Predictiveness and Drivers of Avian Influenza in Europe

*Michael Rogo Opata1,2 Andrea Lavarello Schettini1, Jan C. Semenza1, Joacim Rocklöv1,2,3\**  
*Interdisciplinary Center for Scientific Computing, Heidelberg University, Im Neuenheimer Feld 205, D-69120 Heidelberg, Germany.*

*\*Corresponding author. E-mail :* [*joacim.rockloev@iwr.uni-heidelberg.de*](mailto:joacim.rockloev@iwr.uni-heidelberg.de)*, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany*

*1. Heidelberg Institute of Global Health, Heidelberg University, Heidelberg, Germany*

*2. Interdisciplinary Centre for Scientific Computing, Heidelberg University, Heidelberg, Germany*

*3. Department of Public Health and Clinical Medicine, Section of Sustainable Health, Umeå University, Sweden*

**SUPPLEMENTARY INFORMATION**

# **Data Collection**

The outbreak data used in the study was obtained from WAHIS. All the 183 WOAH Members, through their veterinary authorities, are required to submit an immediate notification anytime an exceptional epidemiological event concerning high impact animal diseases (WOAH-Listed diseases) unfolds in their territory. This immediate notification must be followed by weekly follow-up reports until the event is resolved, the disease is eradicated or, until the disease is considered established and endemic in the country. An exceptional epidemiological event refers to any introduction or recurrence of WOAH-Listed diseases into the country’s territory.  A complete description of the variables is shown in (Table SI 2).

**Supplementary Files**

The files provided in this section can be used to reproduce the results reported in the main article. The description provides a brief explanation of the contents and the purpose they served.

|  |  |  |
| --- | --- | --- |
| **Table SI 1**: Table of supplementary files used for constructing the AIV models | | |
|  | **File** | **Description** |
| 1. | ranked\_wild\_bird\_species.xlsx | Contains ranked wild bird score according to their respective impacts on the model as generated from SHAP score. |
| 2. | m1train.xlsxm1test.xlsx | Training and test dataset for Model 1 |
| 3. | m2train.xlsxm2test.xlsx | Training and test dataset for Model 2 |
| 4. | m3train.xlsxm3test.xlsx | Training and test dataset for Model 3 |

|  |  |  |
| --- | --- | --- |
| **Table SI 2**: Spatial-temporal features and targets used in constructing AIV models for the years 2006 to 2021. The features were engineered based on the question answered by the model. | | |
| Feature category | Target variables | Target names |
| Environmental | Vegetation index | lai\_hv\_\* = high vegetation index |
| lai\_lv\_\* = low vegetation index |
|  | Normalized deviation water index | ndwi\_\* |
|  | Normalized deviation vegetation index | ndvi\_\* |
| Climate | Temperature | min\_temp\_\*  mean\_temp\_\*  maximum\_temp\_\* |
| Precipitation | total\_rain\_\*  total\_rain\_\*  total\_rain\_\* |
| Bioclimatic | Temperature-related | BIO1 = Annual Mean Temperature  BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))  BIO3 = Isothermality (BIO2/BIO7) (×100)  BIO4 = Temperature Seasonality (standard deviation ×100)  BIO5 = Max Temperature of Warmest Month  BIO6 = Min Temperature of Coldest Month  BIO7 = Temperature Annual Range (BIO5-BIO6)  BIO8 = Mean Temperature of Wettest Quarter  BIO9 = Mean Temperature of Driest Quarter  BIO10 = Mean Temperature of Warmest Quarter  BIO11 = Mean Temperature of Coldest Quarter |
| Precipitation-related | BIO12 = Annual Precipitation  BIO13 = Precipitation of Wettest Month  BIO14 = Precipitation of Driest Month  BIO15 = Precipitation Seasonality (Coefficient of Variation)  BIO16 = Precipitation of Wettest Quarter  BIO17 = Precipitation of Driest Quarter  BIO18 = Precipitation of Warmest Quarter  BIO19 = Precipitation of Coldest Quarter |
| Trade | Goods traded/year | Loading = Annual trading loading  Unloading = Annual trading unloading |
| Economic | Income/year (in Million Euros) | GDP\_Annual = Annual mean regional income |
| Demographic | Female population /year | F\_TOTAL = Total Female population |
| F\_Y\_15-64 = Female 15 – 64 Years |
| F\_Y\_GE65 = Female > 65 Years |
| F\_Y\_LT15 = Female < 15 years |
| Male population / year | M\_TOTAL = Total Male population |
| M\_Y\_15-64 = Male 15 – 64 Years |
| M\_Y\_GE65 = Male > 65 Years |
| M\_Y\_LT15 = Male < 15 years |
| Birds | Wild bird population | Wild bird species excluding other mammals |
| Avian Influenza outbreak events | Response feature | Label = NUTS3 Avian Influenza virus outbreak events\*\* |

*\* The respective features were aggregated at quarterly intervals to represent the seasonality in Europe. Each value constitutes q1, q2, q3, and q4 variables. \*\* Labels were also manipulated by switching them on and off to investigate phenomena of interest.*

# **Methods**

Models were tuned to generalize well on unseen data. In general, at each node, the ‘learnable’ are the choice of decision. We therefore use them to set numeric thresholds that help in informing why we take the left or the right branch. Ideally the algorithms can be made more conservative or non-conservative through these values. The set of the hyperparameters we use in this work can be broadly explained as:

1. Eta (learning rate): Used to shrink the weights on each step hence making the model more robust.
2. scale\_pos\_weight: Used for scaling unbalanced data for faster convergence.
3. num\_boost\_round: Represents the number of rounds for boosting.
4. max-depth: Controls overfitting by preventing learning relations very specific to a particular sample.
5. gamma: Makes the model more conservative in the sense that a node is split only when the resulting split gives a positive reduction in the loss function.
6. subsample: Gives the fraction of observations to be random samples for each tree.
7. colsample\_bytree: Represents the fraction of columns to be random samples for each tree.
8. min\_child\_weight: Used to control overfitting in the sense that higher values restrict the model from learning relations that are specific to the sample of a selected tree.
9. reg\_alpha: represents L1 regularization term on weights.
10. reg\_lambda: represents L2 regularization term on weights.

We employ the python version of XGBoost with a combination of the scikit-learn API.1 The algorithm has previously been tested to be superior to support vector machine (SVM), random forest, näive model, and lightGBM (light gradient boosted method) in a similar study at NUTS3 level for the West Nile Virus (WNV).2 The main advantage XGBoost algorithm has over other methods is its novel sparsity-aware and weighted quantile nature which enables handling of huge dataset that can fit into memory.3 In the context of our current study, classification trees can be used to provide prediction of dependent classes such as AIV presence versus AIV absence as influenced by the given independent variables. Since our data has been geocoded into specific regions, a regression mathematical relationship for predicting the outbreak events was derived. The logistic regression task involved splitting nodes and pruning trees thus one can represent a region with observations in node as:

Here the features space has been partitioned for each region m. Typically, the training data is numerically represented as and their respective targets ; where and is equivalentto in the above equation.This constitutes the basis of the so-called objective function (training loss and regularization terms).4 The XGBoost internal workings classify a tree model as

Where is the number of trees and is a function which determines, for each iteration, the trees generated in the functional space . An objective function such as this one can then be optimized to produce an efficient generalized model. In principle, we combine many ‘weak’ learners/classifiers to produce better learning trees than using a single classifier.5 This is achieved by greedily adding trees that mostly improve the model using second-order optimization approximations. See ref [3] for further details of XGBoost equations and derivations. We equally performed cross-validation studies using XGBoost internal k-fold strategy. Finally, we construct our models with the resulting tuned parameters.

# **Results**

**Evaluation metrics**

The optimized parameters for each of the models are summarized in *Table SI 3* for reproducibility purposes. The respective graphs of the training versus the cross-validation experiments are also shown in Figure SI 1. A convergence trend is observed in the optimization steps signaling stability of the models. Consequently, the performances of the models on the test dataset were evaluated using these parameters.

|  |  |
| --- | --- |
| ***Table SI 3****: Optimal values for each of the models we constructed. Respective features in the models were engineered hence the subtle differences. These computations were performed on a GPU cluster.* | |
| Model | Objective function |
| M1 | {'verbosity': 0, 'booster':'gbtree', 'objective': 'binary:logistic', 'scale\_pos\_weight': 3.6075131113550434,'tree\_method': 'gpu\_hist','eval\_metrics': 'logloss', 'learning\_rate': 0.09963558437961703, 'num\_boost\_round': 191.58714451312179, 'max-depth': 4, 'gamma': 6.140489891496016e-05, 'subsample': 0.7329970919203579, 'reg\_alpha': 1.5492015574694834e-05, 'reg\_lambda': 0.002889192981174229, 'colsample\_bytree': 0.7341652208091972, 'min\_child\_weight': 0, 'n\_estimators': 443} |
| M2 | {'verbosity': 0, 'booster': 'gbtree', 'objective': 'binary:logistic','scale\_pos\_weight': 3.6063414634146342, 'max\_delta\_step': 1, 'tree\_method': 'gpu\_hist', 'eval\_metrics': 'logloss', 'learning\_rate': 0.09813841335627332, 'num\_boost\_round': 281.7702180112571, 'max-depth': 2, 'gamma': 3.8914577758283693e-05, 'subsample': 0.7341873457611429, 'reg\_alpha': 0.22771113527284387, 'reg\_lambda': 1.6884029915870075e-07,  'colsample\_bytree': 0.850855026540212, 'min\_child\_weight': 3, 'n\_estimators': 473} |
| M3s | {'verbosity': 0, 'booster': 'gbtree',  'objective': 'binary:logistic', 'scale\_pos\_weight': 8.125875815414352, 'max\_delta\_step': 1, 'tree\_method': 'gpu\_hist', 'eval\_metrics': 'logloss',  'learning\_rate': 0.09880726959534625, 'num\_boost\_round': 198.863880711211,  'max-depth': 5, 'gamma': 1.5604867919189752e-08, 'subsample': 0.6569083845212979, 'reg\_alpha': 7.461449079675969e-06, 'reg\_lambda': 8.072448854162746e-08, 'colsample\_bytree': 0.9103345902075206, 'min\_child\_weight': 0, 'n\_estimators': 446} |

|  |  |  |
| --- | --- | --- |
| M1 | M2 | M3 |
|  |  |  |

**Figure SI 1**: Log loss plots for the three constructed models. A convergence trend observed in the epochs shows the stability of the models.

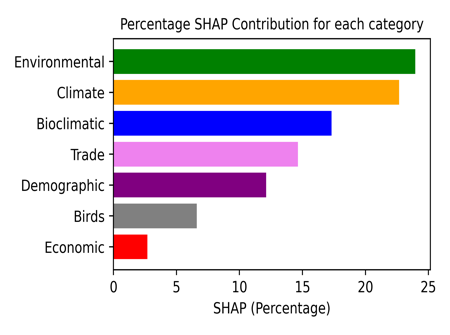
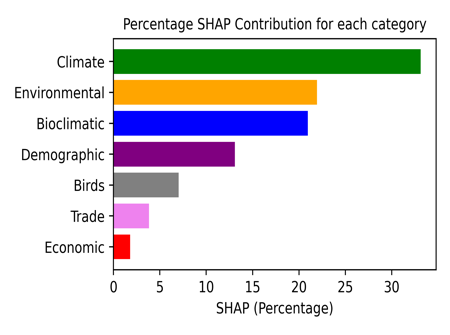
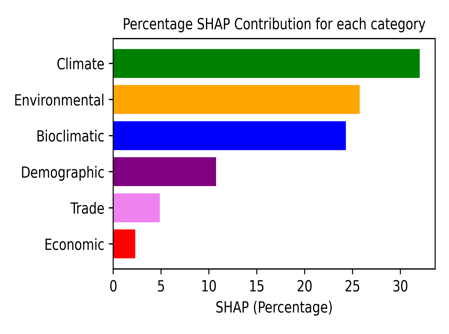
**Key drivers of AIV outbreaks for M1 and M3**

The top global ranking features with a positive impact on M1 are the minimum temperature of the third quarter (min\_temp\_q3), mean temperature of the second quarter (mean\_temp\_q2), water index of quarter one (ndwi\_q1), and high vegetation index of quarter one (lai\_hv\_q1). Disabling the wild bird labels In M3 resulted in a number of observable differences from the M2 model. The validation and the training log loss shift signifying a reduction in the model’s performance on the test data (See Figure 2 in the main text). We also observe differences in the global ranking of SHAP values with loading being the top ranked. The min\_temp\_q3, unloading, ndvi\_q4, ndwi\_q1 and annual GDP follow respectively (Figure SI 2).

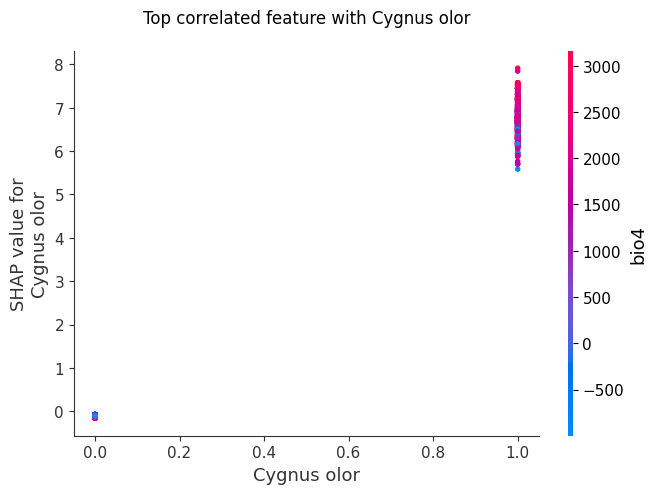
|  |  |
| --- | --- |
| **M1** |  |
| **M3** |  |

**Figure SI 2**: Summary SHAP plots for models M1 and M3. In M1, positive outbreaks are largely driven by climatic factors while in M3, they are driven by a combination of economic and environmental factors.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Table SI 4****: Comprehensive performance report for all the models at thresholds (0.1 to 0.5). As explained in the text, M2 had a well-balanced performance when adjusting the thresholds.* | | | |
| Model | | % True Positives | % True Negatives |
| M1 | |  |  |
|  | 0.1 | 98.72 | 73.58 |
|  | 0.2 | 98.40 | 77.36 |
|  | 0.3 | 97.97 | 81.51 |
|  | 0.4 | 97.76 | 83.77 |
|  | 0.5 | 97.55 | 86.79 |
| M2 |  |  |  |
|  | 0.1 | 99.05 | 72.72 |
|  | 0.2 | 98.11 | 81.34 |
|  | 0.3 | 98.11 | 86.36 |
|  | 0.4 | 98.11 | 89.58 |
|  | 0.5 | 97.64 | 90.73 |
| M3 |  |  |  |
|  | 0.1 | 99.47 | 73.26 |
|  | 0.2 | 99.36 | 79.07 |
|  | 0.3 | 99.26 | 83.33 |
|  | 0.4 | 98.94 | 86.05 |
|  | 0.5 | 98.94 | 89.53 |



**Figure SI 3**: SHAP contribution for each class/category obtained as percentages from the individual features for M1, M2, and M3 respectively.



**Figure SI 4**: Bi-variate SHAP analysis of Cygnus olor with bio4 as the main interaction.

# Discussion

**Evaluation metrics**

From the evaluation plot shown in (Figure 2 main text), it is evident that the introduction of different features meaningfully impacts the predicted outcome of our models. The logloss of our model (M2) drastically reduces i.e., accuracy of prediction increases when the wild birds are used as features. A change in the thresholds confirms its sensitivity to the shifts with a difference of up to 26% (*Table SI 4*). Changing thresholds is a way of creating artificial biasness in cases where the data is imbalanced guaranteeing correct class distributions.6 Sensitivity enables the model to capture any changes in feedback loops either through deviations in the outbreaks or outright bias in the models. In other words, our model is capable of producing consistent results when predicting the presence of AIVs in the event of external data variation. The consistency affirms our model’s reliability to distinguish between two classes of unseen AIV outbreaks from a set of time-series dataset.

M1 served as a generic template for predicting AIV outbreaks. In quarter one, the environmental features, water index and vegetation index, positively drive the outbreaks. During cold seasons, these habitats could contribute to migration of wild birds and result in more interactions with other mammals. This finding is in line with literature where migrating birds, especially those moving over short distances, are reported to easily shed AIV viruses.7 Temperature plays a role in driving AIV events during quarter two as well as quarter three. Like M2, we observe complex relationships where the top 2 ranking variables exhibit a U-shaped pattern (see the shifted blue and yellow plots in Figure SI 2).

Model M3 was constructed by switching off the targets of the wild birds and represents an association between wild birds with highly pathogenic AIVs and domesticated birds while they (wild birds) remain unaffected. This model is akin to the prediction of domesticated outbreaks after being infected by wild birds. Trade (Loading feature) emerged as the most influential global feature. As seen in (Figure SI 2 M3.), this feature negatively impacts the model away from the base and has the interpretation; As trade decreases, there is less likelihood of observing AIVs events. An opposite view is held, implying that more trade would result in increased events. Similar to the above discussed models, the min\_temp\_q3 is the dominant climate variable ranked second after trade. Other top features include Unloading, normalized deviation vegetation index in quarter four (ndvi\_q4) and normalized deviation water index (ndwi\_q1). The mean\_temp\_q2 had the highest local impact. Compared to the previous model, the negative influence of this feature is higher whereby a decrease in mean temperature for this quarter increases the chances of AIV events in most regions.

The introduction of trade as a leading variable highlights the difference between M3 and the previous two models. This influence is confirmed by the overall SHAP percentage contribution as this category ranks above the demographic category (Figure SI 3). Additionally, the environmental category ranks top in the SHPA percentage contribution and not climate as observed in M1 and M2. A clear implication of this model is that one would use trade policies as well as regulations to minimize the impact of AIV outbreak events once detected in wild birds to avoid wide spreads. Recently, efforts have been geared towards reducing invasive species of both wild and domesticated birds within Europe via trade restrictions, quarantine, and black and white blacklistings for non-qualifying companies.8 Interventions can be introduced early enough to mitigate the extent of impact of AIV using available data. Insights form this model introduce an opportunity to strengthen wild bird trade policies which have been implemented in Europe.9

References

(1) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12* (85), 2825–2830.

(2) Farooq, Z.; Rocklöv, J.; Wallin, J.; Abiri, N.; Sewe, M. O.; Sjödin, H.; Semenza, J. C. Artificial Intelligence to Predict West Nile Virus Outbreaks with Eco-Climatic Drivers. *The Lancet Regional Health – Europe* **2022**, *17*. https://doi.org/10.1016/j.lanepe.2022.100370.

(3) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016; pp 785–794. https://doi.org/10.1145/2939672.2939785.

(4) Raschka, S.; Mirjalili, V. *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2*, Third edition.; Expert insight; Packt: Birmingham Mumbai, 2019.

(5) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* **2001**, *29* (5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

(6) Provost, F. Machine Learning from Imbalanced Data Sets 10. **2001**.

(7) Martin, V.; Forman, A.; Lubroth, J. *Preparing for Highly Pathogenic Avian Influenza*, rev. ed.; FAO animal production and health manual; Food and Agriculture Organization of the United Nations: Rome, 2009.

(8) Keller, R. P.; Geist, J.; Jeschke, J. M.; Kühn, I. Invasive Species in Europe: Ecology, Status, and Policy. *Environ Sci Eur* **2011**, *23* (1), 23. https://doi.org/10.1186/2190-4715-23-23.

(9) Cardador, L.; Tella, J. L.; Anadón, J. D.; Abellán, P.; Carrete, M. The European Trade Ban on Wild Birds Reduced Invasion Risks. *Conservation Letters* **2019**, *12* (3), e12631. https://doi.org/10.1111/conl.12631.