



OPEN

Predicting chronic wasting disease in white-tailed deer at the county scale using machine learning

Md Sohel Ahmed^{1,22✉}, Brenda J. Hanley¹, Corey I. Mitchell^{2,23}, Rachel C. Abbott¹, Nicholas A. Hollingshead¹, James G. Booth³, Joe Guinness³, Christopher S. Jennelle⁴, Florian H. Hodel⁵, Carlos Gonzalez-Crespo⁶, Christopher R. Middaugh⁷, Jennifer R. Ballard⁷, Bambi Clemons⁸, Charlie H. Killmaster⁹, Tyler M. Harms¹⁰, Joe N. Caudell¹¹, Kathryn M. Benavidez Westrich¹¹, Emily McCallen¹¹, Christine Casey¹², Lindsey M. O'Brien¹³, Jonathan K. Trudeau¹³, Chad Stewart¹⁴, Michelle Carstensen¹⁵, William T. McKinley¹⁶, Kevin P. Hynes¹⁷, Ashley E. Stevens¹⁷, Landon A. Miller¹⁷, Merrill Cook¹⁸, Ryan T. Myers¹⁸, Jonathan Shaw¹⁸, Michael J. Tonkovich¹⁹, James D. Kelly⁸, Daniel M. Grove²⁰, Daniel J. Storm²¹ & Krysten L. Schuler¹

Continued spread of chronic wasting disease (CWD) through wild cervid herds negatively impacts populations, erodes wildlife conservation, drains resource dollars, and challenges wildlife management agencies. Risk factors for CWD have been investigated at state scales, but a regional model to predict locations of new infections can guide increasingly efficient surveillance efforts. We predicted CWD incidence by county using CWD surveillance data depicting white-tailed deer (*Odocoileus virginianus*) in 16 eastern and midwestern US states. We predicted the binary outcome of CWD-status using four machine learning models, utilized five-fold cross-validation and grid search to pinpoint the best model, then compared model predictions against the subsequent year of surveillance data. Cross validation revealed that the Light Boosting Gradient model was the most reliable predictor given the regional data. The predictive model could be helpful for surveillance planning. Predictions of false positives emphasize areas that warrant targeted CWD surveillance because of similar conditions with counties known to harbor CWD. However, disagreements in positives and negatives between the *CWD Prediction Web App* predictions and the on-the-ground surveillance data one year later underscore the need for state wildlife agency professionals to use a layered modeling approach to ensure robust surveillance planning. The *CWD Prediction Web App* is at <https://cwd-predict.streamlit.app/>.

Chronic wasting disease (CWD) is a transmissible spongiform encephalopathy that infects members of the Cervidae family¹. The disease stems from the misfolding of prion proteins, leading to neurodegeneration, weight

¹Wildlife Health Lab, Cornell University, Ithaca, NY, USA. ²Desert Centered Ecology, LLC, Tucson, AZ, USA. ³Department of Statistics and Data Science, Cornell University, Ithaca, NY, USA. ⁴Minnesota Department of Natural Resources, Nongame Wildlife Program, Saint Paul, MN, USA. ⁵Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI, USA. ⁶Center for Animal Disease Modelling and Surveillance, University of California, Davis, CA, USA. ⁷Arkansas Game and Fish Commission, Little Rock, AR, USA. ⁸Florida Fish and Wildlife Conservation Commission, Gainesville, FL, USA. ⁹Georgia Department of Natural Resources, Social Circle, GA, USA. ¹⁰Iowa Department of Natural Resources, Ames, IA, USA. ¹¹Indiana Department of Natural Resources, Bloomington, IN, USA. ¹²Kentucky Department of Fish and Wildlife Resources, Frankfort, KY, USA. ¹³Maryland Department of Natural Resources, Annapolis, MD, USA. ¹⁴Michigan Department of Natural Resources, Grand Rapids, MI, USA. ¹⁵Minnesota Department of Natural Resources, Wildlife Health Program, Forest Lake, MN, USA. ¹⁶Mississippi Department of Wildlife, Fisheries, and Parks, Jackson, MS, USA. ¹⁷New York State Department of Environmental Conservation, Delmar, NY, USA. ¹⁸North Carolina Wildlife Resources Commission, Raleigh, NC, USA. ¹⁹Ohio Department of Natural Resources, Athens, OH, USA. ²⁰University of Tennessee, Nashville, TN, USA. ²¹Wisconsin Department of Natural Resources, Madison, WI, USA. ²²Present address: Texas A & M Transportation Institute, Austin, TX, USA. ²³Present address: U.S. Fish and Wildlife Service, Tucson, AZ, USA. ✉email: sohelcu06@gmail.com

loss, altered behavior, and eventual death². Since first detected in the 1960s, CWD continues to spread through wild and captive cervids across North America³. To date, 34 United States (US) state wildlife agencies and four Canadian provincial wildlife agencies have detected CWD in at least one wild cervid herd³.

Wildlife agencies in North America have established surveillance programs to detect CWD in wild cervid populations⁴. Such programs focus on identifying locations most likely to harbor CWD and provide the best opportunity to manage the disease while prevalence is low⁵; however, these programs constitute an enormous monetary and human resource cost to agencies⁶. Accordingly, post hoc evaluation of existing surveillance data has focused on pinpointing variables in association with the emergence and spread of CWD to further inform the next year of surveillance⁷.

Anthropogenic factors such as transport and captivity^{5,8,9} of cervids and natural movements⁸ of cervids can contribute to initial introduction of CWD. Persistence of prions in the environment¹⁰, soil types¹¹, baiting and feeding¹², forest cover¹³, water¹⁴, cervid density¹⁵, and natural movements⁸ contribute to disease spread. Authority for non-imperiled terrestrial wildlife, including most deer species, resides with state and provincial governments^{16,17}; as a result, management and surveillance efforts for CWD are highly variable between jurisdictions.

Important and complex questions are driving rapid development, refinement, and use of technology in ecology^{18,19}. Among these technologies are machine learning (ML) techniques²⁰, which are already revolutionizing analyses in wildlife conservation^{21,22}. For example, deep learning has used wildlife imagery to propel detection, inventory, and classification of animals²³. Full implementation of ML technologies into wildlife science, however, is slowed by our limited ability to rapidly generate high-resolution and standardized data across complex ecologies²⁴. Nevertheless, ML is a promising tool for detecting or tracking diseases^{25,26}.

A branch of ML is classification, where the goal is to appropriately sort phenomena into categories. Well known classifiers include random forest (RF), decision tree (DT), gradient boosting (GB), and light gradient boosting (LGB) algorithms. A RF is an ensemble of decision trees, where each tree classifies the phenomenon, then votes on the final classification²⁷. A DT uses decision rules to divide data further and further into ultimate classifications²⁸. The GB is another tree-based ensemble classifier that uses a gradient descent optimization much like binary regression problems²⁹. Finally, the LGB functions like GB but with faster computing and improved accuracy³⁰.

Statisticians compare ML classifiers using a host of performance summaries. A confusion matrix illustrates the distribution of true negatives (TN), true positives (TP), false negatives (FN), and false positives (FP). Subsequent metrics to assess the performance of ML classifiers use information from the confusion matrix, including accuracy, sensitivity, specificity, precision, recall, F1-score, the receiver-operating-characteristic-area-under-curve (ROC), and area-under-the-curve (AUC)^{31,32}.

Our goal was to apply ML classifiers to a regional CWD surveillance dataset to develop a novel model that predicts CWD incidence in wild white-tailed deer (*Odocoileus virginianus*) in counties of 16 states in the midwestern and eastern US. Our objectives were to (1) fit ML classifiers to historical surveillance data, (2) use performance metrics to identify the best classifier, (3) assess which cofactors contribute to the prediction of CWD-status at the county level, and (4) program a user-friendly website application containing the predictive model.

Results

The Pooled Dataset consisted of 31,636 combinations of counties (1438) and season-years (22), spanning over two decades (1 July 2000–30 June 2022). The Pooled Dataset included variables depicting disease introduction risk (*Cervid_facilities*, *Taxidermists*, *Processors*, *Captive_status*), regulations surrounding disease introduction risk (*Breeding_facilities*, *Hunting_enclosures*, *Interstate_import_of_live_cervids*, *Intrastate_movement_of_live_cervids*, *Whole_carcass_importation*), disease establishment risk (*Buck_harvest*, *Doe_harvest*, *Total_harvest*), environmental variables (*Latitude*, *Longitude*, *Area*, *Forest_cover*, *Clay_percent*, *Streams*, *Stream_Length*), diagnostic tallies (*Tests_positive*, *Tests_negative*), and regulations surrounding both introduction and establishment risk (*Baiting*, *Feeding*, *Urine_lures*). Details of each variable appear in the data readme³³. Of the 31,636 records, 1.98% (626/31,636) depicted counties with at least one case of CWD in deer (CWD-positive) and 98.02% (31,010/31,636) depicted counties where CWD had not been detected (CWD-non detect).

The Orthogonal Dataset consisted of 1438 combinations of counties (1438) and season-years (1) spanning the time period from 1 July 2019–30 June 2020. The Orthogonal Dataset included variables depicting disease introduction risk (*Cervid_Facilities*, *Captive_status*), regulations surrounding introduction risk (*Hunting_enclosures*, *Whole_carcass_importation*), disease establishment risk (*Total_harvest*), environmental variables (*Forest_cover*, *Clay_percent*, *Streams*), and regulations surrounding both introduction risk and establishment risk (*Baiting*, *Feeding*, *Urine_lures*). Details of each variable appear in Table 1. Of the 1,438 records, 5.91% (85/1438) depicted CWD-positive counties, 94.09% (1353/1438) depicted CWD-non detect counties (Fig. 1).

The Balanced Orthogonal Dataset consisted of a subset of 158 counties depicting conditions in the 2019–20 season-year. Of the 158 counties, 50% (79/158) represented CWD-positive counties and 50% (79/158) represented randomly selected CWD-non detect counties. All counties in the Balanced Orthogonal Dataset contained values for hunter harvest (although that value could have been zero). [Note that of the 85 total positive counties available in the Orthogonal Dataset, six counties in the US state of Mississippi were excluded from the Balanced Dataset due to missing *Total_harvest* values.] The Training Dataset consisted of 126 (80%) records randomly selected from the Balanced Orthogonal Dataset while the Testing Dataset consisted of the remaining 32 (20%) records of the Balanced Orthogonal Dataset. Summary statistics for each variable in the Pooled, Orthogonal, Balanced Orthogonal, Training, and Testing Datasets are provided in the Supplement.

The Balanced Orthogonal Dataset set contained non-linear data and outliers, so we analyzed the Training and Testing Datasets using four supervised ML algorithms: Random Forest (RF), Decision Tree (DT), Gradient

Details	Variable name	Definition
Positive Number $0 \leq x \leq \infty$	<i>Cervid_Facilities</i>	Number of privately owned premises in the management area that maintain cervids in captivity for the purposes of breeding, farming, display, and/or enclosed (fenced) hunting operations known to the wildlife agency in the given season-year or in the most recent prior season-year for which the number is known ³³
	<i>Total_harvest</i>	Number of wild white-tailed deer harvested by hunters in the management area in the season-year or in the most recent prior season-year for which the number is known and reported to the wildlife agency ³³
	<i>Streams</i>	Average distance in meters from any location within a management area to the nearest high order stream (Strahler order 4 or greater) as defined in the National Hydrography Dataset (NHD) Plus; derived by intersecting the NHD Plus and USCB Cartographic Boundary File ³³
Proportion $0 \leq x \leq 1$ Where 0 indicates no forest, and 1 indicates all forest	<i>Forest_cover</i>	Proportion of the area (including inland waters) classified as forest in the North American Land Change Monitoring System ³³ . Land cover classes included as “forest” include: class 1 (temperate or sub-polar needleleaf forest), class 2 (sub-polar taiga needleleaf forest), class 3 (tropical or sub-tropical broadleaf evergreen forest), class 4 (tropical or sub-tropical broadleaf deciduous forest), class 5 (temperate or sub-polar broadleaf deciduous forest), class 6 (mixed forest), class 7 (tropical or sub-tropical shrubland), class 8 (temperate or sub-polar shrubland), and class 14 (wetland) ³³
Percentage $0 \leq x \leq 100$ Where 0 indicates no clay, and 100 indicates all clay	<i>Clay_percent</i>	Average percentage of clay in the upper 50 cm of soil, as calculated using the clay content at standard depths ³³
Ordinal $x \in \{0, 0.5, 1\}$ 0: Governing officials restrict this activity 0.5: Governing officials partially restrict this activity 1: Governing officials allow this activity	<i>Hunting_enclosures</i>	Indicates whether captive cervid hunting is allowed in the administrative area ³³
	<i>Feeding</i>	Indicates whether feeding of cervids, defined as the placement of substances, including grains, minerals, hay, or other food materials, used to attract cervids for non-hunting purposes is allowed in the administrative area ³³
	<i>Whole_carcass_importation</i>	Indicates whether the administrative area allows importation of whole cervid carcasses from other administrative areas ³³
	<i>Baiting</i>	Indicates whether baiting of cervids, defined as the placement of substances, including grains, minerals, hay, or other food materials, used to attract cervids for the purpose of hunting is allowed in the administrative area ³³
	<i>Urine_lures</i>	Indicates whether the use of urine lures, defined as natural or synthetic urine-based attractants for hunting purposes, is allowed in the administrative area ³³
Binary $x \in \{0, 1\}$ 0: CWD has never been confirmed 1: CWD has been confirmed	<i>Captive_status</i>	Indicates whether CWD has been detected at one or more captive cervid facilities in the management area in the given season-year or in any prior season-year and reported to the USDA APHIS ³³

Table 1. Definitions of variables in the Orthogonal Dataset, borrowed with permission³³. Definitions of variables in the Pooled Dataset are in the Supplement.

Boosting (GB), and Light Gradient Boosting (LGB)^{27–30}. We used k-fold validation to determine the best hyper-parameters and avoid overfitting the model (see Supplement). We found that the LGB was the best model among those evaluated due to its balance between training and testing performance across multiple validations. While RF and GB initially appeared strong due to high training performance and testing performance, the overfitting concern diminished their appeal when compared to LGB, which demonstrated a more balanced performance and superior capacity for generalization. Light Gradient Boosting achieved correct classification of CWD-positive counties in 71.88% of the records in the Testing Dataset (with the highest average accuracy across the fivefold validation of 76.25%; see Supplement). Similarly, the LGB achieved a F1-score of 68.75%, precision of 73.33%, recall of 64.71%, and ROC of 78.82%, implying strong consistency, specificity, sensitivity, and discriminative power (see Supplement). Due to its superior performance in the cross validation, we deemed the LGB to be the strongest performer in predicting the status of CWD at the county level in the midwestern and eastern US given these data. Further assessment of the LGB model revealed that the most influential variables included in the model for these predictions of CWD (Fig. 2) included regulations surrounding risk of anthropogenic introduction of infectious materials (use of *urine lures* and importation of *whole carcasses*) and natural deer movement to reach water (distance to *streams*; see Supplement).

Investigation of the LGB model revealed good accuracy when we compared *CWD Prediction Web App* predictions to the results of the field-based surveillance from the subsequent year (i.e., the season-year 2020–21). Relative to the CWD-status from on-the-ground surveillance in 2020–21, the *CWD Prediction Web App* predictions contained 75% accuracy, 82% sensitivity, 74% specificity, 29% F1-score, 82% recall, and 78% ROC. The *CWD Prediction Web App* showed 946 TNs, 70 TPs, 15 FNs, and 325 FPs relative to known data from the 2020–21 season-year (Table 2; Fig. 2).

The *CWD Prediction Web App* had 70 TPs for the 2020–21 season-year, 66 of which constituted counties already known to be CWD-positive in white-tailed deer from the 2019–20 surveillance data. The remaining four TPs depicted counties that indeed turned positive in white-tailed deer for the first time in the 2020–21 season-year, just as the model predicted (Dakota county, Minnesota; Shawano, Washington, and Wood counties, Wisconsin). The *CWD Prediction Web App* had 325 FPs relative to surveillance data from the 2020–21 season-year.

The *CWD Prediction Web App* had 946 TNs for the 2020–21 season-year. The *CWD Prediction Web App* had 15 FNs for the 2020–21 season-year, 13 of which were counties the *CWD Prediction Web App* knew were positive from the 2019–20 but incorrectly assigned to be negative in the 2020–21 season-year. The remaining two counties (Wyandot county, Ohio; Lauderdale county, Tennessee) were negative in 2019–20 and detected a positive

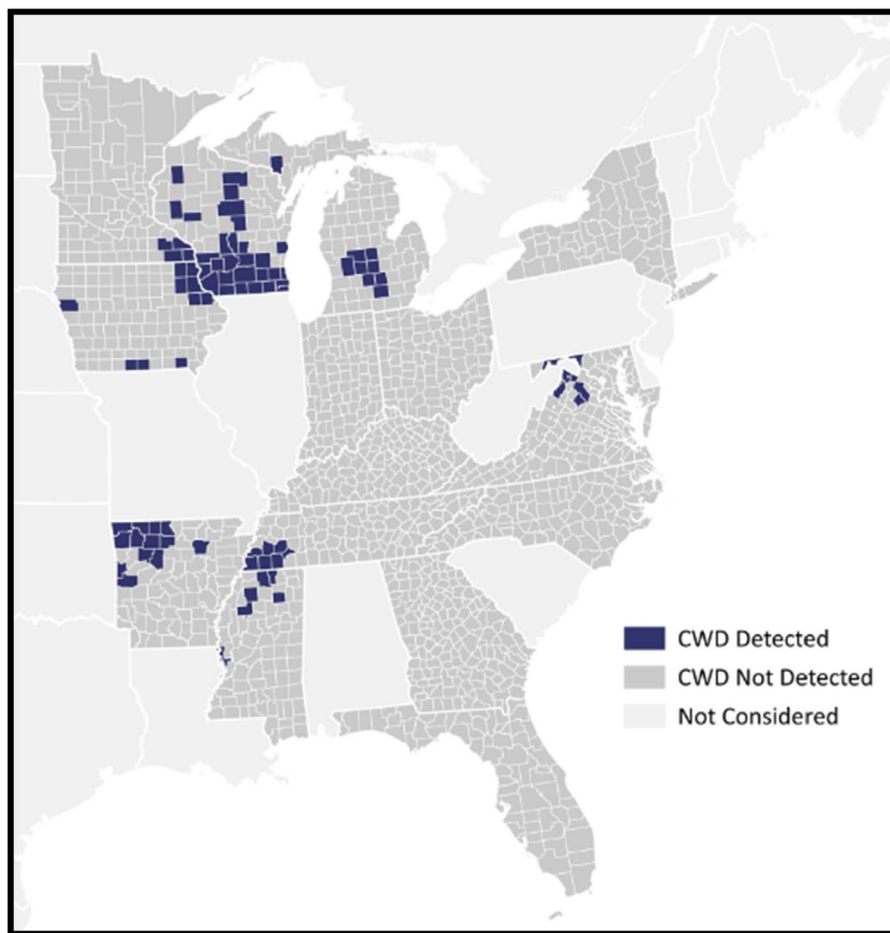


Figure 1. The known status of chronic wasting disease (CWD) in wild white-tailed deer by county in the 2019–20 season according to the results of surveillance testing by US state wildlife agencies³³. CWD Detected represents counties where governing wildlife officials confirmed at least one CWD-positive case in wild, white-tailed deer in the 2019–20 season. CWD Not Detected represents counties where governing wildlife officials conducted CWD testing in 2019–20 in wild, white-tailed deer, but did not confirm CWD in any subject. Not Considered represents counties that did not exist in the Pooled Dataset³³. Map was created in QGIS (version 3.32.2-Lima)⁶⁰.

in 2020–21, but the *CWD Prediction Web App* did not successfully predict that transition in CWD-status. The *CWD Prediction Web App* is at <https://cwd-predict.streamlit.app/>. The code is available at <https://github.com/sohel10/lgbm>.

Discussion

Despite the governing autonomy of management agencies, free-ranging wildlife spans jurisdictional boundaries. Consequently, wildlife agencies across North America would benefit from cooperative efforts designed to understand shared risk factors of disease. Our study was the first to use regional data that represent a single species exposed to diverse management goals, herd dynamics, habitat types, and regulations spanning 16 US states. As well, our cutting-edge application of ML techniques to wildlife health data enabled us to identify counties that contain characteristics similar to counties around the midwestern and eastern US with confirmed CWD.

Our results from the LGB algorithm revealed that regulations have a bearing on the CWD predictions shown in Fig. 2. Indeed, wildlife professionals have long pointed to risk factors for CWD introduction from human-assisted movement of prions via live cervids, carcasses, trophy heads, deer parts, and urine lures^{8,9}. Consequently, wildlife agencies have installed a variety of regulatory measures to limit or extinguish avenues for introduction from anthropogenic sources³⁴. Our results from the LGB algorithm further corroborates prior knowledge that natural movements of deer³⁵ (here specifically to visit water sources) is an important feature driving the predictions of CWD-status. However, we strongly caution that these features and their importances may be phenomena of the data and not absolute. Afterall, the other three candidate algorithms performed similarly with these data (see Table S2 in the Supplement), and their results hinged on entirely different sets and ranks of factor importances. Specifically, the RF algorithm ranked hunter harvest (a proxy for deer density)³⁶, clay-based soils^{37–41}, forest cover¹³, and then distance to streams³⁵ as the most important features driving its predictions of

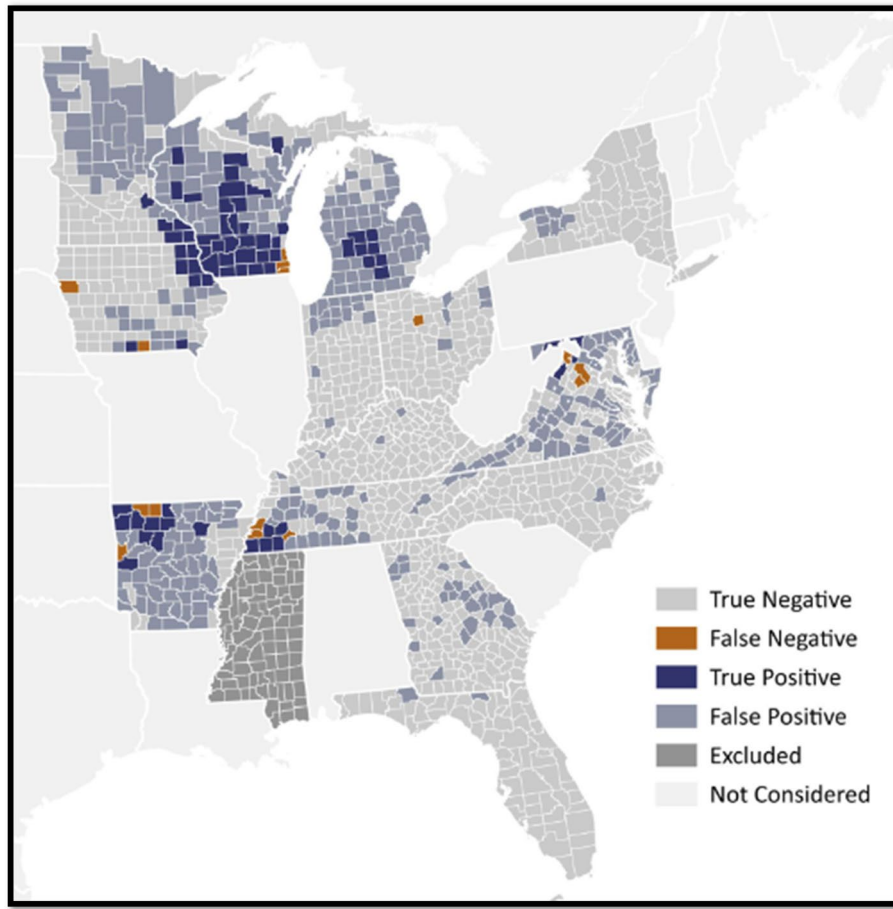


Figure 2. Comparison of chronic wasting disease (CWD) status in free-ranging white-tailed deer in season-year 2020–21 between the *CWD Prediction Web App* and state surveillance data³³. True Negatives (TNs) occurred when the *CWD Prediction Web App* prediction and the surveillance data agreed that CWD-status was CWD-non detect for the county in the season-year 2020–21. True Positives (TPs) occurred when the *CWD Prediction Web App* prediction and the surveillance data agreed that CWD-status was CWD-positive for the county in the season-year 2020–21. False Negatives (FNs) occurred when the *CWD Prediction Web App* predicted CWD-non detect, but the surveillance data declared CWD-positive for the county in season-year 2020–21. False Positives (FPs) occurred when the *CWD Prediction Web App* predicted CWD-positive, but the surveillance data declared CWD-non detect for the county in season-year 2020–21. Excluded represents counties omitted from predictions because harvest data was either not collected or could not be approximated by-county. Not Considered represents areas omitted from the Pooled Dataset³³. Two sources of known error can cause predictions to deviate from reality: (1) model classification error and/or (2) error in CWD-status from surveillance. Specific to Minnesota, a third known error could cause predictions to deviate from reality: (3) error arising from the conversion of harvest data collected in Deer Permit Areas into county-approximations (see the Supplement for specific details). Map was created in QGIS (version 3.32.2-Lima)⁶⁰.

		Model predictions	
		CWD-non detect	CWD-positive
On-the-ground surveillance	CWD-non detect	True negative 70% (946/1,356)	False positive 24% (325/1,356)
	CWD-positive	False negative 1% (15/1,356)	True positive 5% (70/1,356)

Table 2. The confusion matrix of the best Light Gradient Boosting (LGB) model when *CWD Prediction Web App* predictions were compared against on-the-ground surveillance in white-tailed deer in the season-year 2020–21.

CWD-status, in that order. The DT algorithm ranked hunter harvest³⁶, distance to streams³⁵, clay-based soils^{37–41} and then forest cover¹³ as the top features of importance driving its predictions of CWD-status, in that order. And finally, the GB algorithm ranked hunter harvest³⁶, distance to streams³⁵, forest cover¹³, and then clay-based soils^{37–41} as the top features driving its predictions of CWD-status, in that order. With every algorithm, there is some way to corroborate the importances using prior research. These seemingly similar results beg the question: if accuracy was similar across LGB, RF, DT, and GB algorithms, then how did we pick the LGB algorithm to present in Fig. 2? The answer lies in the underlying mathematics: we recognized that we do not yet have enough data for the obvious superior predictor to emerge, so we chose the predictor with the highest current *average accuracy* (even when other algorithms outperformed LGB by random chance in any given singular instance). As well, the LGB demonstrated a more optimal balance between training and testing accuracy than the RF, DT, and GB options. As more data are incorporated into the future fitting of these ML models (see additional discussion below), performance *averages* will settle into the asymptotic means according to the Law of Large Numbers⁴², and any of these four algorithms (with their corresponding feature importances and ranks) could emerge as the superior predictor of CWD-status.

Factor importances from the LGB, RF, DT, and GB algorithms arose from a spatially diverse dataset, and therefore, results offer additional insights relative to those obtained using more localized data. However, these factors emerged as important to these algorithms only out of the factors assessed, and other factors that may be intrinsic pathobiological properties of CWD were omitted from this study. For example, the Pooled Dataset³³ did not contain data on potentially relevant drivers of CWD-status like weather, prion strain, diagnostic test type, deer genetics, explicit dispersal of deer³⁵, management strategies⁸, existence of sympatric susceptible species⁴³, illegal activities such as unapproved movement or release of captive cervids from CWD-positive herds^{44,45}, or geographical proximity to infections in neighboring areas⁵.

Results from the LGB algorithm applied to the Pooled Dataset³³ revealed that regulations matter in predicting CWD-status. However, the two specific regulations pinpointed by the LGB model (*urine lures* and *whole carcass import*) are confounded with the other regulations that we removed due to high correlation (*breeding facilities*, *interstate import of live cervids*, *intrastate movement of live cervids*). Due to covarying regulations (and our selection procedure regarding the variable to remove and the variable to retain, see methods), rather than taking variable names at face value, we recommend interpreting the importance of urine lures and whole carcasses regulations as proxy for regulations depicting *general human activities that could introduce contamination into the reservoir*.

There are numerous potential improvements to this model. The North American Model of Wildlife Conservation recognizes science as the appropriate tool for directing wildlife resource management⁴⁶; however, it is within the purview of state wildlife agencies to determine the scientific methods that best meet their needs^{16,17}. Thus, the first challenge of this work was to find the spatial unit that was the ‘least common denominator’ across all states. Because many agencies represented in the Pooled Dataset³³ recorded county in their CWD testing (surveillance) data (and ancillary spatial data was collected at a unit such that we could confidently infer county from their reported locations), we elected to conduct our analysis at the county-scale. However, we acknowledge county may not be ecologically relevant to either the biology of cervid herds or the spatial unit of interest to wildlife managers. In addition, our selection of county presented problems for predictions in Minnesota (see discussion below). Nevertheless, there were several advantages to using county. First, the decision enabled us to leverage the power in the largest set of existing CWD surveillance data to create the first-ever regional model depicting predictions of CWD-status in North America. Second, the decision enabled us to compare CWD-status across myriad local configurations (i.e., management and policies) to pinpoint potential intrinsic properties of CWD. While there remains work to pinpoint the best algorithm for predicting CWD-status in North America, our results thus far suggest that regulations, hunter harvest (as a proxy for deer density), and habitat variables (forest, clay, and distance to streams) may play a role in CWD-status regardless of local management decisions and policies. Finally, county is the scale of interest to public health departments⁴⁷ who share interest in tracking CWD in wild herds. The ML method requires a single year of pooled data to train the model and the next year of pooled data to assess predictions. Accordingly, if other scales are of interest in surveillance planning, we suggest that agencies coordinate to collect information at the scale of interest for two consecutive years.

Disagreements in CWD-status between the *CWD Prediction Web App* predictions and surveillance data of the 2020–21 season-year are explainable for all participating states in one of two ways: (Case 1) the *CWD Prediction Web App* predicted CWD-positive, the surveillance data reported CWD-non detect, and CWD truly did not exist in white-tailed deer in that county (and therefore the error was on the part of the model) and (Case 2) the *CWD Prediction Web App* predicted CWD-positive, the surveillance data reported CWD-non detect, but CWD truly existed in white-tailed deer in that county (and therefore the error was on the part of surveillance data). Disagreements specific to Minnesota are explainable in a third known way: (Case 3) the *CWD Prediction Web App* predicted CWD-positive or CWD-non detect status for each county in Minnesota using harvest estimates that themselves deviated from reality. [Despite the lack of information to confidently convert harvest data across spatial scales in Minnesota, proportional allocation was used³³ to make county-based approximations of harvest from harvest tallies by Deer Permit Areas (DPAs). Sensitivity analysis of *CWD Prediction Web App* predictions relative to alterations in harvest revealed vulnerabilities in binary predictions. Specifically, 100% (52/52) of the predicted CWD-non detect counties and 94.3% (33/35) of the predicted CWD-positive counties in Minnesota hinged on the value of harvest obtained through the county-approximation. There is no way to know if or to what extent county approximations differ from reality. Nevertheless, the Supplement contains the county approximation value of hunter harvest used in predictions as well as the bifurcation point differentiating a CWD-positive prediction from a CWD-non detect prediction for each county in Minnesota.] Error reduction in (Case 1) is attainable by rerunning the model for a single season-year containing all the counties herein plus counties from additional states that have both CWD-positive and CWD-non detect herds (the model cannot be

improved by adding additional years of data from counties in states already depicted and cannot be improved by adding counties from new states that do not have CWD). Error reduction in (Case 2) is attainable by ensuring that sufficient samples are taken in each county to be 95% confident that CWD-non detect counties in the data are indeed free-from-disease⁴⁸. Error reduction in (Case 3) case is attainable by pooling regional records with outright comparable units (or spatial scales) or using only records containing sufficient information for one-to-one transformations between units (or spatial scales).

Despite a large dataset and powerful modeling tools, the data underlying the *CWD Prediction Web App* are wrought with statistical and ecological complications. For instance, the Pooled Dataset³³ reported presence and absence of CWD in a county directly from sample testing data, but did not account for sampling effort, latent introduction time, deer population growth rates, disease transmission, or detection probability⁴⁹. While the Pooled Dataset³³ constituted the best available regional information regarding CWD-status by county/season-year, we acknowledge that counties deemed to be CWD-free may consist of too few samples to support such a declaration. Should this analysis be repeated with more agency partners, which we recommend, we suggest using data from counties for which there were sufficient samples taken to ensure statistical confidence in the CWD-status. As well, there exists standardized diagnostics for CWD in captive cervid herds⁵⁰, but similar standards do not exist for wild cervids and CWD designation is made by state wildlife authorities. We further suggest the adoption of standardized terminology and definitions surrounding all CWD topics to facilitate comparability of data in future regional studies.

The *CWD Prediction Web App* constitutes an important new tool for CWD surveillance planning, especially when managers overseeing vast areas do not know where to begin testing for the disease. However, we caution the use of the *CWD Prediction Web App* in three ways. First, it might be tempting to use this tool to predict CWD-status in geographical areas smaller than counties, such as Game Management Units. We do not recommend this use until the model underlying the *CWD Prediction Web App* is validated using a known dataset containing true positives and negatives at this geographical scale. Instead, we currently recommend using the Habitat Risk model⁵¹ for such analyses, should the surveillance data in the area of interest have exact geographical locations. Second, due in part to our findings regarding FNs, the *Web App* should not be used in isolation to determine a sampling strategy nor to replace the collection and testing of tissues conducted by agencies each year in the field. And third, due to our findings of similar predictive performance yet differing feature importances among the four ML algorithms, we do not recommend interpreting the LGB feature importances as absolute truth in CWD-predictions.

The Pooled Dataset³³ did not contain data on distance to infection, yet the regional map revealed that many predictions of CWD-positive status are largely contiguous to known infections (Fig. 2). While agencies may already be searching for CWD in areas contiguous to core infections, the *CWD Prediction Web App* may be particularly helpful in illuminating counties vulnerable to CWD in non-obvious places. In noncontiguous counties predicted by the *CWD Prediction Web App* to be CWD-positive, we suggest using the *CWD Prediction Web App* in conjunction with other models that pinpoint conditions for in situ outbreaks^{7,51,52} for surveillance planning. In addition to the error reductions recommended above, we recommend that future ML models better characterize the spread of disease across the landscape by incorporating geographical proximity data or information from diffusion models⁵³ which we did not do.

Conclusion

The *CWD Prediction Web App* produced 325 FPs relative to the subsequent season-year of surveillance. Ostensibly, this may appear to be too much inaccuracy. However, these FPs are quite helpful in understanding regional patterns and vulnerabilities to change in CWD-status. Specifically, the preponderance of FPs signals the counties that warrant increased CWD surveillance in upcoming years, as they share conditions with counties around the region known to harbor CWD. Alternatively, the *CWD Prediction Web App* should not be used in isolation for surveillance planning because it produced 15 FNs relative to the subsequent season-year of surveillance data. Hence, we recommend using the *CWD Prediction Web App* in conjunction with other models to ensure surveillance does not miss introduction in assumed 'low-risk' counties. Indeed, a true measure of the accuracy of the *CWD Prediction Web App* will emerge as predictions are followed through time.

This research simultaneously demonstrates the opportunity and limitations of integrating ML into disease surveillance planning. While the first of its kind to rely on such a large initial dataset (31,636 records), by the time we transformed these data for use in the ML algorithms, usable records had diminished to 'small data'⁵⁴ (158 records). Despite this limitation, we illustrated that it is still possible to build a predictive ML system to predict CWD occurrence across a vast geographical region. We recommend iterative improvements to this model through the inclusion of additional data as ML processes are recursive and responsive to added information. Continued enhancement of the *CWD Prediction Web App* via incorporation of additional data will hone predictions, improve surveillance, and reduce costs for all.

Methods

We used CWD surveillance and ancillary data from the midwestern and eastern US³³. Here we refer to this data as the Pooled Dataset. The Pooled Dataset contains multivariate records in white-tailed deer from the US states of Arkansas, Florida, Georgia, Indiana, Iowa, Kentucky, Maryland, Michigan, Minnesota, Mississippi, New York, North Carolina, Ohio, Tennessee, Virginia, and Wisconsin, and spans the season-years 2000–01 to 2021–22³³. Definitions for each variable appear in the data documentation³³. Minnesota collected harvest data at the Deer Permit Area (DPA) spatial scale, so proportional allocation was used to convert their recorded harvest data into county-scale approximations³³.

We checked all variable pairs for multicollinearity and high correlation, then removed one of the offending variables with correlation exceeding 0.7⁵⁵. When applicable, we weighed which variable to remove based on the total number of missing values or if one variable had a higher difficulty of collection in on-the-ground efforts. We removed linearly inseparable data⁵⁶ by retaining only records for the 2019–20 season-year. We chose the 2019–20 season-year, because it was the period for which we had complete data for the largest number of unique counties. We called this subset of the Pooled Dataset the Orthogonal Dataset.

We deemed our response variable in the Orthogonal Dataset to be whether or not the source agency reported at least one wild deer to be CWD-positive in the county during the 2019–20 season-year (i.e., the *Management_area_positive* variable). Imbalances in the binary outcomes (1 means the county is CWD-positive and 0 means the county is CWD-non detect) are known to skew predictions and introduce inaccuracies due to insufficient information about the minority class⁵⁷. We therefore checked for an imbalance in the number of CWD-positive and CWD-non detect counties in *Management_area_positive*, and if present, applied resampling techniques for the majority class (CWD-non detect) to balance the number of CWD-positive and CWD-non detect counties. We created the Balanced Orthogonal Dataset by taking all full records of CWD-positive counties and adding them to the same number of randomly selected CWD-non detect counties. We instructed the computer to randomly partition the Balanced Orthogonal Dataset into two subsets: a Training Dataset comprising 80% of the records [regardless of CWD-status] and a Testing Dataset with the remaining 20% of the records.

We built four ML models to predict the binary outcome of CWD in a county⁵⁸. We selected candidate ML algorithm(s) that aligned with the dataset's characteristics. We used the Training Dataset to create a prediction classifier, then the Testing Dataset to assess the model's performance in predicting the presence of CWD. We used k-fold cross-validation accuracy to select the hyperparameters of each model⁵⁶.

We used the sci-kit-learn (version 1.4.2)⁵⁷ to assess the performance of each classifier by considering accuracy, F1-score, precision, recall, and ROC simultaneously. We chose the model that demonstrated the best balance between training and testing data, then used the predictor gain method⁵⁹ to evaluate the importance of variables contained in the model. We generated its confusion matrix relative to the subsequent season-year (2020–21) of surveillance data. We programmed the top model into the *CWD Prediction Web App* to predict CWD-status in each county.

Data availability

The data are publicly available at <https://doi.org/10.7298/7txw-2681.2>. The CWD Prediction Web App is at <https://cwd-predict.streamlit.app/>.

Received: 31 October 2023; Accepted: 15 June 2024

Published online: 22 June 2024

References

- Williams, E. & Young, S. Chronic wasting disease of captive mule deer: a spongiform encephalopathy. *J. Wildl. Dis.* **16**, 89–96 (1980).
- Poggiolini, I., Saverioni, D. & Parchi, P. Prion protein misfolding, strains, and neurotoxicity: An update from studies on mammalian prions. *Int. J. Cell Biol.* **2013**, 24. <https://doi.org/10.1155/2013/910314> (2013).
- United States Geological Survey (USGS). *Distribution of chronic wasting disease in North America*. <https://www.usgs.gov/media/images/distribution-chronic-wasting-disease-north-america-0>. (2024).
- Association of Fish and Wildlife Agencies (AFWA). *Best management practices for surveillance, management, and control of chronic wasting disease*. fishwildlife.org/application/files/1315/7054/8052/AFWA_CWD_BMP_First_Supplement_FINAL.pdf. (Washington, DC, USA, 2018).
- Schuler, K., Hollingshead, N., Kelly, J., Applegate, R., & Yoest, C. *Risk-based surveillance for chronic wasting disease in Tennessee*. Tennessee Wildlife Resources Agency (TWRA) Wildlife Technical Report 18–4 (2018).
- Chiavacci, S. J. The economic costs of chronic wasting disease in the United States. *PLoS One* **17**(12), e0278366. <https://doi.org/10.1371/journal.pone.0278366> (2022).
- Hanley, B. *et al.* Informing surveillance through the characterization of outbreak potential of chronic wasting disease in white-tailed deer. *Ecol. Model.* **471**, 110054. <https://doi.org/10.1016/j.ecolmodel.2022.110054> (2022).
- Miller, M., & Fischer, J. The first five (or more) decades of chronic wasting disease: Lessons for the five decades to come. *Transactions of the North American wildlife and natural resources conference*. Vol. **81** 110–120 (2016).
- Miller, M. & Williams, E. Chronic wasting disease of cervids. *Curr. Top. Microbiol. Immunol.* **284**, 193–214 (2004).
- Miller, M., Williams, E., Hobbs, N. & Wolfe, L. Environmental sources of prions transmission in mule deer. *Emerg. Infect. Dis.* **10**(6), 1003–1006 (2004).
- Johnson, C., Pederson, J., Chappell, R., McKenzie, D. & Aiken, J. Oral transmissibility of prion disease is enhanced by binding to soil particles. *PLoS Pathog.* **3**(7), 0874–0881 (2007).
- New York State Interagency. *CWD Risk Minimization Plan*. https://extapps.dec.ny.gov/docs/wildlife_pdf/cwdpreventionplan2018.pdf (2018).
- Ruiz, M., Kelly, A., Brown, W., Novakofski, J. & Mateus-Pinilla, N. Influence of landscape factors and management decisions on spatial and temporal patterns of the transmission of chronic wasting disease in white-tailed deer. *Geospat. Health* **8**(1), 215–227 (2013).
- Nichols, T. *et al.* Detection of protease-resistant cervid prion protein in water from a CWD-endemic area. *Prion* **3**(3), 171–183 (2009).
- Storm, D. *et al.* Deer density and disease prevalence influence transmission of chronic wasting disease in white-tailed deer. *Ecosphere* **4**, 1–14 (2013).
- Baier, L. E. Federalism, preemption, and the nationalization of American wildlife management: the dynamic balance between state and federal authority. (Rowman & Littlefield, Lanham, Maryland, USA, 2022).
- Hilton, C. D. & Ballard, J. R. Wildlife regulations that affect veterinarians in the United States. in *Fowler's Zoo and Wild Animal Medicine Current Therapy*, Vol. 10. 43–46. <https://doi.org/10.1016/B978-0-323-82852-9.00008-3> (2023).
- Recknagel, F. Applications of machine learning to ecological modelling. *Ecol. Model.* **146**(1–3), 303–310 (2001).
- Allan, B. *et al.* Futurecasting ecological research: The rise of technoecology. *Ecosphere* **9**(5), e02163. <https://doi.org/10.1002/ecs2.2163> (2018).

20. Ahmed, M., Ishikawa, F., & Sugiyama, M. Testing machine learning code using polyhedral region. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 1533–1536 (2020).
21. Dietterich, T. Machine learning in ecosystem informatics and sustainability. In *International Joint Conference on Artificial Intelligence (IJCAI)* (Pasadena, CA, USA, 2009).
22. Tuia, D. *et al.* Perspectives in machine learning for wildlife conservation. *Nat. Commun.* **13**, 792–807 (2022).
23. Nakagawa, S. *et al.* Rapid literature mapping on the recent use of machine learning for wildlife imagery. *Peer Community J.* <https://doi.org/10.24072/pcjournal.261> (2023).
24. Besson, M. *et al.* Towards the fully automated monitoring of ecological communities. *Ecol. Lett.* **25**, 2753–2775 (2022).
25. Robles-Fernández, Á., Santiago-Alarcon, D. & Lira-Noriega, A. Wildlife susceptibility to infectious diseases at global scales. *Proc. Natl. Acad. Sci.* **119**(35), e2122851119. <https://doi.org/10.1073/pnas.2122851119> (2022).
26. Pillai, N., Ramkumar, M. & Nanduri, B. Artificial intelligence models for zoonotic pathogens: A survey. *Microorganisms* **10**(10), 1911–1931 (2022).
27. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
28. Quinlan, J. Induction of decision trees. *Mach. Learn.* **1**(1), 81–106 (1986).
29. Friedman, J. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**(5), 1189–1232 (2001).
30. Ke, G., *et al.* LightGBM: A highly efficient gradient boosting decision tree. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* 3149–3157 (2017).
31. Staartjes, V. & Kernback, J. Foundations of machine learning-based clinical predictions modeling: Part III—model evaluation and other points of significance. In *Machine learning in clinical neuroscience. Acta Neurochirurgica Supplement* Vol. 134 (eds Staartjes, V. *et al.*) (Springer, 2022). https://doi.org/10.1007/978-3-030-85292-4_4.
32. R Statistical Computing (R). R package: Confusion matrix. [https://search.r-project.org/CRAN/refmans/qwraps2/html/confusion_matrix.html#:~:text=sensitivity%20%3D%20TP%20%2F%20\(TP%20%2B,TN%20%2F%20\(TN%20%2B%20FN\)\)](https://search.r-project.org/CRAN/refmans/qwraps2/html/confusion_matrix.html#:~:text=sensitivity%20%3D%20TP%20%2F%20(TP%20%2B,TN%20%2F%20(TN%20%2B%20FN))). (2023).
33. Schuler, K., *et al.* North American wildlife agency CWD testing and ancillary data (2000–2022) [Dataset]. Cornell University Library eCommons Repository (2024). <https://doi.org/10.7298/7txw-2681.2>
34. Michigan Department of Natural Resources (MIDNR). *CWD and Cervidae regulations in North America NEW*. <https://michigan.gov/dnr/managing-resources/wildlife/cwd/hunters/cwd-and-cervidae-regulations-in-north-america>. (2022).
35. Green, M. L., Manjerovic, M. B., Mateus-Pinilla, N. & Novakofski, J. Genetic assignment tests reveal dispersal of white-tailed deer: Implication for chronic wasting disease. *J. Mammal.* **95**(3), 646–654 (2014).
36. Conner, M. M. *et al.* The relationship between harvest management and chronic wasting disease prevalence trends in western mule deer (*Odocoileus Hemionus*) herds. *J. Wildl. Dis.* **57**(4), 831–843 (2021).
37. Kuznetsova, A., McKenzie, D., Baner, P., Siddique, T. & Aiken, J. Potential role of soil properties in the spread of CWD in western Canada. *Prion* **8**(1), 92–99 (2014).
38. Walter, W., Walsh, D., Farnsworth, M., Winkelman, D. & Miller, M. Soil clay content underlies prion infection odds. *Nat. Commun.* **2**, 200–206 (2011).
39. Dorak, S. *et al.* Clay content and pH: Soil characteristic associations with the persistent presence of chronic wasting disease in northern Illinois. *Sci. Rep.* **7**(1), 18062–18072 (2017).
40. Wyckoff, A. *et al.* Clay components in soil dictate environmental stability and bioavailability of cervid prions in mice. *Front. Microbiol.* **7**, 1–11 (2016).
41. Booth, C., Lichtenberg, S., Chappell, R. & Pedersen, J. Chemical inactivation of prions is altered by binding to the soil mineral montmorillonite. *ACS Infect. Dis.* **7**(4), 859–870 (2021).
42. Ibe, O. C. Laws of large numbers. Basic concepts in probability. In *Markov processes for stochastic modeling* 2nd edn (Elsevier, 2013). <https://doi.org/10.1016/C2012-0-06106-6>.
43. Cunningham, C., Peery, R., Dao, A., McKenzie, D. & Coltman, D. Predicting the spread-risk potential of chronic wasting disease to sympatric ungulate species. *Prion* **14**(1), 56–66 (2020).
44. Tidd, J. *Trophy-hunting business owner admits to illegally importing deer to Kansas* (Accessed 10 August 2018); <https://www.kansas.com/sports/outdoors/article198543619.html>
45. Fitzgerald, R. *They smuggled deer to Forrest County, feds say. But that wasn't the only problem* (Accessed 10 August 2018); <https://www.sunherald.com/news/local/crime/article173323226.html>
46. Organ, J., *et al.* *The North American Model of Wildlife Conservation*. The Wildlife Society Technical Review 12–04. The Wildlife Society. (Bethesda, Maryland, USA, 2012).
47. Centers for Disease Control (CDC). *Chronic wasting disease (occurrence)* (Accessed 15 April 2024); <https://www.cdc.gov/prions/cwd/occurrence.html>
48. Booth, J. G. *et al.* Sample size for estimating disease prevalence in free-ranging wildlife: A Bayesian modeling approach. *J. Agric. Biol. Environ. Stat.* <https://doi.org/10.1007/s13253-023-00578-7> (2023).
49. Schuler, K., Jenks, J., Klaver, R., Jennelle, C. & Bowyer, R. Chronic wasting disease detection and mortality sources in a semi-protected deer population. *Wildl. Biol.* **1**, 1–7 (2018).
50. United States Department of Agriculture and Animal and Plant Health Inspection Service Veterinary Services. *Chronic wasting disease program standards*. www.aphis.usda.gov/sites/default/files/cwd-program-standards.pdf (2019).
51. Walter, W. D. *et al.* Predicting the odds of chronic wasting disease with Habitat Risk software. *Spat. Spatio-temporal Epidemiol.* **49**, 100650. <https://doi.org/10.1016/j.sste.2024.100650> (2024).
52. Winter, S. N., Kirchgessner, M. S., Frimpong, E. A. & Escobar, L. E. A landscape epidemiological approach for predicting chronic wasting disease: A case study in Virginia, US. *Front. Vet. Sci.* <https://doi.org/10.3389/fvets.2021.698767> (2021).
53. Hefley, T., Hooten, M., Russell, R., Walsh, D. & Powell, J. When mechanism matters: Bayesian forecasting using models of ecological diffusion. *Ecol. Lett.* **20**(5), 640–650 (2017).
54. Todman, L., Bush, A. & Hood, A. “Small data” for big insights in ecology. *Trends Ecol. Evol.* **38**(7), 615–622 (2023).
55. Albon, C. *Machine learning with Python cookbook* (O'Reilly Media Inc., 2018).
56. Shalev-Shwartz, S. & Ben-David, S. *Understanding machine learning: From theory to algorithms* (Cambridge University Press, 2014).
57. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
58. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: Data mining, inference and prediction* 2nd edn. (Springer, 2009). <https://doi.org/10.1007/978-0-387-84858-7>.
59. Pudjihartono, N., Fadason, T., Kempa-Liehr, A. & O'Sullivan, J. A review of feature selection methods for machine learning-based disease risk prediction. *Front. Bioinform.* **2**, 1–17 (2022).
60. QGIS.org, QGIS Geographic Information System. QGIS Association. <http://www.qgis.org> (2024).

Acknowledgements

We thank W. Kozłowski, J. Peaslee, M. Cosgrove, J. Dennison, T. DeRosia, C. Faller, D. Howlett, M. Keightley, E. Kessel, J. Kessler, E. Larson, M. McCord, A. Nolder, D. Stanfield, S. Stura, J. Sweaney, G. Timko, B. Wallingford, B. Wojcik, C. Yoest, A. J. Riggs, and 5 anonymous individuals, Arkansas Game and Fish Commission Wildlife Management Division staff and F23AF01862-00, Georgia Wildlife Resources Division, Indiana Department

of Natural Resources Division of Fish and Wildlife, Iowa DNR Wildlife Bureau staff, Minnesota DNR staff, Mississippi Department of Wildlife, Fisheries, and Parks staff, New York State Department of Environmental Conservation Wildlife Health Unit staff, the North Carolina Wildlife Resources Commission staff, Ohio DNR Division of Wildlife staff. Data collection was funded in part by Arkansas's Wildlife Restoration funds, 'State Wildlife Health'; Florida's State Game Trust Fund Deer Management Program; Georgia's Wildlife and Sport Fish Restoration Program; Indiana DNR and Fish and Wildlife F18AF00484, W38R05 White-tailed Deer Management, F20AF10029-00, Monitoring Wildlife Populations and Health W-51-R-01, F21AF02467-01, Monitoring Wildlife Populations and Health W-51-R-02; Iowa's Fish and Wildlife Trust Fund and U. S. Fish and Wildlife Service Wildlife and Sport Fish Restoration Program; Maryland's award of the U. S. Fish and Wildlife Service Wildlife and Sport Fish Restoration Program, W 61-R-29; Minnesota DNR; New York's Wildlife Health Unit and New York's award for Federal Aid Wildlife Restoration Grant W-178-R; North Carolina's award for Federal Aid in Wildlife Restoration; Tennessee's award for the Wildlife Restoration Program; Virginia's Pittman-Robertson Federal Aid; Wisconsin's award for the Federal Aid in Wildlife Restoration; Multistate Conservation Grant Program F21AP00722-01. The Michigan Disease Initiative RC109358, Alabama Department of Conservation and Natural Resources, Florida Fish and Wildlife Conservation Commission, Tennessee Wildlife Resources Agency, and New York State Department of Environmental Conservation contributed funding to the overall project. The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the U.S. Fish and Wildlife Service.

Author contributions

Conceptualization: MSA; BJH; KLS. Funding: JGB; JG; BJH; CSJ; KLS. Literature review: BJH; MSA; RCA; KLS. Wrote software: MSA; BJH. Conducted the analysis: MSA; BJH. Provided data and applied agency expertise CRM; JRB; BC; CHK; TMH; JNC; KMBW; EM; CC; LMO; JKT; MC; WTM; CS; KPH; AES; LAM; MC; RTM; JS; MJT; JDK; DMG; DJS. Wrote the draft: BJH; MSA. All authors provided critical edits and gave approval for submission.

Competing interests

MSA, BJH, RCA, JGB, JG, NAH, CGC, CRM, JRB, BC, CHK, TMH, JNC, KMBW, EM, CC, LMO, JKT, MC, WTM, KPH, AES, LAM, CS, MC, RTM, JS, MJT, JDK, DMG, DJS, and KLS declare no competing interests. A portion of this work was completed when CIM was a consultant at Desert Centered Ecology, LLC; CSJ was a consultant at Christopher S. Jennelle; and FHH was a consultant at Florian H. Hodel.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-65002-7>.

Correspondence and requests for materials should be addressed to M.S.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024