PLOS ONE

RESEARCH ARTICLE

A framework for the risk prediction of avian influenza occurrence: An Indonesian case study

Samira Yousefinaghani[1], Rozita Dara[1]*, Zvonimir Poljak[2], Fei Song[1], Shayan Sharif[3]

1 School of Computer Science, University of Guelph, Guelph, Ontario, Canada, **2** Department of Population Medicine, Ontario Veterinary College, University of Guelph, Guelph, Ontario, Canada, **3** Department of Pathobiology, University of Guelph, Guelph, Ontario, Canada

* drozita@uoguelph.ca

Citation: Yousefinaghani S, Dara R, Poljak Z, Song F, Sharif S (2021) A framework for the risk prediction of avian influenza occurrence: An Indonesian case study. PLoS ONE 16(1): e0245116. https://doi.org/10.1371/journal.pone.0245116

Editor: Alessandro Rizzo, Politecnico di Torino, ITALY

Received: February 27, 2020

Accepted: December 23, 2020

Published: January 15, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0245116

Data Availability Statement: The data underlying this study is accessible at https://figshare.com/articles/dataset/Indonesia_data/13519088.

Abstract

Avian influenza viruses can cause economically devastating diseases in poultry and have the potential for zoonotic transmission. To mitigate the consequences of avian influenza, disease prediction systems have become increasingly important. In this study, we have proposed a framework for the prediction of the occurrence and spread of avian influenza events in a geographical area. The application of the proposed framework was examined in an Indonesian case study. An extensive list of historical data sources containing disease predictors and target variables was used to build spatiotemporal and transactional datasets. To combine disparate sources, data rows were scaled to a temporal scale of 1-week and a spatial scale of 1-degree × 1-degree cells. Given the constructed datasets, underlying patterns in the form of rules explaining the risk of occurrence and spread of avian influenza were discovered. The created rules were combined and ordered based on their importance and then stored in a knowledge base. The results suggested that the proposed framework could act as a tool to gain a broad understanding of the drivers of avian influenza epidemics and may facilitate the prediction of future disease events.

Introduction

Avian Influenza (AI) disease is caused by influenza type A viruses, which can infect domestic poultry, wild birds and mammalian species, including humans. Despite researchers' efforts to eradicate and control this disease, it has continuously caused significant losses to poultry and has threatened human lives. To mitigate the impact of AI outbreaks, it is necessary to understand the extent to which different risk factors and their interactions contribute to the introduction and spread of outbreaks. To date, an extensive array of studies have reported on spatiotemporal surveillance and control of AI using approaches, including logistic regression, boosted regression tree, cluster analysis and maximum entropy in different geographical scales. Studies have mapped the distribution of risk [1, 2] and identified important risk factors for disease occurrence [3, 4]. Some studies have performed spatiotemporal surveillance on a country scale such as those in Bangladesh [5], China [6], Indonesia [7], India [8], Thailand [1]

Funding: This work was funded by Egg Farmers of Canada, Chicken Farmers of Saskatchewan, and the Canadian Poultry Research Council. This research is supported in part by the University of Guelph's Food from Thought initiative, thanks to funding from the Canada First Research Excellence Fund.

Competing interests: The authors have declared that no competing interests exist.

and Vietnam [3] while others have focused on regional [9] or global [10] scales. Such disease risk-profiling approaches could assist in understanding the predictors of disease occurrence and preparing for future events.

Environmental conditions [11, 12], waterfowl [10, 13, 14], poultry farming and trading activities [15, 16], agricultural activities [17] and land cover [10, 18] are identified as major factors of introduction and dispersion of AI occurrence.

The impact of environmental factors and climate change on the spread and geographical distribution of AI outbreaks is documented in the literature [9, 19, 20]. For example, annual precipitation is introduced as an important predictive variable for the risk of highly pathogenic avian influenza (HPAI) in China [19]. Also, in the Middle East, the precipitation in the warmest quarter of a year is positively connected with HPAI H5N1 outbreaks [9]. In contrast, in Europe [20] and Bangladesh [4, 5], precipitation is negatively associated with H5N1 outbreaks in wild birds and poultry. Another important factor is the temperature that is positively associated with H5N1 outbreaks in wild birds in Europe [20] while an opposite pattern is found for poultry in Bangladesh [4, 5].

Moreover, the role of waterfowl density in the distribution of AI outbreaks is highlighted in a number of studies. In Asia [2, 21], domestic waterfowl density appears to be an important risk factor for H5N1 occurrence in poultry. Moreover, a positive association between duck density and H5N1 occurrence is found in Vietnam, Thailand [1, 22], India [8] and global scale [15].

Similarly, poultry density is considered as one of the factors associated with AI outbreaks. A strong association between HPAI outbreaks and densities of chickens was found in California [16], the Middle East [9] and globally [15]. Also, poultry market density in China is considered an important predictor of the risk of AI H7N9. In another study, Henning *et al.* [3] found that medium poultry density is associated with the risk of H5N1 outbreaks in Vietnam. Contrary to aforementioned studies, Yupiana *et al.* [7] found a negative association between H5N1 outbreaks and poultry density.

The spatial distribution of H5N1 outbreaks and its transmission to various regions have been associated with wild bird flyways [10, 23]. Moreover, the introduction of H5N1 to poultry in Europe, Asia and Africa may be partly through wild bird migration [24]. However, in a number of studies [25–27], a limited or a negative association has been found between migratory waterfowl sites and outbreaks of H5N1.

Despite the efforts that have been made to determine the essential predictor variables and suitable areas of AI presence, there are still some research gaps that need to be filled.

Motivation

The present paper is aimed to obtain insights on the prediction of AI events using an extensive spatiotemporal dataset. For this, we identified gaps in the existing research concerning applied predictor variables and methodologies.

Despite the highlighted importance of risk predictor variables mentioned earlier, the precision and completeness of explanatory data have received limited attention in the literature. For example, a global climate system [28] has been frequently used in AI surveillance studies [9, 11, 29, 30]. The WorldClim website provides monthly average amount of climatic variables for various spatial resolutions. Clearly, the system only provides historical information and the monthly average of climatic variables is a low temporal resolution. Moreover, the geographic distribution of wild migratory birds has not been included in some existing work [15, 21, 31].

The existing approaches for determining how AI events occur in a region usually rely on regression [2, 6, 8, 19, 32] or boosted regression tree models [8, 15, 17, 18, 25]. The existing

work has aimed to find the most important predictors of disease and then profile the risk of outbreaks. In the aforementioned studies, the average impact of individual risk factors on the output is assessed. However, the identification of subgroups with different risk profiles is overlooked. This can consequently, ignore important information and produce biased results [17, 33]. Analysis relying on only one or a few numbers of risk factors could ignore important information and produce biased results.

The application of rule-based prediction models has been limited to a few studies for Dengue [12], Depression [34] and Diabetes [35, 36]. We are aware of only one study which used rule-based models aimed at analyzing AI outbreaks [37]. Xu *et al.* [37] constructed a data cube model with OLAP (Online Analytical Processing) actions. Then, geographical and temporal insights into disease spread with various abstraction levels were extracted. Moreover, sequential pattern mining and association rule mining were applied to provide understandings of potential serial spread routes and linkage between outbreak sites [37]. The researchers in this study Xu *et al.* [37] used the disease occurrence data regardless of the importance of explanatory variables.

The present study exploits historical data sources to extract predictive patterns of AI. The approach used here is complementary to prior research with three main contributions: (1) It uses several data sources with high temporal and spatial resolutions. (2) It employs rule-discovery models rather than focusing on predictive regression models commonly used in the relevant existing studies. Rule-discovery models generate patterns that can link the risk of disease presence with subsets of risk factors. (3) It contributes to automation and transparency of predictions. Although transparency of predictions is essential in application areas of epidemiology [38], the transparency of surveillance systems and their outcomes has received less attention in the literature. Since the proposed framework used a collection of rules as a high-level description of data, the explanation capability of predictions can be enhanced. This means public health officials can find the reasons behind the predictions made by the system.

The proposed framework was implemented and tested for an Indonesian case study. Indonesia was selected as a case study as this country has had a high number of reporoted AI outbreaks over the years and, importantly, it provides accessible explanatory data sources. This framework can form a basic model for risk prediction of AI events.

Methodology

The main goal here is to extract prediction patterns of AI occurrence from a set of disparate data sources. We designed, implemented and tested a framework with four main parts including data collection, data aggregation and pre-processing, data analysis, and prediction. An overview of the main framework is presented in Fig 1.

In the first step (Data Collection), independent variables were identified and their respective data sources were collected. These variables were identified by the help of subject matter experts and from relevant literature. In the next step (Data Aggregation and Pre-processing), a relational database containing time and geographic information along with several covariates and outcome variables was designed. In the third step (Data Analysis), we applied rule discovery algorithms to the labelled dataset (training dataset) and extracted hidden patterns. These patterns indicated which combination of risk factors had led to high or low risk of disease occurrence and what linkage between event sites had been observed. Moreover, an experiment was conducted to evaluate the performance of predictions. In the last step (Prediction), end-users can communicate with the system through a user interface. Here, the user interface can include mapping and monitoring of the risk, given a current spatiotemporal dataset.

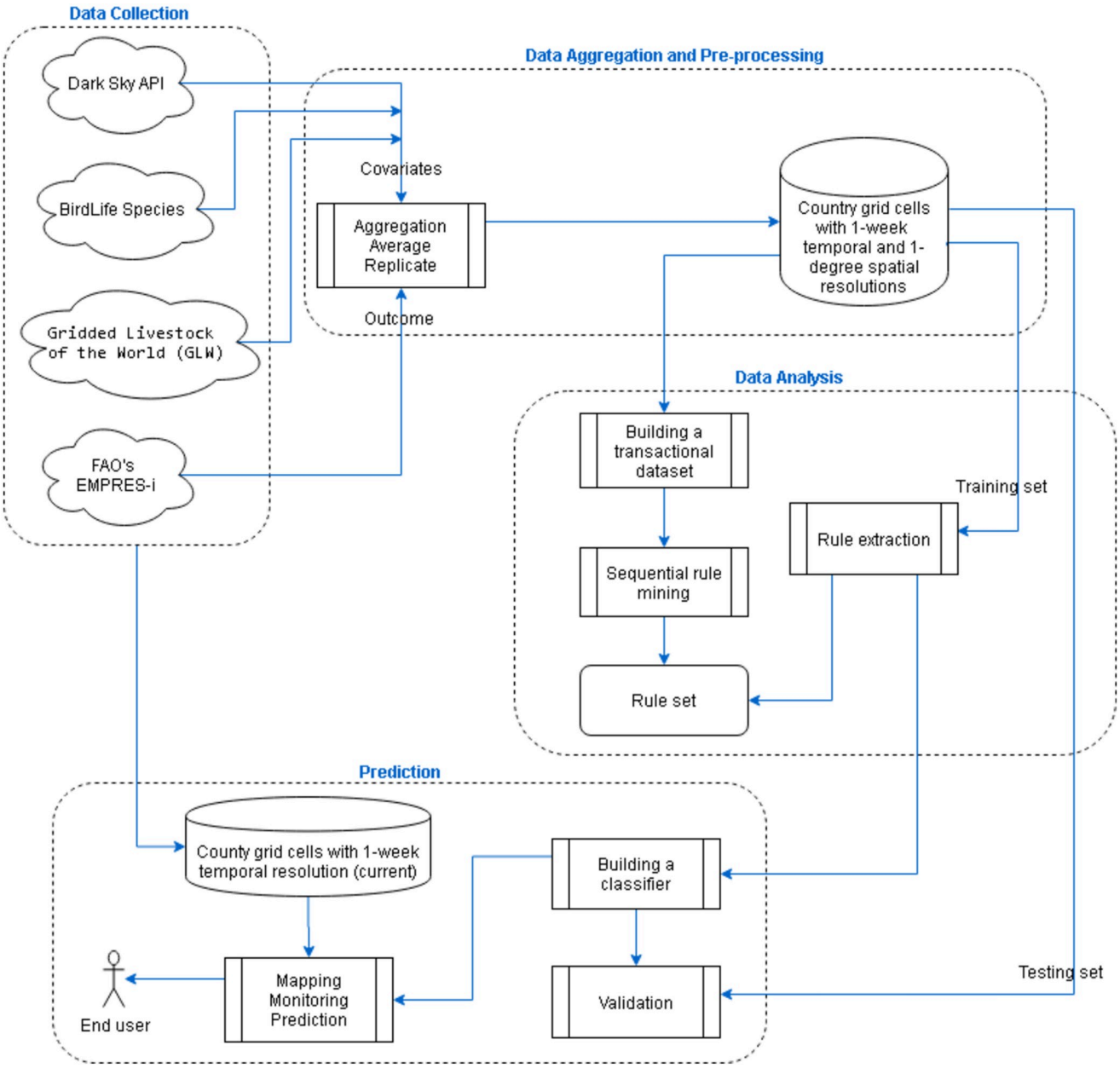

Fig 1. Overall framework.

https://doi.org/10.1371/journal.pone.0245116.g001

Data collection

A number of data sources have been downloaded and stored in their respective tables in a relational database that is visualized in Fig 2. Table 1 presents a summary of data sources used in the database construction. These data sources included climatic variables, geographical distribution of migratory bird species, distribution of poultry and AI historical records. More detail information on data tables is given in S1–S3 Tables in S1 File.

The risk factors obtained from these sources have been shown in the previous studies to correlate with AI outbreaks [12, 15, 16]. A list of risk factors used in the study along with their

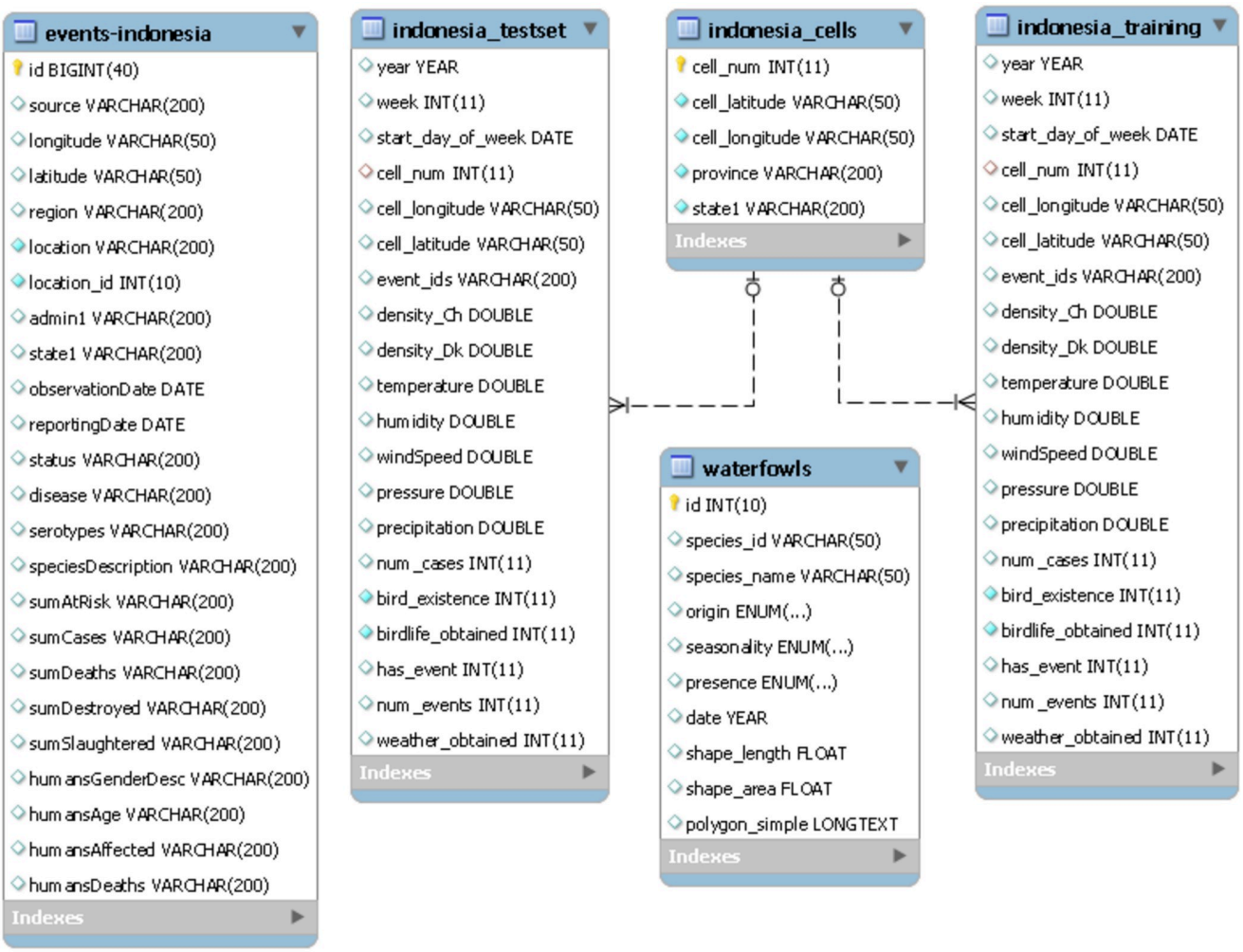

Fig 2. Database schematic.

https://doi.org/10.1371/journal.pone.0245116.g002

Table 1. Sources of data.

Data Source	Description
Dark Sky API	The API offers several climatic variables including temperature, humidity and wind speed. We automatically collected the variables that have been frequently used as risk factors of AI. The 'Time Machine Requests' API offered by Dark Sky [39] was used to retrieve weather information given latitude, longitude and time parameters.
BirdLife-species	The data provides geographic extents of species distribution ranges and is available in the Environmental Systems Research Institute (ESRI) Geodatabase formats [40].
Gridded Livestock of the World (GLW3)	Food and Agriculture Organization (FAO) has developed the GLW3, in which the global distribution of chickens and ducks in 2010 is expressed by the total number of birds per pixel (5 minutes of arc) [41].
EMPRES-i	FAO's Emergency Prevention System (EMPRES) offers a web-based application in order to facilitate the organization and access to disease data in various geographical scales which supports veterinary services [42].

https://doi.org/10.1371/journal.pone.0245116.t001

Table 2. Attributes.

Attribute	Type	Unit	Resolution
temperature	numerical	Fahrenheit	point
precipitation	numerical	millimetre	point
relative humidity	numerical	between 0 and 1	point
wind speed	numerical	miles per hour	point
pressure	numerical	sea-level air pressure in millibars	point
chicken density	numerical	density	5-minute arc
duck density	numerical	density	5-minute arc
waterfowl	numerical	Boolean	point

https://doi.org/10.1371/journal.pone.0245116.t002

type, unit and resolution is shown in Table 2. These attributes are also visualized in 'indonesia_training' and 'indonesia_testset' data tables in Fig 2.

Data aggregation and pre-processing

To build the basis of the model, we divided Indonesia land mass into rectangle cells, each with size 1-degree × 1-degree (equal to 60-minutes arc). A visualization of sample cell centres is provided in Fig 3. In addition, the temporal resolution of 1-week was considered. This resolution was selected as it offers a good balance between the precision of decision-making and the time required for data processing. The response variable for each cell (i.e. each row of dataset) was classified as zero if there were no AI events within the cell and during the specified

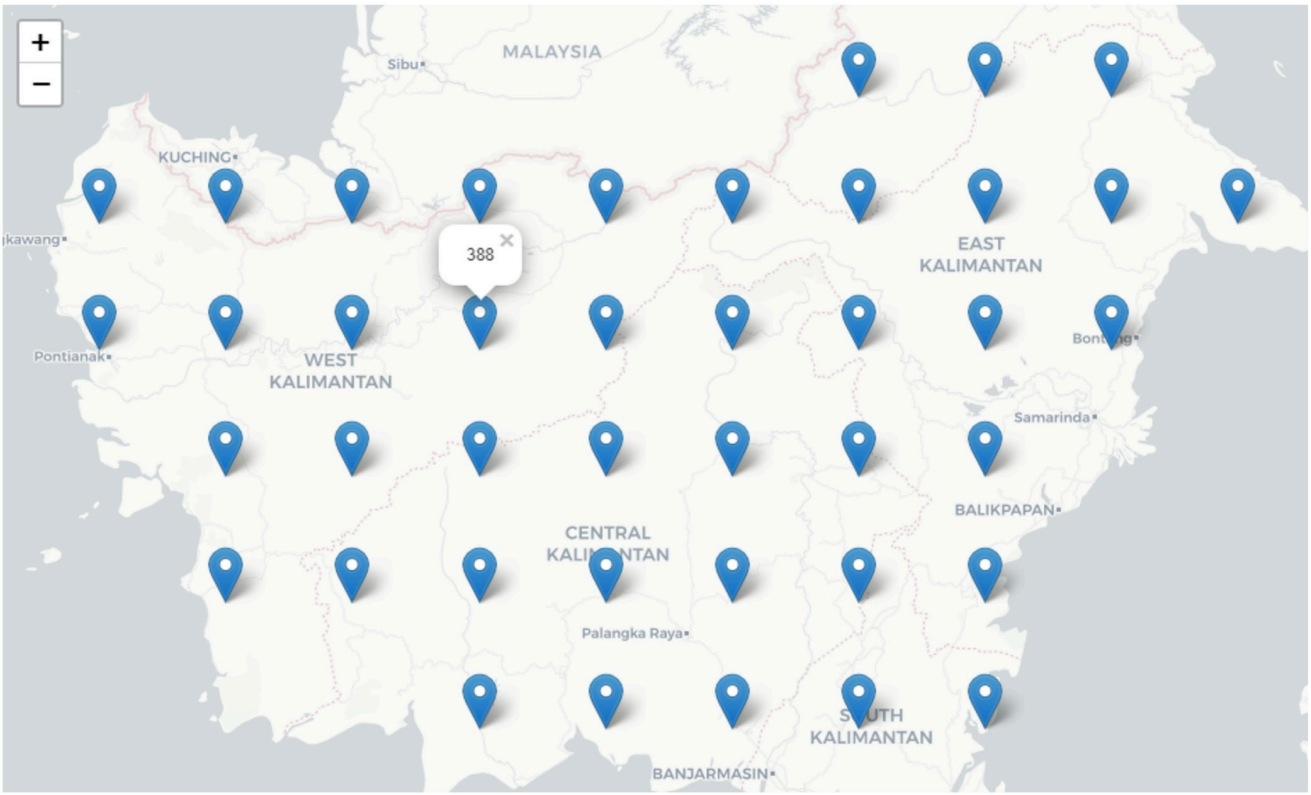

Fig 3. Sample spatio scale (Indonesia). Base map and data from OpenStreetMap and OpenStreetMap Foundation.

https://doi.org/10.1371/journal.pone.0245116.g003

temporal scale. Conversely, if there was at least one disease event within the spatiotemporal scale, the response variable was classified as one. The response variable is demonstrated by 'has_event' field in the database schema (Fig 2).

The data aggregation was performed in a Structured Query Language table. The response variable describes a disease event that occurred within a certain time interval (week number t) and a spatial compartment with a center of (x,y), where x represents longitude and y represents latitude. The spatiotemporal information in "Indonesia" data tables in Fig 2 are demonstrated by "week", "cell_longitude" and "cell_latitude" attributes. Similarly, for other covariates, preprocessing techniques were applied to transform the records into the defined grid-based table. Since the table included coordinate information for each record, a weekly timeline of information could be visualized in maps using GIS software. Moreover, extracted patterns from the grid-based data could be simple and easy to understand.

The information on AI events in Indonesia was obtained from the Emergency Prevention System for Animal Health (EMPRES) [42, 43]. The reporting date and geographical coordinates of events in the EMPRES-i platform was used to scale the disease events to specified cells.

To make a classifier that could predict the response variable based on predictor variables, we divided the data into training and testing sets. The data from 2009 to 2016 was used for training and the data containing 2006, 2007, 2008 and 2017 was used for testing purposes. Overall, the number of observations in the training dataset was 61,152 (147 [number of cells] × 8 [number of years] × 52 [number of weeks per year]) and in the testing dataset was 30,576.

In total, the training dataset contained 1,860 rows with target variable of one and 59,292 of zero while the testing dataset contained 168 of class one and 30,408 of zero.

In addition, the explanatory variables were obtained for each spatial and temporal dimension. The "Time Machine Request" type of Darksky API was used to retrieve the climatic information. This returns the observed daily weather conditions given a specified date in the past and a location point. The API responses consist of a JSON-formatted object, from which we selected a set of predictors that have been previously known as factors contributing to outbreaks of AI [44–46]. These predictor variables included temperature, precipitation, humidity, wind speed and pressure.

The chicken and duck distribution data for Indonesia were computed using The Gridded Livestock of the World (GLW) [47]. The data offered GeoTIFF format files that were converted to longitude-latitude-value format using the Rasterio library in Python [48] and then imported to a designed database. The spatial resolution of GLW data (a pixel) was higher than the defined spatial resolution of the present study. Thus, the density points inside a cell have been averaged. However, GLW provides low temporal resolution (2010 only) and therefore the densities have been replicated for all data points with a particular spatial resolution.

Birdlife species data included shapefiles that could be visualized by geographical information system (GIS) software such as ArcGIS. We filtered polygons related to 133 duck species. In addition, due to the very large size of the data, we simplified the polygons. This enabled us to decrease the processing time. Finally, in the field called 'bird_existence' in the database, we specified whether each cell was inside a bird polygon or not.

In addition to the aforementioned explanatory variables, we defined winter, spring, summer, fall seasons by dividing weeks into 48-12, 12-24, 24-36 and 36-48, respectively. Indonesia is passed by the equator and the weather can be split into dry (May-September) and rainy (October-April) seasons. In the present study, fall and winter divisions represent the rainy season while spring and summer represent the dry season.

Predictor variables coming from disparate data sources had different spatial and temporal resolutions. Therefore, the variables were arranged with respect to the defined spatial and temporal resolution. When the spatial resolution was higher than a cell or temporal resolution was

higher than a week, we averaged the values. Conversely, when the resolution was lower than a cell or a week, we repeated the same values for all the cells that fit into that resolution. Finally, various data sources were assembled into a database with a uniform spatiotemporal resolution.

Data analysis

Given the created dataset, we employed RuleFit, Frequent Pattern Growth (FP-Growth) and Prefix-projected Sequential Pattern Mining (PrefixSpan) models to discover hidden rules that might be predictive or indicate dispersion paths of the risk of AI occurrence.

Patterns were extracted in the form of "IF-THEN" rules. The general form of "IF-THEN" rules is demonstrated as follows (Eq 1). Where X is called an antecedent and Y is called a consequent of the rule. The outcome variable (Y) is true if the condition variable (X) is satisfied.

$$IF\ (X\ is\ A)\quad THEN\quad (Y\ is\ B) \tag{1}$$

A rule consists of several interacting risk factors and their ranges. A combination of the extracted rules was used to build the final rule-based classifier.

Given the prepared training set, a supervised ensemble rule learner (RuleFit) was trained to induce rules. RuleFit [49] is a computational algorithm for rule discovery from a large number of candidate risk factors [36]. It generates rules by first exhaustively searching for candidate rules over the potential risk factors in the "rule generation" phase. Rules are generated automatically by traversing each path through a decision tree. Subsequently, the redundant and irrelevant rules are pruned out in the "rule pruning" phase [49, 50].

Among the advantages of the RuleFit algorithm, several points are of note: 1) This algorithm can rank features by their importance. 2) It outputs interpretable rules. 3) RuleFit relies on a non-parametric model, i.e. Gradient Boosting, with fewer modelling assumptions. Moreover, studies comparing rule extraction methods have shown a competitive accuracy of RuleFit [51, 52].

To address the disparity of explanatory variables, these variables were discretized. The categories of very low (VL), low (L), medium (M), high (H) and very high (VH) were calculated based on the histograms of explanatory variables. Moreover, since the dataset was imbalanced, i.e. the number of negative classes was 30 times more than positive classes in the training set, we under-sampled instances of the majority class (one-to-zero ratio of 0.2). Additionally, some data points were discarded due to the high number of missing values. Subsequently, the model was trained with a 5-fold cross-validation. For each subset, the training set was used to learn the rules and the remaining part to evaluate the model.

In each round, we calculated sensitivity, specificity, precision and F-score metrics. Specificity (Eq 3) measures the proportion of actual negatives that have been correctly identified while sensitivity represents the proportion of actual positives that have been accurately identified (Eqs 4 and 2). Precision represents how many selected cases are relevant and the F-score (Eq 5) is a weighted average of the precision and recall. For two-class classifications, there are four possible cases: For a positive class, if the prediction is positive, this is a called true positive (TP) and if negative, it is a false negative (FN). For a negative example, if the prediction is negative, it is called true negative (TN) and if positive, it is a false positive (FP).

Next, we used the unsupervised FP-Growth algorithm for mining the rules from the training set. The algorithm was first proposed by Han *et al.* [53] and it mines the frequent itemsets without candidate generation [54]. The algorithm first compresses the database into a frequent-pattern tree (FP-tree). Then, FP-tree is divided into a set of conditional databases [54]. The FP-Growth algorithm has proven to be time efficient and to consume less memory than the Apriori Algorithm for mining frequent itemset [12, 55].

Extracted rules can be representative of relations between variables in the dataset. The rules indicating a relationship between predictor and response variables were obtained from the FP-Growth algorithm applied to the training set. Similar to the RuleFit, we discretized explanatory variables and additionally, assigned high risk (HR) and low risk (LR) to the response variable for values one and zero, respectively. Following that, the support and confidence criteria were used to select the most important rules. The support of a rule is the number of instances in the dataset that endorses that rule and the confidence indicates the number of times the "IF-THEN" statements are found true.

$$Recall\ or\ Sensitivity = \frac{TP}{(TP + FN)} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$Precision\ or\ Positive\ Predictive\ Value\ (PPV) = \frac{TP}{(TP + FP)} \tag{4}$$

$$F_1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

$$F_\beta - score = \frac{(1 + \beta^2) * Precision * Recall}{(\beta^2 * Precision) + Recall} \tag{6}$$

Mining sequential rules: Event linkage sites. We performed an additional analysis to understand paths by which disease could be transmitted. Such findings could contribute to gaining a better understanding of the risk of AI occurrence. For this, we prepared transactional datasets for each year from 2010 to 2016. Given a year, the sequence of cells containing events along with ranges of their associated risk factors were calculated for each month of the year. An example of the transactional dataset of 2011 prepared for this analysis is provided in Table 3. For example, in January, a time-line of cells with events starting with cell number 253

Table 3. A sample of transactional dataset (2011).

Month	Sequence of Cells with disease presence
Jan	253: dC_H, dD_L, t_H, h_H, ws_L, pc_H, $pr_L \rightarrow$
	376: dC_H, dD_H, t_H, h_M, ws_H, pc_H, $pr_L \rightarrow$
	294: dC_H, Dk_H, t_H, h_M, ws_H, pc_H, $pr_M \rightarrow$
	355: dC_H, dD_H, t_H, h_M, ws_H, pc_H, $pr_L \rightarrow$
	...
Feb	232: dC_L, dD_L, t_H, h_H, ws_L, pc_M, $pr_M \rightarrow$
	210: dC_M, dD_L, t_H, h_M, ws_H, pc_M, $pr_M \rightarrow$
	231: dC_M, dD_L, t_H, h_L, ws_L, pc_M, $pr_L \rightarrow$
	415: dC_M, dD_H, t_L, h_H, ws_L, pc_M, $pr_M \rightarrow$
	...
Mar	210: dC_M, dD_L, t_H, h_H, ws_L, pc_M, $pr_M \rightarrow$
	167: dC_M, dD_L, t_H, h_H, ws_L, pc_M, $pr_M \rightarrow$
	253: dC_H, dD_L, t_H, h_H, ws_H, pc_M, $pr_M \rightarrow$
	376: dC_H, dD_H, t_H, h_M, ws_H, pc_H, $pr_M \rightarrow$
	...

https://doi.org/10.1371/journal.pone.0245116.t003

is produced. In the Table, symbols of 'dC', 'dD', 't', 'h', 'pc' and 'pr' denote 'density of chickens', 'density of ducks', 'temperature', 'humidity', 'precipitation' and 'pressure', respectively. Moreover, the subscripts 'L', 'M', 'H' denote 'very low or low', 'medium', 'high or very high', respectively.

These datasets were then fed to the PrefixSpan algorithm. PrefixSpan is a well-known sequential pattern mining algorithm [56]. Studies have shown that PrefixSpan, in most cases, outperforms the Apriori-based algorithms such as the GSP (generalized sequential pattern algorithm), FreeSpan (frequent pattern-projected sequential pattern mining), and SPADE (sequential pattern discovery using equivalence classes) [56, 57]. This is because it finds the frequent items after scanning the sequence data for a single time.

The outcome of the analysis, i.e. serial paths of disease spread, can be added to the final knowledge base and contribute to calculating the risk of AI occurrence.

The discovered patterns of this part of the framework were then used in the prediction part as illustrated in Fig 1 to predict the risk of AI presence and evaluate these predictions.

Prediction

The extracted rules from the RuleFit and FP-Growth algorithms were ordered using their scores and then two groups of rules were combined to be used for defining a classifier. The process of rule extraction and risk prediction is depicted in Fig 4. Due to the highly imbalanced nature of the data set (a ratio of 1:56 of positives to negatives), unseen data points were undersampled with a positive to negative label ratio of 0.2. Subsequently, for each data point, we searched for up to ten first matching rules. Matching rules are those that their antecedent covers the given data point. We used a variable called 'determinant' to determine the class label of data points as shown in Fig 4. Each time, if a matching rule was a risk-increasing type, we added one unit to the defined variable and if it was risk-decreasing type, we subtracted one unit from the variable. Finally, data points with an amount more than an integer threshold ranged from minus four to four were determined as 'at risk of an disease'. Conversely, those with a value less than the threshold were labelled as 'not at risk of disease'. Comparing these labels with actual ones, we evaluated the results with several measures defined in Eqs 2–5.

Results and discussion

Underlying patterns in form of "IF-THEN" rules along with their respective importance were identified by RuleFit and FP-Growth models. Ordered lists of rules are shown in Tables 4 and 5, respectively.

We used the measures of support, confidence and coefficient to calculate the degree of importance of the rules. Coefficients represent the change in the response variable for one unit of change in the predictor variable. The importance measure for RuleFit is calculated by the multiplication of coefficient and support measures. The rules with positive coefficient are denoted by "increasing" and negative coefficient by "decreasing" risk types as shown in Table 5. The evaluation of outcomes of the RuleFit algorithm with 5-fold cross-validation showed F-score of 64%.

To extract rules using FP-Growth, we separated the dataset based on their target label. For each group, given generated rules from the FP-Growth algorithm, relevant rules were selected using a comparative support criterion, in which we took into account the ratio of instances in high and low risk groups. The rules with the measure greater than a threshold were considered relevant. The extracted rules of the form "risk factors → event occurrence risk" are given in Table 5. Low and high risks are denoted by 'LR' and 'HR' in the table, respectively.

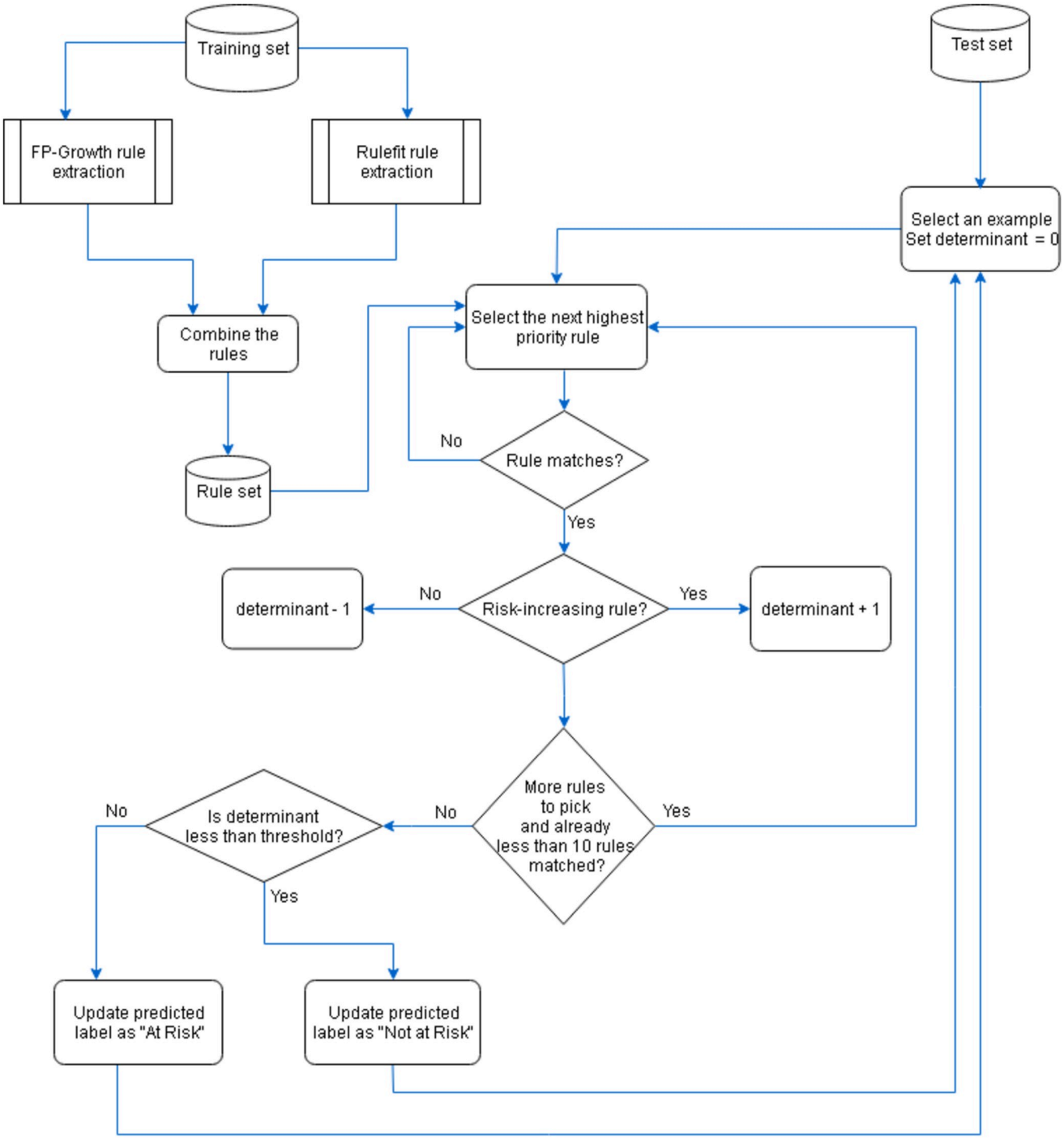

Fig 4. Rule-based prediction.

https://doi.org/10.1371/journal.pone.0245116.g004

From the most relevant rules obtained from the RuleFit and FP-Growth, we discovered that the FP-Growth and RuleFit algorithms agreed on the impact of several predictors such as chicken density, duck density, season and temperature. Furthermore, these rules were consistent with similar studies [2, 15, 16, 21, 58], which validated our rule-based analysis.

Table 4. The top risk rules identified by RuleFit.

Rule	Effect	Risk Type
not in the fall season	2.4045	decreasing
not in the fall season and chicken density not in [H,VH]	1.5648	decreasing
not in the fall season and duck density not in [H,VH] and temperature is not VH	0.3915	decreasing
VL duck density	0.3011	decreasing
duck density not VL and temperature is not VH	0.2519	increasing
duck density not VL and temperature is not VH and pressure is not in [H,VH]	0.1819	increasing
not in the fall season and duck density is VL	0.1680	decreasing
chicken density in [H,VH] and precipitation not in [H,VH]	0.1371	increasing
duck density not VL and season is not Winter and precipitation not in [H,VH]	0.1222	increasing
Winter season and chicken density not in [H,VH]	0.1187	increasing
M chicken density and precipitation is not VH and pressure not in [H,VH] and temperature is not VH	0.0923	increasing

https://doi.org/10.1371/journal.pone.0245116.t004

Looking at the extracted rules, it is evident that both algorithms agree on the direct relationship between chicken/duck densities and the risk of AI occurrence. This is consistent with the previous studies [15, 16], in particular the same results have been obtained for developing countries [59] and Indonesia [60].

A low or medium amount of precipitation (less than 300 millimetre per month) was associated with AI occurrence, which was detected by both algorithms. This pattern was aligned with the findings of other studies [12, 19, 20]. Also, both algorithms agreed with regard to finding a connection between rainy season (September-March) and AI occurrence. This might be similar to findings by Loth *et al.* [4] who outlined that wet summers can have a negative association with AI occurrence.

Table 5. Mined rules by FP-Growth algorithm.

Rule	Comparative Support
no waterfowls → LR	824
L chicken density → LR	634
M pressure and no waterfowls → LR	555
H temperature and no waterfowl → LR	537
L chicken density and no waterfowls → LR	536
L duck density → LR	519
M precipitation and no waterfowls → LR	511
VL wind speed and no waterfowls → LR	467
H chicken density → HR	433
H temperature and M precipitation → LR	427
VH humidity → LR	421
H duck density → HR	410
VL wind speed and M pressure → HR	376
Winter season → HR	367
VH chicken density → HR	359
M humidity → HR	335
H chicken density and M pressure → HR	334
Spring season → HR	330
H chicken density and H temperature → HR	324

https://doi.org/10.1371/journal.pone.0245116.t005

Table 6. Sample extracted sequential rules by PrefixSpan algorithm.

Year	Frequent sequences of cells with disease events
2010	356: $dC_H, dD_H, t_H, h_H, ws_L, pc_L, pr_M \rightarrow$ 253: $dC_H, dD_L, t_H, h_H, ws_L, pc_L, pr_M$
	253: $dC_H, dD_L, t_H, h_H, ws_L, pc_L, pr_M \rightarrow$ 356: $dC_H, dD_H, t_H, h_H, ws_L, pc_L, pr_M$
2011	315: $dC_H, dD_H, t_M, h_H, ws_L, pc_H, pr_M \rightarrow$ 356: $dC_H, dD_H, t_H, h_H, ws_L, pc_H, pr_M$
	335: $dC_H, dD_H, t_H, h_H, ws_L, pc_H, pr_M \rightarrow$ 315: $dC_H, dD_H, t_M, h_H, ws_L, pc_H, pr_M$
	356: $dC_H, dD_H, t_H, h_H, ws_L, pc_L, pr_M \rightarrow$ 253: $dC_H, dD_L, t_H, h_H, ws_L, pc_L, pr_M$
	253: $dC_H, dD_L, t_H, h_H, ws_L, pc_L, pr_M \rightarrow$ 356: $dC_H, dD_H, t_H, h_H, ws_L, pc_L, pr_M$
2012	315: $dC_H, dD_H, t_M, h_H, ws_L, pc_H, pr_M \rightarrow$ 356: $dC_H, dD_H, t_H, h_H, ws_L, pc_H, pr_M$
	335: $dC_H, dD_H, t_H, h_H, ws_L, pc_H, pr_M \rightarrow$ 315: $dC_H, dD_H, t_M, h_H, ws_L, pc_H, pr_M$
	315: $dC_H, dD_H, t_M, h_H, ws_L, pc_M, pr_M \rightarrow$ 356: $dC_H, dD_H, t_H, h_H, ws_L, pc_L, pr_M$
	356: $dC_H, dD_H, t_H, h_M, ws_M, pc_L, pr_M \rightarrow$ 315: $dC_H, dD_H, t_M, h_H, ws_L, pc_L, pr_M$
	356: $dC_H, dD_H, t_H, h_H, ws_L, pc_L, pr_M \rightarrow$ 253: $dC_H, dD_L, t_H, h_H, ws_L, pc_L, pr_M$
	253: $dC_H, dD_L, t_H, h_H, ws_L, pc_L, pr_M \rightarrow$ 356: $dC_H, dD_H, t_H, h_H, ws_L, pc_L, pr_M$
	253: $dC_H, dD_L, t_H, h_H, ws_L, pc_L, pr_M \rightarrow$ 355: $dC_H, dD_H, t_H, h_M, ws_L, pc_L, pr_M$

https://doi.org/10.1371/journal.pone.0245116.t006

While the RuleFit algorithm was able to detect the negative association between temperature and AI presence that has been previously outlined in the literature [5, 58], this association was not found using the FP-Growth algorithm.

In an additional analysis, we explored frequent sequences of cells with disease occurrence. A sample of outcome results for 2010 to 2012 is given in Table 6 and also visualized in Fig 5. Outcomes indicate the path between the regions of Lampung, West Java, Central Java, Yogyakarta and East Java. The most frequent paths were between the Lampung and Yogyakarta, West Java and Yogyakarta, East Java to Central Java, Lampung to Central Java, Central Java to West Java and West Java to Lampung.

Risk prediction

The classifier was parametrized with a range of thresholds as explained earlier and a precision-recall curve was generated. The curve is visualized in Fig 6 and represents the trade-off

Fig 5. Linkage sites of disease events.

https://doi.org/10.1371/journal.pone.0245116.g005

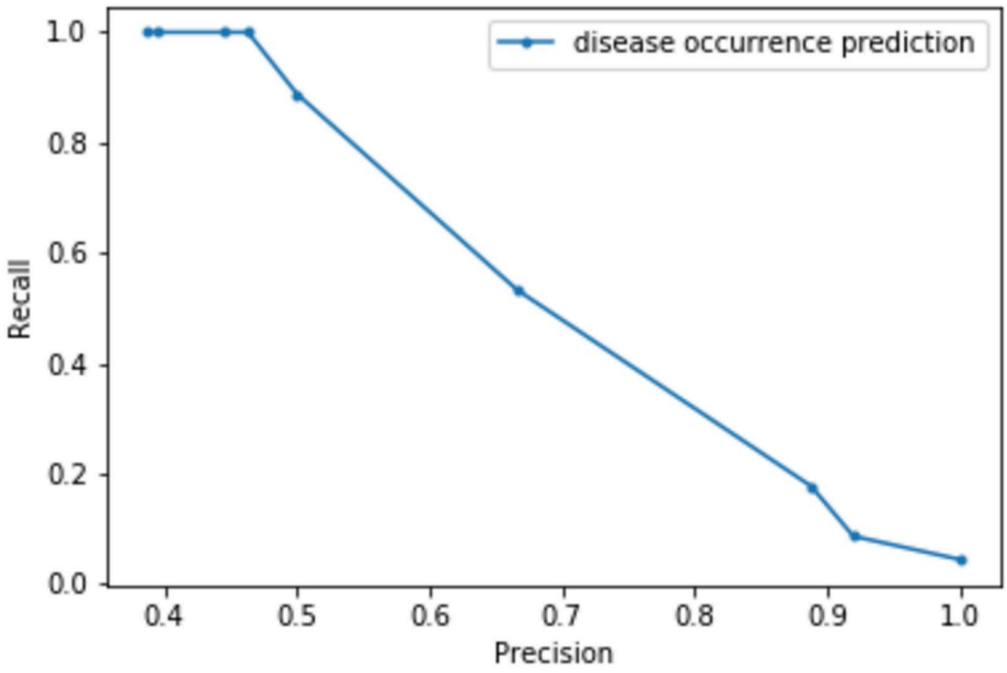

Fig 6. Precision-recall curve.

https://doi.org/10.1371/journal.pone.0245116.g006

between precision and recall measures of predictions. The blue dot points show the thresholds starting from minus four at the left and ending to four at the right of the graph. As the threshold is increased, the sensitivity decreases and the precision increases.

Since the goal of the classifier defined here is to predict disease event occurrence, having a high sensitivity is important. This is because, in such prediction systems, having less false negatives is more desirable than having less false positives. Therefore, a threshold of zero was selected.

The graph visualized in Fig 6 represents the trade-off between precision and recall measures. Based on Fig 6, at the threshold of zero, the classifier gains a sensitivity of about 88% and a positive predictive value of 50%. The high sensitivity means that the classifier is strong for correctly predicting the disease presence. The system also predicts actual disease absent points with a probability of 82.22%. These results are summarized in the Table 7.

Since the dataset here contains more negative than positive classes, precision and recall are better metrics to look at. A refined version of F_1-score called F_β-score (Eq 6) is more practical for imbalanced data since it allows for higher weighting of either precision or recall. Given the importance of sensitivity over precision in the current study, in addition to traditional F_1-score, F_β-score with $\beta = 2$ was reported.

Table 7. Evaluation measures of the rule-based classifier (threshold = 0).

Measure	Value
Sensitivity	88.88%
Specificity	82.22%
Positive predictive value	50%
F_1-score	64%
F_β-score	76.92%

https://doi.org/10.1371/journal.pone.0245116.t007

Table 8. Evaluation measures of the a Random Forest classifier (number of trees = 20).

Measure	Value
Sensitivity	56.8%
Specificity	82.22%
Positive predictive value	80%
F_1-score	63.3%
F_β-score	58.9%

https://doi.org/10.1371/journal.pone.0245116.t008

To compare the performance results with a basic classifier, we applied the Random Forest algorithm with 10-fold cross-validation on unseen data. The result reported in Table 8 shows that the performance of the proposed classifier is comparable with Random Forest. Although both classifiers gained the same F-measure, giving more importance to sensitivity (i.e. $\beta = 2$), the proposed model obtained a higher F_β-score. It should be noted that different resolutions and case studies in previous studies impede us to make a direct comparison of performance with them [61].

Moreover, we generated the cumulative gains curve, which was used to assess the performance of the prediction. It shows the percentage of targets reached when considering a certain percentage of the population. First, all the observations were ordered according to the output of the model. Therefore, observations with the highest rank were placed on the left-hand side of the horizontal axis. The vertical axis of the curve indicates which percentage of true positives included in the curve.

The chart can tell how much the model predicts better compared to a random selection. According to Fig 7, if we consider the 20% of the observations, the model will ensure that 88.89% of the true positives are in this group, while the random pick would provide only the 20% of the targets.

Conclusion and future work

Here, we have proposed a framework to discover hidden patterns from an extensive list of data sources using rule-discovery techniques. This framework facilitates the understanding of how AI predictors and occurrence data can be aggregated and pre-processed as input for rule-discovery techniques. Subsequently, a classifier was built from extracted rules to predict the disease presence in new circumstances. This approach is complementary to existing AI risk profiling methods. A rule-set used here can offer easier interpretations of predictions for end-users. This means that users can understand how predictions are made. Also, it is easy to identify which factors have contributed to the predictions and whether the predictions are reasonable. An understanding of risk-increasing factors and their interactions and building risk-profiling maps, can be useful in emergency preparedness. For instance, authorities may use such information to prioritize and target areas for interventions.

The outcome patterns in this study were consistent with earlier studies. For example, the positive association between disease presence and chicken/duck densities, waterfowl density and the negative association between disease presence and precipitation were aligned with previous studies. Nevertheless, the impact of temperature on disease occurrence showed contradictory results, which might be due to the tropical climate of Indonesia. In Indonesia, the temperature is usually high and does not change much during a year.

An important limitation of the present study is that the change in the distribution of data through the process of under-sampling without considering the impact of imbalanced data on the classification output might be misleading [62]. This is because the removal of examples

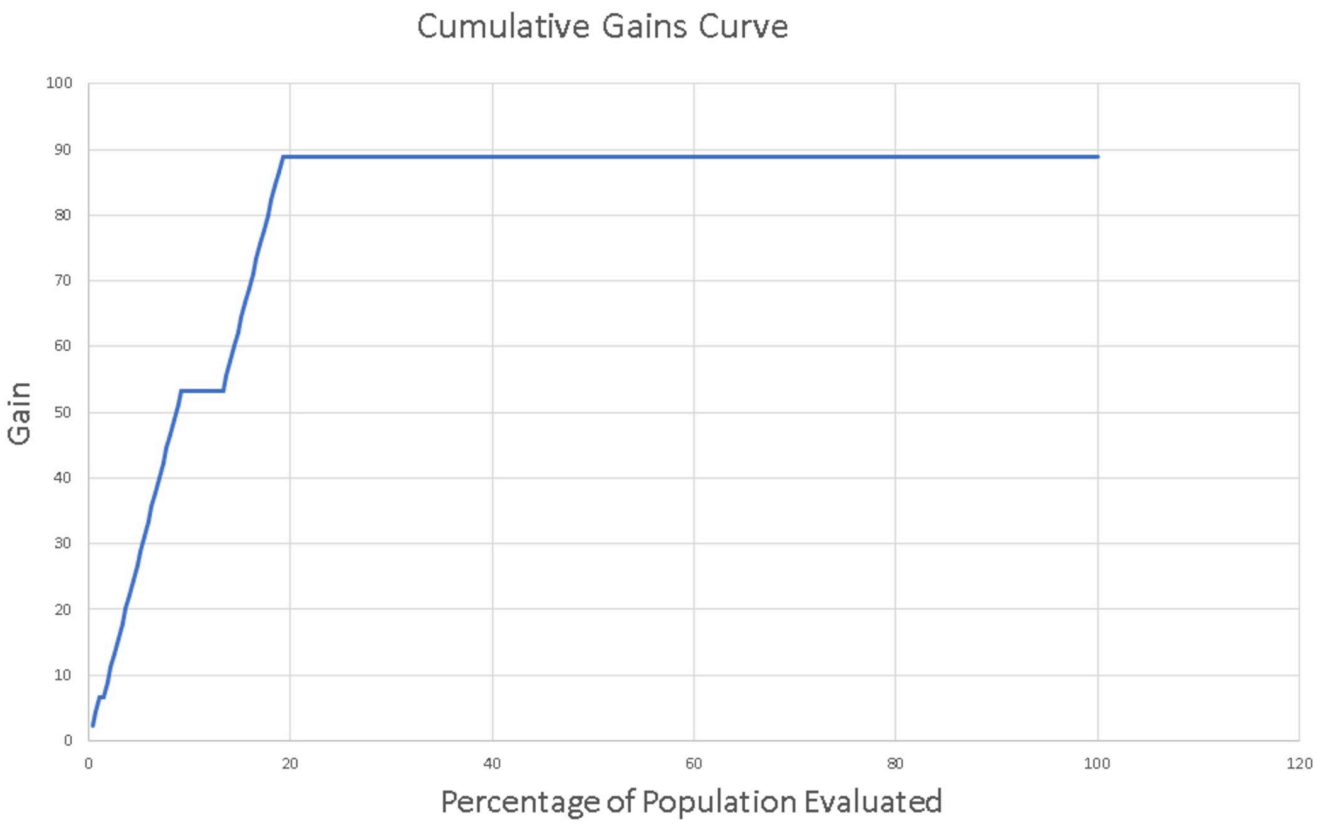

Fig 7. Cumulative gains curve.

https://doi.org/10.1371/journal.pone.0245116.g007

from the majority class could lead to the loss of potentially important information about the class [63]. Moreover, further work is required to improve the performance of predictions. One approach is the collection of additional data sources. For example, live bird trades information could be taken into consideration. Trading of live birds is known to be a major pathway of AI transmission that can happen through the movements of contaminated traders [64]. To improve the timeliness of predictions, a continuous pipeline from data collection to analysis is required. It means that during specified time intervals, data is automatically collected, pre-processed, integrated and analysed. The patterns in the rule-base will be updated in each interval, which ensures real-time predictions.

The proposed framework may provide public health officials and animal health authorities with warnings that can be used for identifying areas with a high risk of disease presence. Such information can potentially be used for response in high priority areas and executing interventions.

Supporting information

S1 File.
(ZIP)

Author Contributions

Conceptualization: Rozita Dara, Shayan Sharif.

Data curation: Samira Yousefinaghani.

Formal analysis: Rozita Dara.

Methodology: Samira Yousefinaghani, Rozita Dara, Shayan Sharif.

Supervision: Rozita Dara, Shayan Sharif.

Validation: Samira Yousefinaghani, Rozita Dara, Zvonimir Poljak, Fei Song.

Writing – original draft: Samira Yousefinaghani.

Writing – review & editing: Rozita Dara, Zvonimir Poljak, Fei Song, Shayan Sharif.

References

1. Gilbert M, Xiao X, Pfeiffer DU, Epprecht M, Boles S, Czarnecki C, et al. Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. Proceedings of the National Academy of Sciences. 2008; 105(12):4769–4774. https://doi.org/10.1073/pnas.0710581105 PMID: 18362346
2. Martin V, Pfeiffer DU, Zhou X, Xiao X, Prosser DJ, Guo F, et al. Spatial distribution and risk factors of highly pathogenic avian influenza (HPAI) H5N1 in China. PLoS Pathogens. 2011; 7(3):e1001308. https://doi.org/10.1371/journal.ppat.1001308 PMID: 21408202
3. Henning J, Pfeiffer DU, et al. Risk factors and characteristics of H5N1 Highly Pathogenic Avian Influenza (HPAI) post-vaccination outbreaks. Veterinary Research. 2009; 40(3):1. https://doi.org/10.1051/vetres:2008053 PMID: 19081006
4. Loth L, Gilbert M, Osmani MG, Kalam AM, Xiao X. Risk factors and clusters of highly pathogenic avian influenza H5N1 outbreaks in Bangladesh. Preventive Veterinary Medicine. 2010; 96(1-2):104–113. https://doi.org/10.1016/j.prevetmed.2010.05.013
5. Biswas PK, Islam MZ, Debnath NC, Yamage M. Modeling and roles of meteorological factors in outbreaks of highly pathogenic avian influenza H5N1. PloS One. 2014; 9(6):e98471. https://doi.org/10.1371/journal.pone.0098471
6. Wang X, Wang Q, Cheng W, Yu Z, Ling F, Mao H, et al. Risk factors for avian influenza virus contamination of live poultry markets in Zhejiang, China during the 2015–2016 human influenza season. Scientific Reports. 2017; 7:42722. https://doi.org/10.1038/srep42722 PMID: 28256584
7. Yupiana Y, de Vlas SJ, Adnan NM, Richardus JH. Risk factors of poultry outbreaks and human cases of H5N1 avian influenza virus infection in West Java Province, Indonesia. International Journal of Infectious Diseases. 2010; 14(9):e800–e805. https://doi.org/10.1016/j.ijid.2010.03.014
8. Dhingra MS, Dissanayake R, Negi AB, Oberoi M, Castellan D, Thrusfield M, et al. Spatio-temporal epidemiology of highly pathogenic avian influenza (subtype H5N1) in poultry in eastern India. Spatial and Spatio-temporal Epidemiology. 2014; 11:45–57. https://doi.org/10.1016/j.sste.2014.06.003 PMID: 25457596
9. Alkhamis M, Hijmans RJ, Al-Enezi A, Martínez-López B, Perea AM. The use of spatial and spatiotemporal modeling for surveillance of H5N1 highly pathogenic avian influenza in poultry in the Middle East. Avian Diseases. 2016; 60(1s):146–155. https://doi.org/10.1637/11106-042115-Reg
10. Sun L, Ward MP, Li R, Xia C, Lynn H, Hu Y, et al. Global spatial risk pattern of highly pathogenic avian influenza H5N1 virus in wild birds: A knowledge-fusion based approach. Preventive Veterinary Medicine. 2018; 152:32–39. https://doi.org/10.1016/j.prevetmed.2018.02.008 PMID: 29559103
11. Bui CM, Gardner L, MacIntyre R, Sarkar S. Influenza A H5N1 and H7N9 in China: A spatial risk analysis. PloS One. 2017; 12(4):e0174980. https://doi.org/10.1371/journal.pone.0174980
12. Dizon FSV, Farinas SKR, Mahinay Jr RJTH, Pardo HS, Delfinado CJA. Learning of high dengue incidence with clustering and FP-Growth algorithm using WHO historical data. arXiv preprint arXiv:190111376. 2019;.
13. Mu JE, McCarl BA, Wu X, Ward MP. Climate change and the risk of highly pathogenic avian influenza outbreaks in birds. British Journal of Environment and Climate Change. 2014; 4(2):166. https://doi.org/10.9734/BJECC/2014/8888
14. Tian H, Zhou S, Dong L, Van Boeckel TP, Pei Y, Wu Q, et al. Climate change suggests a shift of H5N1 risk in migratory birds. Ecological Modelling. 2015; 306:6–15. https://doi.org/10.1016/j.ecolmodel.2014.08.005
15. Dhingra MS, Artois J, Robinson TP, Linard C, Chaiban C, Xenarios I, et al. Global mapping of highly pathogenic avian influenza H5N1 and H5Nx clade 2.3. 4.4 viruses with spatial cross-validation. Elife. 2016; 5:e19571. https://doi.org/10.7554/eLife.19571 PMID: 27885988

16. Belkhiria J, Hijmans RJ, Boyce W, Crossley BM, Martínez-López B. Identification of high risk areas for avian influenza outbreaks in California using disease distribution models. PloS One. 2018; 13(1): e0190824. https://doi.org/10.1371/journal.pone.0190824

17. Gilbert M, Golding N, Zhou H, Wint GW, Robinson TP, Tatem AJ, et al. Predicting the risk of avian influenza A H7N9 infection in live-poultry markets across Asia. Nature Communications. 2014; 5:4116. https://doi.org/10.1038/ncomms5116 PMID: 24937647

18. Artois J, Jiang H, Wang X, Qin Y, Pearcy M, Lai S, et al. Changing geographic patterns and risk factors for avian influenza A (H7N9) infections in humans, China. Emerging Infectious Diseases. 2018; 24 (1):87. https://doi.org/10.3201/eid2401.171393 PMID: 29260681

19. Fang LQ, de Vlas SJ, Liang S, Looman CW, Gong P, Xu B, et al. Environmental factors contributing to the spread of H5N1 avian influenza in mainland China. PloS One. 2008; 3(5):e2268. https://doi.org/10.1371/journal.pone.0002268 PMID: 18509468

20. Si Y, Wang T, Skidmore AK, de Boer WF, Li L, Prins HH. Environmental factors influencing the spread of the highly pathogenic avian influenza H5N1 virus in wild birds in Europe. Ecology and Society. 2010; 15(3):26–26. https://doi.org/10.5751/ES-03622-150326

21. Stevens KB, Gilbert M, Pfeiffer DU. Modeling habitat suitability for occurrence of highly pathogenic avian influenza virus H5N1 in domestic poultry in Asia: a spatial multicriteria decision analysis approach. Spatial and Spatio-temporal Epidemiology. 2013; 4:1–14. https://doi.org/10.1016/j.sste.2012.11.002

22. Gilbert M, Xiao X, Chaitaweesub P, Kalpravidh W, Premashthira S, Boles S, et al. Avian influenza, domestic ducks and rice agriculture in Thailand. Agriculture, Ecosystems & Environment. 2007; 119(3-4):409–415. https://doi.org/10.1016/j.agee.2006.09.001 PMID: 18418464

23. Peterson AT, Benz BW, Papeş M. Highly pathogenic H5N1 avian influenza: entry pathways into North America via bird migration. PLoS One. 2007; 2(2):e261. https://doi.org/10.1371/journal.pone.0000261

24. Kilpatrick AM, Chmura AA, Gibbons DW, Fleischer RC, Marra PP, Daszak P. Predicting the global spread of H5N1 avian influenza. Proceedings of the National Academy of Sciences. 2006; 103 (51):19368–19373. https://doi.org/10.1073/pnas.0609227103

25. Ward M, Maftei D, Apostu C, Suru A. Association between outbreaks of highly pathogenic avian influenza subtype H5N1 and migratory waterfowl (family Anatidae) populations. Zoonoses and Public Health. 2009; 56(1):1–9. https://doi.org/10.1111/j.1863-2378.2008.01150.x

26. Feare CJ. Role of wild birds in the spread of highly pathogenic avian influenza virus H5N1 and implications for global surveillance. Avian Diseases. 2010; 54(s1):201–212. https://doi.org/10.1637/8766-033109-ResNote.1

27. Soliman A, Saad M, Elassal E, Amir E, Plathonoff C, Bahgat V, et al. Surveillance of avian influenza viruses in migratory birds in Egypt, 2003–09. Journal of Wildlife Diseases. 2012; 48(3):669–675. https://doi.org/10.7589/0090-3558-48.3.669 PMID: 22740532

28. WorldClim. Global Climate Data; Accessed December 2019.

29. Si Y, de Boer WF, Gong P. Different environmental drivers of highly pathogenic avian influenza H5N1 outbreaks in poultry and wild birds. PloS One. 2013; 8(1):e53362. https://doi.org/10.1371/journal.pone.0053362

30. Young S, Carrel M, Malanson G, Ali M, Kayali G. Predicting avian influenza co-infection with H5N1 and H9N2 in Northern Egypt. International Journal of Environmental Research and Public Health. 2016; 13 (9):886. https://doi.org/10.3390/ijerph13090886

31. Guerrini L, Paul MC, Leger L, Andriamanivo HR, Maminiaina OF, Jourdan M, et al. Landscape attributes driving avian influenza virus circulation in the Lake Alaotra region of Madagascar. Geospatial Health. 2014; 8(2):445–453. https://doi.org/10.4081/gh.2014.33 PMID: 24893021

32. Wu T, Perrings C. The live poultry trade and the spread of highly pathogenic avian influenza: Regional differences between Europe, West Africa, and Southeast Asia. PloS One. 2018; 13(12):e0208197. https://doi.org/10.1371/journal.pone.0208197

33. Gaidet N, Cappelle J, Takekawa JY, Prosser DJ, Iverson SA, Douglas DC, et al. Potential spread of highly pathogenic avian influenza H5N1 by wildfowl: dispersal ranges and rates determined from large-scale satellite telemetry. Journal of Applied Ecology. 2010; 47(5):1147–1157. https://doi.org/10.1111/j.1365-2664.2010.01845.x

34. Lin Y, Huang S, Simon GE, Liu S. Data-based decision rules to personalize depression follow-up. Scientific Reports. 2018; 8(1):5064. https://doi.org/10.1038/s41598-018-23326-1

35. Lin Y, Qian X, Krischer J, Vehik K, Lee HS, Huang S. A rule-based prognostic model for type 1 diabetes by identifying and synthesizing baseline profile patterns. PloS One. 2014; 9(6):e91095. https://doi.org/10.1371/journal.pone.0091095

36. Haghighi M, Johnson SB, Qian X, Lynch KF, Vehik K, Huang S, et al. A comparison of rule-based analysis with regression methods in understanding the risk factors for study withdrawal in a pediatric study. Scientific Reports. 2016; 6:30828. https://doi.org/10.1038/srep30828 PMID: 27561809

37. Xu Z, Lee J, Park D, Chung Y. Multidimensional analysis model for highly pathogenic avian influenza using data cube and data mining techniques. Biosystems Engineering. 2017; 157:109–121. https://doi.org/10.1016/j.biosystemseng.2017.03.004

38. Huysmans J, Baesens B, Vanthienen J. Using rule extraction to improve the comprehensibility of predictive models. Available at SSRN 961358. 2006; p. 1–56.

39. Darksky. Dark Sky API; Accessed August 2019.

40. BirdLife. International and Handbook of the Birds of the World (2018) Bird species distribution maps of the world. Version 2018.1.; Accessed August 2019.

41. GLW 3. Gridded Livestock of the World; Accessed August 2019.

42. EMPRES-i. Global Animal Disease Information System (EMPRES-i) of the Food and Agriculture Organization of the United Nations (FAO); Accessed August 2019.

43. Welte VR, Terán MV. Emergency prevention system (EMPRES) for transboundary animal and plant pests and diseases. the EMPRES-livestock: an FAO initiative. Annals of the New York Academy of Sciences. 2004; 1026(1):19–31. https://doi.org/10.1196/annals.1307.003

44. Erraguntla M, Ramachandran S, Wu CN, Mayer RJ. Avian influenza datamining using environment, epidemiology, and etiology surveillance and analysis toolkit (E3SAT). In: Proceedings of the 2010 43rd Hawaii International Conference on System Sciences. Washington, DC, USA; 2010. p. 1–7.

45. Belkhiria J, Alkhamis MA, Martínez-López B. Application of species distribution modeling for avian influenza surveillance in the United States considering the North America migratory flyways. Scientific Reports. 2016; 6:33161. https://doi.org/10.1038/srep33161

46. He F, Hu Zj, Zhang Wc, Cai L, Cai Gx, Aoyagi K. Construction and evaluation of two computational models for predicting the incidence of influenza in Nagasaki Prefecture, Japan. Scientific Reports. 2017; 7(1):7192. https://doi.org/10.1038/s41598-017-07475-3

47. Wint G, Robinson T. Gridded livestock of the world 2007. Rome: Food and Agricultural Organization of the United Nations. Animal Production and Health Division. 2007; p. 131.

48. rasterio 1 0 25. Rasterio Python library; Accessed August 2019.

49. Friedman JH, Popescu BE, et al. Predictive learning via rule ensembles. The Annals of Applied Statistics. 2008; 2(3):916–954. https://doi.org/10.1214/07-AOAS148

50. Molnar C, et al. Interpretable machine learning: A guide for making black box models explainable. E-book at < https://christophmgithubio/interpretable-ml-book/>, version dated. 2018; 10.

51. Bologna G, Hayashi Y. A comparison study on rule extraction from neural network ensembles, boosted shallow trees, and SVMs. Applied Computational Intelligence and Soft Computing. 2018; 2018(1):1–20.

52. Bénard C, Biau G, Da Veiga S, Scornet E. SIRUS: making Random Forests interpretable; 2019.

53. Han J, Cheng H, Xin D, Yan X. Frequent pattern mining: current status and future directions. Data mining and knowledge discovery. 2007; 15(1):55–86. https://doi.org/10.1007/s10618-006-0059-1

54. Said AM, Dominic P, Abdullah AB. A comparative study of fp-growth variations. International journal of computer science and network security. 2009; 9(5):266–272.

55. Molnar C. Interpretable machine learning. Lulu, 1st edition; eBook (GitHub, 2019-06-19); 2019.

56. Pei J, Han J, Mortazavi-Asl B, Wang J, Pinto H, Chen Q, et al. Mining sequential patterns by pattern-growth: The prefixspan approach. IEEE Transactions on Knowledge and Data Engineering. 2004; 16 (11):1424–1440. https://doi.org/10.1109/TKDE.2004.77

57. Chiu DY, Wu YH, Chen AL. An efficient algorithm for mining frequent sequences by a new strategy without support counting. In: Proceedings of 20th International Conference on Data Engineering. Boston, MA, USA; 2004. p. 375–386.

58. Brown JD, Goekjian G, Poulson R, Valeika S, Stallknecht DE. Avian influenza virus in water: infectivity is dependent on pH, salinity and temperature. Veterinary Microbiology. 2009; 136(1-2):20–26. https://doi.org/10.1016/j.vetmic.2008.10.027

59. Pavade G, Awada L, Hamilton K, Swayne D, et al. The influence of economic indicators, poultry density and the performance of veterinary services on the control of high-pathogenicity avian influenza in poultry. Revue Scientifique et Technique-OIE. 2011; 30(3):661. https://doi.org/10.20506/rst.30.3.2064 PMID: 22435180

60. Ge E, Haining R, Li CP, Yu Z, Waye MY, Chu KH, et al. Using knowledge fusion to analyze avian influenza H5N1 in East and Southeast Asia. PLoS One. 2012; 7(5):e29617. https://doi.org/10.1371/journal.pone.0029617 PMID: 22615729

61. Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography. 2008; 17(2):145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x

62. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence. 2016; 5(4):221–232. https://doi.org/10.1007/s13748-016-0094-0

63. Ali A, Shamsuddin SM, Ralescu AL, et al. Classification with class imbalance problem: a review. Int J Advance Soft Compu Appl. 2015; 7(3):176–204.

64. Fournié G, Tripodi A, Nguyen TTT, Tran TT, Bisson A, Pfeiffer DU, et al. Investigating poultry trade patterns to guide avian influenza surveillance and control: a case study in Vietnam. Scientific reports. 2016; 6:29463. https://doi.org/10.1038/srep29463 PMID: 27405887