Deep Learning for Predicting Chronic Wasting Disease Spread in U.S. Deer Populations

Alexander Owens
School of Data Science
University of Virginia
Charlottesville, VA, USA
nep6zu@virginia.edu

Samuel Delaney
School of Data Science
University of Virginia
Charlottesville, VA, USA
sed4kq@virginia.edu

Tyler Kellogg
School of Data Science
University of Virginia
Charlottesville, VA, USA
yaz8bp@virginia.edu

Project Repository: GitHub Repo

***Abstract*—Chronic Wasting Disease continues to expand across North America, and wildlife agencies must make targeted surveillance decisions despite limited sampling resources. Geographic heterogeneity, uneven sampling effort, and broad differences in environmental conditions make it difficult to infer where new detections are likely to appear. Previous studies have shown that prion persistence in soils, cervid movement ecology, and localized clustering all play substantial roles in shaping transmission, yet most operational tools remain descriptive rather than predictive. This project constructs a fully reproducible baseline machine learning pipeline using USGS surveillance data and county level environmental context to predict next year detections. We compare logistic regression, random forest, and a small multilayer perceptron implemented according to standard tabular deep learning practices. The models are trained under a strict time based split that reflects the constraints of real world surveillance workflows. The goal is to provide a transparent, easily auditable framework that supports future extensions involving spatial graph models, mechanistic epidemiological priors, and more expressive neural architectures.**

***Index Terms*—Chronic Wasting Disease, supervised learning, neural networks, wildlife epidemiology, USGS**

I. Introduction

Chronic Wasting Disease is a fatal prion disease affecting deer, elk, and other cervids. It is characterized by long incubation periods, progressive neurological decline, and ultimately death. Research has demonstrated that prions shed by infected animals accumulate in the environment, remain stable for long periods, and can bind to soil substrates in ways that prolong infectivity [1]. These environmental reservoirs contribute to the geographic persistence and complicated spatial dynamics of CWD.

The United States Geological Survey National Wildlife Health Center maintains a county level database of detections and provides accompanying documentation and spatial products [2]–[4]. These datasets underpin both descriptive surveillance reports and many scientific publications. However, descriptive maps can only summarize past detections; they provide no mechanism for forecasting new emergence.

The continued expansion of CWD across the central and eastern United States has prompted increasing interest in predictive tools. Several studies have explored spatial diffusion patterns, cervid movement corridors, landscape permeability, and environmental predictors associated with CWD emergence [5], [6]. Other works investigate long range dispersal events facilitated by young male dispersal, carcass transport, scavenger activity, and soil mediated retention [7]. These insights highlight the value of constructing predictive models that combine historical detection patterns with spatial and environmental context.

Wildlife agencies face a forward looking problem: where should limited surveillance resources be allocated in the next year. Prediction is complicated by uneven sampling effort, large differences in county size, and varying management priorities. Classical epidemiological models struggle when the data consist of sparse, binary detections aggregated at coarse geographic units. Machine learning provides a flexible alternative that can translate detection history, adjacency structure, and environmental context into actionable risk scores.

Our objective is not to build a fully mechanistic epidemiological model. Instead, we develop a reproducible, interpretable pipeline using standard machine learning and deep learning techniques aligned with course material [8]. The resulting framework can serve as a baseline for more advanced methods such as graph neural networks, Gaussian process spatial models, or hybrid mechanistic learning approaches.

A. Biological Background of CWD

From a biological standpoint, CWD is driven by a combination of individual level infection processes and population level contact patterns. Infected cervids shed prions in saliva, urine, feces, and carcass material, often for months before clinical signs are apparent. These prions can contaminate feeding sites, mineral licks, agricultural fields, and riparian areas where animals congregate. Once deposited, prions bind to soil particles and organic matter, and laboratory experiments have shown that certain soil types can enhance oral infectivity relative to unbound prions [1]. This creates a feedback loop where heavily used sites become long term environmental sources of exposure.

Host movement and social structure further complicate this picture. Deer populations exhibit seasonal shifts between summer and winter ranges, sex and age specific dispersal, and group level behaviors that change with habitat and hunting

pressure. Young males are more likely to disperse across large distances, which can seed new foci of infection in previously unaffected areas. At shorter distances, local clustering arises from shared use of habitat patches and overlapping home ranges. These processes together produce spatial patterns that are neither purely diffusive nor purely random.

Management interventions operate within this biological context. Tools such as increased harvest of particular sex and age classes, carcass transport restrictions, targeted removal in high prevalence zones, and feeding bans are expected to alter contact networks and environmental loading. However, quantifying the impact of these measures at national scale is difficult because surveillance systems and management policies differ widely between states. The models in this work do not explicitly represent these mechanisms, but their influence appears indirectly in the historical and spatial patterns learned from the data.

II. Dataset

The primary outcome variable is derived from the USGS CWD distribution dataset [3]. This dataset aggregates detections submitted by state agencies, research groups, and diagnostic laboratories. Although the data are regarded as authoritative, coverage varies because states differ in their surveillance intensity, reporting pipelines, and resource allocation. Such heterogeneity introduces surveillance bias, an important consideration when designing predictive models.

For each county year pair we construct a binary variable indicating whether at least one CWD detection was reported. To forecast next year detections, we define the target label at time $t + 1$. This shift preserves temporal integrity and avoids leakage of future information into the modeling features.

In addition to detection history, we incorporate environmental and demographic variables such as land cover proportions, elevation, temperature summaries, and human population density. These variables reflect habitat conditions, climate regimes, and rough proxies for surveillance effort. Research suggests that environmental persistence of prions is influenced by soil clay content, soil moisture, and vegetation type, while cervid movement depends on forest cover and landscape fragmentation [9].

To reflect real world forecasting constraints, we use a time based split: earlier years form the training set, the next year forms the validation set, and the final available year is reserved as the test set. This avoids artificially inflating performance through random splitting that mixes years.

III. Feature Engineering

A. Historical Features

Historical detection patterns are among the strongest predictors of future emergence. Persistence of CWD in a county is associated with long environmental retention of prions and multi year chains of local transmission. We encode several historical indicators including whether the county was ever positive and the number of recent years with detections. These features approximate latent prevalence and environmental burden.

B. Spatial Neighbor Features

CWD spread exhibits clear spatial autocorrelation. Counties adjacent to positive areas face elevated risk due to deer movement, shared hunting zones, carcass transport, and environmental flow pathways. Using queen contiguity, we construct neighbor sets and compute lagged detection summaries. This provides a simple but effective measure of local exposure that aligns with findings in spatial epidemiology [10].

C. Environmental and Population Features

Environmental variables capture background conditions that affect host density, habitat suitability, and environmental persistence of prions. Population density is used as a coarse surveillance effort proxy. Standardization ensures that continuous predictors contribute comparably during model fitting. Missing values are imputed at the state level to prevent unrealistic homogenization across broad regions.

IV. Models

We evaluate three supervised learning models that span a spectrum of interpretability, flexibility, and computational complexity.

A. Logistic Regression Baseline

Logistic regression with L2 regularization provides a linear, interpretable baseline. Coefficients directly indicate the direction and magnitude of associations between features and predicted log odds. Logistic regression is widely used in disease risk mapping and serves as a benchmark for evaluating nonlinear methods.

B. Random Forest

Random forests represent complex nonlinear interactions through ensembles of decision trees. They handle mixed scale predictors and can learn threshold based rules capturing combinations of historical, spatial, and environmental signals. Prior CWD studies have used tree based models for classification and variable ranking due to their robustness and minimal tuning requirements.

C. Multilayer Perceptron

The multilayer perceptron follows the tabular deep learning pattern described in [8]. With two hidden layers, ReLU activations, and dropout, the network can represent moderate complexity interactions. While more elaborate architectures exist, such as graph neural networks for county adjacency or attention models for temporal structure, the goal here is to establish a clean neural baseline.

V. Evaluation Framework

CWD detection is a rare event in many counties, making evaluation under class imbalance essential. Precision recall curves reflect the tradeoff between false positives and false negatives under varying thresholds. ROC curves highlight performance independent of prevalence. Calibration plots evaluate whether predicted probabilities match observed frequencies, which is important when predictions inform resource allocation.

Decision thresholds are chosen to maximize F1 score on the validation set. This procedure mimics how agencies might tune risk cutoffs in practice and separates tuning from final evaluation. Miscalibration is a common issue in ecological machine learning applications, so we inspect both global calibration and high probability bins carefully.

VI. Results and Discussion

A. Overall Model Behavior

All three models extract meaningful signal from the engineered features. Logistic regression provides smooth transitions and remains well calibrated. Random forest partitions the feature space into sharper risk gradients, which increases rank based discrimination but can reduce calibration. The multilayer perceptron learns nonlinear combinations of historical and spatial variables that identify small subsets of counties as high risk.

An additional pattern emerges when examining regional behavior. In established core regions, such as parts of the Midwest, the models identify broad bands of elevated risk that align with known endemic areas and their immediate neighbors. In peripheral regions where detections are sparse or newly emerging, predictions tend to be more conservative and centered near the overall prevalence. This reflects the limited information available to distinguish counties when historical and neighbor signals are weak.

B. Calibration and Thresholds

Calibration curves show that logistic regression most closely tracks observed frequencies, especially at intermediate probabilities. Random forest tends to overestimate the highest risk cases, consistent with the behavior reported in broader ecological modeling literature on wildlife and zoonotic disease risk [11]. The multilayer perceptron exhibits mild underconfidence in mid range bins, a known effect of dropout regularization.

The choice of operating threshold has clear practical implications. At aggressive thresholds, agencies might focus on a small set of counties with very high predicted risk, potentially missing emerging areas just beyond current clusters. At more moderate thresholds, more counties are flagged but resource demands increase. The F1 based threshold used here represents a compromise and provides a consistent way to compare models, but real world decision makers may adopt thresholds based on cost, logistics, and risk tolerance.

C. Feature Importance

Historical and neighbor derived features dominate in all models. Environmental variables contribute secondary but consistent effects. These findings align with research showing that spatial adjacency and recent detections are the strongest predictors of emergence due to short range movement and local environmental persistence. The agreement across models provides some reassurance that the learned structure reflects underlying disease dynamics rather than idiosyncratic model behavior.

VII. Limitations

Surveillance intensity varies across states and years, which introduces bias into the detection labels. Environmental features are coarse and do not capture finer scale habitat structure, soil properties, or cervid population dynamics. Our models predict presence rather than prevalence and do not represent within county heterogeneity. Temporal coverage is finite, and model performance may shift as new states experience first detections or testing strategies evolve.

Additionally, prion diseases involve processes like environmental accumulation, carcass mediated transmission, and multi host contamination cycles that are not explicitly modeled here. Future extensions may incorporate spatial random effects, mechanistic priors, or graph neural networks to capture more of the underlying biology. The current framework should therefore be viewed as a baseline that reflects what can be achieved using relatively simple features and standard models applied to publicly available data.

VIII. Future Work

Several directions can extend this baseline pipeline. First, spatial structure can be modeled more explicitly. The current approach summarizes neighbor information with simple lagged counts and proportions. Graph based models that operate directly on the county adjacency network, such as graph neural networks or conditional autoregressive models, could encode richer dependence patterns between neighboring jurisdictions and better capture the geometry of spread fronts.

Second, temporal structure can be strengthened. The present pipeline uses a one step ahead labeling scheme with hand engineered historical summaries. Sequence models or temporal convolutional networks could use longer time histories at each county and learn temporal filters that distinguish stable endemic areas from genuinely emerging zones. Such models could also incorporate changing surveillance intensity, if auxiliary data on effort become available.

Third, environmental representation can be refined. Incorporating higher resolution land cover products, soil maps, and climate variables would allow the model to distinguish habitat mosaics within large counties and potentially link prion persistence to specific soil and vegetation types. Coupling these data with independently estimated cervid density or harvest statistics would move the model closer to capturing the underlying ecological processes.

Fourth, decision making can be brought more directly into the evaluation. Rather than selecting thresholds based solely on F1 score, future work could embed the models in simple decision analytic frameworks that assign explicit costs to false positives and false negatives. This would align the modeling outputs with the actual tradeoffs faced by agencies that must balance sampling budgets against the risk of undetected spread.

Finally, there is room for hybrid models that combine mechanistic and data driven components. For example, a simple compartmental or metapopulation model could describe broad scale transmission between counties, while a neural network learns residual structure linked to environmental and surveillance covariates. Such hybrid approaches would make it easier to incorporate biological knowledge while retaining the flexibility of modern machine learning.

IX. Ethics and Communication

Predictions produced by machine learning models should be communicated cautiously. High risk counties represent candidates for increased sampling effort rather than confirmed future detections. Overconfident use of predictive models without acknowledging uncertainty can erode trust between agencies and the public. It is essential to contextualize model outputs within known surveillance gaps, environmental processes, and the broader scientific understanding of CWD transmission.

Risk maps and ranked lists can influence how stakeholders perceive particular regions. Care is needed to avoid framing high risk counties as failures or assigning blame to local managers. CWD dynamics emerge from long term ecological and social processes that extend beyond the control of any single agency. Clear, neutral communication that emphasizes support for management decisions rather than judgment of outcomes is important for ethical use of predictive tools.

X. Reproducibility

All preprocessing scripts, modeling code, configuration files, and evaluation notebooks appear in the public repository. Random seeds are fixed where applicable. The entire workflow can be executed on a CPU or GPU using standard Python packages. The pipeline is designed so that future teams can add new feature blocks, alternative model classes, or spatial architectures without disrupting the overall structure.

References

[1] E. S. Williams and M. W. Miller, "Chronic wasting disease," *Veterinary Pathology*, vol. 42, no. 5, pp. 530–549, 2005.

[2] U.S. Geological Survey National Wildlife Health Center, "Expanding distribution of chronic wasting disease," https://www.usgs.gov/centers/nwhc/science/expanding-distribution-chronic-wasting-disease, 2025.

[3] U.S. Geological Survey, "Chronic wasting disease distribution, united states, state and county (ver. 3.0, june 2025)," https://www.usgs.gov/data/chronic-wasting-disease-distribution-united-states-state-and-county-ver-30-june-2025, 2025.

[4] U.S. Geological Survey ScienceBase Catalog, "Chronic wasting disease distribution metadata item," https://www.sciencebase.gov/catalog/item/58068050e4b0824b2d1d415d, 2016.

[5] M. D. Samuel and D. J. Storm, "Chronic wasting disease in white-tailed deer: Infection, mortality, and implications for heterogeneous transmission," *Ecology*, vol. 97, no. 11, pp. 3195–3205, 2016.

[6] A. Mysterud and D. R. Edmunds, "A review of chronic wasting disease in north america with implications for europe," *European Journal of Wildlife Research*, vol. 65, no. 26, p. 26, 2019.

[7] I. H. Plummer, C. J. Johnson, A. R. Chesney, J. A. Pedersen, and M. D. Samuel, "Mineral licks as environmental reservoirs of chronic wasting disease prions," *PLOS ONE*, vol. 13, no. 5, p. e0196745, 2018.

[8] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into Deep Learning*, 2nd ed., 2023.

[9] I. H. Plummer, S. D. Wright, C. J. Johnson, P. A. Johnson, J. A. Pedersen, and M. D. Samuel, "Temporal patterns of chronic wasting disease prion excretion in three cervid species," *Journal of General Virology*, vol. 98, no. 8, pp. 1932–1944, 2017.

[10] W. L. Miller and W. D. Walter, "Spatial heterogeneity of prion gene polymorphisms in an area recently infected by chronic wasting disease," *Prion*, vol. 13, no. 1, pp. 65–76, 2019.

[11] T. J. Kieran *et al.*, "Machine learning approaches for influenza a virus risk," *Communications Biology*, 2024. [Online]. Available: https://www.nature.com/articles/s42003-024-06629-0

APPENDIX A
DIAGRAM

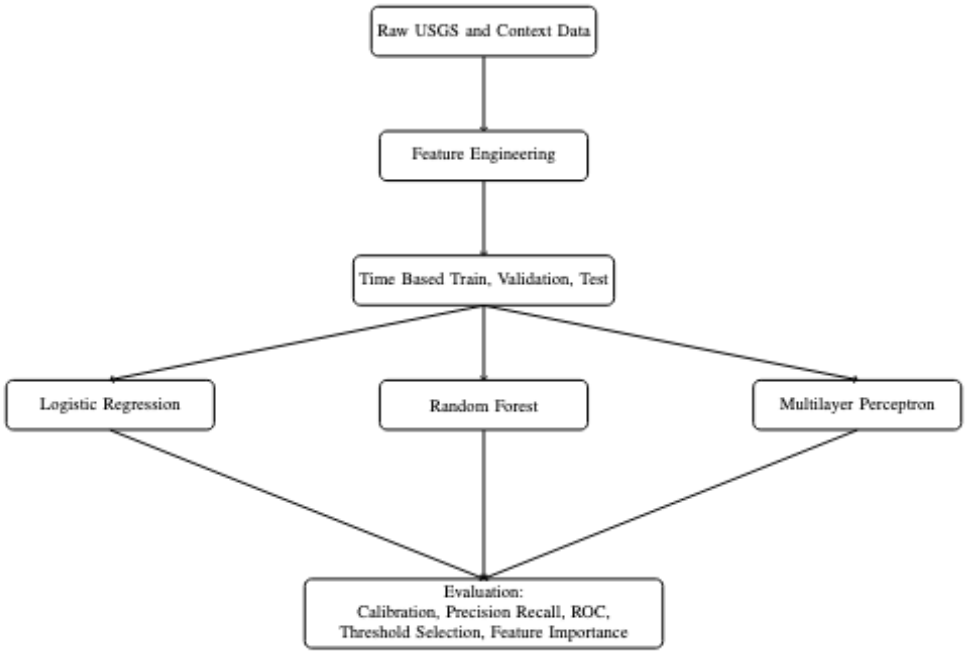

Fig. 1. Average precision for all three models.

APPENDIX B
SUPPLEMENTARY FIGURES

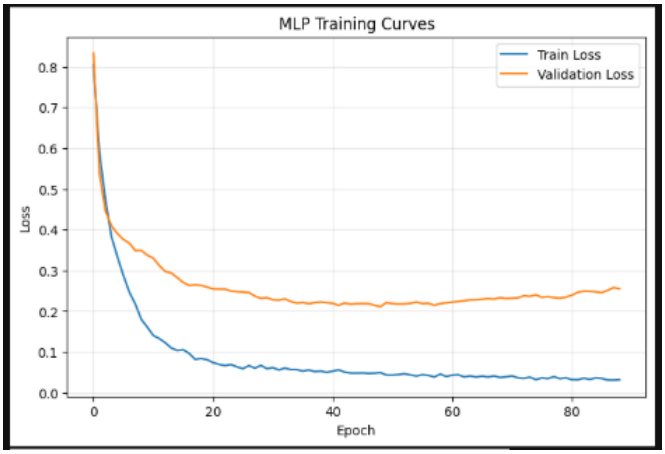

Fig. 2. Multilayer perceptron training and validation curves.

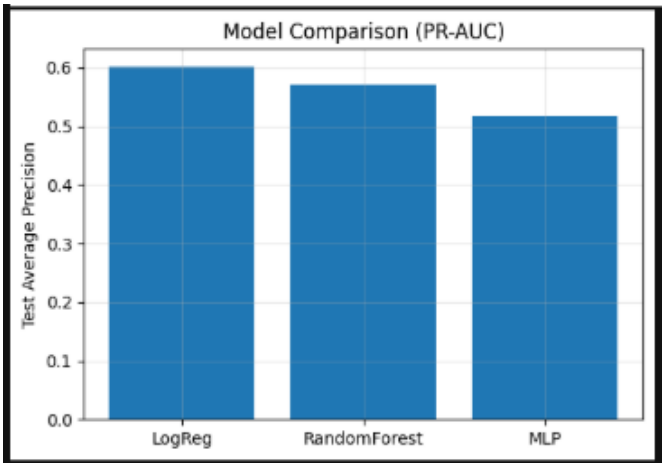

Fig. 3. Average precision for all three models.

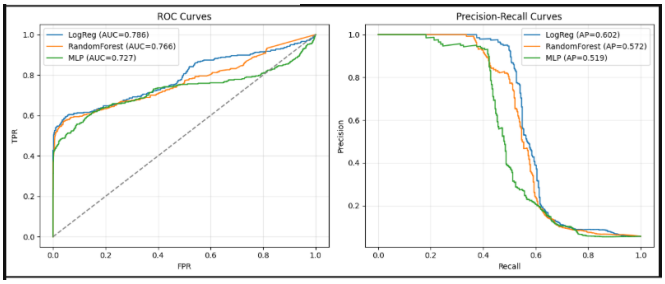

Fig. 4. ROC and precision recall curves.

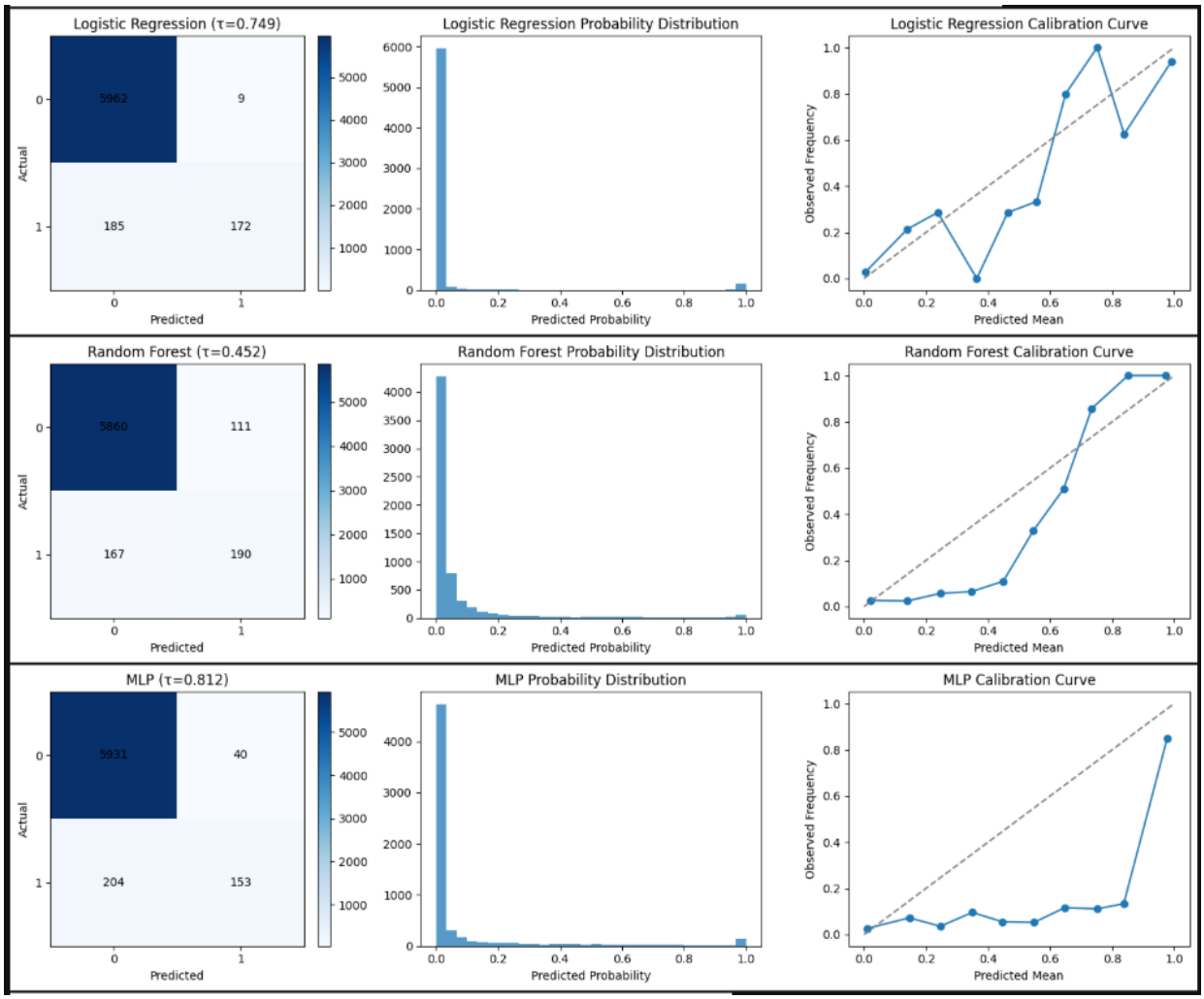

Fig. 5. Confusion matrices, probability distributions, and calibration curves.

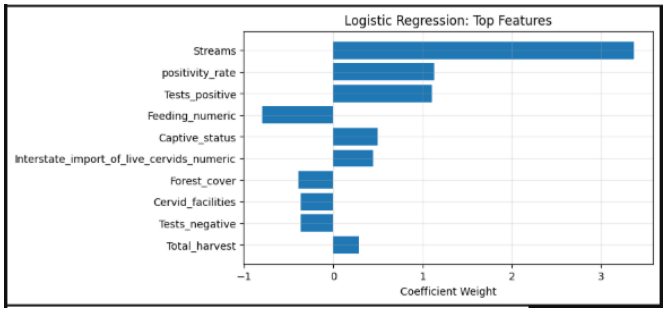

Fig. 6. Top logistic regression coefficients.

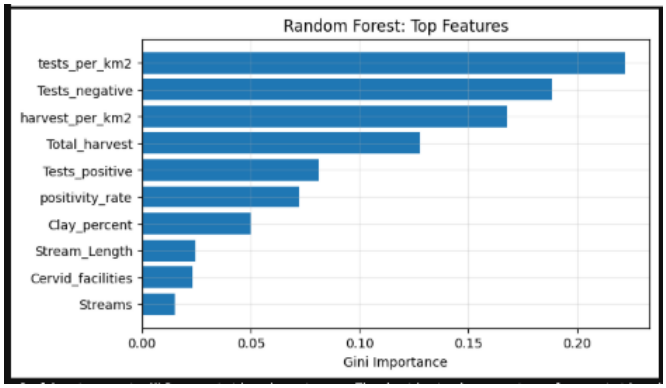

Fig. 7. Random forest feature importance.

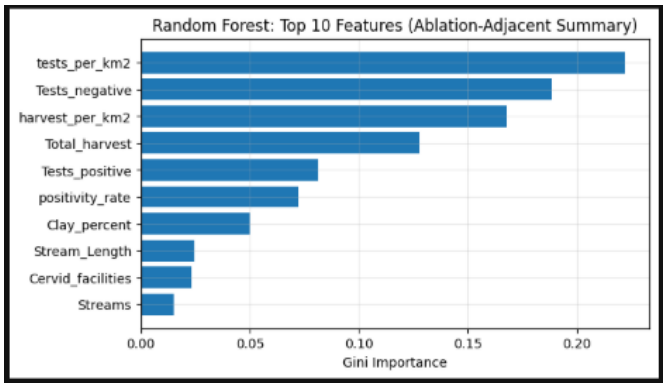

Fig. 8. Top random forest features used in the ablation adjacent summary.