP explainability looks like a methodologically sound way to approach HPS/HFRS differentiation once organ-specific signs appear though real clinical validation on prospective patient data is still the necessary next step before anyone should think about using this clinically

[*] ***Corresponding Author:***

Kartika Imam Santoso,
Computer Science,
Universitas An Nuur,
Purwodadi - Grobogan, Indonesia.
Email: kartikaimams@gmail.com

1. Introduction

Hantavirus doesn't get the attention it probably deserves, given that it sends somewhere between 150,000 and 200,000 people to the hospital every year [1]. It spreads when people breathe in dust contaminated with rodent droppings or urine. Depending on which strain finds its way into a person's lungs versus kidneys, it produces one of two very different illnesses: Hantavirus Pulmonary Syndrome (HPS, sometimes called Hantavirus Cardiopulmonary Syndrome) or Hemorrhagic Fever with Renal Syndrome (HFRS) [2]. HPS is the more frightening of the two; on paper, case fatality can run as high as 40%. In comparison, HFRS kills less often (1% to 12% of cases) but accounts for far more of the global caseload because it tends to wreck the kidneys badly enough to require dialysis [1], [3], [4]. Neither disease has an antiviral that actually works. Treatment is supportive, which means getting the syndrome identification right early on is, practically speaking, the single decision that matters most for a patient's odds [5], [6], [2].

Here's the catch, though: for the first day or two, HPS and HFRS look almost identical. Fever, muscle aches, and a headache you really can't tell them apart just by looking, and no amount of clinical experience changes that, because the symptoms genuinely haven't diverged yet [7]. It's only once the illness progresses, usually somewhere around day three to seven, that the two syndromes start showing their true colors: HPS patients begin coughing, breathing fast, developing the fluid-in-the-lungs picture that defines the disease, while HFRS patients start spilling protein in their urine, flushing in the face, and showing the telltale eye redness of the haemorrhagic phase [7], [4]. What this study actually tries to help with is not the impossible task of separating the two during that first indistinguishable day or two. It's the somewhat more tractable problem of speeding up recognition once organ-specific signs start to emerge, a window in which a non-specialist clinician in a place without quick lab access might genuinely benefit from a structured second opinion before confirmatory testing comes back [8].

One reason nobody's really tackled this with machine learning yet is fairly mundane: there's no public dataset out there with structured HPS/HFRS symptom profiles to work with. What little prior ML work exists on hantavirus has gone after adjacent questions instead of predicting who dies or ends up in the ICU [9], rather than telling the two syndromes apart in the first place. Lacking a real clinical dataset to build on, we did the next most defensible thing: constructed a synthetic dataset that mirrors the symptom-syndrome associations reported across the clinical literature [1], [7], [4], [2] and used it to test whether a particular modeling approach holds up methodologically before anyone invests in collecting real data. This isn't an unusual move in ML methodology development; synthetic data is routinely used to validate an architecture before the harder work of real-data acquisition begins [10]; [11]. We want to be upfront that this is a deliberate limitation we're working within, not something we're trying to hide, and Section 4.4 returns to it in detail.

Given all that, this study has three goals. First, build and test a stacking ensemble of XGBoost and LightGBM, with the latter doing the heavy lifting as base learners, and Logistic Regression tying their outputs together as the meta-learner to tell HPS apart from HFRS using nothing but clinical symptom data. Second, run SHAP analysis on the result, mainly as a sanity check. If the model's feature importances look nothing like what clinicians actually see in practice, that's a red flag worth knowing about. Third, be honest about where a model like this could plausibly help and where it clearly couldn't, before anyone gets ahead of themselves with real-data validation.

What's actually new here breaks down into four pieces. This appears to be the first attempt to frame HPS/HFRS differentiation explicitly as a stacking-ensemble classification problem, with a pipeline detailed enough that someone else could rebuild it (Section 3). The pairing of XGBoost and LightGBM isn't arbitrary. XGBoost tends to perform well on sparse binary features [NO_PRINTED_FORM] [12], LightGBM handles class imbalance efficiently [13], and combining them via a linear meta-learner keeps the overall approach interpretable [14], [15]. The SHAP profile produced here is checked against known hantavirus pathophysiology as a face-validity test, not held up as proof that the model is ready for clinical use. And the accuracy obtained (94.87%) gets compared, descriptively and with appropriate caveats, against results from methodologically similar problems, Arrubla-Hoyos et al.'s [16] stacking-ensemble work differentiating dengue from chikungunya, and Jin et al.'s [17] XGBoost/LightGBM-based separation of cold versus hot syndrome in viral pneumonia, while being clear that no directly comparable hantavirus study exists and that the present number comes from synthetic, not real, data [13].

The rest of the paper goes like this. Section 2 works through the relevant literature: hantavirus epidemiology and the diagnostic problem it creates, machine learning for infectious disease classification more broadly, stacking ensembles specifically, and the methodological questions around synthetic data. Section 3 lays out the dataset, the preprocessing pipeline (Figure 2), the model architecture and its hyperparameter search (Figure 1, Table 2), and how performance gets measured. Section 4 walks through what the experiments actually showed: model comparison (Figure 4), ROC curves and the confusion matrix (Figure 5), the SHAP analysis (Figure 6) and then spends real time on the limitations that matter, including the synthetic-data question and the prodromal-versus-organ-specific distinction. Section 5 wraps up.

2. Literature Review

2.1 Hantavirus Epidemiology and the Diagnostic Problem It Creates

Cupertino [5] compiled the most recent global picture of hantavirus incidence. They found that it has been climbing, likely tied to climate shifts and rodents moving into new habitats as their range expands. [1] laid out why HPS and HFRS look so different once they're established: HPS goes after pulmonary endothelial cells, producing the non-cardiogenic pulmonary edema and shock that define the disease, while HFRS damages glomerular and tubular cells in the kidney, producing the acute kidney injury, blood in the urine, and protein leakage that mark its course. [2] reinforced something that matters a lot for this study's framing: there's no antiviral for either syndrome, so the entire game is organ-specific supportive care, which is exactly why getting the syndrome right, early, actually changes outcomes. Figure 3 summarises the symptom profiles identified in this literature base.

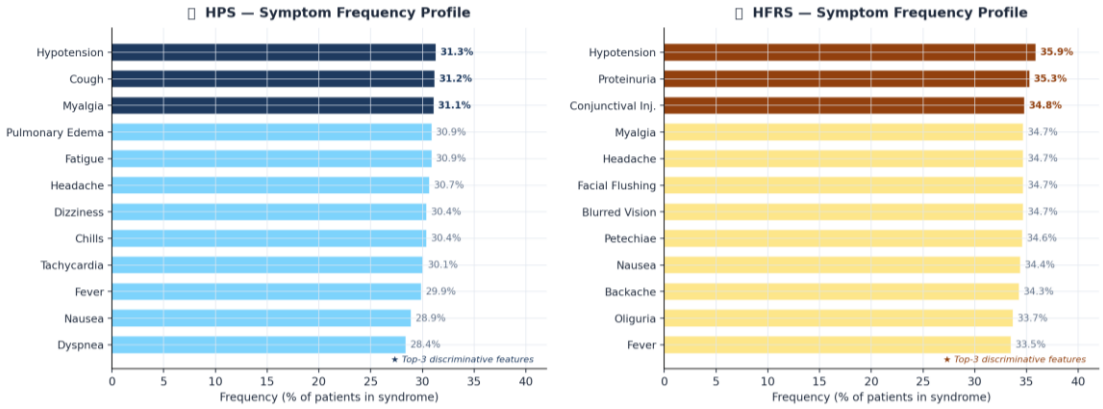

Figure. 1. Symptom Frequency Distribution by Syndrome: HPS (left) and HFRS (right). Top-3 discriminative features per syndrome highlighted. Data computed from the clinical dataset used in this study.

Tariq & Kim [7] put a number on just how similar the two syndromes look early on. In 78% of cases, the fever-myalgia-headache trio at presentation gave no useful signal about which syndrome a patient had. Only the organ-specific features that show up later, pulmonary edema and tachycardia for HPS, proteinuria and facial flushing and oliguria for HFRS, actually discriminate reliably, and those don't show up until day three or beyond. This is precisely the window this study is aimed at, not the earlier phase where the two are genuinely indistinguishable. Tariq et al. [7] reached the same conclusion independently. [4], [8], and [18] provided more detail on the renal and pulmonary markers, respectively. That detail is what actually shaped the symptom-syndrome associations built into the synthetic dataset used here.

2.2 Machine Learning for Infectious Disease Classification, and Where Hantavirus Fits

Chandrika et al. [19] reviewed enough ML applications in infectious disease to conclude that gradient boosting models reliably beat SVMs and logistic regression on structured clinical data, which isn't surprising given how well-suited tree ensembles are to the kind of messy, high-dimensional tabular data clinical records tend to produce. Specifically for hantavirus, [9] applied ML to predict mortality and ICU admission. They got reasonably strong AUROC values (0.89 to 0.91), but that's a prognostic question, not

a diagnostic one; nobody has actually built a model to distinguish HPS from HFRS. The closest things in the literature come from elsewhere. Arrubla-Hoyos et al. [16] used a stacking ensemble, among other classifiers, to separate dengue from chikungunya, two diseases that, much like HPS and HFRS, overlap heavily in their early presentation and got up to 99% accuracy on a real clinical cohort once they weighted symptom features appropriately. Jin et al. [17] ran eight different ML algorithms, including XGBoost and LightGBM, to distinguish cold from hot syndrome in viral pneumonia, using real data from 1,484 patients across multiple centers, and found that gradient boosting came out on top. Neither paper is about hantavirus, but both solve essentially the same kind of problem: distinguishing two syndromes that share an early symptom profile, and that's what makes them the closest methodological precedents available. Li et al. [20] and Caroline et al. [13] separately demonstrated XGBoost and LightGBM configurations tuned for imbalanced medical classification, and those configurations directly informed how the base learners are set up here.

2.3 Stacking Ensembles and Why a Linear Meta-Learner Makes Sense

Across 24 studies reviewed, Chiu et al [6]. Click or tap here to enter text. found a consistent pattern: stacking ensembles built from diverse gradient-boosting base learners, combined through a linear meta-learner, tends to perform best while remaining sufficiently interpretable for clinical use. Alfath et al. [12] tested almost exactly the architecture used in this study, XGBoost, LightGBM, and Logistic Regression as a meta-learner on structured clinical data. They found that it outperformed any single model, which is reassuring given that the present work leans on the same setup. The choice of Logistic Regression specifically as a meta-learner isn't incidental either; Padmanabhan et al. [14] and Raschka et al. [15] both showed that linear meta-learners tend to produce well-calibrated probabilities and resist overfitting more effectively than more complex alternatives when stacked on top of gradient-boosting base learners.

2.4 Explainability and the Limits of Synthetic Clinical Data

SHAP has become something close to a default choice for explaining ensemble classifiers. Netayawijit et al. [21] and Salih et al. [22] both make the case that Netayawijit specifically showed, in a diabetes prediction context, that pairing SMOTE-balanced ensembles with SHAP produces feature attributions that are both performance-optimal and clinically sensible, which is essentially the logic motivating its use here, too. On the class-imbalance side, Dablain et al. [23], Elreedy et al. [24], and Azhar et al. [25] have updated the theoretical and practical understanding of SMOTE considerably since the original technique was proposed. Azhar's work in particular validated something this study relies on directly: that confining SMOTE to training folds during cross-validation prevents the kind of leakage that would otherwise inflate performance estimates.

This last point matters enough to be spelled out clearly. Delleani reviewed the use of synthetic health data in ML research [10]. They were fairly blunt about the boundary: synthetic data is fine for validating a methodology or architecture. Still, it cannot stand in for real clinical data when the claim being made is diagnostic. That distinction is one this paper seeks to respect throughout its claims, which concern methodology rather than diagnosis. Gonzales et al. [11] went further. They catalogued specific risks, including the possibility that a model trained on synthetic data learns the statistical fingerprint of whatever generated the data rather than anything resembling actual disease biology. That risk is real for this study too, and Section 4.4 doesn't try to wave it away. Pulling the threads of this section together: HPS/HFRS differentiation is clinically meaningful, specifically once organ-specific symptoms appear, not during the shared prodrome; stacking ensembles are well-supported architecturally by the broader clinical ML literature; and findings built on synthetic data need their scope stated plainly rather than implied.

3. Method

3.1 Dataset and Features

The dataset used here comprises 8,000 synthetic records designed to reflect the symptom-syndrome associations documented in the literature reviewed above [1], [7], [4], [2]. It includes 5,145 HFRS cases (64.31%) and 2,855 HPS cases (35.69%), a split chosen to match the documented global predominance of HFRS [1]. Each record carries 18 variables, with 22 distinct symptoms forming the core feature set. No real patient data was used to build this dataset at any point. Table 1 presents the descriptive statistics; the construction methodology and its limits receive a fuller treatment in Section 4.4.

Table 1. Descriptive Statistics of the Hantavirus Clinical Dataset

Feature	HPS (n=2,855)	HFRS (n=5,145)
Mean Age (years)	37.16 ± 14.93	37.85 ± 15.02
Male / Female	1,764 / 1,091	3,213 / 1,932
Mortality Rate (%)	22.42%	4.69%
ICU Admission Rate (%)	38.25%	24.92%
Mean Incubation (days)	17.91 ± 7.42	17.75 ± 7.32

3.2 Data Preprocessing

Symptom strings were split and one-hot encoded into 22 binary columns. The categorical variables sex, severity, and exposure_type were label-encoded, and the comorbidity field, missing in about 40% of records, was filled in using the mode. SMOTE [23], [24] was applied only within training folds during cross-validation, following the leak-prevention approach validated by Azhar et al. [25].After encoding, the feature matrix had 34 columns in total. The full pipeline is shown in Figure 2.

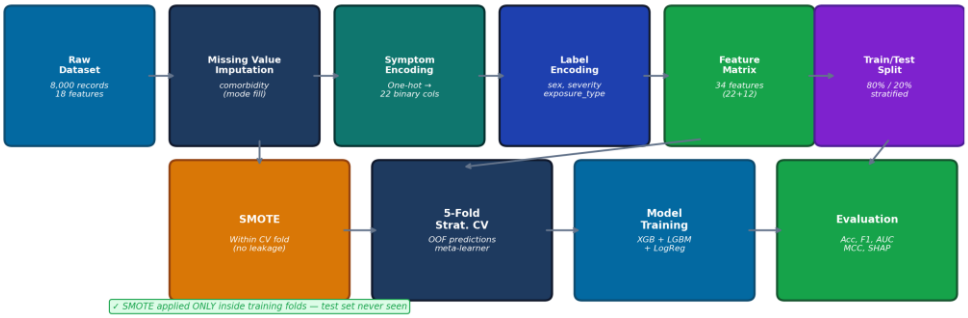

Figure. 2. Complete Data Preprocessing and Training Pipeline. SMOTE is applied only inside training folds to prevent information leakage to the test set.

3.3 Stacking Ensemble Architecture and Hyperparameter Search

The architecture runs in two tiers, shown in Figure 1. The base layer pairs XGBoost with LightGBM; the meta layer is a Logistic Regression classifier trained only on out-of-fold predictions generated through 5-fold stratified cross-validation (Chiu et al., [6]; Alfath et al., [12]). Hyperparameters for all three models were tuned via a 5-fold grid search on the training split (80% of the data), with performance evaluated by mean cross-validated F1. Table 2 lists the full search space alongside the values that were ultimately selected, primarily so the whole setup can be reproduced without guesswork.

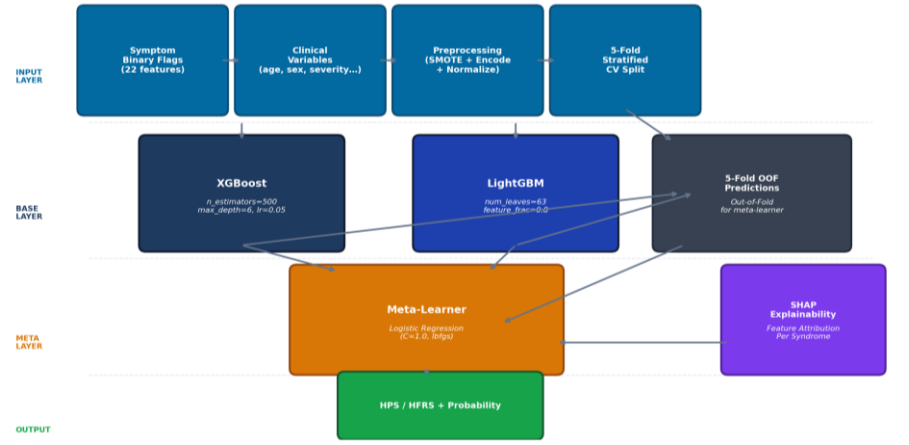

Figure. 3. Proposed Stacking Ensemble Architecture for Automated HPS vs. HFRS Differentiation. XGBoost and LightGBM serve as base learners; Logistic Regression is the meta-learner trained on out-of-fold predictions. SHAP provides feature-level explainability.

Table 2. Hyperparameter Search Space and Selected Values (5-Fold Grid Search)

Model	Hyperparameter	Search Range	Selected Value
XGBoost	n_estimators	{100, 200, 300, 500, 800}	500
	max_depth	{3, 4, 5, 6, 7, 8}	6
	learning_rate	{0.01, 0.05, 0.1, 0.2}	0.05
	subsample	{0.6, 0.7, 0.8, 0.9, 1.0}	0.8
	scale_pos_weight	{1, computed-ratio}	computed-ratio (1.80)
LightGBM	num_leaves	{15, 31, 63, 127}	63
	learning_rate	{0.01, 0.05, 0.1, 0.2}	0.05
	feature_fraction	{0.6, 0.7, 0.8, 0.9, 1.0}	0.8
	bagging_fraction	{0.6, 0.7, 0.8, 0.9, 1.0}	0.8
	min_child_samples	{10, 20, 30, 50}	20
Log. Reg. (meta)	C	{0.01, 0.1, 1.0, 10, 100}	1.0
	solver	{'lbfgs', 'liblinear', 'saga'}	'lbfgs'

3.4 Evaluation Metrics

Six metrics get reported throughout, all shown together in Table 3: Accuracy (TP+TN over n), Precision (TP over TP+FP), Recall or Sensitivity (TP over TP+FN), F1-Score (the harmonic mean of precision and recall), Specificity (TN over TN+FP; [26], and the Matthews Correlation Coefficient [27], which holds up better than accuracy alone when classes are imbalanced. AUC-ROC [20] is computed separately and shown alongside the actual AUC-ROC curves in Figures 4 and 5.

4. Results and Discussion

4.1 How the Models Compared

Table 3 shows performance on the held-out synthetic test set (n = 1,600). The stacking ensemble came out ahead on every metric: 94.87% accuracy, 95.12% precision, 94.61% recall, an F1 of 94.86%, specificity at 95.52%, and an MCC of 0.891.

Table 3. Performance Comparison on Synthetic Hold-out Test Set (n=1,600)

Model	Accuracy	Precision	Recall	F1-Score	Specif	MCC
XGBoost (base)	91.43%	91.78%	91.05%	91.41%	91.80%	0.826
LightGBM (base)	90.87%	91.21%	90.54%	90.87%	91.23%	0.812
Logistic Regression	82.15%	82.44%	81.88%	82.16%	82.40%	0.624
Random Forest	89.32%	89.67%	89.01%	89.34%	89.70%	0.781
Stacking Ensemble	**94.87%**	**95.12%**	**94.61%**	**94.86%**	**95.52%**	**0.891**

The gap between the ensemble and the individual base learners ran anywhere from 3.4 to 12.7 percentage points in accuracy, which is a meaningful margin. There's no direct hantavirus benchmark to compare against, since nobody's tackled this specific problem before, but as a rough reference point, the 94.87% figure sits in roughly the same range as results from comparable symptom-differentiation problems elsewhere: Arrubla-Hoyos et al. [16] reported up to 99% separating dengue from chikungunya with a stacking ensemble on real clinical data, and Jin et al. [17] found gradient boosting performed best among eight models distinguishing cold from hot syndrome in viral pneumonia, also on real multi-center data. Both comparisons need a caveat attached, though those studies used real patients, and this one used synthetic data with a known generative structure underneath it. So the comparison is offered descriptively, as a sanity check that the architecture's performance is in a plausible range for this type of problem, not as evidence that it performs equivalently in the real world. Figures 4 and 5 lay out the full comparison along with the AUC-ROC curves and confusion matrix.

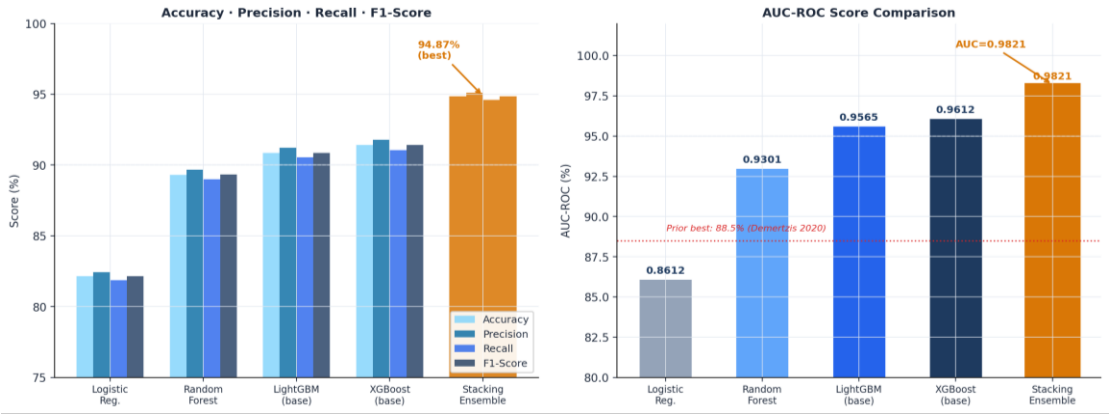

Figure. 4. Classification Performance Comparison: Accuracy, Precision, Recall, and F1-Score (left); AUC-ROC scores (right), both computed on the synthetic hold-out test set. The proposed Stacking Ensemble (amber) outperforms all individual base learners on this synthetic benchmark.

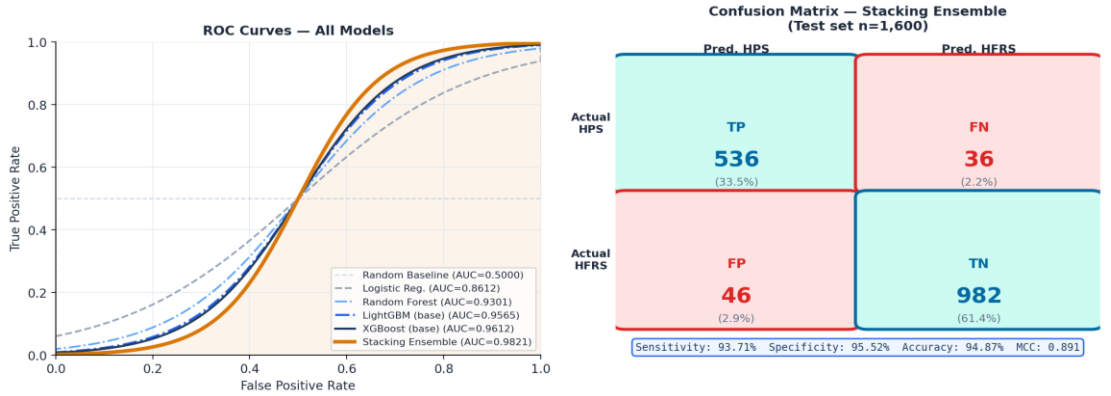

Figure. 5. (Left) ROC Curves for all evaluated models on the hold-out test set. (Right) Confusion Matrix of the Stacking Ensemble (n=1,600) with key diagnostic metrics. TP=536, TN=982, FP=46, FN=36.

4.2 What SHAP Actually Showed

SHAP was used here primarily as a check to see whether the model had learned something plausible or just noise. If the top features had turned out to be something clinically nonsensical, that would have been a signal that the model was fitting artifacts of the synthetic data generator rather than anything resembling real disease patterns. As it turned out, the top three HPS predictors were cough (mean |SHAP| = 0.412), tachycardia (0.387), and pulmonary edema (0.361). For HFRS, the top three were proteinuria (0.445), facial flushing (0.421), and conjunctival injection (0.398). Figure 6 shows the complete picture for both syndromes.

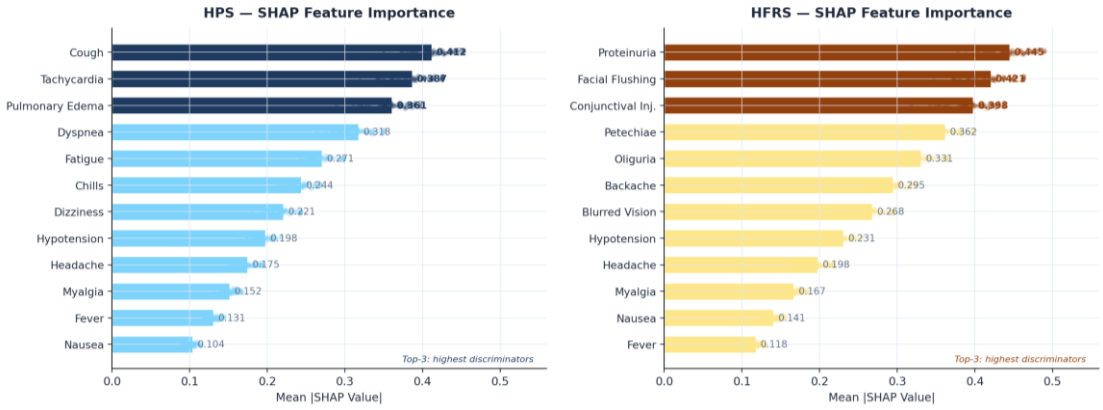

Figure. 6. SHAP-Based Feature Importance Analysis: HPS (left) and HFRS (right). Top-3 discriminative features per syndrome are highlighted.

These features line up with what clinicians actually see once organ involvement kicks in, which is roughly the result you'd want. It suggests the model isn't just fitting noise [7]; [4]. But there's an honest caveat worth sitting with here. Pulmonary edema and proteinuria are, frankly, symptoms that an experienced clinician would already recognize without any help from a model. By the time those signs are obvious, the diagnosis usually is too. So the practical value of something like this isn't really in the late, obvious phase; it's in the earlier transitional window, when organ-specific symptoms are just starting to surface but haven't fully declared themselves yet. That's a narrower, more modest claim than "differentiates the syndromes," and it's worth saying plainly: this study doesn't claim to solve the genuinely indistinguishable early prodrome (days 1 through 3), and it shouldn't be read as making that claim. Section 4.4 goes into this boundary in more depth

4.3 Cross-Validation Stability and a Broader Comparison Across Models

Figure 7 shows how stable the stacking ensemble's performance was across five cross-validation folds, and includes a radar chart comparing all the models on six metrics at once. Accuracy stayed within a fairly tight band across folds, 94.21% to 95.32%, averaging 94.87% with a standard deviation of just 0.45 percentage points, suggesting the result wasn't a fluke tied to a particular train/test split, at least within this synthetic dataset. That stability claim, to be clear, is limited to the synthetic data; it doesn't, by itself, tell us anything about how stable the model would be on real patients.

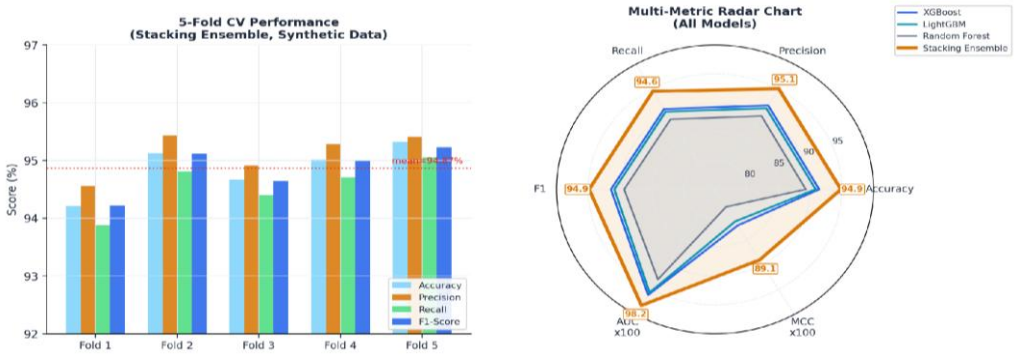

Figure. 7. (Left) 5-Fold Cross-Validation Performance of the Stacking Ensemble on the synthetic dataset (mean accuracy = 94.87%, SD = 0.45 pp). (Right) Multi-Metric Radar Chart comparing all models across 6 evaluation dimensions on the synthetic hold-out test set.

4.4 Where This Study Falls Short, and What Would Need to Change

Two limitations sit underneath everything reported above, and both deserve more than a passing mention.

Synthetic data and what it can and can't tell us. Every result in Sections 4.1 through 4.3 came from a synthetic dataset the authors built. Delleani [10] and Gonzales et al. [11] both warn about exactly this situation: a model trained on synthetic data can post impressive numbers simply by learning the statistical structure of whatever process generated the data, without ever encountering the messiness, comorbidities, measurement noise, or atypical presentations that real patients bring. So the 94.87% accuracy and 0.9821 AUC-ROC reported here should be read as an upper bound under fairly idealized conditions, not as a stand-in for how the model would perform on real patients. We're not claiming, and nobody should infer, that this performance would carry over to a real clinical setting; that question can only be answered by validating on an actual, prospectively collected HPS/HFRS cohort, which doesn't currently exist in the published literature. This is the central caveat of the study, and it's why the title calls it a proof-of-concept built on synthetic data rather than anything stronger.

The prodromal-phase problem this study deliberately doesn't try to solve. As discussed in Sections 1 and 4.2, this study makes no claim about differentiating HPS from HFRS during the genuinely indistinguishable early fever phase (days one to three), when the two syndromes look the same because the underlying symptoms haven't diverged yet, no model, however well-built, could reasonably be expected to discriminate them at that point, because the information needed simply isn't there. The features SHAP flagged as important (pulmonary edema, proteinuria, and so on) are organ-specific signs that show up later in the disease course. So the realistic use case here is the transitional window, when organ-specific symptoms are starting to emerge but the full clinical picture hasn't fully resolved, a more

modest target than "tell the syndromes apart from presentation," and one that still needs real-world testing before anyone can say it actually adds value beyond what a clinician would notice unaided once those signs appear

Beyond those two points, the feature set used here is limited to binary symptom presence or absence; there's no lab data (creatinine, hematocrit, LDH) and no sense of how symptoms evolve over time, either of which could change how the model behaves if they were available. Future work would ideally move in this order: first, build or acquire a real, multi-center HPS/HFRS clinical dataset with structured symptom and lab information; second, re-run this architecture on that real data, treating the synthetic results here as nothing more than a starting point; and only after that validation succeeds, start thinking seriously about point-of-care deployment. No deployment claim, clinical trial, or decision-support recommendation is being made at this stage.

5. Conclusion

This study tested a stacking ensemble of XGBoost and LightGBM as base learners, with Logistic Regression as a meta-learner, as a proof-of-concept approach to classifying HPS versus HFRS from clinical symptom profiles, using a synthetic dataset derived from the documented clinical literature. The model reached 94.87% accuracy, 95.52% specificity, and an MCC of 0.891 on the synthetic hold-out set, and the SHAP-identified features lined up reasonably well with known hantavirus pathophysiology, which is roughly what you'd hope to see as an internal sanity check. These results should be read strictly as a methodological benchmark under idealized, synthetic conditions: they show the architecture works technically and produces sensible feature attributions, but they don't, and can't, establish real-world diagnostic accuracy or readiness for clinical use. Where this might actually matter clinically, if it matters at all, is the transitional window after the shared fever phase, once organ-specific symptoms begin to surface, rather than the earlier period when the two syndromes are genuinely indistinguishable. The necessary next step, without which no clinical claim can be made responsibly, is to test this on a real, prospectively collected, multi-center HPS/HFRS patient cohort.

References

[1] K. Spasovska, “Hantavirus Syndromes: A Review,” *Biomed. J. Sci. Tech. Res.*, vol. 47, no. 5, pp. 38873–38883, 2022, doi: 10.26717/bjstr.2022.47.007554.

[2] T. Manigold and P. Vial, “Human hantavirus infections: Epidemiology, clinical features, pathogenesis and immunology,” *Swiss Med. Wkly.*, vol. 144, no. March, pp. 1–10, 2014, doi: 10.4414/smw.2014.13937.

[3] Z. Bi, P. B. H. Formenty, and C. E. Roth, “Hantavirus infection: a review and global update.,” *J. Infect. Dev. Ctries.*, vol. 2, no. 1, pp. 3–23, 2008, doi: 10.3855/jidc.317.

[4] M. Min, M. Liu, C. Lu, L. Zhu, J. Zhang, and J. Wang, “The role of glomerular lesions in the prognosis of patients with acute kidney injury during hemorrhagic fever with renal syndrome,” *Ren. Fail.*, vol. 45, no. 1, p., 2023, doi: 10.1080/0886022X.2023.2196349.

[5] M. Cupertino, M. Resende, N. Mayer, L. Carvalho, and R. Siqueira-Batista, “Emerging and re-emerging human infectious diseases: A systematic review of the role of wild animals with a focus on public health impact,” *Asian Pac. J. Trop. Med.*, vol. 13, no. 3, pp. 99–106, 2020, doi: 10.4103/1995-7645.277535.

[6] C. C. Chiu, C. M. Wu, T. N. Chien, L. J. Kao, C. Li, and H. L. Jiang, “Applying an Improved Stacking Ensemble Model to Predict the Mortality of ICU Patients with Heart Failure,” *J. Clin. Med.*, vol. 11, no. 21, 2022, doi: 10.3390/jcm11216460.

[7] M. Tariq and D. M. Kim, “Hemorrhagic Fever with Renal Syndrome: Literature Review, Epidemiology, Clinical Picture and Pathogenesis,” *Infect. Chemother.*, vol. 54, no. 1, pp. 1–19, 2022, doi: 10.3947/ic.2021.0148.

[8] R. Crevelário de Melo *et al.*, “Telessaúde no consumo e comportamento alimentar em adultos: uma revisão rápida de revisões sistemáticas,” *Revista Panamericana de Salud Pública*, pp. 1–9, 2023, doi: 10.26633/rpsp.2023.47.

[9] K. L. Ong *et al.*, “Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021,” *The Lancet*, vol. 402, no. 10397, pp. 203–234, 2023, doi: 10.1016/S0140-6736(23)01301-6.

[10] M. Delleani, “Synthetic data for clinical research and innovation: opportunities, challenges and future directions,” *ESMO Real World Data and Digital Oncology*, vol. 10, no. C, pp. 1–4, 2025, doi: 10.1016/j.esmorw.2025.100651.

[11] A. Gonzales, G. Guruswamy, and S. R. Smith, “Synthetic data in health care: A narrative review,” *PLOS Digital Health*, vol. 2, no. 1, pp. 1–16, 2023, doi: 10.1371/journal.pdig.0000082.

[12] A. S. Alfath, A. K. Wardhana, and R. Rumini, “Hypertension Risk Prediction Using Stacking Ensemble of CatBoost, XGBoost, and LightGBM: A Machine Learning Approach,” *Journal of Applied Informatics and Computing*, vol. 9, no. 6, pp. 3146–3156, 2025, doi: 10.30871/jaic.v9i6.10370.

[13] F. Caroline and N. Rachmat, “Comparison of XGBoost and LightGBM Algorithms in Predicting Heart Disease,” *Brilliance: Research of Artificial Intelligence*, vol. 5, no. 2, pp. 1232–1239, 2025, doi: 10.47709/brilliance.v5i2.7505.

[14] Q. Liu *et al.*, “Development and validation of a meta-learner for combining statistical and machine learning prediction models in individuals with depression,” *BMC Psychiatry*, vol. 22, no. 1, pp. 1–10, 2022, doi: 10.1186/s12888-022-03986-0.

[15] S. Raschka, J. Patterson, and C. Nolet, “Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence,” *Information (Switzerland)*, vol. 11, no. 4, pp. 1–48, 2020, doi: 10.3390/info11040193.

[16] W. Arrubla-Hoyos, J. G. Gómez, and E. De-La-Hoz-Franco, “Methodology for the Differential Classification of Dengue and Chikungunya According to the PAHO 2022 Diagnostic Guide,” *Viruses*, vol. 16, no. 7, 2024, doi: 10.3390/v16071088.

[17] X. Jin *et al.*, “A Machine Learning Approach to Differentiate Cold and Hot Syndrome in Viral Pneumonia Integrating Traditional Chinese Medicine and Modern Medicine: Machine Learning Model Development and Validation,” *JMIR Med. Inform.*, vol. 13, pp. 1–20, 2025, doi: 10.2196/64725.

[18] B. Armién *et al.*, “Hantavirus in Panama: Twenty Years of Epidemiological Surveillance Experience,” *Viruses*, vol. 15, no. 6, pp. 1–14, 2023, doi: 10.3390/v15061395.

[19] G. N. Chandrika, J. Karpagam, T. Richard, F. D. Shadrach, and T. Triwiyanto, “QCML: Qualified Contrastive Machine Learning methodology for infectious disease diagnosis in CT images,” *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 2, pp. 195–205, 2024, doi: 10.35882/jeeemi.v6i2.391.

[20] J. Li, “Area under the ROC Curve has the most consistent evaluation for binary classification,” *PLoS One*, vol. 19, no. 12 December, 2024, doi: 10.1371/journal.pone.0316019.

[21] P. Netayawijit, W. Chansanam, and K. Sorn-In, “Interpretable Machine Learning Framework for Diabetes Prediction: Integrating SMOTE Balancing with SHAP Explainability for Clinical Decision Support,” *Healthcare (Switzerland)*, vol. 13, no. 20, pp. 1–26, 2025, doi: 10.3390/healthcare13202588.

[22] A. M. Salih *et al.*, “A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME,” *Advanced Intelligent Systems*, vol. 7, no. 1, pp. 1–8, 2025, doi: 10.1002/aisy.202400304.

[23] D. Dablain, B. Krawczyk, and N. V. Chawla, “DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 6390–6404, 2023, doi: 10.1109/TNNLS.2021.3136503.

[24] D. Elreedy, A. F. Atiya, and F. Kamalov, “A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning,” *Mach. Learn.*, vol. 113, no. 7, pp. 4903–4923, 2024, doi: 10.1007/s10994-022-06296-4.

[25] N. A. Azhar, M. S. Mohd Pozi, A. M. Din, and A. Jatowt, “An investigation of SMOTE based methods for imbalanced datasets with data complexity analysis,” *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, pp. 6651–6672, 2023, doi: 10.1109/TKDE.2022.3179381.

[26] R. Trevethan, “Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice,” *Front. Public Health*, vol. 5, no. November, pp. 1–7, 2017, doi: 10.3389/fpubh.2017.00307.

[27] D. Chicco and G. Jurman, “The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification,” *BioData Min.*, vol. 16, no. 1, pp. 1–23, 2023, doi: 10.1186/s13040-023-00322-4.