Predictive Modeling of H5N1 Bird Flu in United States of America: A 2022-2023 Analysis

Weilin Cheng
Department of Statistics
University of California, Davis
Davis, CA 95616
wncheng@ucdavis.edu

Hengyuan Liu
Department of Statistics
University of California, Davis
Davis, CA 95616
hyliu@ucdavis.edu

Kathy Mo
Department of Statistics
University of California, Davis
Davis, CA 95616
kamo@ucdavis.edu

Sida Tian
Department of Economics
University of Michigan, Ann Arbor
Ann Arbor, MI 48109
startian@umich.edu

Li Yuan
Department of Statistics
University of Michigan, Ann Arbor
Ann Arbor, MI 48109
leeyuan@umich.edu

January 30, 2024

Abstract

This research uniquely focuses on predicting the likelihood of H5N1 outbreaks in the United States at the county level. Unlike previous studies, which either excluded the United States or used outdated data, we utilized diverse statistical techniques and publicly available H5N1-related data from January 2022 to March 2023. Employing logistic regression, regularization methods, cross-validation, and eXtreme Gradient Boosting (XGBoost), our models demonstrated remarkable predictive efficacy. Notably, the XGBoost model, trained with 10-fold cross-validation, outperformed others in terms of ROC-AUC. This research provides valuable epidemiological insights, proposes intervention strategies for H5N1 in the United States, and suggests future research directions.

Keywords H5N1 · avian influenza · risk assessment · classification · regularization · xgboost

1 Introduction

Since its first recognition in the 1880s, avian influenza has significantly influenced human health, as exemplified by its role in approximately 50 million deaths in 1918 (Taubenberger and Morens 2006) and the infection of around 1,500 individuals between 2013 and 2019 (World Health Organization 2018).

This study acknowledges the historical significance of avian influenza and highlights the necessity for current research methodologies to adapt to the virus's evolution. Extensive research exists on models predicting H5N1 outbreaks in regions such as the Middle East, West Africa (Williams and Peterson 2009), Egypt (Kane et al. 2014), and Southeast Asia (Gilbert et al. 2008). However, there is a noticeable scarcity of recent studies within the United States, especially post-2019. Most studies on avian influenza infection prediction in America have relied on older datasets, limiting their current applicability. In the study "Artificial intelligence and avian influenza: Using machine learning to enhance active surveillance for avian influenza viruses," (Walsh et al. 2019) the reliance on wild bird infection data from 2006 to 2011 to train the model raises concerns about its current relevance. Despite the research being conducted in 2019, the outdated nature of the data could limit the model's applicability to contemporary scenarios.

The economic impact of the highly pathogenic avian influenza (HPAI) H5N1 outbreak in the United States has been substantial. Since 2022, the H5N1 virus have been detected across various bird populations, affecting 1,064 counties (Centers for Disease Control and Prevention 2022b). Over 58.7 million poultry in 419 counties and approximately 7,105 wild birds in 1,023 counties were infected (Centers for Disease Control and Prevention

2022b). In 2022, the H5N1 strain led to an estimated economic loss of $2.5 to $3 billion, with around 40 million animal deaths (Farahat et al. 2023). This has brought attention to the United States' high meat consumption, particularly poultry, which was 115 pounds per capita in 2022 (World Animal Foundation 2023). Additionally, the significant rise in egg prices in 2022 (Iacurci 2023) has disrupted American dietary patterns, indicating broader challenges for the environment and poultry industry.

In response to these challenges, this study presents a new model that utilizes recent data (2022 - 2023) from the United States, aiming to fill the gaps in avian influenza surveillance and improve H5N1 outbreak predictions. We will analyze avian influenza cases to forecast future incidents and suggest mitigation strategies. Our methodology encompasses logistic regression, model regularizations, cross-validation, and eXtreme Gradient Boosting (XGBoost), focusing on identifying United States counties at possible risk of H5N1 infection. This identification process will guide the allocation of preventive measures. Our data sources include the Centers for Disease Control and Prevention (CDC), United States Department of Agriculture (USDA), United States Census Bureau, and the Bureau of Labor Statistics.

The primary objective of this study is to develop and evaluate machine learning models that identify areas in the United States at high risk for H5N1 outbreaks, generating predictive insights that can accurately forecast future occurrences in different regions. This approach provides insights to future research on the impact of the H5N1 virus on public health and safety.

2 Data

The formulation of a prognostic framework aimed at discerning counties predisposed to H5N1 infection necessitates a meticulous comprehension of the inherent structure and constituents of the dataset. Our methodological approach entails the amalgamation of four disparate datasets combined to a singular dataset, pertaining to documented instances of H5N1 manifestation within each county during the delineated temporal span from January 2022 to March 2023.

This meticulously curated dataset will serve as the substrate for the training of our classification model, a pivotal constituent in forecasting the prospective counties susceptible to H5N1 infection in the ensuing month. Through an incisive scrutiny of discernible patterns within the dataset, we aim to elucidate salient variables intricately correlated with an escalated vulnerability to infection, encompassing geographical locale, ambient temperature, and the typology of avian populations. This analytical expedition is poised to yield insights instrumental in the formulation of precisely targeted interventions and stratagems within the ambit of public health, thereby assuaging the dissemination of H5N1 in locales characterized by an augmented proclivity for contagion.

2.1 Overview

2.1.1 United States Counties Database

The United States Counties Database, sourced from Pareto Software, LLC (2023), is a comprehensive compilation of data from authoritative entities such as the U.S. Census Bureau and the Bureau of Labor Statistics. It encompasses all 3,143 counties across the United States, including Washington D.C., and provides essential details such as Federal Information Processing Standard (FIPS) codes, geographical coordinates, and more.

Table 1: First Five Observations of US Counties Database

FIPS Code	State	County	Latitude	Longitude
6037	CA	Los Angeles County	34.3209	-118.2247
17031	IL	Cook County	41.8401	-87.8168
48201	TX	Harris County	29.8578	-95.3936
4013	AZ	Maricopa County	33.3490	-112.4915
6073	CA	San Diego County	33.0343	-116.7350

Modifications were applied to this dataset to streamline future analyses. These include the conversion of state names into their respective abbreviations, as shown in Table 1, facilitating a more efficient data handling process. This dataset is instrumental in generating intricate geographical reports of H5N1 cases in each county by correlating with subsequent datasets from the CDC.

2.1.2 H5N1 Bird Flu Detections across the United States (Backyard and Commercial)

The second dataset, procured from Centers for Disease Control and Prevention (2022a), focuses on H5N1 bird flu outbreaks in various types of bird populations across the United States, including commercial poultry facilities, backyard poultry, and hobbyist bird flocks. To enhance its utility for our analyses, we refined this dataset by transforming the date format to separate year, month, and day components and excluded records post-March 31st, 2023, due to incomplete data.

Furthermore, we simplified the classification of flock types into "Poultry" and "Non-Poultry", as per the definitions provided by World Organisation for Animal Health (2022). This categorization assimilates 15 distinct commercial flock types under the "Poultry" umbrella, aligning with WOAH's definition. The dataset, post-modification, consists of 817 observations across 6 variables, forming a foundational component of our subsequent analyses.

Table 2: Initial and Final Five H5N1 Backyard and Commercial Outbreaks in the United States until March 31st, 2023

State	County	Year_Month	Day	Type	Cases
Indiana	Dubois County	2022_02	08	Poultry	29000
Virginia	Fauquier County	2022_02	12	Non-Poultry	90
Kentucky	Fulton County	2022_02	12	Poultry	231400
Kentucky	Webster County	2022_02	15	Poultry	53300
Indiana	Dubois County	2022_02	16	Poultry	26600
⋮	⋮	⋮	⋮	⋮	⋮
Michigan	Lapeer County	2023_03	23	Poultry	950
Colorado	Arapahoe County	2023_03	24	Non-Poultry	10
Kansas	Ellsworth County	2023_03	24	Non-Poultry	50
Colorado	Yuma County	2023_03	28	Poultry	310
Oregon	Umatilla County	2023_03	30	Non-Poultry	50

Table 2 delineates the initial and final five instances of H5N1 outbreaks in backyard and commercial settings across the United States, up to March 31st, 2023. The data reveals that the inaugural outbreak occurred on February 8th, 2022, in Dubois, Indiana, with a significant impact of 29,000 cases, classified under the category "Poultry". Conversely, the most recent outbreak recorded on March 30th, 2023, in Umatilla, Oregon, involved a comparatively smaller scale of 50 cases, categorized as "Non-Poultry".

2.1.3 H5N1 Bird Flu Detections across the United States (Wild Birds)

The dataset on the detection of highly pathogenic avian influenza (HPAI) A(H5) viruses in wild birds across the United States, also provided by Centers for Disease Control and Prevention (2022b), has been meticulously revised to enhance its utility for our analysis. Similar to the modifications applied to the dataset discussed in Section 2.1.2, we have reformatted the date entries and omitted records post-March 31st, 2023. Additionally, standardization of column names has been implemented to facilitate seamless data cleaning and analysis. Post-modification, this dataset encompasses 2,752 observations across 6 variables, representing a crucial resource for our future analytical endeavors.

Table 3: Initial and Final Five H5N1 Wild Bird Outbreaks in the United States until March 31st, 2023

State	County	Year_Month	Day	Type	Cases
North Carolina	Hyde County	2022_01	12	Wild bird	2
South Carolina	Colleton County	2022_01	13	Wild bird	2
North Carolina	Hyde County	2022_01	16	Wild bird	2
North Carolina	Hyde County	2022_01	20	Wild bird	3
North Carolina	Pamlico County	2022_01	20	Wild bird	34
⋮	⋮	⋮	⋮	⋮	⋮
Alaska	Sitka County	2023_03	31	Wild bird	1
Maryland	Harford County	2023_03	31	Wild bird	1

State	County	Year_Month	Day	Type	Cases
Minnesota	Wright County	2023_03	31	Captive wild bird	1
Utah	Millard County	2023_03	31	Wild bird	1
Washington	Benton County	2023_03	31	Wild bird	1

It is imperative to note the classification of outbreak types in this context: "Wild bird" refers to birds exhibiting natural phenotypes, living independently without human supervision, and "Captive wild bird" denotes birds with minimal human-induced phenotypic changes yet requiring human supervision or control (World Organisation for Animal Health 2022).

Table 3 provides a chronological overview of the first and last five H5N1 wild bird outbreaks in the United States, up to March 31st, 2023. Analysis reveals that the initial outbreak was recorded on January 12th, 2022, in Hyde, North Carolina, involving two cases, identified as "Wild bird" in type. The most recent outbreak, as of March 31st, 2023, occurred in Benton, Washington, with a single case, also classified under the "Wild bird" category.

2.1.4 Compilation and Refinement of the Monthly Average Temperature Dataset

This dataset, an amalgamation of data from National Centers for Environmental Information (2023) and Cedar Lake Ventures, Inc (2023), provides the monthly average temperature in Fahrenheit degrees (°F) for U.S. counties. The NCEI dataset covers all counties except those in Hawaii from January 2022 to March 2023, while the Cedar dataset specifically addresses the average temperature for Hawaii's five counties during the same period.

To optimize the dataset for our analysis, we have standardized the formats and column names for state, county, and month. Furthermore, we have rectified discrepancies in county names based on the dataset outlined in Section 2.1.1. Notably, due to the unavailability of Hawaii's average temperature data in offline formats, we have manually inputted these values. For analytical efficiency, the months from January 2022 to March 2023 have been converted to a month index ranging from 1 to 15 respectively.

Table 4: First Five Recorded Observations of County-Level Average Temperatures in Fahrenheit Across the United States

State	County	Month Index	Average Temperature
AL	autauga county	1	45.1
AL	baldwin county	1	50.1
AL	barbour county	1	45.4
AL	bibb county	1	43.2
AL	blount county	1	41.6

Post-modification, the dataset comprises 47,160 observations across 4 variables, proving to be a critical component for our analysis and predictive modeling. Table 4 presents the initial five observations from this dataset.

2.1.5 Curated Dataset of Monthly H5N1 Cases by County

The amalgamation of data sets expounded upon in sections 2.1.1 through 2.1.4 is intricately executed through the fusion of state and country nomenclature. This methodological approach is necessitated by the nuanced geographical landscape, where certain counties bear identical names yet are dispersed across distinct states. The resultant refined dataset, thereby achieved, stands as the foundational wellspring for subsequent analytical endeavors. Comprising an expansive repository of 188,580 observations spanning 10 variables as shown in Table 5, this dataset delineates the monthly distribution of H5N1 cases by county within the United States, covering the temporal expanse from January 2022 to March 2023.

Table 5: Descriptive Summary of Variables in the Curated Dataset

Variable	Description
`fips`	The FIPS code, serving as a unique identifier for each county within the United States, manifests as a categorical variable. It embraces a total of 3,143 distinctive values, each with a frequency of 60 entries, thereby elucidating the spatial granularity inherent in the dataset.
`state`	The `state` variable, denoting the abbreviated nomenclature for each state in the United States, constitutes a categorical entity with 51 unique values. Significantly, the individual states exhibit variability in the number of counties, emphasizing the hierarchical structure of administrative divisions.
`county`	Representing the appellations of counties, independent cities, census areas, and administrative counterparts at an equivalent level within the United States, the county variable emerges as a categorical feature encompassing 3,143 distinct values.
`lat`	The latitude of each county.
`lng`	The longitude of each county.
`month.index`	The `month.index` parameter serves as an ordinal indicator denoting the chronological sequence of avian influenza outbreaks, ranging from 1 (January 2022) to 15 (March 2023). Noteworthy is the consistent count of 12,572 entries per month.index, underscoring the inclusivity of all counties irrespective of their H5N1 case counts, with zero cases explicitly recorded.
`type`	The `type` variable, encapsulating the nature of outbreaks in specific counties and months, assumes categorical form with four distinct values: `poultry`, `non-poultry`, `wild bird`, and `captive wild bird`. This classification, comprising 47,145 entries for each type, underscores the exhaustive nature of the cleaned dataset, wherein all counties' H5N1 situations are encompassed, regardless of case counts, with zero cases explicitly represented.
`avg.temp`	The `avg.temp` variable signifies the average temperature in a specific county and month, quantified in Fahrenheit degrees, thereby integrating meteorological factors into the analytic framework.
`cases`	Reflecting the number of H5N1 cases detected in a given county and month, the cases variable serves as a cardinal metric, capturing the epidemiological dimension of the dataset.
`binary.case`	The `binary.case` variable, operating as a binary indicator, assumes the value of `uninfected` (185,907 entries) when the case count for a specific type of outbreak in a county and month is 0. Conversely, it adopts the designation `infected` (2,673 entries) when cases are detected, thereby providing a dichotomous perspective on the outbreak status within the dataset.

2.2 Visualization

In this section, we embark on a visual exploration of the meticulously curated dataset detailed in the preceding section. Visualization serves as a pivotal instrument in our analytical arsenal, offering an intuitive and insightful comprehension of complex patterns, trends, and relationships latent within the expansive dataset.

Figure 1 illustrates a progressive increase in new H5N1 virus cases over time, with the majority of these cases predominantly occurring in the western and midwestern regions of the United States. Notably, there was a noticeable fluctuation in new cases during March, April, May, and from August to December 2022. The color coding in the visualization confirms that new cases did not surpass 1,000,000 in March 2023, and there has been a discernible decline in monthly cases since the onset of 2023.

Figure 1.1: January 2022

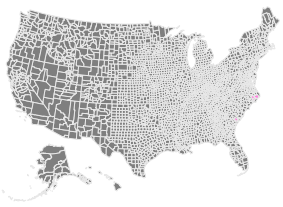


Figure 1.2: February 2022

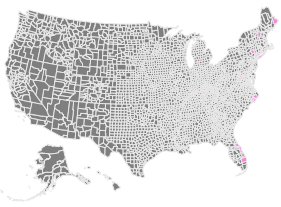


Figure 1.3: March 2022

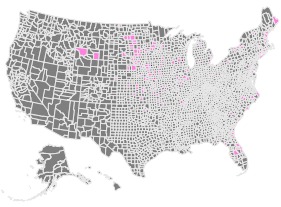
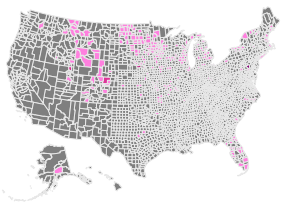


Figure 1.4: April 2022



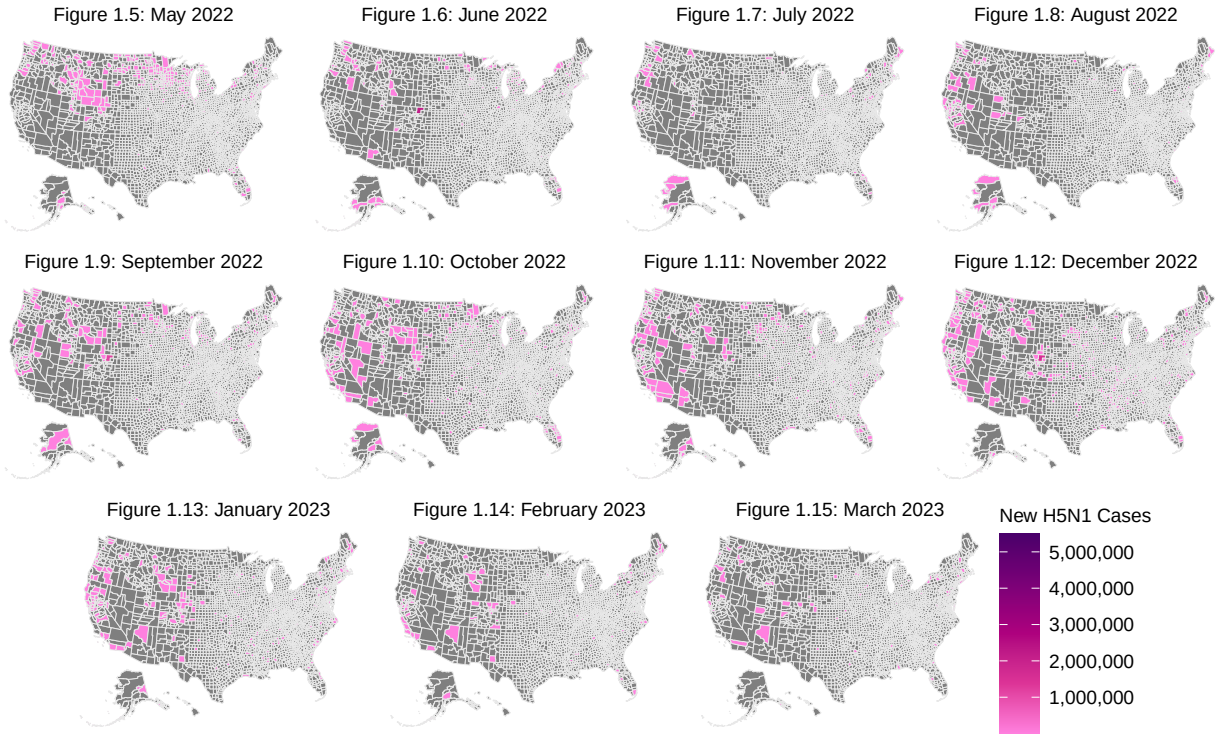


Figure 1: Monthly Distribution of New H5N1 Cases by County in the United States from January 2022 until March 2023

Table 6: Top and Bottom Five Monthly Case Counts by County in the United States from January 2022 until March 2023

FIPS code	State	County	Month Index	Type	Cases
19021	IA	buena vista county	3	poultry	5486700
19143	IA	osceola county	3	poultry	5011700
42071	PA	lancaster county	4	poultry	3782700
39039	OH	defiance county	9	poultry	3748500
55055	WI	jefferson county	3	poultry	2750700
⋮	⋮	⋮	⋮	⋮	⋮
56039	WY	teton county	10	wild bird	1
56039	WY	teton county	6	wild bird	1
56039	WY	teton county	9	wild bird	1
56043	WY	washakie county	13	wild bird	1
56043	WY	washakie county	14	wild bird	1

Table 6 presents a comparative analysis of the five counties with the highest and lowest monthly case counts in the United States as of March 31st, 2023. The data reveals that Buena Vista County, Iowa, recorded the highest number of monthly cases, amounting to 5,486,700 in March 2022. The primary outbreak type in this region was associated with poultry. In contrast, the five counties reporting the lowest monthly case numbers are all located in Wyoming, each registering only a single case related to wild birds.

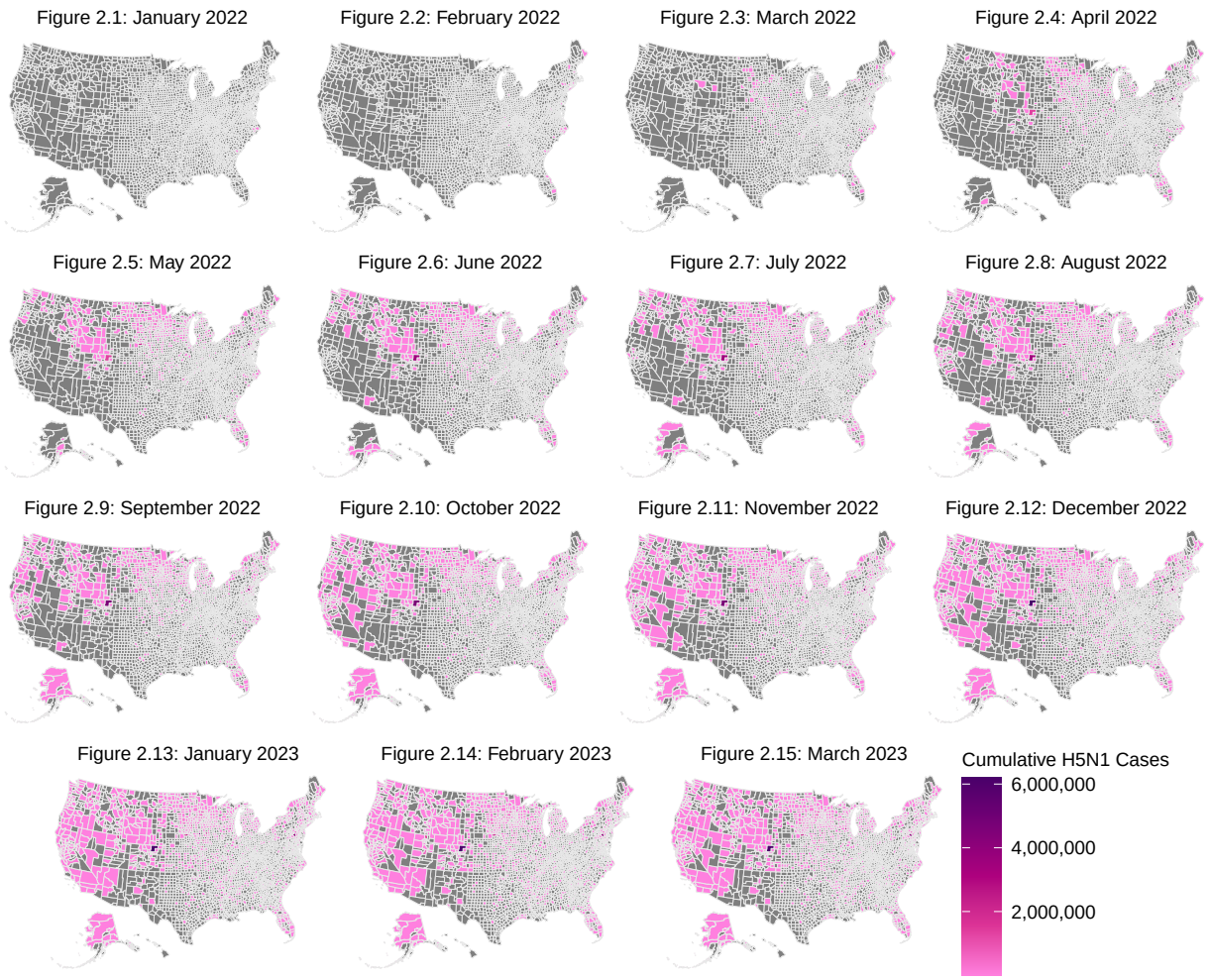


Figure 2: Monthly Cumulative H5N1 Case Totals by County in the United States from January 2022 until March 2023

Figure 2 illustrates the cumulative incidence of H5N1 virus cases across the United States from January 2022 to March 2023. As of March 2023, a widespread occurrence of the virus was observed throughout the country. Nevertheless, it is noteworthy that in most regions, the cumulative case count remained below 2,000,000.

Table 7: Counties with the Highest and Lowest Five Cumulative Case Counts in the United States until March 31st, 2023

FIPS code	State	County	Cases
8123	CO	weld county	6188790
19021	IA	buena vista county	5606301
19143	IA	osceola county	5011700
42071	PA	lancaster county	3855188
39039	OH	defiance county	3748500
⋮	⋮	⋮	⋮
55127	WI	walworth county	1
55135	WI	waupaca county	1
56003	WY	big horn county	1
56037	WY	sweetwater county	1
56043	WY	washakie county	1

Table 7 delineates the counties with the highest and lowest cumulative case counts in the United States as of March 31st, 2023. The analysis indicates that Weld County, Colorado, has the highest cumulative number of cases, totaling 6,188,790. Notably, Buena Vista and Osceola counties in Iowa also report significantly high cumulative cases, with counts of 5,606,301 and 5,011,700, respectively. Conversely, the counties with the lowest cumulative case counts, each reporting only a single case, are situated in Wisconsin and Wyoming.

In Figure 3, the x-axis represents the column variables, while the y-axis corresponds to the row variables.

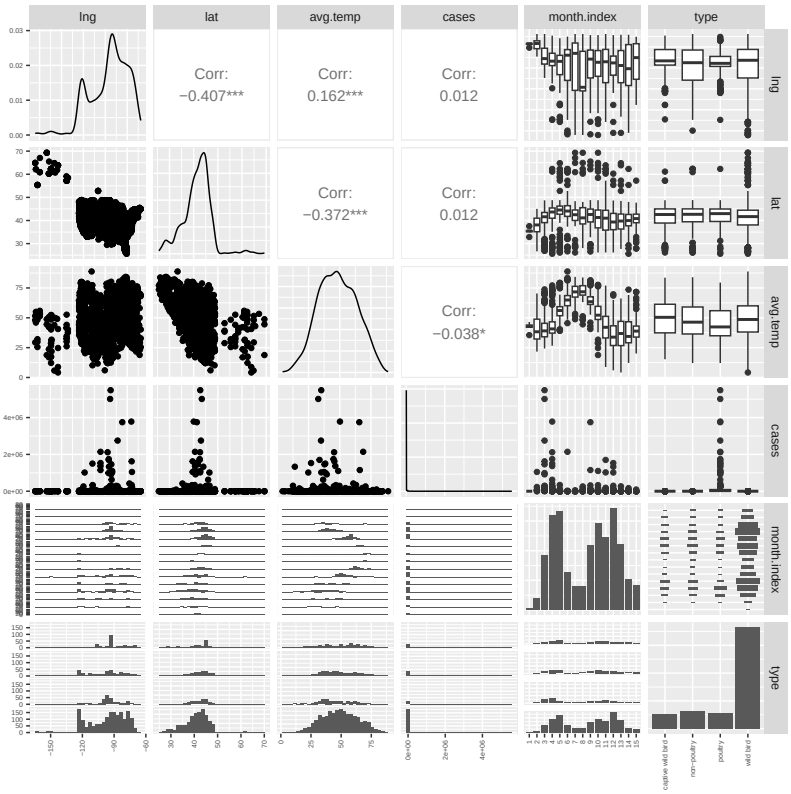


Figure 3: Comprehensive Scatterplot Matrix Depicting Relationships Among Variables in H5N1 Case Data

A notable finding is the correlation between latitude and longitude, which is -0.407, indicating a moderate negative relationship. Specifically, a surge in cases is observed when the latitude ranges between 35 and 45. Similarly, an increase in cases is markedly evident when the longitude falls between -120 and -70, pinpointing regions in the United States with higher case numbers.

Furthermore, the correlation between average temperature and latitude is -0.372. This correlation is logical: as the latitude increases (approaching the North Pole), the average temperature tends to decrease.

The variable `type` highlights that H5N1 outbreaks are more prevalent among wild birds. However, this does not necessarily imply a higher incidence of H5N1 cases in this group.

Additionally, the `cases` variable suggests a relatively low case count per outbreak, yet with numerous outliers. This pattern may be attributed to the fact that, although wild birds constitute the majority of the dataset, poultry are often found in larger groups. Given that viruses spread more efficiently in closely packed populations, the bulk of the cases involve poultry.

An intriguing pattern emerges in the scatterplot with `avg.temp` on the x-axis and `cases` on the y-axis. The distribution of cases against average temperature resembles a normal distribution, with a notable increase in cases when the average temperature lies between 30 and 60 degrees Fahrenheit.

The bar chart presented in Figure 4 below illustrates the monthly distribution of H5N1 cases.

Notably, March 2022 recorded the highest number of cases, aligning with the data in Table 6, which indicates that Buena Vista County, Iowa, experienced the peak monthly case count. Beginning in 2023, there has been a consistent decline in the incidence of H5N1 cases.

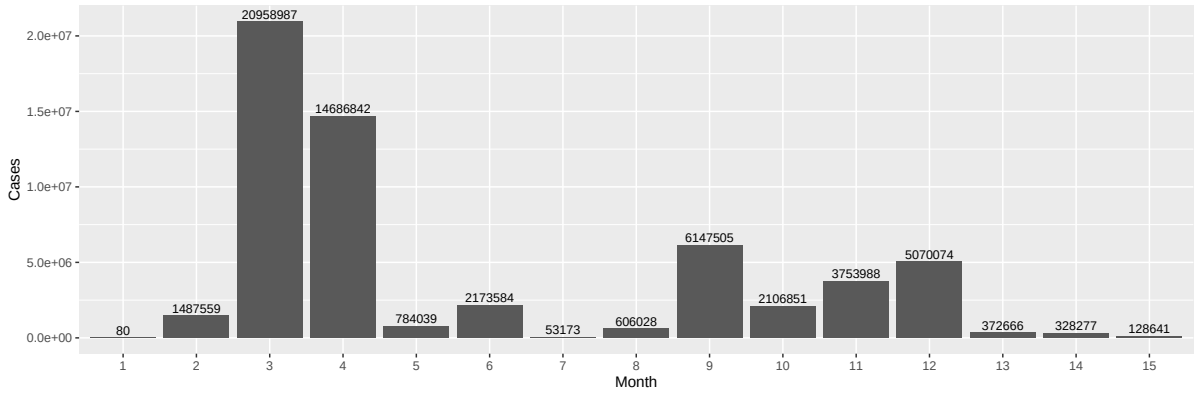


Figure 4: Monthly Distribution of New H5N1 Cases Visualized in a Bar Chart Format

3 Methodology

3.1 Overview

Figure 5, illustrated below, encapsulates a schematic representation delineating the procedural framework employed in this investigation. The diagram distinctly elucidates the successive stages integral to the formulation, evaluation, and selection of the optimal predictive model. The commencement and termination points are signified by red ovals, serving as pivotal markers in the overall sequence. Each procedural step within the study is distinctly demarcated by yellow rectangles, embodying discrete elements of the analytical process. Noteworthy components are highlighted through the incorporation of three blue parallelograms, symbolizing the curated dataset, training set, and testing set essential for model development and assessment.

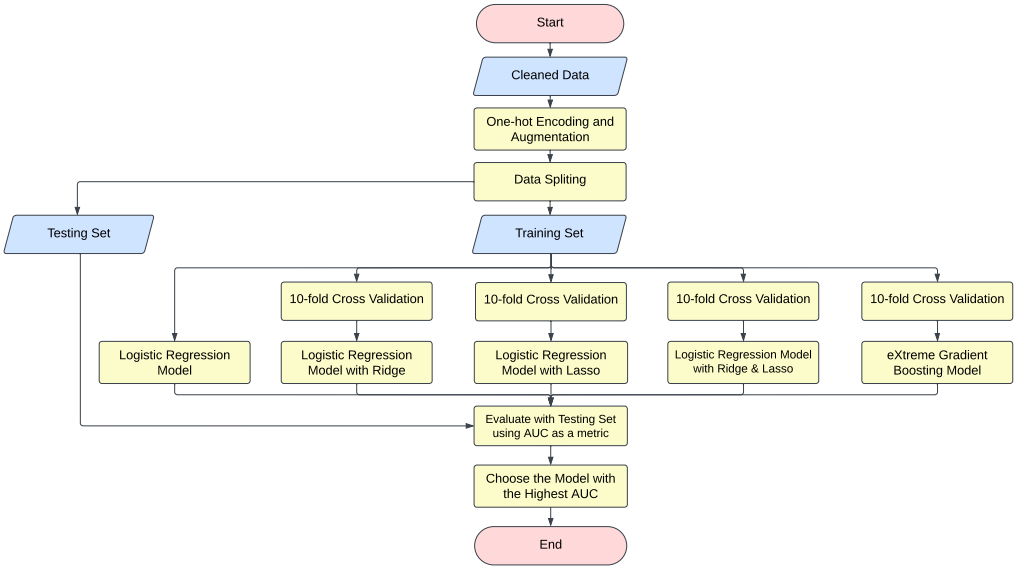


Figure 5: Detailed Flowchart Depicting the Methodology for Building, Evaluating, and Selecting Optimal Models for Predicting Upcoming H5N1 Outbreaks by County

It is crucial to emphasize that the choice of the Area Under the Receiver Operating Characteristic Curve (ROC-AUC score) as the evaluation metric in this study is driven by the presence of unbalanced distribution of `uninfected` and `infected` within the response variable, `binary.case`, as indicated in Table 5. As Brabec et al. (2020) says, "A convenient property of evaluating classifier by ROC-AUC is that it's value is invariant to class imbalance." This metric, by considering the entire range of true positive rates and false positive rates,

offers a comprehensive assessment of a model's performance across various classification thresholds, making it particularly well-suited for situations characterized by imbalanced class distributions.

The primary goal is to discern the model with the highest ROC-AUC score that can effectively prognosticate the likelihood of H5N1 outbreaks in diverse counties for the forthcoming month with precision.

3.2 Data Preparation

The transformation of categorical predictors into a numerical format is often necessitated to facilitate compatibility with various machine learning algorithms. Specifically, the technique of one-hot encoding was employed in the present study to convert the categorical predictor variable labeled as `type`, which encompasses four distinct categories. This strategic utilization of one-hot encoding ensures the preservation of the categorical nature of the variables under consideration throughout the analytical process.

In addition to the aforementioned predictors, our models will incorporate an interaction term denoted as `lat * lng`, representing the product of latitude and longitude. This interaction predictor aims to capture potential joint effects and relationships between geographic coordinates, offering a nuanced perspective on the spatial dynamics associated with the incidence of H5N1 cases. Through the inclusion of this interaction term, we endeavor to enhance the models' capacity to discern location-specific patterns and associations, thereby contributing to a more comprehensive understanding of the factors influencing the predicted outcomes.

Following the implementation of one-hot encoding on the categorical predictor and data augmentation of `lat` and `lng`, the resulting data frame incorporates a collective total of eight columns that represent the predictor variables.

For the purposes of this investigation, our analysis incorporated data encompassing outbreaks that transpired between January 2022 and February 2023 as the training set, constituting a dataset comprising $n = 176,008$ observations. Furthermore, outbreaks occurring in March 2023, totaling 12,572 observations, were utilized as the testing set. This methodological choice was made to facilitate the assessment of the models' efficacy in predicting the incidence of H5N1 cases within specific counties.

3.3 Logistic Regression Model

Logistic regression was selected as the baseline model due to its demonstrated effectiveness in accommodating diverse data characteristics. Its appropriateness in the present context lies in its well-established suitability for binary classification tasks, specifically discerning between instances of `infected` and `uninfected` status of H5N1 in each county. The mathematical representation of the model is as follows:

$$\begin{aligned} Y_i =& \beta_0 + \beta_1 \cdot \texttt{lat} + \beta_2 \cdot \texttt{lng} + \beta_3 \cdot \texttt{month.index} + \beta_4 \cdot \texttt{avg.temp} + \beta_5 \cdot \texttt{type(non-poultry)} \\ &+ \beta_6 \cdot \texttt{type(poultry)} + \beta_7 \cdot \texttt{type(wild bird)} + \beta_8 \cdot \texttt{lat * lng}. \end{aligned}$$

In the logistic regression model, the dependent variable Y_i represents the log-odds of the probability of a county experiencing an infected status of H5N1. The log-odds quantify the natural logarithm of the odds, serving as a measure that facilitates the interpretation of the model's coefficients in the context of binary classification. Moreover, β_0 is the intercept of the model, β_1 to β_8 are the coefficients of independent variables `lat` to `lat * lng`.

Furthermore, the utilization of the sigmoid function is integral to our logistic regression model and is expressed as follows:

$$\begin{aligned} \Pr(Y_i = \texttt{infected}|X_i) &= p(X_i) = \frac{e^{\beta_0+\beta_1\cdot\texttt{lat}+\ldots+\beta_8\cdot\texttt{lat * lng}}}{1+e^{\beta_0+\beta_1\cdot\texttt{lat}+\ldots+\beta_8\cdot\texttt{lat * lng}}} \\ \Pr(Y_i = \texttt{uninfected}|X_i) &= 1 - p(X_i) = 1 - \left(\frac{e^{\beta_0+\beta_1\cdot\texttt{lat}+\ldots+\beta_8\cdot\texttt{lat * lng}}}{1+e^{\beta_0+\beta_1\cdot\texttt{lat}+\ldots+\beta_8\cdot\texttt{lat * lng}}}\right). \end{aligned}$$

Here, X_i represents the vector of predictor variables associated with the i-th observation, including `lat`, `lng`, `month.index`, `avg.temp`, `type(non-poultry)`, `type(poultry)`, `type(wild bird)`, and `lat * lng`.

This sigmoid transformation ensures that the predicted probabilities fall within the interval $(0, 1)$, thereby enhancing the model's capability to generate plausible predictions. Serving as a binary classifier, the model is specifically designed to assess the likelihood of H5N1 cases in a county based on the type of outbreak

encountered. The application of the sigmoid function contributes to the interpretability and calibration of the model's predictions, aligning them with the probabilistic nature of the logistic regression framework.

In the pursuit of parameterizing our logistic regression model and elucidating the factors influencing the likelihood of avian influenza in specific counties, we turn our attention to the log-likelihood function, which is defined as follows:

$$\ell(\beta) = \sum_{i=1}^{n} Y_i \log(p(X_i)) + (1 - Y_i) \log(1 - p(X_i)).$$

To obtain the estimates of the logistic regression model parameters (β_i), we employ the Maximum Likelihood Estimate (MLE) method. Solving the following objective function maximizes the log-likelihood:

$$\hat{\beta} = \arg\max_{\beta} \Big[\ell(\beta) \Big].$$

This optimization process enhances our model's capacity to capture and interpret the intricate relationships between the predictor variables and the likelihood of avian influenza in each county.

3.4 Logistic Regression Model with Regularizations

To mitigate data bias and prevent overfitting, three regularization techniques will be applied to our fundamental logistic model discussed in Section 3.3.

3.4.1 Ridge (L2) Regularization

For Ridge (L2) regularization, the objective function is defined as follows:

$$\hat{\beta} = \arg\max_{\beta} \left[\ell(\beta) + \lambda \sum_{i=1}^{n} \beta_i^2 \right],$$

where $\ell(\beta)$ represents the loss function of the basic logistic regression model, and $\lambda \sum_{i=1}^{n} \beta_i^2$ is the penalty term.

3.4.2 Lasso (L1) Regularization

When employing Lasso (L1) regularization, the objective function takes the form:

$$\hat{\beta} = \arg\max_{\beta} \left[\ell(\beta) + \lambda \sum_{i=1}^{n} |\beta_i| \right],$$

with $\ell(\beta)$ denoting the loss function of the original logistic regression model and $\lambda \sum_{i=1}^{n} |\beta_i|$ as the penalty term.

3.4.3 Combination of Ridge (L2) and Lasso (L1) Regularizations

Given the relatively stronger penalty of Ridge regularization ($\lambda \sum_{i=1}^{n} \beta_i^2$) compared to Lasso regularization ($\lambda \sum_{i=1}^{n} |\beta_i|$), potentially leading to extreme feature selection, a combination of Ridge and Lasso regularizations might perform great. The objective function is then expressed as:

$$\hat{\beta} = \arg\max_{\beta} \left[\ell(\beta) + \lambda \left(\frac{1}{2} \sum_{i=1}^{n} \beta_i^2 + \frac{1}{2} \sum_{i=1}^{n} |\beta_i| \right) \right],$$

where $\ell(\beta)$ is the loss function of the original logistic regression model, and $\lambda \left(\frac{1}{2} \sum_{i=1}^{n} \beta_i^2 + \frac{1}{2} \sum_{i=1}^{n} |\beta_i| \right)$ is the penalty term.

To determine the optimal tuning parameter λ and rigorously evaluate our model, a 10-fold cross validation strategy will be implemented for all the discussed regularization methods in this section. This approach furnishes a robust estimate of the model's performance through its log-likelihood.

3.5 eXtreme Gradient Boosting (XGBoost) Model

In contemporary times, XGBoost has emerged as a robust ensemble learning technique, garnering widespread acclaim for its exceptional predictive prowess. The algorithm methodically assembles a set of weak learners, typically manifested as decision trees, and amalgamates their predictions to bolster accuracy and exhibit strong generalization performance on unseen data.

The XGBoost algorithm, as detailed by Chen and Guestrin (2016), leverages gradient tree boosting for ensemble learning. For our training data with n observations and $m = 8$ features, the tree ensemble model is expressed as follows:

$$\hat{Y}_i = \phi(X_i) = \sum_{k=1}^{K} f_k(X_i), \quad f_k \in F$$

for each $F = \{f(X) = w_q(X)\}$, where $q : \mathbb{R}^m \to T$ and $w \in T$ (T is the number of leaves in the tree), represents the space of regression trees, with q denoting the tree structure mapping an example to the corresponding leaf index, and w being the leaf weights.

The learning objective is formulated with regularization as:

$$L(\phi) = \sum_i l(\hat{Y}_i, Y_i) + \sum_k \Omega(f_k).$$

Here, l is a differentiable convex loss function measuring the difference between predicted $\hat{Y}_i$ and target Y_i, and $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ penalizes the complexity of the model.

Gradient tree boosting is employed to optimize the model in an additive manner. Given the prediction $\hat{Y}_i^{(t-1)}$ at the t-th iteration, a new function f_t is added to minimize the objective:

$$L^{(t)} = \sum_{i=1}^{n} l(Y_i, \hat{Y}_i^{(t-1)} + f_t(X_i)) + \Omega(f_t).$$

The second-order approximation is utilized for optimization:

$$L^{(t)} \approx \sum_{i=1}^{n} [l(Y_i, \hat{Y}_i^{(t-1)}) + g_i f_t(X_i) + \frac{1}{2} h_i f_t^2(X_i)] + \Omega(f_t).$$

Here, $g_i = \partial_{\hat{Y}_i^{(t-1)}} l(Y_i, \hat{Y}_i^{(t-1)})$ is the first-order gradient statistics, and $h_i = \partial^2_{\hat{Y}_i^{(t-1)}} l(Y_i, \hat{Y}_i^{(t-1)})$ is the second-order gradient statistics.

Given the objective function and approximation above, the optimal weight w_j^* of leaf j is computed as:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}.$$

Here $I_j = \{i | q(X_i) = j\}$ is the instance set of leaf j.

The loss reduction after a split is given by:

$$L_{\text{split}} = \frac{1}{2} \left(\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda} \right) - \gamma.$$

Here I_L represents the instance set of left node after split, and I_R represents the instance set of right node after split. Moreover, I_L and I_R should satisfy $I = I_L \cup I_R$.

This formulation is utilized for evaluating split candidates during the tree construction process. The XGBoost model integrates these mathematical representations for effective optimization and predictive performance.

In the implementation of our XGBoost model, we will leverage the exact greedy algorithm. As described by Chen and Guestrin (2016), the exact greedy algorithm, characterized by its thorough exploration of all potential splits across features, aligns with our objective of meticulous feature selection and optimal decision tree construction.

The performance of the XGBoost model will be assessed based on area under the curve (ROC-AUC). Employing 10-fold cross-validation ensures robust estimation of ROC-AUC across different data subsets, enhancing the model's reliability and generalization capabilities.

4 Results

4.1 Logistic Regression Model

Utilizing the Maximum Likelihood Estimate (MLE) method in our logistic regression model yielded the estimated log-odds as follows:

$$\begin{aligned}\hat{Y}_i =& 18.215 - 0.284 \cdot \texttt{lat} + 0.08 \cdot \texttt{lng} - 0.057 \cdot \texttt{month.index} - 0.004 \cdot \texttt{avg.temp} - 0.203 \cdot \texttt{type(non-poultry)} \\ & - 0.055 \cdot \texttt{type(poultry)} - 1.997 \cdot \texttt{type(wild bird)} - 0.002 \cdot \texttt{lat * lng}.\end{aligned}$$

Upon applying the model to the sigmoid function and inputting the test set data, we obtain a test accuracy of 99% and a corresponding test ROC-AUC of 0.74987.

4.2 Logistic Regression Model with Ridge (L2) Regularization

Here the logistic regression model was trained with various ridge regularization strengths (λ), spanning a range from low to high values, using a 10-fold cross-validation strategy. The model's deviance was documented for each regularization strength.

Figure 6 helps identify the optimal λ for the ridge penalty term of the logistic regression model. We aim to get the λ which minimizes the deviance. By looking at the figure, we got the smallest deviance when $\lambda = 0.0014659$ and all 8 predictors were selected by the ridge penalty, which is indicated by the vertical dash line on the left.

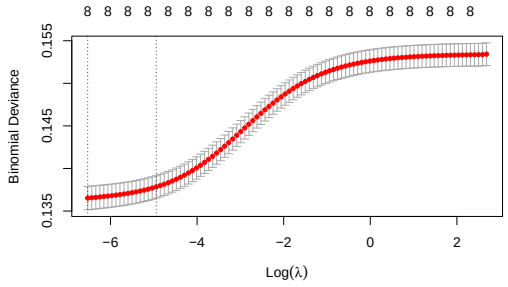


Figure 6: 10-fold Cross-Validation for Optimal λ in Ridge Regularization

The estimated log-odds of the logistic regression model with ridge regularization is as follows:

$$\begin{aligned}\hat{Y}_i =& 9.092 - 0.085 \cdot \texttt{lat} + 0.006 \cdot \texttt{lng} - 0.048 \cdot \texttt{month.index} - 0.001 \cdot \texttt{avg.temp} + 0.043 \cdot \texttt{type(non-poultry)} \\ & + 0.168 \cdot \texttt{type(poultry)} - 1.693 \cdot \texttt{type(wild bird)} - 5.818 \times 10^{-5} \cdot \texttt{lat * lng}.\end{aligned}$$

Upon applying this model to the sigmoid function and inputting the test set data, we attain a test accuracy of 99% and a corresponding test ROC-AUC of 0.74727. However, it is noteworthy that the test ROC-AUC is lower than that of the logistic regression model. This discrepancy suggests that the model may be underfitting the data, likely due to the imposition of a robust ridge penalty.

4.3 Logistic Regression Model with Lasso (L1) Regularization

Here the logistic regression model was trained with various lasso regularization strengths (λ), spanning a range from low to high values, using a 10-fold cross-validation strategy. The model's deviance was documented for each regularization strength.

Figure 7 helps identify the optimal λ for the ridge penalty term of the logistic regression model. We aim to get the λ which minimizes the deviance. By looking at the figure, we got the smallest deviance when $\lambda = 1.807284 \times 10^{-5}$ and all 8 predictors were selected by the lasso penalty, which is indicated by the vertical dash line on the left.

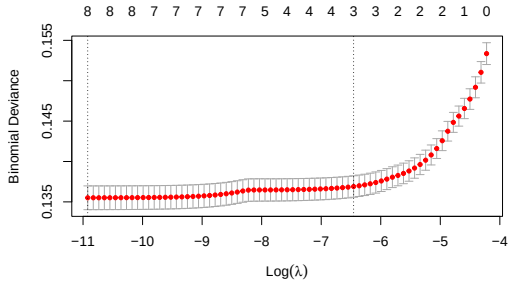


Figure 7: 10-fold Cross-Validation for Optimal λ in Lasso Regularization

The estimated log-odds of the logistic regression model with lasso regularization is as follows:

$$\begin{aligned}\hat{Y}_i =& 17.243 - 0.262 \cdot \texttt{lat} + 0.072 \cdot \texttt{lng} - 0.056 \cdot \texttt{month.index} - 0.004 \cdot \texttt{avg.temp} - 0.197 \cdot \texttt{type(non-poultry)} \\ & - 0.049 \cdot \texttt{type(poultry)} - 1.99 \cdot \texttt{type(wild bird)} - 0.002 \cdot \texttt{lat * lng}.\end{aligned}$$

Upon application of this model to the sigmoid function and feeding it with the test set data, we achieve a test accuracy of 99% and a test ROC-AUC of 0.75073. Notably, the test ROC-AUC exceeds that of the logistic regression model.

4.4 Logistic Regression Model with Ridge (L2) and Lasso (L1) Regularizations

Here the logistic regression model was trained with various combinations of ridge and lasso regularization strengths (λ), spanning a range from low to high values, using a 10-fold cross-validation strategy. The model's deviance was documented for each regularization strength.

Figure 8 helps identify the optimal λ for the combinations of ridge and lasso penalty term of the logistic regression model. We aim to get the λ which minimizes the deviance. By looking at the figure, we got the smallest deviance when $\lambda = 1.7172127 \times 10^{-5}$ and all 8 predictors were selected by the combinations of ridge and lasso penalty, which is indicated by the vertical dash line on the left.

The estimated log-odds of the logistic regression model with the combination of ridge and lasso regularization is as follows:

$$\begin{aligned}\hat{Y}_i =& 17.278 - 0.263 \cdot \texttt{lat} + 0.072 \cdot \texttt{lng} - 0.056 \cdot \texttt{month.index} - 0.004 \cdot \texttt{avg.temp} - 0.197 \cdot \texttt{type(non-poultry)} \\ & - 0.05 \cdot \texttt{type(poultry)} - 1.991 \cdot \texttt{type(wild bird)} - 0.002 \cdot \texttt{lat * lng}.\end{aligned}$$

Upon application of the model to the sigmoid function and inputting the test set data, we achieve a test accuracy of 99% and a corresponding test ROC-AUC of 0.75072. Notably, the test ROC-AUC is slightly

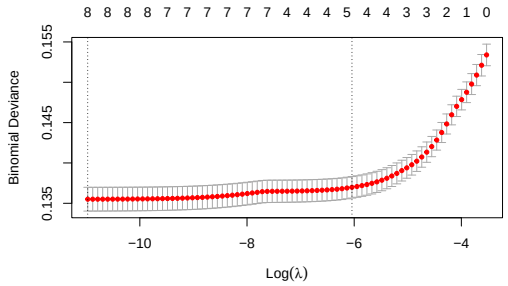


Figure 8: 10-fold Cross-Validation for Optimal λ in Combination of Ridge and Lasso Regularizations

inferior to that of logistic regression with lasso regularization, but it surpasses the test ROC-AUC obtained from logistic regression alone.

4.5 eXtreme Gradient Boosting (XGBoost)

The XGBoost model was trained utilizing the exact tree method, and its performance was systematically monitored through the negative log-likelihood loss at each iteration of the training process. In order to ensure both performance and robustness, a 10-fold cross-validation strategy was implemented at each training round, and certain hyperparameters were systematically adjusted and applied as delineated below:

- The learning rate (η) was designated as 0.02, serving as the step size shrinkage to mitigate overfitting during the training process.
- The subsample ratio was established at 0.75, indicating that the XGBoost algorithm would stochastically select 75% of the training data for each tree's growth.
- The column subsample ratio for each tree was configured to 0.8, signifying that the XGBoost algorithm would randomly consider 80% of predictors when developing a new tree.
- The maximum depth of the trees was restricted to 10.
- The number of training rounds was predetermined to be 700.

Figure 9 illustrates the progression of training and validation losses across 700 rounds. The training loss, depicted by the blue line, exhibits a consistent and gradual decrease over the course of the training iterations. Similarly, the validation loss, represented by the orange line, also demonstrates a continuous and steady decline. Notably, at the conclusion of the 700 training rounds, the validation negative log-likelihood loss converges to a value of 0.05019, indicating the model's improved generalization performance on unseen data.

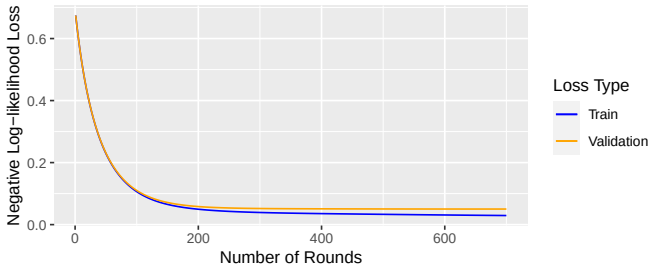


Figure 9: Training and Validation Losses of XGBoost Model over 700 Rounds of Training

Upon applying the model to the test set, we observe a test accuracy of 99% and a test ROC-AUC of 0.87928. Notably, the test ROC-AUC surpasses that of any logistic regression model, whether with or without regularizations.

An inherent characteristic of XGBoost lies in its capability to furnish insightful analyses regarding feature importance. This is accomplished by computing importance scores assigned to each predictor, employing

Gain as the metric. Gain signifies the enhancement in accuracy attributed to a particular feature across the ensemble of the model's trees (Chen and Guestrin 2016), thereby facilitating the identification of the most influential predictors.

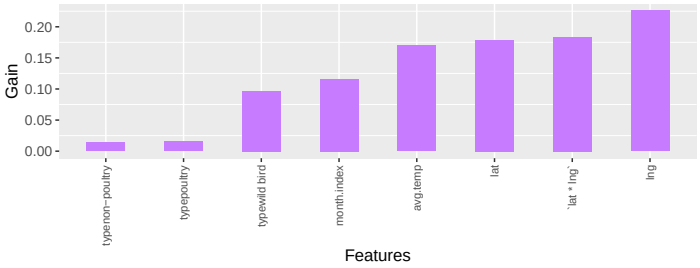


Figure 10: Feature Importance Plot for XGBoost Model

Figure 10 depicts the Gain scores associated with the predictors employed in the XGBoost model. A taller bar, indicative of a higher Gain score, signifies greater importance of the corresponding predictor. Remarkably, pivotal predictors such as `lng` (longitude), `lat * lng` (interaction between latitude and longitude), and `lat` (latitude) have emerged as noteworthy contributors to the predictive efficacy of the model.

4.6 Model Selection and Evaluation

This section undertakes a rigorous evaluation and comparative analysis of diverse models employed from section 4.1 to 4.5. Given the test accuracy of these models are all 99% and the fact that the dependent variable `binary.cases` has imbalanced classes, the test ROC-AUC is the preferred metric here.

Figure 11 encompasses a series of Receiver Operating Characteristic (ROC) Curves corresponding to the models developed in Sections 4.1 to 4.5. Notably, the XGBoost model demonstrates exceptional efficacy, as evidenced by a remarkable test ROC-AUC score of 0.87928. This superiority is discernible when juxtaposed against traditional models such as Logistic Regression, Logistic Regression with Lasso Regularization, Logistic Regression with Ridge Regularization, and their amalgamated approach, as illustrated in Figure 10.6. The ROC curve for the XGBoost model distinctly approaches the upper-left corner, indicative of a robust fit to the test set data and exemplary predictive performance.

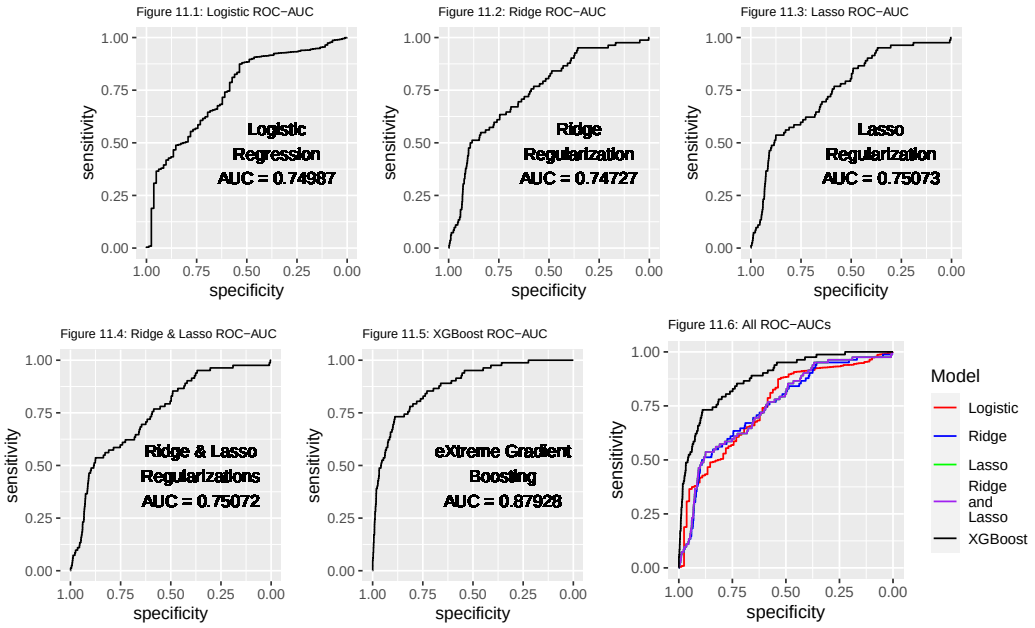


Figure 11: Receiver Operating Characteristic (ROC) Curves for Model Discrimination

Given that the XGBoost model attains the highest test ROC-AUC score, we acknowledge the superior classification efficacy of the XGBoost model among the five considered models.

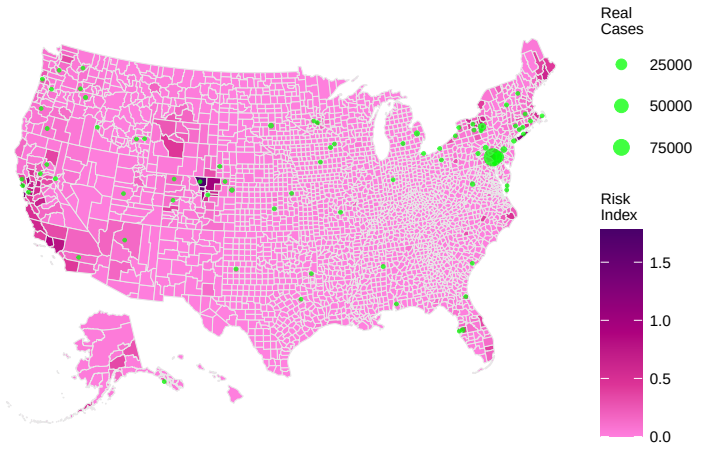


Figure 12: Comprehensive Risk Index Map Derived from the XGBoost Model

In Figure 12, a cartographic representation is presented, depicting the cumulative predicted probabilities for all four potential H5N1 outbreak types (`poultry`, `non-poultry`, `wild bird`, and `captive wild bird`) across all counties in March 2023. These probabilities are computed through the application of the XGBoost model. The amalgamated probability across the four distinct outbreak types for each county may be construed as its risk index. This risk index serves as a composite measure encapsulating the cumulative probabilities associated with the aforementioned outbreak types collectively. A higher value ascribed to this index implies a correspondingly greater overall risk of encountering an outbreak. It is imperative to underscore that a heightened darkness in the visual representation corresponds to an elevated perceived risk of an outbreak.

Furthermore, the identification of green dots in the visualization designates counties that indeed experienced H5N1 cases in March 2023. The magnitude of the green dots is indicative of the extent of outbreak cases, where a larger green dot conveys a higher incidence of outbreaks.

Figure 13 provides an in-depth examination through a series of risk index maps, which divide the United States into four distinct regions: West, South, Midwest, and Northeast. The presence of documented H5N1 cases is represented by green dots within these regions, and the size of each green dot is scaled to reflect its contextual significance within the respective region, aiming to amplify visual impact. Additionally, blue triangles within each region pinpoint counties with reported cases and a risk index exceeding 0.25.

Figure 13.1: West Region
including Alaska and Hawaii

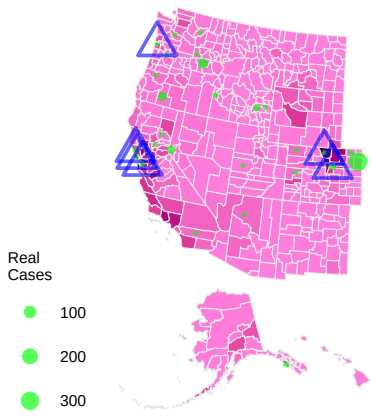
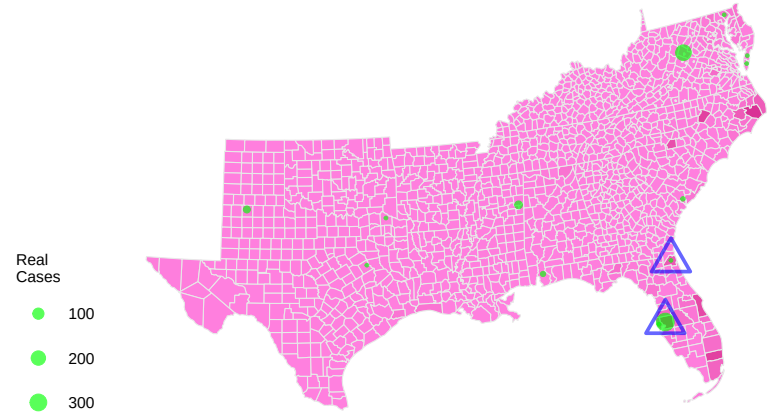


Figure 13.2: South Region



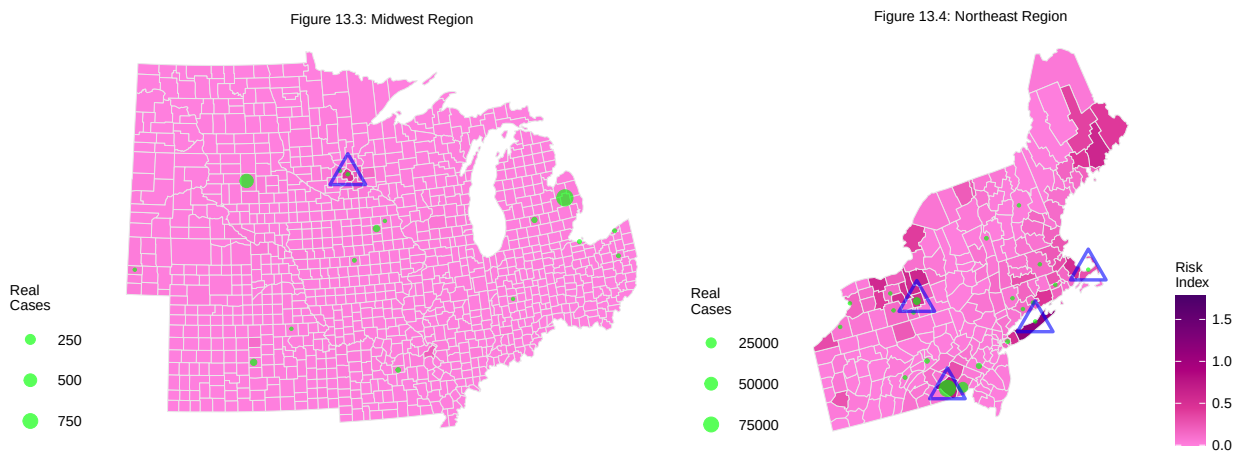


Figure 13: Risk Index Map by Regions Derived from the XGBoost Model

Upon careful examination of Figures 13.1 and 13.2, it becomes evident that these regions experienced the fewest outbreaks in March 2023 among the four delineated regions. Specifically, within the West region, six counties are marked by blue triangles - three in California, two in Colorado, and one in Oregon. The South region exhibits a total of two counties with blue triangles, all located in Florida. Figure 13.3 reveals a solitary county, situated in Minnesota, marked by a blue triangle. However, it is noteworthy that the county highlighted is not necessarily the one with the highest incidence of H5N1 cases, because the size of the green dot is not the largest in the region. Conversely, Figure 13.4 unequivocally illustrates that the Northeast region endured the most severe H5N1 outbreak in March 2023 among the four regions. Four counties are identified by blue triangles, with particular emphasis on the county in Pennsylvania, which recorded the highest number of H5N1 outbreak cases.

Table 8: Counties with Confirmed Cases in March 2023 and High H5N1 Outbreak Risk Index Predicted by XGBoost Model

FIPS	State	County	Risk Index	Real Cases	Region
8069	CO	larimer county	1.7793775	1	West
36103	NY	suffolk county	1.1598013	1	Northeast
6041	CA	marin county	0.8736894	1	West
6001	CA	alameda county	0.8572757	2	West
42071	PA	lancaster county	0.7268378	92813	Northeast
8005	CO	arapahoe county	0.5431374	11	West
36109	NY	tompkins county	0.5326815	6503	Northeast
27053	MN	hennepin county	0.4637383	1	Midwest
12057	FL	hillsborough county	0.4396641	341	South
6097	CA	sonoma county	0.3607736	3	West
25001	MA	barnstable county	0.3071986	1	Northeast
12031	FL	duval county	0.2581048	1	South
41007	OR	clatsop county	0.2504931	1	West

Table 8 further strengthens the portrayal of the geographical distribution of H5N1 risk and actual cases, providing additional depth to the visual insights obtained from the accompanying risk index maps in Figure 12 and 13. Noteworthy is the commendable performance of the XGBoost model, which accurately identifies the Northeast region as particularly risky. This is substantiated by elevated risk indices and confirmed cases in specific counties, such as Suffolk County, New York, and Lancaster County, Pennsylvania, underscoring the effectiveness of the XGBoost model in risk prediction.

5 Conclusion and Suggestions

In this study, logistic regression, model regularization techniques, 10-fold cross validation, and eXtreme Gradient Boosting (XGBoost) have been methodically utilized to predict the probability of H5N1 outbreaks in the United States. The study's results demonstrate a notably high ROC-AUC score of 0.87928 for the XGBoost model, underscoring its effectiveness in forecasting future H5N1 cases. This model has accurately pinpointed counties in the United States, particularly in the Northeast region, with an elevated risk of H5N1 outbreaks in March 2023, which corresponds closely with the actual outbreak patterns observed. Within this context, geographic factors emerged as the most critical features in the XGBoost model. Additionally, the variable of temperature was identified as a subsequent pivotal feature, aiding in the guideline of H5N1 risk control across various U.S. counties.

However, the study acknowledges certain limitations in its methodology. The predictive models predominantly considered variables such as temperature, types of outbreaks, time, and population density. Critical factors like existing control zones, wild waterfowl migration patterns, agricultural practices, off-site daily mortality disposal methods, and wild bird access to feed (Green et al. 2023), which could significantly influence H5N1 spread, were not incorporated.

To augment the robustness and precision of our models, future studies should incorporate more comprehensive datasets, including factors like bird migration, human travel patterns, and agricultural data. Exploring advanced machine learning techniques, such as Convolutional Neural Networks (CNNs) and time series analysis, is recommended to elucidate the intricate interactions among these variables. The efficacy of CNNs in related fields is exemplified in the work of Scarafoni et al. (2019), who adeptly applied deep CNN models to predict the host tropism of the Influenza A virus based on protein sequences, thereby demonstrating the potential of this method in virological studies. Furthermore, integrating time series techniques with tree models such as Random Forest and XGBoost, as demonstrated by Kane et al. (2014) in their analysis of H5N1 outbreak predictions in Egypt, could yield low mean square error (MSE) in both retrospective (MSE = 6.3195) and prospective (MSE = 24.8101) contexts. Additionally, the application of Long Short-Term Memory (LSTM) networks, as shown in the study "Prediction of COVID-19 using long short-term memory by integrating principal component analysis and clustering techniques," (Ilu, Rajesh, and Mohammed 2022) underscores the viability of LSTM in H5N1 research. This study reported both sensitivity and specificity rates of 98% for the LSTM model, indicating its commendable performance in disease forecasting.

Our study identifies geographical location and temperature as key predictors in distinguishing between infected and uninfected cases in our model. "Technical Report: Highly Pathogenic Avian Influenza A(H5N1) Viruses" from Centers for Disease Control and Prevention (2023) highlights the current low public health risk posed by A(H5N1) viruses. However, the widespread geographic distribution of infected birds and poultry, along with the potential for human and mammalian exposures, increases the likelihood of viral evolution or reassortment events, potentially altering the risk assessment. This underscores the connection between H5N1 infection and geographic factors, corroborating our model's emphasis on these variables. Additionally, the study "Avian Influenza Virus (H5N1); Effects of Physico-Chemical Factors on Its Survival" conducted by Shahid et al. (2009) reveals a significant relationship between Avian Influenza Virus H5N1 infection and temperature, demonstrating that the virus maintains infectivity at 4 degrees Celsius (39.2 Fahrenheit) for over 100 days, although hemagglutinin (HA) activity diminishes. In contrast, the virus loses its infectivity after 24 hours at room temperature 28 degrees Celsius (82.4 Fahrenheit).

Future research, building on our study's insights, should delve deeper into the intricate relationships among spatial, time, and temperature factors. Emphasizing the geographical location and temperature as pivotal predictors, researchers could explore the spatial dynamics of H5N1 outbreaks with a finer resolution, considering not only county-level data but also incorporating more granular information to capture localized patterns. Additionally, investigating the interactions among spatial, time, and temperature variables can provide a more nuanced understanding of the dynamics influencing the spread of the virus. Integrating advanced spatial analysis techniques, such as Geographic Information System (GIS) mapping, can help uncover spatial patterns that may not be immediately apparent, contributing to a more comprehensive predictive model.

In addition to research directions, comprehensive datasets would enable a more holistic analysis of the multifaceted aspects influencing H5N1 transmission. Collaborative efforts among researchers, governmental agencies, and international organizations can facilitate the sharing of standardized data, fostering a collective approach to understanding and combating the potential threats posed by H5N1 outbreaks.

By focusing on these research directions, scientists can contribute to the development of more robust and accurate predictive models, ultimately aiding in proactive measures for H5N1 risk management and control.

6 Computational Details

The analysis was conducted using R version 4.3.2 for Windows, with the utilization of various R libraries from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/ to facilitate data manipulation, statistical modeling, and visualization. The following R libraries were employed in this study:

- car: Provides functions for linear model diagnostics and fits linear and non-linear models.
- dplyr: A grammar of data manipulation, providing a set of verbs to efficiently manipulate data.
- GGally: Extends ggplot2 with several functions to handle common multivariate data visualizations.
- ggplot2: A powerful and flexible plotting system for creating elegant and complex graphics.
- ggpubr: Enhances ggplot2 for creating publication-ready plots and statistical summaries.
- glmnet: Fits generalized linear models with elastic net regularization.
- grid: A graphics system for R, used for creating and arranging graphical elements.
- gridExtra: Provides functions to arrange multiple grid-based figures on one page.
- imputeTS: Implements time series imputation techniques for missing data.
- knitr: Enables dynamic report generation in R Markdown, integrating code, text, and output.
- patchwork: Offers a flexible system for combining and arranging multiple ggplots.
- pROC: Analyzes ROC curves and calculates area under the curve (AUC) for model evaluation.
- psych: Contains functions for psychological and psychometric research, including factor analysis.
- textstem: Stems and lemmatizes English words for text analysis.
- tidyr: Assists in tidying and reshaping data for analysis.
- usmap: Provides a simple plotting interface for creating US maps with ggplot2.
- xgboost: Fits eXtreme Gradient Boosting models, a scalable tree boosting system.

The analyses were conducted in the RStudio integrated development environment (IDE) version "Mountain Hydrangea" Release (583b465e, 2023-06-05) for Windows. RStudio can be downloaded at https://posit.co/.

An Intel-compatible 64-bit platform is preferred. At least 2048 MB of Random Access Memory (RAM) is recommended to run the whole script.

7 Reproducibility

Ensuring the reproducibility of this study is of utmost importance. The entire analysis, including data preprocessing, model development, and result generation, is encapsulated in an RMarkdown document. A GitHub repository is also created to store relevant materials: https://github.com/GitData-GA/h5n1.

The RMarkdown file, necessary BibTeX and style files can be downloaded at the following link:

https://h5n1.gd.edu.kg/code/download.zip

To reproduce the findings and generate the same results presented in this paper, follow these steps:

1. Download the Necessary Files:

- Navigate to the provided link in your browser.
- Unzip the downloaded file to a directory of your choice.

2. Open RMarkdown in RStudio:

- Ensure you have R and RStudio installed on your machine.
- Open RStudio and navigate to the directory where you unzipped the files.
- Open the RMarkdown file (main.Rmd) in RStudio.

3. Install Required Packages:

- If not already installed, install the required R packages from the CRAN.

4. Knit the Document:

- Knit the RMarkdown file to reproduce the analysis. This will execute the code chunks, perform the analysis, and generate the final document.

By following these steps, you can recreate the entire analysis and verify the results presented in this paper.

8 Acknowledgement

We are grateful to our project advisor, Professor Ambuj Tewari, for his support in the publication process. We would also like to thank Professor Huaijun Zhou for providing background knowledge about avian influenza. GPT-4 was exclusively employed to enhance the linguistic quality of some sections within this paper.

During this journey, we express our profound gratitude to our loved ones and cherished friends. Their unwavering love and support have been instrumental in enabling us to reach the culmination of this endeavor. Without their uplifting encouragement and unwavering motivation, our accomplishment would not have been attainable.

Reference

Brabec, Jan, Tomáš Komárek, Vojtěch Franc, and Lukáš Machlica. 2020. "On Model Evaluation Under Non-Constant Class Imbalance." In *Computational Science – ICCS 2020*, edited by Valeria V. Krzhizhanovskaya, Gábor Závodszky, Michael H. Lees, Jack J. Dongarra, Peter M. A. Sloot, Sérgio Brissos, and João Teixeira, 74–87. Cham: Springer International Publishing.

Cedar Lake Ventures, Inc. 2023. "Weather Spark." Cedar Lake Ventures, Inc. https://weatherspark.com/map?id=145043.

Centers for Disease Control and Prevention. 2022a. "H5N1 Bird Flu Detections Across the United States (Backyard and Commercial) | Avian Influenza (Flu)." www.cdc.gov. https://www.cdc.gov/flu/avianflu/data-map-commercial.html.

———. 2022b. "H5N1 Bird Flu Detections Across the United States (Wild Birds) | Avian Influenza (Flu)." www.cdc.gov. https://www.cdc.gov/flu/avianflu/data-map-wild-birds.html.

———. 2023. "Technical Report: Highly Pathogenic Avian Influenza a(H5N1) Viruses." Centers for Disease Control; Prevention. https://www.cdc.gov/flu/avianflu/spotlights/2022-2023/h5n1-technical-report_october.htm.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 785–94. https://doi.org/10.1145/2939672.2939785.

Farahat, Ramadan Abdelmoez, Sheharyar Hassan Khan, Ali A Rabaan, and Jaffar A Al-Tawfiq. 2023. "The Resurgence of Avian Influenza and Human Infection: A Brief Outlook." *New Microbes and New Infections* 53 (June): 101122–22. https://doi.org/10.1016/j.nmni.2023.101122.

Gilbert, M., X. Xiao, D. U. Pfeiffer, M. Epprecht, S. Boles, C. Czarnecki, P. Chaitaweesub, et al. 2008. "Mapping H5N1 Highly Pathogenic Avian Influenza Risk in Southeast Asia." *Proceedings of the National Academy of Sciences* 105 (March): 4769–74. https://doi.org/10.1073/pnas.0710581105.

Green, Alice L, Matthew A Branan, Victoria Fields, Kelly A Patyk, Stephanie K Kolar, Andrea Beam, Katherine L Marshall, et al. 2023. "Investigation of Risk Factors for Introduction of Highly Pathogenic Avian Influenza H5N1 Virus onto Table Egg Farms in the United States, 2022: A Case–Control Study." *Frontiers in Veterinary Science* 10 (July). https://doi.org/10.3389/fvets.2023.1229008.

Iacurci, Greg. 2023. "Wholesale Egg Prices Have 'Collapsed.' Why Consumers May Soon See Relief." CNBC. https://www.cnbc.com/2023/02/07/wholesale-egg-prices-have-collapsed-from-record-highs-in-december.html.

Ilu, Saratu Yusuf, Prasad Rajesh, and Hassan Mohammed. 2022. "Prediction of COVID-19 Using Long Short-Term Memory by Integrating Principal Component Analysis and Clustering Techniques." *Informatics in Medicine Unlocked* 31: 100990. https://doi.org/10.1016/j.imu.2022.100990.

Kane, Michael J, Natalie Price, Matthew Scotch, and Peter Rabinowitz. 2014. "Comparison of ARIMA and Random Forest Time Series Models for Prediction of Avian Influenza H5N1 Outbreaks." *BMC Bioinformatics* 15 (August). https://doi.org/10.1186/1471-2105-15-276.

National Centers for Environmental Information. 2023. "Climate at a Glance County Mapping." National Centers for Environmental Information. https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/county/mapping.

Pareto Software, LLC. 2023. "United States Counties Database." Pareto Software, LLC. https://simplemaps.com/data/us-counties.

Scarafoni, Dan, Brian A. Telfer, Darrell O. Ricke, Jason R. Thornton, and James Comolli. 2019. "Predicting Influenza a Tropism with End-to-End Learning of Deep Networks." *Health Security* 17 (December): 468–76. https://doi.org/10.1089/hs.2019.0055.

Shahid, Muhammad, Muhammad Abubakar, Sajid Hameed, and Shamsul Hassan. 2009. "Avian Influenza Virus (H5N1); Effects of Physico-Chemical Factors on Its Survival." *Virology Journal* 6: 38. https://doi.org/10.1186/1743-422x-6-38.

Taubenberger, Jeffery K., and David M. Morens. 2006. "1918 Influenza: The Mother of All Pandemics." *Emerging Infectious Diseases* 12 (January): 15–22. https://doi.org/10.3201/eid1201.050979.

Walsh, Daniel P., Ting Fung Ma, Hon S. Ip, and Jun Zhu. 2019. "Artificial Intelligence and Avian Influenza: Using Machine Learning to Enhance Active Surveillance for Avian Influenza Viruses." *Transboundary and Emerging Diseases* 66 (August): 2537–45. https://doi.org/10.1111/tbed.13318.

Williams, Richard AJ, and A Townsend Peterson. 2009. "Ecology and Geography of Avian Influenza (HPAI H5N1) Transmission in the Middle East and Northeastern Africa." *International Journal of Health Geographics* 8: 47. https://doi.org/10.1186/1476-072x-8-47.

World Animal Foundation. 2023. "The Meaty Truth: A Deep Dive into US Meat Consumption Trends." WorldAnimalFoundation.org. https://worldanimalfoundation.org/advocate/us-meat-consumption/.

World Health Organization. 2018. "Influenza (Avian and Other Zoonotic)." Who.int; World Health Organization: WHO. https://www.who.int/news-room/fact-sheets/detail/influenza-(avian-and-other-zoonotic).
World Organisation for Animal Health. 2022. "Terrestrial Animal Health Code - Glossary." World Organisation for Animal Health. https://www.woah.org/fileadmin/Home/eng/Health_standards/tahc/current/glossaire.pdf.