BMC Genomics

RESEARCH Open Access

Machine learning assessment of zoonotic potential in avian influenza viruses using PB2 segment

Sangwook Kim[1†], Min-Ah Kim[2†], Bitgoeul Kim[2], Jisu Lee[2], Se-Kyung Jung[2], Jonghong Kim[3], Ho-Young Chung[4], Chung-Young Lee[2,5*] and Sungmoon Jeong[4,6*]

Abstract

Background Influenza A virus (IAV) is a major global health threat, causing seasonal epidemics and occasional pandemics. Particularly, Influenza A viruses from avian species pose significant zoonotic threats, with PB2 adaptation serving as a critical first step in cross-species transmission. A comprehensive risk assessment framework based on PB2 sequences is necessary, which should encompass detailed analyses of specific residues and mutations while maintaining sufficient generality for application to non-PB2 segments.

Results In this study, we developed two complementary approaches: a regression-based model for accurately distinguishing among risk groups, and a SHAP-based risk assessment model for more meaningful risk analyses. For the regression-based risk models, we compared various methodologies, including tree ensemble methods, conventional regression models, and deep learning architectures. The optimized regression model, combined with SHAP value analysis, identified and ranked individual residues contributing to zoonotic potential. The SHAP-based risk model enabled intra-class analyses within the zoonotic risk assessment framework and quantified risk yields from specific mutations.

Conclusion Experimental analyses demonstrated that the Random Forest regression model outperformed other models in most cases, and we validated the target value settings for risk regression through ablation studies. Our SHAP-based analysis identified key residues (271A, 627K, 591R, 588A, 292I, 684S, 684A, 81M, 199S, and 368Q) and mutations (T271A, Q368R/K, E627K, Q591R, A588T/I/V, and I292V/T) critical for zoonotic risk assessment. Using the SHAP-based risk assessment model, we found that influenza A viruses from *Phasianidae* showed elevated zoonotic risk scores compared to those from other avian species. Additionally, mutations I292V/T, Q368R, A588T/I, V598A/I/T, and E/V627K were identified as significant mutations in the *Phasianidae*. These PB2-focused quantitative methods

†Sangwook Kim and Min-Ah Kim these authors contributed equally to this work.

*Correspondence:
Chung-Young Lee
cylee87@knu.ac.kr
Sungmoon Jeong
jeongsm00@gmail.com

Full list of author information is available at the end of the article

provide a robust and generalizable framework for both rapid screening of avians' zoonotic potential and analytical quantification of risks associated with specific residues or mutations.

Keywords Influenza A virus, Avian influenza virus, PB2, Artificial intelligence, Machine learning, SHAP, Mutation analysis

Background

Influenza A virus (IAV) is a major global health threat, responsible for causing seasonal epidemics and occasional pandemics. This virus, belonging to the Orthomyxoviridae family, possesses a segmented RNA genome that facilitates genetic reassortment and rapid evolution. Adaptive mutations in viral proteins facilitate cross-species transmission of influenza viruses. The hemagglutinin (HA) glycoprotein mediates receptor binding and membrane fusion, and host range shifts have been classically attributed to HA mutations altering sialic acid binding specificity [1]. Non-structural proteins such as NS1 and PA-X, which antagonize host antiviral responses, also modulate host adaptation. Crucially, the viral RNA-dependent RNA polymerase complex—composed of PB1, PB2, and PA—plays a central role in host-specific replication [2]. Among these, PB2 is a key determinant of host range, with specific residues modulating polymerase activity, nuclear import, and interaction with host cofactors in a species-dependent manner. In particular, PB2 interacts with host factors such as ANP32A analogues ensuring efficient replication and transcription of the viral genome within host cells [3–6]. Adaptation of PB2 to a novel host is often considered a key prerequisite for zoonotic transmission. Numerous studies have demonstrated that specific amino acid substitutions in avian-origin PB2, such as Q591R or E627K, can enhance polymerase activity in mammalian cells, thereby enabling the virus to replicate efficiently in mammalian hosts [1, 2]. Therefore, the acquisition of mammalian-adaptive mutations in PB2 is often regarded as the critical first step for avian-origin IAVs to overcome host-specific barriers and establish infection in mammalian hosts.

To classify protein sequences of influenza viruses, various machine learning techniques, including deep learning models, have been extensively applied. One approach involves embedding protein sequences into vector representations [7–10]. This methodology aims to construct meaningful vector representations with respect to elements' semantic similarities and can be utilized without retraining once established; however, it necessitates the construction of an additional machine learning model to perform specific tasks such as classification or prediction.

Support vector machine (SVM) [11] is a traditional machine learning algorithm that separates data points in high-dimensional space using optimal hyperplanes and often requires feature extraction techniques such as principal component analysis (PCA) [12]. In predicting viral hosts of influenza A viruses, researchers sequentially applied feature extraction and selection methods to virus strains, and trained SVM models to predict transmission from avian to human hosts [13]. In another study, SVM was effectively constructed with position specific scoring matrix (PSSM) [14], an informative feature extraction technique for protein sequences, although it showed slightly lower performance compared to convolutional neural networks (CNNs) [15].

Random Forest (RF) is an ensemble learning method that consists of multiple decision trees working collectively [16]. For the classification of avian and human influenza protein sequences, RF models were constructed and demonstrated high efficacy, utilizing specialized feature vectors that incorporated both amino acid sequences and their physicochemical properties [17, 18]. CatBoost represents another ensemble learning approach, specifically a gradient boosting algorithm [19]. As its name suggests, CatBoost can process categorical variables directly without requiring conversion to numerical representations. In a study classifying SARS-CoV-2 genome sequences, CatBoost outperformed other models including SVM, Random Forest, and logistic regression, while maintaining competitive training speeds [20].

Logistic regression is a fundamental statistical method suitable for binary classification problems where the goal is to predict one of two possible outcomes [21]. It is widely implemented across various fields, particularly in medicine for disease diagnosis. Notably, a comprehensive review found no significant performance advantages of complex machine learning models, including Random Forest and SVM, over logistic regression (including its regularized variants such as LASSO and ridge regression) for clinical prediction models [22]. Several studies have employed logistic regression to evaluate risk factors associated with influenza viruses [23, 24]. *k*-Nearest Neighbors (KNN) algorithm is used for classification and regression tasks, and known to be asymptotically Bayes-optimal as the dataset size increases [25, 26]. This model operates without a learning phase since it relies on pairwise metric comparisons against stored data samples and requires storing the entire dataset in memory. Due to this property, it naturally accepts additional data points while its prediction phase demands relatively long computation time for point-by-point comparisons. For the host classification of avian influenza viruses, KNN outperformed Naïve Bayes, decision trees, and SVM classifiers in F1 score evaluation [27].

CNNs, by contrast, represent one of the most prominent deep learning approaches [28]. The inductive biases of CNNs are particularly well-suited for tasks involving translation-invariant signals with local structures, such as image recognition. However, CNNs have limited application in direct protein sequence-based risk assessment, as the positional information of proteins is more critical than local sequence structures, despite some successful attempts in predicting local structures such as protein folding [29, 30].

Recurrent Neural Networks (RNNs), which specialize in processing sequential data, have been widely implemented in time-series predictions and natural language processing [31–33]. While RNNs such as Long Short-Term Memory (LSTM) excel at detecting latent patterns within sequences, they do not explicitly incorporate positional information, which is crucial in protein sequence analysis. Nevertheless, RNNs have demonstrated significant success in various protein analysis applications [34, 35].

Transformer architectures, characterized by their self-attention mechanisms and parallel processing capabilities, have been successfully implemented across diverse domains, including text generation and image recognition [36]. These models incorporate positional encoding to preserve the sequential information of input tokens. Transformers have been successfully applied to protein sequence representations [37]. Another notable example is AlphaFold 3, a transformer-based model that has achieved unprecedented accuracy in protein structure prediction from amino acid sequences, which was awarded the Nobel Prize in Chemistry in 2024 [38].

Explainable AI is one of the critical topics in the machine learning field, particularly essential for sensitive domains such as medical applications. In line with this need, analysis based on SHAP (SHapley Additive exPlanation) values provides consistent and theoretically grounded feature attribution for predicted outputs, based on Shapley value estimation from game theory [39]. Specifically, Tree SHAP was proposed for tree ensembles like CatBoost [19] and Random Forest [16]. Tree SHAP is an efficient algorithm for estimating SHAP values for tree ensembles, and it maintains consistency whereas existing attribution methods for tree ensembles are considered inconsistent [40].

In this study, we collected PB2 amino acid sequences from both avian and human influenza A viruses to assess the zoonotic potential of avian strains. We developed two risk assessment models. Initially, we defined the risk assessment as a regression problem with three ordinal risk groups consisting of low-risk (avian), mid-risk (cross), and high-risk (human) samples. For the regression-based approach, we compared various regression methods, including tree ensemble methods (CatBoost and Random Forest), conventional regression models, and deep learning architectures [41]. Then, we validated the target value of the mid-risk group for risk modeling through an ablation study. The second approach was a risk model based on SHAP value analyses. Using SHAP values, we ranked residues based on their contribution to risk assessment, and we assessed risk for samples and quantified the effects of mutations using aggregated SHAP values. Furthermore, we analyzed avian virus samples by examining the distribution of quantized risk groups within the avian population.

Methods

Regression-based risk modelling

To assess the zoonotic infection risk of influenza A viruses of avian origin, we analyzed PB2 protein sequences from three distinct sources: avian influenza viruses, human cases of avian influenza, and human influenza viruses. We categorized these sequences into three risk groups based on their host origin: PB2 sequences from avian influenza viruses, human cases of avian influenza, and human influenza viruses were classified as low-, mid-, and high-risk groups, respectively. Since the values of each group are not rigorously defined but the risks should maintain a relative order among groups, risk modeling in the three groups can be formalized as an ordinal regression problem. Let R_i represent the risk value of the i-th d-dimensional sequence vector $\mathbf{x}_i$, and let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a function that measures the risk for a given $\mathbf{x}_i$:

$$R_i = \Phi\ (\mathbf{x}_i)\,;\ i = 1,\ ?,\ N, \tag{1}$$

with N being the total number of samples, $R_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^d$, and d denoting the length of the sequence. The ideal function Φ should satisfy the condition:

$$\Phi\ (\mathbf{x}_l) < \Phi\ (\mathbf{x}_m) < \Phi\ (\mathbf{x}_h)\,;\ \forall\ (\mathbf{x}_l, \mathbf{x}_m, \mathbf{x}_h) \in \mathrm{G}_l \times \mathrm{G}_m \times \mathrm{G}_h \tag{2}$$

,

where G_l, G_m, and G_h represent low-, mid-, and high-risk groups, respectively. If we construct a regression model with sequences G_l, G_m, and G_h, and corresponding target values t_l, t_m, and t_h satisfying such that $t_l < t_m < t_h$, a resulting regression model $\Phi\prime$ implementing the following equation will also solve the ordinal regression in Eq. 2:

$$\Phi'\ (\mathbf{x}_i) = R\prime_i = \begin{cases} t_l & \forall\, \mathbf{x}_i \in \mathrm{G}_l \\ t_m & \forall\, \mathbf{x}_i \in \mathrm{G}_m \\ t_h & \forall\, \mathbf{x}_i \in \mathrm{G}_h \end{cases}, \quad s.t.\ t_l < t_m < t_h. \tag{3}$$

In this study, we assign categorical status values t_l, t_m, and t_h as proxy target values 0, 0.5, and 1, respectively, for risk regression. With this configuration, we can apply various machine learning models and ordinary regression techniques. Table 1 summarizes categories of sequences and their risk levels.

Risk modelling with SHAP values

Although regression method is a natural approach for risk modeling, it alone cannot quantify and analyze detailed contributions to risk scores. SHAP values provide consistent and theoretically grounded feature attribution based on Shapley value estimation from game theory. In this study, SHAP values were employed to quantify the contribution of residues and effects of mutations to the risk value. While SHAP values explain local predictions limited to each sequence, properly aggregated SHAP values can be derived to quantify the expected risk values related to each feature or mutation. Although mean absolute values of SHAP values are widely accepted as feature importance measures, they do not provide the expected risk for each feature's values. This subsection investigates how aggregated SHAP values can reconstruct risk values without relying on regression models. To develop risk scores, we calculated the expectation of SHAP values. The risk score with aggregating SHAP values for *j*-th one-hot-encoded sequence, $\Phi'_{\text{shap}}(\mathbf{x}_j)$, can be modeled as Eqs. 4 and 5:

$$w_i^+ = \frac{\sum_{j=1}^{N} v_{ij} \times x_{ij}}{\sum_{j=1}^{N} x_{ij}}, \quad w_i^- = \frac{\sum_{j=1}^{N} v_{ij} \times (1-x_{ij})}{\sum_{j=1}^{N} (1-x_{ij})} \tag{4}$$

.

$$\Phi'_{\text{shap}}(\mathbf{x}_j) = \mathbf{x}_j \bullet \mathbf{w}^+ + (1-\mathbf{x}_j) \bullet \mathbf{w}^- + b \tag{5}$$

.

These equations formulate the risk scoring mechanism, where w_i^+ and w_i^- represent the expected SHAP values for the presence and absence of the *i*-th feature, respectively. Here, v_{ij} denotes the SHAP value of the *i*-th feature for the *j*-th sample, *N* represents the total number of samples, and *b* is the baseline prediction value. Notably, the separate aggregation of SHAP values for feature presence and absence is essential to prevent mutual cancellation effects.

While the regression performance of SHAP-based risk modeling may be inferior to conventional regression models (as SHAP values are derived from a regression model), it offers several distinct advantages. First, although we defined three risk groups with corresponding target values, this classification does not account for relative risks within groups, as the regression model attempts to map each group's distribution to a single target value. In contrast, risk modeling with SHAP values provides aggregated risk values comprising multiple risk components, rather than mapping risks to discrete target points. This approach effectively distinguishes high-risk samples from low-risk samples within the same group, enabling detailed intra-group analyses. Second, once risk terms are established, we can evaluate the risk for any given sample without requiring additional predictions from the regression model structure. Finally, risk contributions can be immediately decomposed without recalculating SHAP values, thus providing enhanced interpretability for the risk assessment.

Table 1 Definition of sequence groups and their risk levels

Source Type	Abbreviation	Risk Group	Target Value
Avian Influenza Viruses	Avian	Low-risk (G_l)	0.0
Avian Influenza Viruses from Human	Cross	Mid-risk (G_m)	0.5
Human Influenza Viruses	Human	High-risk (G_h)	1.0

Assessment of mutation-related risks

Mammalian Pathogenicity-related Mutations (MPMs) are represented using the commonly used colloquial nomenclature, such as T271A [42]. In this notation, T271A indicates that the mutation has changed the amino acid residue threonine (T) at position 271 to alanine (A). We calculate the quantitative measure of mutation risk changes using Eq. 6.

$$\xi_p(r_s, r_d) = w^+_{(p,d)} - w^-_{(p,d)} - w^+_{(p,s)} + w^-_{(p,s)} \tag{6}$$

In this formulation, the position-specific mutation risk score $\xi_p(r_s, r_d)$ quantifies the impact of amino acid substitution at position *p*, where r_s and r_d represent the source and the destination residues, respectively. This score is calculated as the difference between the positive and negative interactions of both the destination (*d*) and source (*s*) residues. Specifically, $w_{(p,d)}{}^+$ and $w_{(p,d)}{}^-$ represent the expected SHAP values for the presence and absence of the destination residue at position *p*, and vice versa for the source residue ($w_{(p,s)}{}^+$ and $w_{(p,s)}{}^-$). Note that subscript *i* in Eq. (4) represents the index in one-hot encoded feature dimensions, where each pair (*p*, *s*) and (*p*, *d*) determines its corresponding index *i* based on the position *p* of a residue in the sequence. This scoring function enables the evaluation of residue mutations by quantifying the differential impact of residue presence versus absence through SHAP analysis.

Table 2 Number of samples and unique sequences by risk group

	Low-risk	Mid-risk	High-risk	Total
Samples	34,896	1751	148,883	185,530
Unique Sequences	16,312	1011	25,144	42,467

Data collection and Preparation

A total of 185,530 full-length PB2 amino acid sequences were obtained from GISAID (Global Initiative on Sharing Avian Influenza Data) [43]. PB2 sequences from human seasonal IAVs, including H1N1(59.8%) and H3N2 (40.2%), were classified as 'human' and high-risk. PB2 sequences from avian IAVs, including H5N1 (28.3%) and H9N2 (12.4%) were classified as 'avian' and low-risk. For a more robust assessment, PB2 sequences from human influenza viruses excluding seasonal IAVs were classified as 'cross' and mid-risk, including H7N9 (64.7%) and H5N1 (23.0%). Subtypes representing less than 10% of their respective categories are detailed in Supplementary Table S1.

Collected PB2 sequences have a length of 759 amino acids, represented by 20 standard single-letter amino acid codes (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y) and X, where X denotes unknown or ambiguous amino acids.

Table 2 shows the number of samples and unique sequences in each category. As this study relies exclusively on sequence information, we created a dataset containing only unique sequences rather than using all samples with their duplicate sequences.

Preprocessing and regression models

For regression modeling of the risk, we used linear regression, LASSO regression, ridge regression, KNN regression, Multi-Layer Perceptron (MLP), Stacked Bi-directional Long Short-Term Memory (Bi-LSTM) with Fully Connected Network (FCN), Random Forest, CatBoost and TabNet. We utilized pre-implemented models from *scikit-learn* (v1.5.2) [44] for most experiments, implemented the Stacked Bi-LSTM with FCN using *TensorFlow 2* (2.18.0), and employed CatBoost through the *catboost* package (v1.2.7) and TabNet through *pytorch-tabnet* (v4.1). To evaluate the generalization performance, 5-fold stratified cross-validation was performed with random initialization and the same folds were used across all models for fair comparison. Full details of model hyper-parameters can be found in the supplementary materials.

Since amino acid sequences consist of categorical letters rather than numerical data, we encoded them as one-hot vectors for all models except CatBoost and Stacked Bi-LSTM with FCN. Given that elements in one-hot encoding are binary (0 or 1), no additional scaling was performed on the input vectors. The resulting feature vectors have a dimension of 15,939 (21×759), where each sequence contains exactly 759 ones with the remaining elements being zeros. For CatBoost regression, since the model can directly handle categorical features without numerical embedding, the amino acid sequences were used as input without transformation, maintaining the original sequence length of 759. For Stacked Bi-LSTM with FCN, we applied an embedding layer for amino acids at the bottom layer of the model, which can be constructed via end-to-end training, instead of encoding input sequences as one-hot vectors.

Table 3 Regression performances

Method	MSE	MSE-clipped
Linear regression	> 10,000	0.0225 ± 0.0018
LASSO regression	0.0056 ± 0.0004	0.0056 ± 0.0004
Ridge regression	0.0043 ± 0.0003	0.0038 ± 0.0003
KNN regression	0.0033 ± 0.0004	0.0033 ± 0.0004
MLP	0.0042 ± 0.0005	0.0036 ± 0.0004
Stacked Bi-LSTM w/ FCN	0.0197 ± 0.0043	0.0196 ± 0.0043
Random Forest	**0.0027 ± 0.0004**	**0.0027 ± 0.0004**
CatBoost	0.0032 ± 0.0004	0.0032 ± 0.0004
TabNet	0.0123 ± 0.0010	0.0122 ± 0.0010

Results

Performance comparison of regression models

We evaluated the performance of the constructed models using Mean Squared Error (MSE) across 5-fold cross-validation. While MSE was not originally designed specifically for ordinal targets, previous research has demonstrated its effectiveness as a meaningful metric for ordinal regression with imbalanced datasets [45]. Since the target risk values are constrained to {0, 0.5, 1}, we clipped the predicted values to the range [0, 1] before calculating MSE (hereafter referred to as MSE-clipped). This clipping operation is justified as any prediction outside this range would be meaningless in our context, where the risk values are naturally bounded. Additionally, this approach prevents the MSE from being dominated by extreme predictions that fall outside the valid range, thereby providing a more interpretable measure of model performance. For instance, while linear regression showed extremely high MSE values (> 10,000) before clipping, the clipped MSE provided more meaningful and interpretable results that better reflected the model's practical performance. The performance results of all models are presented in Table 3. The results show that the Random Forest-based model achieved the best performance, followed by CatBoost in second place. As mentioned in the Background section, tree ensemble models demonstrated superior performance compared to other methods. KNN regression showed slightly worse performance than CatBoost, but better than MLP.

Classification performances can be evaluated by testing risk group membership, which is determined by assigning samples to their closest representative group value. For instance, a sample with a risk value of 0.4 would be allocated to the mid-risk group, as its representative value (0.5) is the nearest to the measured risk value.

Table 4 presents the performance metrics for each model in terms of accuracy, weighted Cohen's Kappa [46], Macro F1 [47], and Custom-weighted F1. While all models achieve nearly perfect accuracy scores (>0.95), this metric alone is insufficient due to significant class imbalance in our dataset, where the mid-risk group constitutes only 2.4% of the total samples. To address this limitation and ensure rigorous evaluation, we employed three complementary metrics, each serving a distinct analytical purpose. Weighted Cohen's Kappa was selected as a evaluation metric for its sensitivity to ordinal relationships between risk groups. The metric incorporates linear weighting where weights are assigned based on the absolute difference between category indices. These weights quantify the importance of agreements and disagreements at different ordinal distances, with larger weights given to disagreements between more distant categories. The metric ranges from ≤0 (chance-level agreement) to 1.0 (perfect agreement). To capture class-specific performance variations, we utilized macro F1, which averages F1 scores across all classes with equal weights regardless of their sample sizes. This provides insights into model performance on minority classes that might be obscured by accuracy or micro-averaged metrics. Additionally, we introduced a custom-weighted F1 score that assigns weights of 0.5 to the mid-risk group and 0.25 to other groups, motivated by both clinical and technical considerations: (1) mid-risk cases require the most nuanced intervention planning, (2) they are underrepresented in the training data, and (3) their feature distributions show substantial overlap with neighboring risk groups, making classification particularly challenging.

As shown in Tables 3 and 4, the Random Forest-based regression model again achieved the highest performance across all metrics, while deep learning models exhibited substantially lower predictive power in this task. Two possible explanations for the limited performance of deep learning models are: (1) insufficient data samples to train the deep architectures effectively, and (2) the classification of PB2 sequences in this task may not necessitate modeling of complex feature interactions, which is one of the key advantages of deep learning models. Interestingly, the shallow neural network (MLP) demonstrated better performance than the deep learning models in this case, supporting the aforementioned explanations. Since TabNet is capable of unsupervised pretraining, there is potential for improving the model's performance if such pretraining becomes feasible. Stacked Bi-LSTM with

Table 4 Classification performances

Method	Accuracy	Weighted Cohen's Kappa	Macro F1	Weighted F1
Linear regression	0.965 ± 0.002	0.944 ± 0.004	0.856 ± 0.011	0.800 ± 0.016
LASSO regression	0.976 ± 0.004	0.974 ± 0.004	0.811 ± 0.017	0.724 ± 0.025
Ridge regression	0.984 ± 0.001	0.983 ± 0.002	0.875 ± 0.013	0.817 ± 0.018
KNN regression	0.986 ± 0.002	0.984 ± 0.002	0.887 ± 0.015	0.835 ± 0.022
MLP	0.987 ± 0.001	0.986 ± 0.001	0.894 ± 0.014	0.845 ± 0.021
Stacked Bi-LSTM w/ FCN	0.925 ± 0.015	0.920 ± 0.017	0.728 ± 0.018	0.617 ± 0.026
Random Forest	**0.989 ± 0.001**	**0.988 ± 0.001**	**0.918 ± 0.008**	**0.880 ± 0.012**
CatBoost	0.987 ± 0.001	0.986 ± 0.002	0.897 ± 0.012	0.850 ± 0.017
TabNet	0.960 ± 0.003	0.968 ± 0.002	0.837 ± 0.007	0.766 ± 0.010

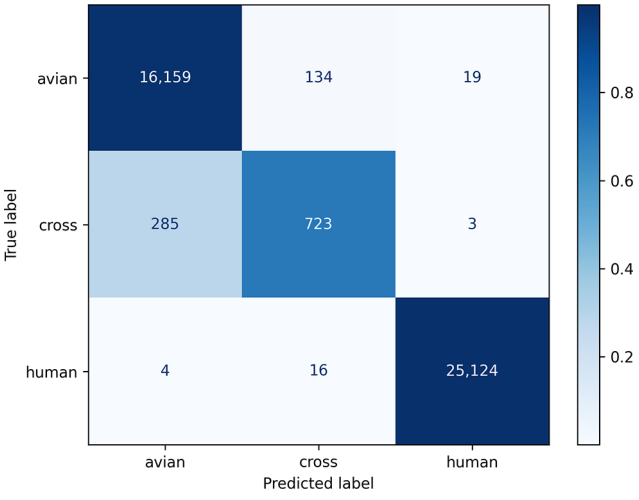

Fig. 1 Confusion matrix of random forest

FCN showed significantly lower performance in distinguishing the minor category (mid-risk group), as shown in F1 and Weighted F1 scores. This underperformance highlights the crucial role of positional information in protein sequence modeling. MLP, furthermore, despite showing worse regression performance than KNN, it demonstrated higher performance across all classification performance measures than KNN.

The aggregated confusion matrix across five cross-validation folds (out-of-folds) from Random Forest for the three groups is shown in Fig. 1. As shown in the results above, classification accuracies for low-risk (avian) and

high-risk (human) groups are nearly perfect, while the model has relatively more difficulty in distinguishing mid-risk (cross) group samples from the low-risk group. Table 5 illustrates the precision, recall, and F1 score for each group.

The Receiver Operating Characteristic (ROC) curve is typically used to represent detection performance. While it can be applied to binary classification by defining a positive class, its direct application to multi-category classification is challenging. Although ROC curves for multi-category classification can be drawn in parallel, using a one-vs-rest approach, this method was not applicable in our study because our categories (risk groups) have ordinal relationships. To address this issue, we designed two separate ROC curve plots: one for low-risk vs. mid-risk and another for mid-risk vs. high-risk. Since our categories are ordinal and risk values are one-dimensional, the threshold distinguishing mid-risk and high-risk groups is independent of the decision threshold between low-risk and mid-risk groups, and vice versa. Let θ_1, and θ_2 be the decision thresholds determining the boundaries between low-risk and mid-risk, and mid-risk and high-risk categories, respectively. For classification metrics shown in Table 5, we simply set $\theta_1 = 0.25$ and $\theta_2 = 0.75$, as these values halve the ranges between $[t_l, t_m]$ and $[t_m, t_h]$. The values θ_1 and θ_2 can be selected independently, since they should satisfy the inequality $t_l < \theta_1 < t_m < \theta_2 < t_h$, in general. In other words, changes in θ_1 cannot affect decisions for values between $[t_m, t_h]$ because θ_1 is bounded above by t_m. Based on this property, we constructed separate ROC curves for low-risk vs. mid-risk and mid-risk vs. high-risk classifications.

Table 5 Precision, recall and F1-score for each group

Group	Precision	Recall	F1 score
Low-risk	0.98	0.99	0.99
Mid-risk	0.83	0.72	0.77
High-risk	1.00	1.00	1.00

Figure 2 shows the ROC curves and their corresponding AUROC (Area Under the Receiver Operating Characteristic) scores for the five models with the highest macro F1 scores. The lines represent the mean values across cross-validation folds, and the colored bands indicate the variability (± standard deviation). In Fig. 2A and B, the positive class was designated as the *cross* (mid-risk) and *human* (high-risk) groups, respectively. While Random Forest showed the best performance in aforementioned regression and classification metrics, CatBoost, MLP and ridge regression models achieved higher AUROC scores than Random Forest. Figure 2A reveals that these three models particularly outperformed Random Forest in regions with high true-positive rates. In Fig. 2B, all models achieved nearly perfect scores for discriminating between cross and human groups.

Ablation study on target values of the mid-risk group

Since risk target values are currently assigned as cardinal numbers without rigorous theory or practical heuristics, an ablation study on the mid-risk group's target values is required to gain insights into their effects and validate the designations. Based on the comprehensive ablation experiments conducted on various target values for the mid-risk group (detailed analysis provided in Supplementary Materials with Fig. S1 and Fig. S2), we found that

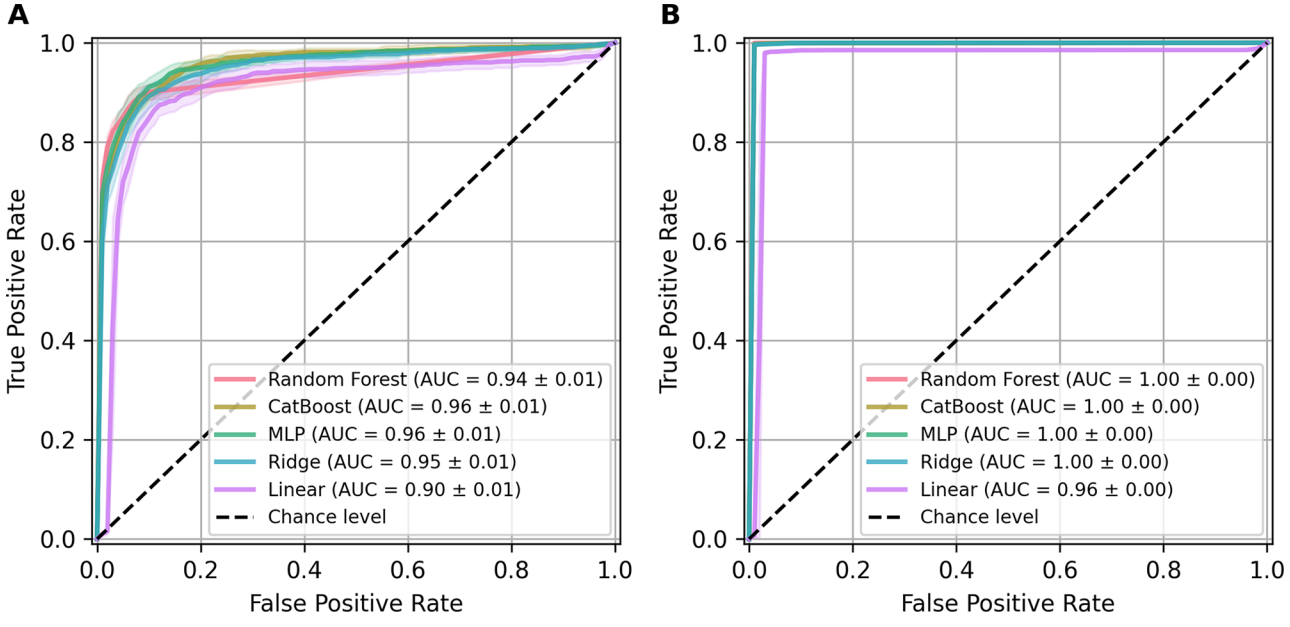

Fig. 2 ROC Curves for Binary Classifications; (**A**) Avian versus Cross; (**B**) Cross versus Human

selecting a mid-risk target value of 0.5 appears to be a reasonable initial benchmark, particularly from the perspective of numerical optimization. Random Forest models demonstrated notably robust and stable performance across different target values compared to other models. However, we observed that higher target values for the mid-risk group tend to result in larger regression errors, primarily due to the challenging nature of discriminating between avian and cross categories. This phenomenon causes models to focus disproportionately on this difficult boundary region at the expense of optimizing the relatively simpler task of distinguishing between cross and human categories. Further theoretical and experimental investigation may be needed for a more rigorous justification of the target value selection.

Impact of retaining sequences with ambiguous residues on model performance

In our dataset preparation process, we made a deliberate decision to retain sequences containing ambiguous residues ('X'), contrary to some conventional approaches that exclude such sequences [15, 17]. This decision was supported by our experimental validation, which demonstrated that the inclusion of these sequences consistently enhances model performance in F1 and weighted F1 scores while several models showed slightly (about 1% point) enhanced performances in metrics not considering class-imbalance. Although ambiguous residues are traditionally considered indicators of poor sequence quality, our analysis revealed that the non-ambiguous portions of these sequences contribute valuable contextual information to the model. This phenomenon aligns with principles observed in transfer learning paradigms [48–50], where partial data can still provide meaningful signals for model training. The value of these contributions is particularly evident in our study, where the limited sample size in *cross* category examples creates a scenario that benefits from leveraging all available information, even if incomplete. Performance metrics of experiments without sequences containing ambiguous residues are provided in Supplementary Table S2.

Impact of subtype hold-outs on model performance

While our initial 5-fold cross-validation demonstrated promising generalization capabilities across randomly partitioned data, we recognized the importance of evaluating our models under conditions that more closely simulate real-world applications. In influenza surveillance and risk assessment, models are frequently required to make predictions on emerging viral subtypes with genetic compositions that may differ from those in the training data. To address this challenge and to further assess the robustness of our models against potential overfitting, we implemented an evaluation strategy based on influenza subtype hold-outs. In this approach, we systematically excluded all PB2 sequences from particular influenza subtypes from the training set of Random Forest and used them exclusively for validation. In this scenario, if a subtype has samples less than ten in a risk group, it is considered as a novel viral subtype and included in the hold-out set. This evaluation method provides a more realistic assessment of model performance in scenarios where predictions must be made on novel viral subtypes not represented in the training data, thereby offering deeper insights into the practical utility of our approach for influenza surveillance systems.

Our analysis encompassed 39 distinct subtypes, yielding MSE, accuracy, macro F1, and weighted F1 scores of 0.001, 0.995, 0.832, and 0.583, respectively. The substantial difference between macro F1 and weighted F1 scores indicates a performance discrepancy across classes. Specifically, the *cross* category was severely impacted by the subtype constraining approach, while the well-represented class *avian* received favorable classification bias, contributing to an overall accuracy score higher than in the cross-validation cases.

The model's ability to maintain low MSE (0.001) and high accuracy (0.995) when tested on previously unseen subtypes suggests resistance to severe overfitting. However, the considerably lower weighted F1 score (0.583) reveals that despite successfully classifying most samples, the model struggles with underrepresented classes when confronted with novel subtypes. These results align with our previous findings regarding the beneficial impact of incorporating incomplete data, including sequences with ambiguous residues, particularly for the minority category *cross*. While our approach demonstrates robust generalization capabilities overall, these findings highlight the ongoing challenge of addressing class imbalance when developing predictive models for emerging viral subtypes.

Distribution of regression-based risk assessments

As shown above, Random Forest outperforms other methods in both risk value regression and risk group classification tasks, although this advantage diminishes in regions of high true positive rates in the ROC curve. Risk measurements from the Random Forest models, evaluated on cross-validation test sets, are shown in Fig. 3. Note that the bars represent probability density rather than histogram counts to account for the significant class imbalance among risk groups. As expected, samples in the *cross* category are frequently misclassified as belonging to the *avian* category, since these samples are avian influenza viruses that have acquired virulence in human hosts. While a small proportion of the *cross* category's risk scores are spread across a wide range, most samples

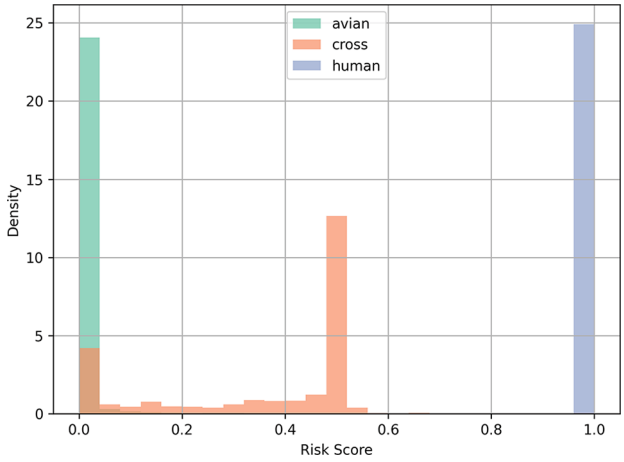

Fig. 3 Risk score distribution by risk group

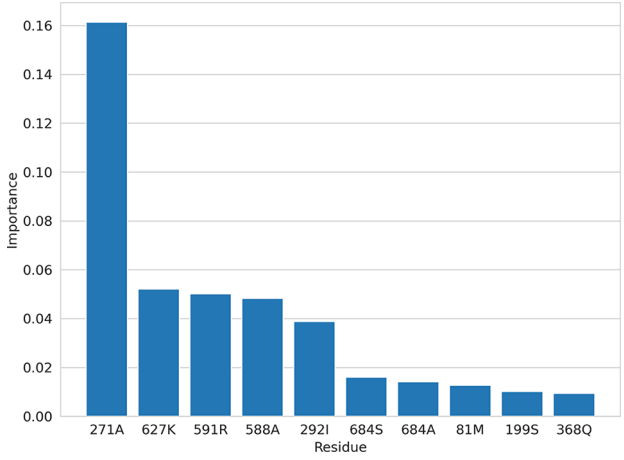

Fig. 4 Ten most important residues identified by SHAP values

from all three categories are concentrated within their corresponding narrow value bands.

Feature attribution with SHAP values

The effectiveness of SHAP values relies on the performance of the regression model. Throughout previous experiments, Random Forest demonstrated superior results compared to other models, and we subsequently calculated SHAP values using this Random Forest model. Fig. S3 in the supplementary materials shows the means and standard deviations of mean absolute SHAP values across cross-validation. Figure 4 presents the ten most important residues in the risk model, which was trained on the complete dataset without a validation set holdout for more accurate analyses. The importance of residues was quantified using mean absolute SHAP values. Notably, the ranking of these top ten residues remained consistent between the cross-validation SHAP analysis and the SHAP analysis conducted on the complete dataset. The amino acid residues identified as the top ten most significant features were 271A, 627K, 591R, 588A, 292I, 684S, 684A, 81M, 199S, and 368Q. Notably, five of these residues– 271A, 627K, 591R, 588A, and 199S– have been previously characterized as critical determinants of viral host specificity through conventional experimental approaches [51]. Figure 5 shows the distributions of SHAP values for all samples, avian, cross, and human groups, with each panel highlighting the ten most important residues for its corresponding category. SHAP values represent the contributions of features to the output, where positive values indicate increased risk values, while negative SHAP values indicate decreased risk. In this Random Forest model, input features were binary (one-hot encoded) vectors, input feature values are colored as blue (0) or red (1). For instance, in Fig. 5A, the presence of 271A (amino acid A at position 271) is shown in red on the positive side, while its absence is blue and located on the negative side. This indicates that the presence or absence of 271A significantly affects the risk value in numerous cases, establishing it as an important risk factor. As shown in Fig. 5B and C, and 5D, the order of feature importance varies among groups, although feature 271A consistently ranks as the most important feature in all cases. The complete ranking of the 100 most important features and their corresponding mean absolute SHAP values is provided in Supplementary Table S3. In Table S4, using the top-ranked 100 features, compact models of Random Forest, MLP, and ridge regression were constructed and evaluated to validate feature importances. Despite the feature dimensionality being reduced from 15,939 to 100 (less than 0.7%), the regression and classification performances of the compact Random Forest were very close to the models with full features. This suggests that the selected top 100 features successfully capture the most essential information for prediction, confirming the effectiveness of our feature importance ranking approach.

Risk model with SHAP values

Individual SHAP values provide local explanations limited to single predictions and offer a theoretical framework for measuring features' contribution to the output. In the above subsections, SHAP values were aggregated to measure feature importances across samples. This subsection, however, focuses on distribution of aggregated SHAP values to understand their ability to reconstruct output values. Figure 6 illustrates the distribution of risk values derived from the aggregation of SHAP values as defined in Eq. 5.

Interestingly, while SHAP values were primarily developed as a framework for interpretability and explainable AI, their careful aggregation also enabled effective modeling of zoonotic risk. Consistent with our previous results, distinguishing between avian and cross groups proved more challenging than identifying human group.

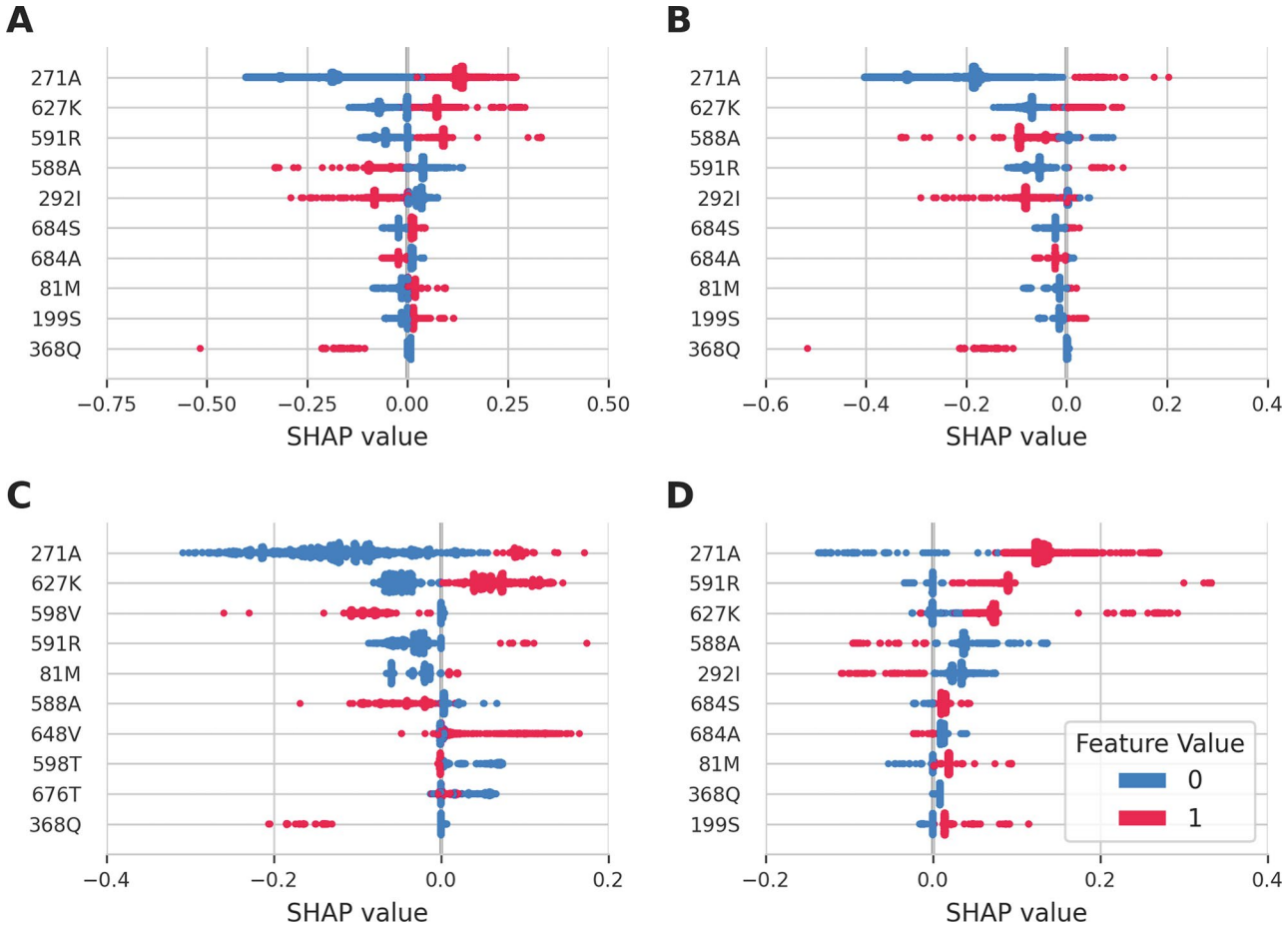

Fig. 5 SHAP values on features; (**A**) All groups; (**B**) Avian; (**C**) Cross; (**D**) Human

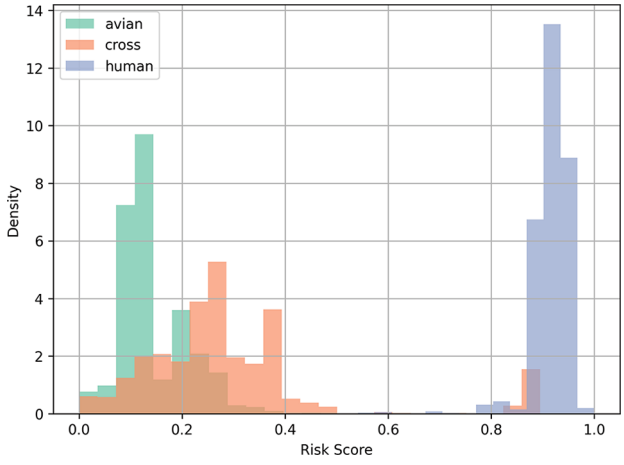

Fig. 6 Distribution of aggregated risk scores derived from SHAP values

The AUROC scores were 0.826 for *avian* versus *cross* classification and 0.997 for cross versus human classification. Although the classification performance across risk categories is comparatively suboptimal compared to high-performing regression-based risk assessment models, this risk model enables both the analysis of relative significance within each risk category and the assessment of sample similarities between groups. This approach facilitates intra-class analyses such as species-specific and family-specific risk assessments within a group, as it does not constrain the risk scores and corresponding SHAP value weightings within each risk group to single target values.

Quantitative analysis of mutation effects

In this experiment, we quantified the risk yield of mutations according to Eq. 6. We specifically focused on missense mutations, excluding other mutation types such as nonsense mutations (e.g., T271X). To ensure statistical reliability, we analyzed only residues that appeared in at least 1% of the total samples ($n \geq 426$) as either source or destination residues. The highest scoring mutations were: T271A, Q368K/R, Q591R, E627K, A588T/I/V, I292V/T. Although D701N is one of the key factors in surpassing the host barrier [51, 52] and had a higher mutation score than I292V/T, it has not been listed because the count of D701N in our dataset was 169, less than 1% of the total samples. Supplementary Table S5 presents the 20 most

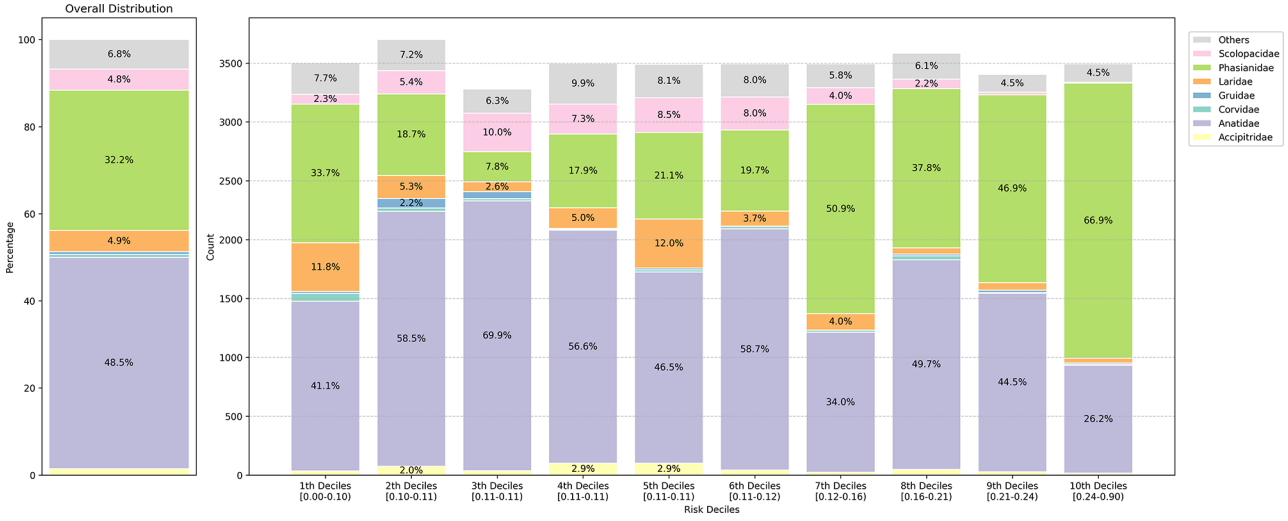

Fig. 7 Distribution of avian families by risk deciles

significant mutations identified in this study, including their mutation scores and occurrence counts.

Distribution of risk scores for avian sequences

To analyze the distribution of risk scores by avian categories, we assigned a scientific family name for each sample of avian category data. If a sample contains insufficient information such as 'wildbird', 'avian', 'bird', 'seabird' and so on, the sample is categorized into 'unknown' family. Typos like 'chichen' (for 'chicken') were manually corrected. Our categorization was conducted based on *The eBird/Clements checklist of Birds of the World*, version of 2024-October [53]. Figure 7 presents the avian family distribution based on risk scores derived from SHAP analyses, and we can observe that *Phasianidae* occupies a large proportion of high-risk deciles compared to other avian families. Based on this observation, we hypothesized that poultry has an important role in zoonotic infections.

Based on our observations, we evaluated risk-associated mutations in *Phasianidae* as given in Eq. 6. Only residues present in at least 1% of *Phasianidae* were considered in the analysis. The mutations identified with the highest risk were I292V/T, Q368R, A588T/I, V598A/I/T, and E/V627K.

Structural modeling of PB2 protein

A consensus PB2 amino acid sequence was generated from avian PB2 sequences and used for protein structure prediction with AlphaFold 2 [54], implemented through ColabFold v1.5.5 [55]. The resulting predicted structure was subsequently visualized and analyzed using UCSF ChimeraX [56]. Key residues identified with SHAP analysis are shown on the predicted structure of the PB2 sequence in Fig. S4. As shown in the figure, all residues are surface-exposed and distributed across the protein, although 588A, 591R, and 627K are spatially clustered. This suggests that multiple mechanisms may underlie their contribution to zoonotic potential.

Discussion

This study evaluated the zoonotic risk of avian influenza viruses through analysis of PB2 sequences from avian, human, and cross-species strains. However, the current model's predictive capacity is limited when confronted with influenza sequences from other hosts, such as swine, potentially leading to unreliable outcomes. Moreover, our current analyses only examined the PB2 segment independently, without considering the potential synergistic effects between multiple genomic segments. These inter-segment interactions likely play crucial roles in determining viral fitness, host adaptation, and zoonotic potential. Future research should expand the scope to encompass influenza viruses from diverse animal hosts, incorporating all eight genomic segments (PB2, PB1, PA, HA, NP, NA, M, and NS). This comprehensive approach will inherently increase the complexity of the analysis, as a single-dimensional risk regression model may prove inadequate for characterizing more heterogeneous viral populations. Additionally, such research should not only analyze individual segments in isolation but also investigate how mutations across different segments collectively influence viral phenotypes through cooperative or compensatory mechanisms.

From a methodological perspective, our comparative analysis of machine learning and statistical approaches revealed that tree ensembles like Random Forest and CatBoost consistently outperformed other methods across multiple metrics, while deep learning models exhibited notably inferior performance. Tree ensemble models have demonstrated superior efficiency on

tabular datasets compared to neural networks such as multi-layer perceptrons or Transformer networks [57], although certain studies suggest that MLP variants or tailored Transformer models can be competitive with tree ensemble models for tabular datasets [58, 59]. Amino acid sequences differ from typical tabular data in their fundamental structure: sequences maintain a critical ordering of elements, while tabular data are generally permutation-invariant. Despite this distinction, we opted for tree ensemble models in our analysis. This choice reflects the established importance of positional information in protein sequence analysis. Notably, many deep learning models exhibit translation-invariance properties that may not effectively capture these positional dependencies, potentially limiting their effectiveness for viral host range risk assessment.

The limitations of deep learning in our context mirror their known constraints in processing tabular data, which closely resembles our residue sequence format. To address these limitations, deep learning models could be enhanced through architectural modifications such as attention mechanisms or positional encodings, as demonstrated by specialized models like TabNet. This approach is particularly relevant given the critical importance of residue position information in sequence analysis. Furthermore, deep learning models typically require larger training datasets due to their extensive parameter space, a challenge that could be addressed through pre-training approaches, including semi-supervised/unsupervised learning or transfer learning methodologies.

Beyond performance considerations, tree ensemble models like Random Forest offer superior explainability, providing an additional advantage for our analysis. Their decision rules are directly interpretable without requiring sophisticated post-hoc analyses, contrasting with neural networks' black-box nature that often demands complex interpretation techniques [39, 40]. Regarding analytical methods, while SHAP analysis provided interpretable, additive assessments of residue-level risk and effects of mutations, it had limited capability as a direct means of capturing complex non-linear interactions. Specifically, the current methodology cannot systematically detect cases where specific residues modulate the effects of residues at other positions. Development of analytical tools capable of elucidating such compositional or conditional effects represents a crucial direction for future research.

Our SHAP-based analysis identified key residues including 81M, 199S, 271A, 292I, 368Q, 588A, 591R, 684A/S, and 627K, and mutations (T271A, I292V/T, Q368R/K, A588T/I/V, Q591R, and E627K) critical for zoonotic risk assessment. While some of these residues form a cluster, many are distributed throughout the PB2 protein, suggesting that multiple mechanisms may contribute to zoonotic potential. Many of these residues and mutations have been previously reported as critical for mammalian adaptation. Among these, E627K and Q591R mutations in PB2 are the most well-documented mutations facilitating human adaptation of animal IAVs [1, 2]. These mutations raise the degree of positive charge in the region, which affects the interaction with several host factors [1, 60]. Additionally, mutations such as A199S, T271A, I292V, A588T/I/V, and A684S are known to increase viral polymerase activity [61–64]. Although residues such as 81M and 368Q were among the top ten identified in our SHAP-based analysis, these residues remain underexplored. The N-terminus of PB2 interacts with PB1, and 81M may modulate this interaction. Meanwhile, residue 368Q is located within the cap-binding domain of PB2, and charge alteration resulting from the R368Q mutation may interfere with binding to host pre-mRNA caps. Future studies employing in vitro polymerase activity assays may help elucidate the functional consequences of these mutations and their potential role in host adaptation.

Phasianidae family, which includes chickens and turkeys, has historically served as a major source of human infections by avian IAVs. Nevertheless, their role as bridge hosts facilitating viral transmission to mammalian hosts remains poorly understood. Our analysis of risk score distribution from the SHAP-based model across avian categories revealed that IAVs from the *Phasianidae* family tend to exhibit higher risk scores compared to those from other avian species. Additionally, we identified mutations, including I292V/T, Q368R, A588T/I, V598I/A/T, and E/V627K as important risk factors in *Phasianidae* compared to other avian families, warranting further investigation. These findings suggest that the transmission of avian IAVs to *Phasianidae* may be associated with frequent spillovers into humans through species in this family, such as chickens, turkeys, and quails [65–67]. However, the contributions of other genomic segments remain unclear and require further investigation. These results underscore the importance of evaluating the role of *Phasianidae* family as a bridge host for IAV transmission to mammalian hosts.

Conclusions

In this study, we developed two complementary approaches for modeling the zoonotic potential of influenza PB2 sequences. We defined regression-based risk models and evaluated various methods, including deep learning models, through cross-validation. The selection of target values for the mid-risk group was validated through a comprehensive ablation study. Using the optimized regression model, we conducted SHAP value analysis and constructed a SHAP-based risk assessment model by aggregating SHAP values. Individual residues based on their contribution to zoonotic risk prediction,

were calculated and yielded 81M, 199S, 271A, 292I, 368Q, 588A, 591R, 627K, and 684A/S as important residues. Furthermore, we developed a SHAP-based metric for quantifying mutational effects on zoonotic risk. Based on our risk assessment, we identified T271A, I292V/T, Q368R/K, A588T/I/V, Q591R, and E627K as the mutations with the highest zoonotic potential. Through these in silico analyses, we identified the *Phasianidae* family as having elevated human infection risk and characterized specific mutations associated with risk.

Abbreviations

AUROC	Area Under the Receiver Operating Characteristic
CNN	Convolutional Neural Network
FCN	Fully Connected Network
IAV	Influenza A Virus
KNN	k-Nearest Neighbors
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron
MPM	Mammalian Pathogenicity-related Mutation
MSE	Mean Squared Error
PCA	Principal Component Analysis
PSSM	Position-Specific Scoring Matrix
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive explanation
SVM	Support Vector Machine

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11589-8.

Supplementary Material 1

Acknowledgements

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

Author contributions

S. Kim, M.-A. Kim, C.-Y. Lee, and S. Jeong contributed to the conception of the study. M.-A. Kim, B. Kim, J. Lee, S.-K. Jung, and C.-Y. Lee provided sequence data. S. Kim, J. Kim, and S. Jeong conducted experiments. S. Kim, M.-A. Kim, J. Lee, and S.-K. Jung carried out the data analysis. S. Kim, M.-A. Kim, and C.-Y. Lee wrote the manuscript. H.-Y. Chung, C.-Y. Lee, and S. Jeong reviewed the manuscript critically. All authors participated in the study and approved the final version of the manuscript.

Funding

This research was partly supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2022-KH130593), the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (grant number: RS-2023-00210169 and RS-2023-00228644), and Global - Learning & Academic research institution for Master's·PhD students, and Postdocs (LAMP) Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (grant number: RS-2023-00301914).

Data availability

The datasets generated and/or analysed during the current study are available in the GISAID's EpiFlu database, with GISAID Identifier EPI_SET_250121zc. To view the contributors of each individual sequence with details such as accession number, Virus name, Collection date, Originating Lab and Submitting Lab and the list of Authors, visit https://doi.org/10.55876/gis8.250121zc.Additional data not included in the supplementary materials are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Bio-medical Research Institute, Kyungpook National University Hospital, Daegu, South Korea
[2]Department of Microbiology, School of Medicine, Kyungpook National University, Daegu, South Korea
[3]Department of Neurology, Keimyung University Dongsan Medical Center, Daegu, South Korea
[4]Department of Medical Informatics, School of Medicine, Kyungpook National University, Daegu, South Korea
[5]Untreatable Infectious Disease Institute, Kyungpook National University, Daegu, South Korea
[6]Research Center for Artificial Intelligence in Medicine, Kyungpook National University Hospital, Daegu, South Korea

Received: 16 January 2025 / Accepted: 9 April 2025
Published online: 23 April 2025

References

1. Long JS, Mistry B, Haslam SM, Barclay WS. Host and viral determinants of influenza A virus species specificity. Nat Rev Microbiol. 2019;17:67–81.
2. Gilbertson B, Duncan M, Subbarao K. Role of the viral polymerase during adaptation of influenza A viruses to new hosts. Curr Opin Virol. 2023;62:101363.
3. Long JS, Giotis ES, Moncorgé O, Frise R, Mistry B, James J, et al. Species difference in ANP32A underlies influenza A virus polymerase host restriction. Nature. 2016;529:101–4.
4. Weber M, Sediri H, Felgenhauer U, Binzen I, Bänfer S, Jacob R, et al. Influenza virus adaptation PB2-627K modulates nucleocapsid Inhibition by the pathogen sensor RIG-I. Cell Host Microbe. 2015;17:309–19.
5. Sheppard CM, Goldhill DH, Swann OC, Staller E, Penn R, Platt OK, et al. An influenza A virus can evolve to use human ANP32E through altering polymerase dimerization. Nat Commun. 2023;14:6135.
6. Peacock TP, Sheppard CM, Lister MG, Staller E, Frise R, Swann OC, et al. Mammalian ANP32A and ANP32B proteins drive differential polymerase adaptations in avian influenza virus. J Virol. 2023;97:e00213–23.
7. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst, 2013; 26.
8. Kimothi D, Biyani P, Hogan JM, Soni A, Kelly W. Learning supervised embeddings for large scale sequence comparisons. PLoS ONE, 2020; 15(3).
9. Asgari E, Morfrad MRK. Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS ONE, 2015; 10(11).
10. ElAbd H, Bromberg Y, Hoarfrost A, Lenz T, Franke A, Wendorff M. Amino acid encoding for deep learning applications. BMC Bioinformatics. 2020;21:1–14.
11. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
12. Greenacre M, Groenen PJF, Hastie T, D'Enza AI, Markos A, Tuzhilina E. Principal component analysis. Nat Reviews Methods Primers. 2022;2(1):100.
13. Wang J, Ma C, Kou Z, Zhou YH, Liu HL. Predicting transmission of avian influenza A viruses from avian to human by using informative physicochemical properties. Int J Data Min Bioinform. 2013;7(2):166–79.
14. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
15. Xu Y, Wojtczak D. Predicting influenza A viral host using PSSM and word embeddings. In2021 IEEE Conference on Computational Intelligence in

Bioinformatics and Computational Biology (CIBCB) 2021 Oct 13 (pp. 1–10). IEEE.
16. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
17. Eng CL, Tong JC, Tan TW. Predicting host tropism of influenza A virus proteins using random forest. BMC Med Genom. 2014;7:1–1.
18. Kwon E, Cho M, Kim H, Son HS. A study on host tropism determinants of influenza virus using machine learning. Curr Bioinform. 2020;15(2):121–34.
19. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst, 2018; 31.
20. Miao M, De Clercq E, Li G. Towards efficient and accurate SARS-CoV-2 genome sequence typing based on supervised learning approaches. Microorganisms. 2022;10(9):1785.
21. Tolles J, Meurer WJ. Logistic regression: relating patient characteristics to outcomes. JAMA. 2016;316(5):533–4.
22. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22.
23. Martin V, Pfeiffer DU, Zhou X, Xiao X, Prosser DJ, Guo F, Gilbert M. Spatial distribution and risk factors of highly pathogenic avian influenza (HPAI) H5N1 in China. PLoS Pathog. 2011;7(3):e1001308.
24. Yoo DS, Chun BC, Hong K, Kim J. Risk prediction of three different subtypes of highly pathogenic avian influenza outbreaks in poultry farms: based on Spatial characteristics of infected premises in South Korea. Front Veterinary Sci. 2022;9:897763.
25. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967:21–7.
26. Uddin S, Haque I, Lu H, Moni MA, Gide E. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Sci Rep. 2022;12(1):6256.
27. Humayun F, Khan F, Fawad N, Shamas S, Fazal S, Khan A, Ali A, Farhan A, Wei DQ. Computational method for classification of avian influenza A virus using DNA sequence information and physicochemical properties. Front Genet. 2021;12:599321.
28. LeCun Y, Yoshua B. Convolutional networks for images, speech, and time series. Handb Brain Theory Neural Networks, 1995; 3361(10).
29. Hou J, Adhikari B, Cheng J. DeepSF: deep convolutional neural network for mapping protein sequences to folds. Bioinformatics. 2018;34(8):1295–303.
30. Luo F, Wang M, Liu Y, Zhao XM, Li A. DeepPhos: prediction of protein phosphorylation sites with deep learning. Bioinformatics. 2019;35(16):2766–73.
31. McCulloch WS, Walter P. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5:115–33.
32. Schmidhuber J, Sepp H. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
33. Amari SI. Learning patterns and pattern sequences by self-organizing Nets of threshold elements. IEEE Trans Comput. 1972;100(11):1197–206.
34. Shen Z, Wenzheng B, De-Shuang H. Recurrent neural network for predicting transcription factor binding sites. Sci Rep, 2018; 8(1).
35. Liu Q, Fang L, Yu G, Wang D, Xiao CL, Wang K. Detection of DNA base modifications by deep recurrent neural network on Oxford nanopore sequencing data. Nat Commun. 2019;10(1):2449.
36. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst, 2017.
37. Cui F, Zhang Z, Zou Q. Sequence representation approaches for sequence-based protein prediction tasks that use deep learning. Brief Funct Genomics. 2021;20(1):61–73.
38. Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. Nature. 2024;630:493–500.
39. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–74.
40. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B,… Lee SI. From local explanations to global understanding with explainable AI for trees. Nature machine intelligence, 2020;2(1):56–67.
41. Arik SÖ, Pfister T, TabNet. Attentive Interpretable Tabular Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 2021; 35(8): 6679–6687.
42. Ogino S, Gulley ML, den Dunnen JT, Wilson RB, Association for Molecular Pathology Training and Education Committee. Standard mutation nomenclature in molecular diagnostics: practical and educational challenges. J Mol Diagn. 2007;9(1):1–6.
43. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges. 2017;1:33–46. https://doi.org/10.1002/gch2.1018.
44. Pedregosa F, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
45. Gaudette L, Japkowicz N. Evaluation methods for ordinal classification. Adv Artif Intell. 2009;22nd Canadian Conference on Artificial Intelligence:207–10.
46. Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213–20.
47. Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv (CSUR). 2002;34(1):1–47.
48. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. InArtificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4–7, 2018, Proceedings, Part III 27 2018 (pp. 270–279). Springer International Publishing.
49. Bozinovski S. Reminder of the first paper on transfer learning in neural networks, 1976. Informatica. 2020;44(3).
50. Bengio Y. Deep learning of representations for unsupervised and transfer learning. InProceedings of ICML workshop on unsupervised and transfer learning 2012 Jun 27 (pp. 17–36). JMLR Workshop and Conference Proceedings.
51. Lee C-Y, An S-H, Choi J-G, Lee Y-J, Kim J-H, Kwon H-J. Rank orders of mammalian pathogenicity-related PB2 mutations of avian influenza A viruses. Sci Rep. 2020;10:5359.
52. Czudai-Matwich V, Otte A, Matrosovich M, Gabriel G, Klenk HD. PB2 mutations D701N and S714R promote adaptation of an influenza H5N1 virus to a mammalian host. J Virol. 2014;88(16):8735–42.
53. Clements JF, Rasmussen PC, Schulenberg TS, Iliff MJ, Fredericks TA, Gerbracht JA, Lepage D, Spencer A, Billerman SM, Sullivan BL, Smith M, Wood CL. The eBird/Clements checklist of Birds of the World: v2024, 2024.
54. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A. Highly accurate protein structure prediction with alphafold. Nature. 2021;596(7873):583–9.
55. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022;19(6):679–82.
56. Meng EC, Goddard TD, Pettersen EF, Couch GS, Pearson ZJ, Morris JH, Ferrin TE. UCSF chimerax: tools for structure Building and analysis. Protein Sci. 2023;32(11):e4792.
57. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Inform Fusion. 2022;81:84–90.
58. Zabërgja G, Kadra A, Grabocka J. Tabular data: is attention all you need?? ArXiv Preprint. 2024. arXiv:2402.03970.
59. Gorishniy Y, Rubachev I, Khrulkov V, Babenko A. Revisiting deep learning models for tabular data. Adv Neural Inf Process Syst. 2021;34:18932–43.
60. Mehle A, Doudna JA. An inhibitory activity in human cells restricts the function of an Avian-like influenza virus polymerase. Cell Host Microbe. 2008;4:111–22.
61. Bussey KA, Bousse TL, Desmet EA, Kim B, Takimoto T. PB2 residue 271 plays a key role in enhanced polymerase activity of influenza A viruses in mammalian host cells. J Virol. 2010;84:4395–406.
62. Foeglein Á, Loucaides EM, Mura M, Wise HM, Barclay WS, Digard P. Influence of PB2 host-range determinants on the intranuclear mobility of the influenza A virus polymerase. J Gen Virol. 2011;92:1650–61.
63. Gao W, Zu Z, Liu J, Song J, Wang X, Wang C, et al. Prevailing I292V PB2 mutation in avian influenza H9N2 virus increases viral polymerase function and attenuates IFN-β induction in human cells. J Gen Virol. 2019;100:1273–81.
64. Hayashi T, Wills S, Bussey KA, Takimoto T. Identification of influenza A virus PB2 residues involved in enhanced polymerase activity and virus growth in mammalian cells at low temperatures. J Virol. 2015;89:8042–9.
65. Pillai SPS, Pantin-Jackwood M, Yassine HM, Saif YM, Lee CW. The high susceptibility of Turkeys to influenza viruses of different origins implies their importance as potential intermediate hosts. Avian Dis. 2010;54:522–6.
66. Perez DR, Lim W, Seiler JP, Yi G, Peiris M, Shortridge KF, et al. Role of quail in the interspecies transmission of H9 influenza A viruses: molecular changes on HA that correspond to adaptation from ducks to chickens. J Virol. 2003;77:3148–56.
67. Lee C-Y. Exploring potential intermediates in the Cross-Species transmission of influenza A virus to humans. Viruses. 2024;16:1129.

Publisher's note