Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning Framework for Porcine Reproductive and Respiratory Syndrome Outbreak Forecasting

Mohammadsadegh Shamsabardeh (mshamsabardeh@ucdavis.edu)
University of California, Davis
Beatriz Martínez-López
University of California, Davis
Kathleen C. O'Hara
University of California, Davis
Jose Pablo Gomez
University of California, Davis
Xin Liu
University of California, Davis

Article

Keywords:

Posted Date: June 29th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1785633/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Porcine Reproductive and Respiratory Syndrome (PRRS) is one of the most challenging and costly viral infectious diseases impacting the swine industry. The disease transmission pathways for PRRS are very complex, requiring a combined approach of intensive surveillance (i.e., testing), biosecurity, and vaccination for control and eradication. This study builds a proactive framework to forecast the risk of having a PRRS outbreak on a farm. This forecasting allows for early detection of disease outbreaks and could direct risk-based, and thus more cost-effective, interventions. Machine learning algorithms were trained using multi-scale data (pig group-, farm-, and area-level data). For the first time, on-farm, between-farm, and environmental variables, including farm location, pig movements, production parameters, diagnostic data, and climatic information, were combined for the prediction of PRRS outbreaks. Multi-scale datasets were merged via feature extraction, followed by the wrapper and filter feature selection, to find those feature subsets with the best forecasting performance. The predictive value of each feature selection mechanism was evaluated in terms of its stability. Numerical results demonstrate good forecasting performance in terms of area under the ROC curve.

1 Introduction

The livestock industry capitalizes on the production of the highest quality animals through the most economically efficient means. The success of a given producer relies on their ability to maintain the health of their herds through good management practices, and the capacity to prevent, detect and control both endemic and epidemic diseases.

The US is the world's second largest pork producer and the second largest meat exporter (North American

Meat Institute, 2016). Within the US, most pigs are raised within multi-site swine production systems (i.e. separate facilities by pig type and age), allowing for specialized housing and feed. However, this multisite system intrinsically requires the frequent movement of live animals between sites, providing a source of disease movement and introduction. Further, these intensive production systems create environments of high pig density, which increases the risk of disease spread. Porcine Reproductive and Respiratory Syndrome Virus (PRRSV) is currently the most challenging and costly viral infectious disease in the US swine industry, accounting for over $660M in losses annually [1]. Among these outbreaks, 55% are associated with growing pigs and 45% with breeding farms. The high viral mutation rate seen in PRRSV results in high levels of sequence variability, making vaccine development and implementation a challenge [2]. The high cost of diagnostic screening tests, biosecurity (e.g., air filtration) and vaccination, as well as the direct losses associated with outbreaks, highlight the need to develop forecasting models to help identify farms at highest risk of having an outbreak. Such models allow more cost-effective and efficient disease mitigation efforts, with risk-based surveillance, vaccination and outbreak response strategies.

The current approach to PRRSV control includes the maintenance of high biosecurity, routine disease surveillance via diagnostic testing, and the use of standard vaccine protocols [3]. Serologic and molecular diagnostic tests are available for use on blood, oral fluids, and tissue samples from live and dead pigs [4]. The shedding (using PCR) and exposure (using ELISA) status of a herd can be determined based on the results of these tests. For breeding herds (sow and nursery farms) there are four disease status categories: (I) positive unstable, (II) positive stable, (III) provisional negative and (IV) negative [5]. Growing herds (finishing herds) are classified as either positive or negative status. The challenge is that untested farms have uncertain status and cannot be easily categorized as positive or negative. Some farm managers accept the risk of an outbreak rather than continuously running tests. Therefore, the level of diagnostic information, as well as the biosecurity and vaccination protocols, may vary by farm.

The aim of this study is to examine different machine learning models and to explore those variables or features that would most effectively forecast and enable the early detection of PRRSV outbreaks [6]. This is a study based on multi-scale data (pig group-, farm-, area- level data). On-farm, between-farm, and environmental variables, including farm location, pig movements [7], production parameters, diagnostic data, and climatic information are evaluated. The ability to forecast high risk farms can inform strategies for more efficient testing and targeted mitigation plans to reduce the impact of PRRSV on the swine industry. This study focuses on finishing farms, which currently have the lowest frequency of disease screening and the lowest standards of immunization and biosecurity. Finishing farms could greatly benefit from a system that helps to forecast outbreaks. Importantly, improving health outcomes at finishing farms would contribute to reducing the burden of disease transmission to breeding herds, thus improving the health status of the entire system.

PRRSV transmission can occur by both direct and indirect contacts. The two main modes of PRRSV between-farm transmission are 1) the transportation of infected live pigs and 2) airborne transmission from nearby infected farms [8]. Other indirect routes of transmission include the use of infected semen, contaminated personnel, tools or materials, or insects which can act as mechanical vectors. In this study, we just considered the two main pathways for disease transmission: direct transmission through the reception of pigs from other farms, and indirect airborne transmission from nearby farms.

Different features are created to represent these disease pathways and other risk factors that may contribute to PRRSV epidemics. In general, adding additional features potentially increases the accuracy of a forecasting model. However, using a large number of features with comparably few data samples can result in overfitting to training data, and consequently, decreases the generalization of the model to new data samples [9]. To combat this issue, feature selection methods are used. Feature selection is the process of selecting a subset of relevant features that are useful for predicting response variables. In this work, filter method feature selection based on correlation [10], and wrapper method based on recursive feature elimination (RFE) [11–13], are used to find the most relevant features influencing PRRSV outbreaks. Furthermore, to compare the robustness of each feature selection algorithm with respect to different training data samples, stability analysis using Tanimoto distance is performed [14].

Overall, this work examines multiple machine learning models for outbreak forecasting and early detection in finishing farms using a combination of diagnostic, production, and pig trade data. This work demonstrates the strength of these techniques and provides the basis for future real-time dashboards that can allow producers to actively monitor and respond to shifting disease dynamics on their farms.

2 Data And Feature Generation

In this section, the data source and structure, data pre-processing steps to build the features, and methods to forecast the probability of having a PRRSV outbreak are explained.

2.1 Data Sources

This study is conducted based on one large-scale swine production system with multiple sow, nursery, and finishing farms in the midwest of the United States.

For the time period 2006–2019, a rich database from this system provided information on the movement of pigs between farms, production of the farms, and PRRSV testing results. During this period, there were over 230,000 movement records to or from farms within the production system. For each movement entry, the source and destination, the farm type, the number of shipped pigs, the total weight of the shipped pigs, and the date of the movement are available.

At each finishing farm, the period of time from the first pig entering the farm to the last one leaving the farm is defined as Finishing Period (FP). Lab results demonstrate that 620 out of the 3770 FP during our study period experienced at least one outbreak. In practice, most of the farms are tested only when there is evidence of health problems on the farm. Thus the lab results are positive for almost all submitted samples and negative samples are not statistically representative of the negative class. To build a machine learning model that can classify negative and positive samples, samples for both classes are needed. In this study, domain knowledge expertise is used to define criteria for an assumed negative classification. A farm is assumed to be negative if it meets two conditions: the mortality rate is in the lowest 10 percent (i.e., in the 10th percentile), and the percentage of exiting pigs with weight in the standard range is in top 10 percent (i.e. at 90th percentile). This results in 5 percent of FPs being negative.

For each farm, climate information was obtained from the closest weather station. The data was obtained using the R package ‘riem’, which queries the data from an online interface to obtain weather information. The location of the weather station is not reported for data confidentiality. Temperature (f_5), relative humidity (f_6), wind speed (f_7) and altitude (f_8) are considered for this study, where $f_k$refers to the feature k in the Supplementary table.

2.2 Data Pre-Processing

2.2.1 Data Cleansing

Identification and correction of inaccurate data is an important step of data analysis. Using incomplete or inaccurate data samples in the training procedure may lead to poor model performance. Hence, the data were extensively analyzed to correct incomplete or inaccurate data samples. Some fields, such as weather information, were missing for several records. Missing fields were assigned the average value of the same period of time in the previous years. Weather related missing fields associated with each location (farm) are replaced with the average taken over previous years of the same period of time (e.g., month). Due to discrepancies in the naming system in production data and lab results we could not associate some production data to the health status of the farm. Records with such inconsistencies were removed from the dataset. Moreover, inaccurate and invalid records were removed by applying a set of rational range constraints. Specifically, in some records the number of dead and survived pigs did not add up to the number of pigs entering the farm. Also, some weights of the pigs were out of the reasonable range for that type of a farm. These records were removed.

2.2.2 Feature Engineering

Machine learning can provide good predictions if it can extract the relevant information from the data. This means that its success depends on both the goodness of the model and the data representation, the transformation of the raw data into feature vectors. The better the data representation, the simpler the deployed model can be for the same performance metrics, meaning less chance of over-fitting and better generalization. Feature engineering is the manual construction of features from raw data. The importance of data representation and feature engineering becomes clearer when the number of data samples are small compared to the model complexity required to capture the relationship between dependent and independent parameters. The feature engineering step is the most time-intensive step of this work.

Domain knowledge is key to the construction of relevant features. One key contribution of this work is to construct features that represent those factors that affect the risk of having an outbreak on a farm. This paper combines data across different scales for better forecasting. After the construction of different features, feature selection methods can be used to evaluate whether a feature is improving the forecasting performance or not. For example, to incorporate the effect of temperature, features representing different seasons (f_{57-60}) and average temperature (f_5) were created based on the expectation that the spread of PRRSV would follow different patterns in warm [15] versus cold [16] temperatures.

First Pathway (Direct Contact) Features

To model the PRRSV transmission through direct contacts, different risk factors were considered. Most pigs are able to clear PRRSV infection after getting infected, but some become persistently infected and can then act as carriers, spreading the virus if that pig is transferred to another farm. To capture this effect, a feature () was created representing the number of entering pigs that are coming from a farm that has had an outbreak during the lifetime of that pig on the farm, i.e. if the nursery that the pig is coming from had an outbreak during that pig's lifetime. In addition, the total number of times that pigs enter a

farm during a given FP (), and the total number of different sources that pigs are coming from (), are additional risk factors. The total number of pigs on a farm () was also considered.

Second Pathway (Airborne) Features

The second pathway is through airborne transmission from nearby farms. To model this pathway, the vicinity of the farm was defined as the circular area around the farm within a defined distance. For each farm, the total number of movements () and number of pigs (), entering or exiting the vicinity were calculated. In addition, different neighborhood sizes (vicinity diameters) of 5km, 10km and 20km were examined in this study. Figure 1 depicts the neighborhood for farm F1 at the center of the circle. The dashed red arrows show the movements that the model counts for airborne effect. The solid red arrows represent direct contact movements.

Each movement feature has two versions based on time period: 1) from the start of the FP up to its forecasting date (all features indicated with FP in the FP/H column in the Supplementary table), and 2) the historical equivalent of this feature for the one year prior to FP start date (all features indicated with H in the FP/H column in the Supplementary table). These two sets of features are highly correlated, but together can indicate how the current FP movements are different from what is expected on average for the farm's neighborhood. Each of the features for current FP were normalized by dividing the feature value by the period of time for which they were calculated.

A farm with a higher density of neighboring farms (f_{9-11}) is at higher risk for having an outbreak. More importantly, the number of outbreaks happening in the neighborhood of the farm one year prior to the start of FP, represent how risky the area is (f_{50}).

Figure 1: The two main PRRSV pathways for farm F3:1) Airborne effect from neighboring farms (movements (orange color) with source or destination to farms that are located in the circle), 2) direct reception of pigs (blue colors). Other movements (grey) and farms are assumed to have no effect.

Production Features

Production data include total feed consumed and exiting weight at the end of the FP. This information cannot be used for the purpose of forecasting an outbreak for that same FP because it will violate causality. However, such data can be used for the evaluation of future FPs, because it is a good indicator of the overall performance/management practices/risk of a farm. Thus, features for historical production data for each farm were built (). The total number of pigs (), and the average weight of pigs entering a farm (), are two features that can be used for the current FP.

The following features based on historical data were used for each FP. Based on the weight of existing pigs, the percentage of sub-standard pigs (f_{23}) was calculated by dividing the number of surviving pigs that were not within the standard weight range by the total number of survived pigs. In addition, the average weight of sub-standard pigs on that farm was determined (f_{24}). Similarly, the percent (f_{18}) and average weight of exiting pigs falling within the standard range (f_{20}) were calculated. The total net

weight survived is the weight difference of survived pigs from entrance to exit and is divided by the number of survived pigs to obtain average net weight survived. Total pig days is the number of days pigs spend on a farm. The total net weight survived can be divided by the total pig days to obtain Average Daily Gain (ADG) (f_{35}). Similarly, the Average Daily Feed (ADF) (f_{39}) was calculated as the ratio of total consumed feed and total pig days. Next, the ratio between ADG and ADF (f_{40}) provides the net weight survived per one pound of food consumed. From the information on the number of days on-farm the following features were calculated for dead and survived pigs: the total live pig days (f_{27}) is the number of days that survived pigs were alive; total dead pig days (f_{28}) is the total number of days that pigs were on-farm before their death. The total live pig days and total dead pig days were divided by the number of survived pigs and the number of dead pigs, respectively ($f_{30,31}$). All these are used as features.

Other Overall Management Practices/Performance Features Additional features that can demonstrate the general management practices of a farm were created. Good management practices include the disinfection of the farm before the start of each FP. Additionally, receiving new animals all together within a few days of the start of the FP (i.e. all-in all-out), is considered a better practice than to allow continual additions throughout the FP. The continuous reception of pigs, versus a single time point population of the farm, results in a staggered cycle of animals leaving the farm at different time points, meaning there is no time at which the entire farm can be disinfected, allowing for the possible retention of infection from previous FPs. To represent this risk factor, a feature was created for the number of days between the first and last reception of new pigs. The percentage of time that a farm has had an outbreak (f_{41}) in the past is a mixed indicator of all of the above mentioned historical factors.

Climate Features In addition to wind speed, relative humidity, and altitude provided in the climate data, the average and lowest temperature of the past 15, 30, 45, 60 and 90 days prior to the forecasting date were built. It was not deemed necessary to have exact temperature measurements for every time point given the expected variability between true on-farm temperature values versus those recorded by weather stations due to on-site thermoregulation, distance from the weather station, and data collection in windows of a minimum of 15 days. Missing weather data were removed, and the average temperature for the past 60 days (f_5) was selected as the best temperature feature for the model. The season of the FP was defined as the season of the forecasting day. It is a categorical feature and was represented using one hot encoding. In other words, four binary features were created, and a single value was assigned to a given FP based on the corresponding season for a given forecasting day.

3 Machine Learning Methods

In this section, the machine learning methods that were used to forecast the PRRSV outbreak probability are explained. As in the other classification problems, the goal was to find the discriminator function that most efficiently mapped features to target labels.

3.1 Standard Machine Learning Models

Various machine learning algorithms including: Logistic Regression (LR), Support Vector Machine (SVM), Gradient Boosting (GB), and Random Forest (RF) were trained to forecast the PRRSV status of a farm. The probability of each farm being classified as positive for PRRSV is obtained from the output of each of these models. A farm is identified as positive by the model if its probability of infection is higher than a given threshold. Thus, metrics such as accuracy, sensitivity and specificity are dependent on this threshold. The Receiver Operating Characteristic (ROC) does not have this issue as it can be computed for every possible threshold. ROC curve shows how true positive rate (sensitivity) changes with false positive rate (1- specificity) for different thresholds. Therefore, the Area Under Curve (AUC) of the ROC is a good metric for comparing different models and is used here.

3.2 Cross Validation and Hyper-parameter Tuning

Given the past and the present observations on the health status of a farm and its features, the goal of this work is to predict the *future* health status of the farm. Therefore, all data about events that occur chronologically after the time of forecasting should be withheld and not used for prediction. The data were therefore split temporally into two non-overlapping parts for training and testing. The first portion was used for model training, while the second half was reserved for performance evaluation of the model.

Each model has different hyper-parameters that govern its complexity. We tuned hyper-parameters, namely, learning rate, tree depth, etc., to find the best fit model, i.e., to prevent the model from both overfitting and underfitting. To achieve this, the training data were further divided temporally into two chunks. The first chunk included 80 percent of the training data and was used to train the models with different sets of hyper-parameters; the remaining 20 percent of training data was used to test these models and find the best hyper-parameters values. First, hyperparameter-tuning was performed using the training data by performing a grid search over a range of values for the hyperparameters in different types of models. Specifically, each of these models was trained using eighty percent of the training data, and then tested over the remaining twenty percent for each collection of values of the hyperparameters to find the best hyper-parameters values. Using these best hyper-parameters values, a performance evaluation was then carried out by repeatedly training the model using a randomly selected subset (60%) of the training set and reporting the area under the receiver operator characteristic curve (AUC-ROC) on a randomly selected subset (80%) of the test set. Note that this approach for model validation was chosen so that all the points used for testing come chronologically after the ones used in training as the standard K-fold cross validation is not appropriate for time series data.

3.3 Feature Selection

The process of selecting features with the highest contribution towards forecasting the output is called feature selection. Having a high-dimensional feature space can cause the training algorithm to have impaired learning performance, be prone to overfitting, and become computationally cumbersome. The main goal of feature selection is either: 1) to find the subset of features that minimizes generalization

error, or 2) to select the smallest possible subset of features that satisfies the performance criterion and allows for better model interpretability. The main approach in this work is the latter.

Prior to the feature selection, we used hierarchical clustering based on within the feature correlation matrix to observe the degree to which the features are correlated. Hierarchical clustering groups the features such that the features constituting one group have more similarity among themselves than features in the other groups.

Feature selection methods are categorized into wrapper, filter and embedded methods. In this study, we used filter and wrapper methods as explained below:

Filter Methods

The filter method performs a feature selection procedure regardless of the type of learning model. A scoring measure based on data characteristics such as distance, information, or correlation, is used as the metric to filter those features that seem more relevant to the response variable. The filtering in this work was done based on Pearson correlation and mutual information.

Pearson correlation measures the similarity between two variables. In a univariate method, the correlation between each feature and the response variable is obtained, and feature selection is done based on the correlation with the response variable (target). Another popular filter method is a mutual information-based feature selection [17] which uses mutual information as the entropy measure to choose the subset of important features. Mutual information is a measure between two random variables, e.g., each input feature and the response variable, that quantifies the amount of information obtained about one, through the other. The drawback of these two approaches is that they do not take the correlation between features into account (only that between the feature and response variable), and may thereby choose two highly correlated features such that one is redundant in presence of the other. To solve this problem, we proposed a metric based on Pearson correlation and mutual information, described in Algorithm 1, to find the desired subset of features. Specifically, the metric, in Eq. 1, denoted as $M$is directly proportional to the mutual information MI and to the correlation with target P_{Target}. It is, however, reversely proportional to the correlation with the previously selected features P_{Feat}, as we want to avoid selecting highly correlated features and choose features that can contribute to the classifier accuracy with different information. We convert the proportionality to equality as:

$$M = \frac{(MI \times P_{Target})}{\alpha + \beta P_{Feat}}$$, Eq. (1)

where α and $\beta$are hyperparameters that control the dependency of the metric M to feature correlation P_{Feat} and correlation with response variable P_{Target}. High ratios of α/β will eliminate the dependency of M on P_{Feat} and high values of α removes the dependency of the M on P_{Target}. We performed a grid-search over the range of values 1, 10 and 100 for both $\alpha$and β to determine their optimal value.

Algorithm 1

Algorithm for proposed filter method

Input

Training set T,

Set of p features $F = (f_1, f_2, \ldots, f_p)$,

Target label y,

Mutual information of features with target $MI(T, f, y)$

The number of features to be selected k

Output

Final subset of features $F_{selected} = (f_1, f_2, \ldots, f_k)$

1: $F_{selected} = ()$

2: **for** $j = 1$ to k **do**:

3: scores = {}

3: **for** f in $F = \{f_1, f_2, \ldots, f_p\}$ **do**:

4: $P_{Feat} = \rho_{Feat}(T, F_{selected}, f)$

5: $P_{Target} = \rho_{Target}(T, f, y)$

6: $MI = MI(T, f, y)$

7: $M = (MI * P_{\text{Target}})/(\alpha + \beta P_{\text{Feat}})$

8: scores[f] = M

9: **end for**

10: scores.sorted(key = M)

11: f^* = feature with highest M score

12: $F_{selected} \leftarrow$ add f^* to the set

13: $F \leftarrow F - f^*$

14: **end for**

15: **return** F_{selected}

The Pearson's correlation function $\rho_{Feat}(T, F_{selected}, f)$ takes the training set T computes the Pearson's correlation of feature f with all the features in set $F_{selected}$ and returns the maximum within feature correlation over features in set $F_{selected}$. The function $\rho_{Target}(T, f, y)$ computes the Pearson's correlation of feature f with target y. Finally, the function $MI(T, f, y)$ take the training set T and computes the Mutual information of feature f with the target y.

Wrapper Methods

Wrapper methods merge feature selection and learning steps allowing the learning algorithm to interact with the bias of the feature selection step, decreasing the total bias. Thus, using wrapper methods, a subset of features that result in better prediction performance will be selected. The Recursive Feature Elimination (RFE) method [13] is a commonly used wrapper model. It is a recursive algorithm that ranks features according to some measure of their importance. For example, SVM-RFE ranks the features based on SVM, (Sanz et al., 2018). In this paper, we use LR-RFE, SVM-RFE, GB-RFE and RF-RFE to eliminate and rank features.

3.4 Stability of Feature Selection

Many feature selection algorithms have been successful at improving the forecasting accuracy of learning models while reducing feature-space dimensionality and model complexity (Khalid et al., 2014). Beyond high accuracy, the stability of feature selection is another important attribute of these algorithms. The stability of a feature selection algorithm is defined as the robustness of the feature set it produces to differences in training sets drawn from the same generating distribution *P*(*X,C*), where *C* is the class label for *X*. Here we use the stability measure proposed in (Kalousis et al., 2007) to compare several feature selection methods, informing selection of the one that best fits our dataset and performance needs. Similarity between two subsets of features using a straightforward S_s takes values in [0,1] with 0 meaning there is no overlap between the two sets and 1 that the two sets are identical. To empirically estimate the stability of a feature selection algorithm for a given dataset, the distribution *P*(*X,C*) from which the training sets are drawn is simulated by using a re-sampling technique

$$S_s(s, s') = 1 - \frac{|s|+|s'|-2|s \cap s'|}{|s|+|s'|-|s \cap s'|} \quad \text{Eq. (2)}$$

where S_s takes values in [0,1] with 0 meaning there is no overlap between the two sets, and 1 that the two sets are identical.

To calculate the stability, *K* data subsets were created by randomly shuffling the data and dividing it into folds. A small *K* does not produce a robust estimation of the variance for stability estimation as there are few instances of its measurement. A large *K*, on the other hand, decreases the number of data points in each fold, and as a result does not yield a reliable AUC-ROC score. Therefore, as a compromise between

accuracy for AUC-ROC and stability, we chose $K = 5$ folds. For each fold, the selected features are computed according to the feature selection method. Then, the similarity of each pair of selected features, i.e. $K(K-1)/2$ pairs, is computed using the similarity measure.

4 Results

In this section, the performance analysis, in terms of test ROC-AUC and stability of the feature selection, regarding the four predictive models on the extracted features are presented.

4.1 Performance and Stability Results

We used 196 data points for the method training and evaluation. Specifically, the training set consists of 157 data points (80%), and the remaining 39 data points (20%) constitute the testing set. A performance evaluation was conducted by repeatedly training the model using 94 randomly selected data points (60% of the training set) and reporting the AUC-ROC on 31 randomly selected data points (80% of the test set).

The various hyper-parameters for each model, alongside the related AUC-ROC score for N = 10, 20, 40 features, are presented in Table 1. The hyper-parameters that we investigated for each of the classifiers are as follows. For the SVC classifier, we assumed a linear kernel and investigate the influence of C, the regularization parameter. The parameter C would control the effect of outliers in your model. A large C enforce the model to choose a smaller-margin hyperplane while attempting to classify all the training points correctly. On the other hand an small C will encourage a larger-margin separating hyperplane while allowing for some outlier points to be misclassified. For the LR classifier, we investigated two regularization techniques: (i) L_1 regularization that penalizes the sum of absolute values of the weights, and L_2 regularization that penalizes the sum of squares of the weights. We also considered different value for inverse regularization parameter C, which controls the strength of our regularization. A high value of C encourages the model to trust the training data and decreases the weight of the complexity penalty. In the RF classifier, we perform a grid-search over the number of trees in the forest (estimators) and the maximum depth of the trees. Finally for the GB classifier, we optimize for the learning rate and the maximum depth of the trees. Learning rate controls how fast the model learns and the slower learning rate helps the model to generalize better.

Table 1
Hyper-parameter tuning for the four classifiers used for PRRSV outbreak prediction. The ROC-AUC for each set of hyper-parameters is reported for different feature sizes *N* = 10, 20, 40, and the best set is shown in bold font.

Models	**Parameters**		**AUC-ROC**		
SVC	Kernel	C	N = 10	N = 20	N = 40
	linear	0.1	0.82 ± 0.014	0.85 ± 0.014	0.81 ± 0.016
	linear	1	0.83 ± 0.014	0.86 ± 0.014	0.81 ± 0.018
	linear	**10**	**0.85 ± 0.014**	**0.86 ± 0.012**	**0.83 ± 0.013**
LR	Penalty	C	N = 10	N = 20	N = 40
	L1	0.1	0.84 ± 0.010	0.84 ± 0.011	0.82 ± 0.010
	L1	1	0.86 ± 0.014	0.85 ± 0.011	0.83 ± 0.013
	L1	10	0.84 ± 0.013	0.85 ± 0.014	0.82 ± 0.012
	L2	**0.1**	**0.86 ± 0.007**	**0.87 ± 0.009**	**0.86 ± 0.010**
	L2	1	0.83 ± 0.021	0.86 ± 0.014	0.85 ± 0.013
	L2	10	0.83 ± 0.019	0.84 ± 0.019	0.84 ± 0.014
RF	Estimators	Maximum depth	N = 10	N = 20	N = 40
	50	4	0.82 ± 0.017	0.86 ± 0.010	0.87 ± 0.009
	50	8	0.82 ± 0.010	0.86 ± 0.008	0.86 ± 0.010
	50	16	0.82 ± 0.020	0.86 ± 0.019	0.86 ± 0.019
	50	32	0.83 ± 0.017	0.87 ± 0.019	0.87 ± 0.008
	100	4	0.82 ± 0.013	0.86 ± 0.012	0.86 ± 0.010
	100	**8**	**0.83 ± 0.018**	**0.87 ± 0.013**	**0.88 ± 0.003**
	100	16	0.83 ± 0.012	0.87 ± 0.006	0.88 ± 0.015
	100	32	0.83 ± 0.014	0.87 ± 0.011	0.88 ± 0.013
GB	Learning rate	Maximum depth	N = 10	N = 20	N = 40
	0.01	3	0.84 ± 0.016	0.84 ± 0.014	0.84 ± 0.012
	0.01	5	0.86 ± 0.014	0.85 ± 0.011	0.83 ± 0.013
	0.1	**3**	**0.85 ± 0.021**	**0.86 ± 0.019**	**0.86 ± 0.020**
	0.1	5	0.84 ± 0.014	0.86 ± 0.011	0.86 ± 0.013

The best hyper-parameter that achieves higher AUC-ROC score, while introducing less complexity in terms of the number of parameters is indicated with bold font. An example of hyper-parameter tuning for the support vector classifier for features length ranging from 1 to 60 can be seen in Fig. 2. A range of three values for the regularization parameter (C) were considered, where the strength of the regularization is inversely proportional to C.

For each model, Fig. 3, demonstrates the AUC-ROC and stability measure mean and standard deviation across the folds. This figure shows the performance of each of these models together with the stability of the corresponding RFE-based feature selection method in terms of AUC-ROC score. According to the results, the two non-linear models, GB and RF, have better AUC-ROC scores than the linear models, SVC and LR (Fig. 3a). In Fig. 3b, it can be seen that the non-linear models, GB and RF, are not as stable as the linear models in selecting a robust subset of features.

Since the tree-based models (GB and RF) have a higher AUC-ROC, we used them as the base classifiers to assess other filter-based feature selection methods (see Fig. 4). In Fig. 4a-b, we show the performance assessment of different feature selection methods using GB as the classifier. Specifically, we compared a RFE-GB feature selection method and three other filter-based feature selection methods: correlation with target, Mutual Information (MI), and our proposed algorithm (Algorithm 1). Similarly, in Fig. 4c-d, we showed the performance assessment of these feature selection methods using RF as the classifier. As demonstrated in these two figures, the filter-based feature selection methods had higher stability, but lower ROC-AUC, in comparison with the RFE-based methods. Algorithm 1 surpassed the stability of RFE-based feature selection methods, while showing a comparable ROC-AUC performance.

4.2 Feature Selection

To identify highly correlated features, hierarchical clustering was performed as shown in Fig. 5. The features are clustered according to correlation-based similarity value between features and the cluster value, shown as a distinct color in the horizontal bar plot. Each feature belonging to a specific cluster is representative of that cluster and has relatively the same contribution in terms of classification performance.

Table 2

Subsets of selected features using Recursive Feature Elimination + Gradient boosting classifier to forecast PRRSV outbreaks as evaluated by AUC-ROC. Historical is defined as a period of one year prior to the start of the current finishing period.

AUC-ROC	Features
0.64	1) Pig population (f_1)
0.70	1) Pig population (f_1) 2) Historical average number of days dead pigs lived in a farm divided by the number of surviving pigs (f_{31})
0.73	1) Pig population (f_1) 2) Historical average number of days dead pigs were alive divided by the number of surviving pigs (f_{31}) 3) Historical average number of days dead pigs were alive divided by number of dead pigs (f_{32})
0.80	1) Pig population (f_1) 2) Historical number of living days of dead pigs (f_{27}) 3) Historical number of dead pigs (f_{28})

AUC-ROC	Features
0.80	1) Pig population (f_1) 2) Historical average number of days dead pigs were alive divided by the number of surviving pigs (f_{31}) 3) Historical average number of days dead pigs were alive divided by number of dead pigs (f_{32}) 4) Historical average daily gain (f_{35}) 5) Average daily feed (f_{39})

According to these sampled selected feature subsets, it was found that the average 60 prior days was the best temperature feature when combined with other features for forecasting, thus these values were used in the model. The prior 60 day average temperature and wind were selected more frequently than relative humidity and altitude. Seasonal features were not selected when a temperature feature was chosen. In addition, it was found that the number of movements in the neighborhood for different radii (the number of movements to/from any farm located in 5km, 10km and 20 km) can be used together to improve prediction. The 20km radius features were more frequently selected than those of the 5km and 10km vicinity. In general, the number of movements were more important than the number of animals being shipped in a given movement. The total number and weight of incoming pigs, and the percent of existing pigs with substandard weight were important. The number of dead pigs and the average number of days that dead pigs have lived on the farm are also important predictive features. Moreover, average daily feed, past outbreak frequency in the farm, and the number of outbreaks in the neighborhood during the current finishing period, were amongst the most important features.

Table 3
Subsets of selected features using Recursive Feature Elimination + Support vector Classifier to forecast PRRSV outbreaks as evaluated by AUC-ROC. Historical is defined as a period of one year prior to the start of the current finishing period.

AUC-ROC	Features
0.70	1) Pig population (f_1)
0.71	1) Pig population (f_1) 2) Number of shipments in any farm located within 20km (f_{53})
0.72	1) Pig population (f_1) 2) Number of shipments in any farm located within 20km (f_{53}) 3) Number of shipped pigs in any farm located within 20km (f_{48})
0.79	1) Total weight of the pigs (f_{13}) 2) Historical total number of days dead and surviving pigs lived in farm (f_{25}) 3) Historical total number of days surviving pigs lived in farm (f_{26}) 4) Historical total number of dead pigs (f_{28}) 5) Past outbreak frequency in the farm (total number of finishing periods with recorded outbreaks divided by total number of finishing periods — since beginning to the current finishing period) (f_{42})

5 Discussion

This study incorporates swine farm- and area-level data in the forecasting of the farm-level PRRSV outbreaks. Using a uniquely rich real-world dataset, obtained from our industry collaborators, we included a level of detail that, to the best of our knowledge, has not been previously considered in the prediction of PRRSV outbreaks. We integrated production data, movement data, and climate information for our predictions. Further, we demonstrate the generation of new features from standard industry variables that

better represent farm-level management practices and risk for use in forecasting models. In this manner, we have addressed the two main PRRSV transmission pathways, direct contact and airborne, as well as onsite disease history and management practices, and the role of near-farm status, on outbreak risk.

Based on the AUC-ROC and stability results, the two tree-based models have superior AUC-ROC scores in comparison with the linear models when used as classifiers. This is expected as the tree-based models are non-linear in nature and, therefore, can capture the nonlinearities in the data. However, due to their inherent randomness, they do not show a reasonable stability when used as RFE-based feature selection methods. As observed in the results section, filter-based feature selection methods demonstrated considerable stability and, hence, were used as the basis for developing a new feature selection algorithm. By combining Algorithm 1 for feature selection, which inherits the superior stability of the filter-based feature selection, and a tree-based model as classifier, we achieve high predictive performance and stability.

Considering Tables 2 and 3, we observe that all different types of data, i.e., shipment, diagnostic, and production, play an important role in improving the prediction performance of the model. Based on the feature selection results, the most predictive feature in the dataset is the pig population. As the most frequently selected features across different methods, features representing movements in a neighborhood during the current FP are strong predictive features. Features related to pig movements to/from farms located in a 20 km neighborhood is a strong predictive feature to capture the effect of airborne disease transmission pathway. Features representing the number of dead pigs and the number of days they lived can represent the magnitude and impact of the PRRSV infection (and associated co-morbidities) on the farm. The historical features, which are the averages over the past measurements, are important because they provide the model with the information about the biosecurity and management practices of the farm over the time, while the current FP features are informative in terms of the recent events. The superior performance of the 60-day climatic period may be due to the fact that it captures seasonality; or, it may outperform other time periods because more shipments happen during the 60 days prior to the forecasting day and thus this average may best represent the temperature that pigs were exposed to during shipment.

Different subsets or combinations of features can yield the same performance. This is due to the fact that a feature in a given subset may be substituted with another feature that is highly correlated with it. This gives flexibility to those wanting to generate their own models. The clusters presented in Figure. 5, and on the list and description of features included in each cluster (Supplement 1), provide alternative features for use in forecasting when all the data fields used in this work are not available. In general, selecting 4 features, one from each of the first 4 clusters in Figure. 5 should provide high predictive power.

6 Conclusion

To the best of our knowledge, this is one of the first attempts to apply multiple machine learning models for PRRSV forecasting using multi-level data. We have demonstrated the strength of these methods for

disease prediction in the swine industry, and believe they could be readily adapted for use on other diseases and for additional livestock species. This approach could save the swine industry millions of dollars through the improved efficiency and reduced economic burden provided by early, targeted, risk mitigation strategies. This work uses a rich, multi-scale (pig group-, farm-, area-level data) feature set - assessing on-farm characteristics at a level previously unreported for farm health analysis in the swine industry. Additionally, the integration of historical data with current cycle data to improve forecasting accuracy is a novel approach applied in this work. Generating an expansive set of features, ranging across farm level, time and space, allowing the evaluation of multiple disease transmission pathways, environmental factors, and management practices as risk factors for disease occurrence, resulted in improved outbreak forecasting ability. Stable feature selection allowed us to identify and represent the most important risk factors for PRRSV outbreaks. These variables can now be further explored by the industry and research community as points for future intervention. These approaches offer a strong basis for ongoing work, and we hope the adaptation of these methods into dynamic dashboards within the Disease BioPortal (https://bioportal.ucdavis.edu) will provide industry users with near real-time information for improved health management decisions.

Declarations

Data Availability

Climate data is publicly available.

Code Availability

Code is available upon request.

Acknowledgment

This project was partially funded by the NSF BIGDATA:IA Award #1838207 and NSF Track-D award #2134901. Authors would like to acknowledge swine industry collaborators for the provision of data.

References

1- D.J. Holtkamp, J.B. Kliebenstein, E.J. Neumann, J.J. Zimmerman, H.F. Rotto, T.K. Yoder, C. Wang, P.E. Yeske, C.L. Mowrer, C.a. Haley. Assessment of the economic impact of porcine reproductive and respiratory syndrome virus on United States pork producers J. Swine Heal. Prod., 21 (2013), pp. 72-84

2- E. Mateu and I. Diaz. The challenge of PRRS immunology. *The Veterinary Journal*, 177(3):345–351, 2008.

3- C. A. Corzo, E. Mondaca, S. Wayne, M. Torremorell, S. Dee, P. Davies, and R. B. Morrison. Control and elimination of porcine reproductive and respiratory syndrome virus. *Virus research*, 154(1-2):185–192, 2010.

4- G. Nodelijk. Porcine reproductive and respiratory syndrome (PRRS) with special reference to clinical aspects and diagnosis: a review. *Veterinary quarterly*, 24(2):95–100, 2002.

5- D. J. Holtkamp, D. D. Polson, M. Torremorell, D. M. Classen, L. Becton, S. Henry, M. T. Rodibaugh, R. R. Rowland, H. Snelson, B. Straw, et al. Terminology for classifying swine herds by porcine reproductive and respiratory syndrome virus status. *Journal of swine health and production*, 19(1):44–56, 2011.

6- M. Shamsabardeh, S. Rezaei, J. P. Gomez, B. Martínez-López, and X. Liu. A novel way to predict PRRS outbreaks in the swine industry using multiple spatio-temporal features and machine learning approaches. *Frontiers in Veterinary Science*, 6, 2019. doi: 10.3389/conf.fvets.2019.05.00085.

7- P. Valdes-Donoso, K. VanderWaal, L. S. Jarvis, S. R. Wayne, and A. M. Perez. Using machine learning to predict swine movements within a regional program to improve control of infectious diseases in the us. *Frontiers in veterinary science*, 4:2, 2017.

8- S. Otake, S. Dee, C. Corzo, S. Oliveira, and J. Deen. Long-distance airborne transport of infectious PRRSV and mycoplasma hyopneumoniae from a swine population infected with multiple viral variants. *Veterinary microbiology*, 145(3-4):198–208, 2010.

9- I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

10- M. A. Hall. Correlation-based feature selection for machine learning. 1999.

11- P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2): 83–90, 2006.

12- A.-C. Haury, P. Gestraud, and J.-P. Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS one*, 6(12):e28210, 2011.

13- I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

14- A. Kalousis, J. Prados, and M. Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 12(1):95–116, 2007.

15- S. Dee, J. Deen, K. Rossow, C. Weise, R. Eliason, S. Otake, H. S. Joo, and C. Pijoan. Mechanical transmission of porcine reproductive and respiratory syndrome virus throughout a coordinated sequence of events during warm weather. *Canadian journal of veterinary research*, 67(1):12, 2003.

16- S. Dee, J. Deen, K. Rossow, C. Wiese, S. Otake, H. S. Joo, and C. Pijoan. Mechanical transmission of porcine reproductive and respiratory syndrome virus throughout a coordinated sequence of events during cold weather. *Canadian Journal of Veterinary Research*, 66(4):232, 2002.

17- G. Brown, A. Pocock, M.J. Zhao and M. Luján. Conditional likelihood maximization: a unifying framework for information theoretic feature selection. *The journal of machine learning research*, 13, pp.27-66, 2012.

Figures

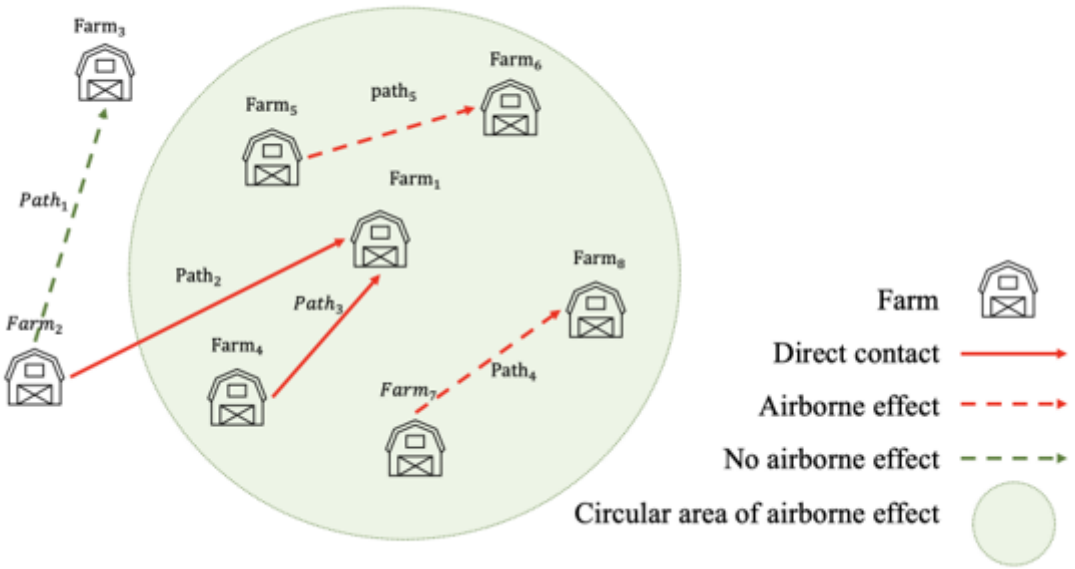
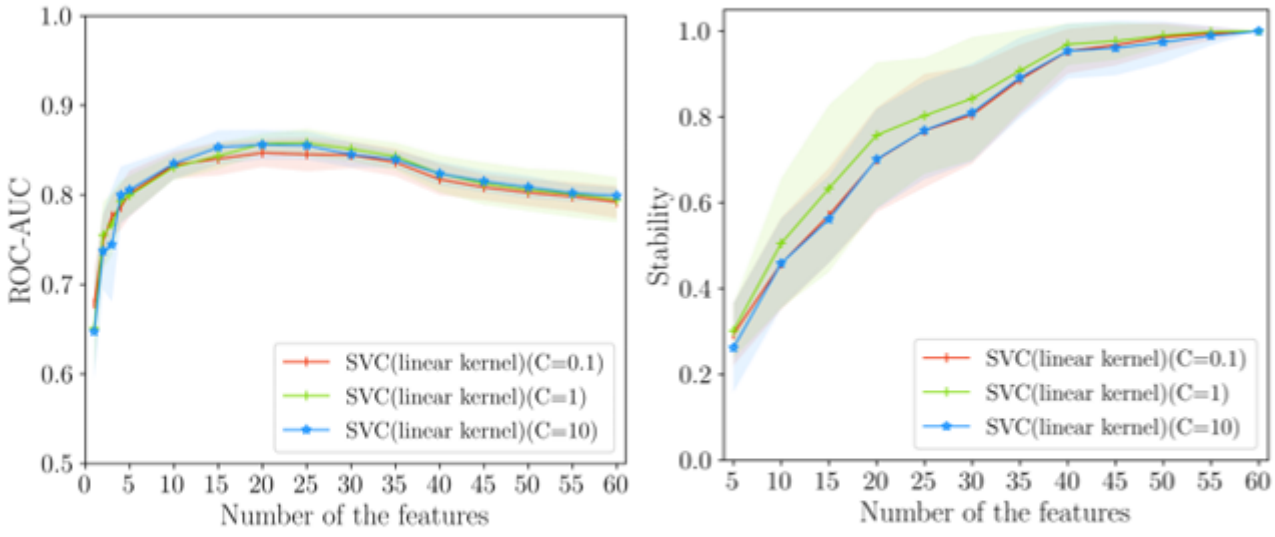


Figure 1

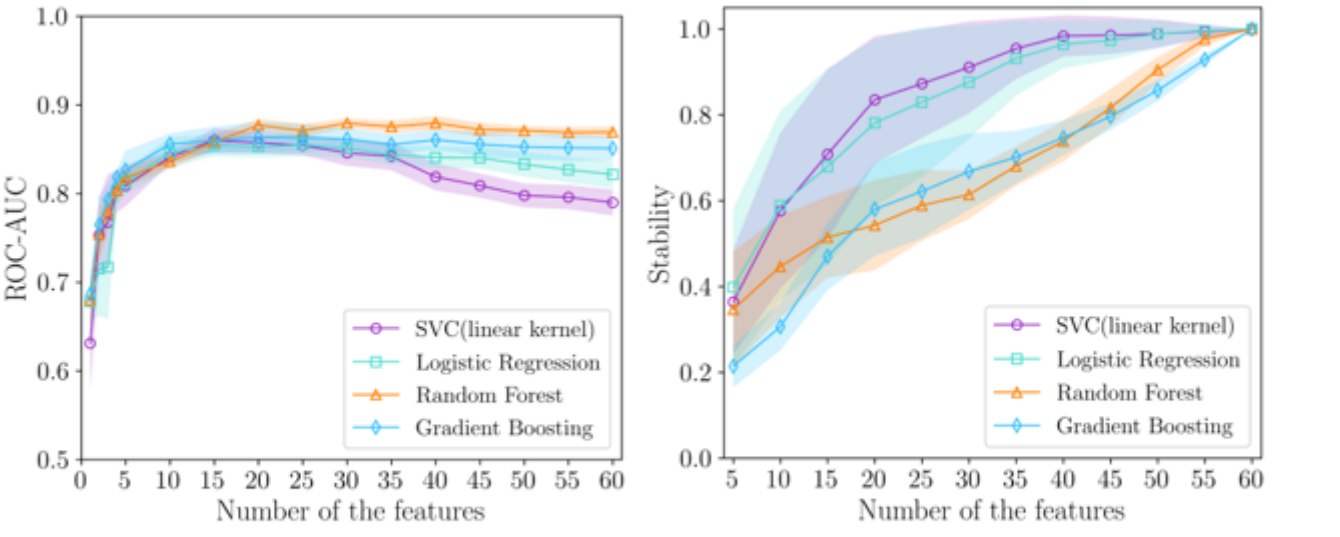
Demonstration of the two main PRRSV transmission pathways for farm F1: airborne and direct contact. The red dashed-arrows P4 and P5 are the airborne effect of the farm F1, thus they are indirect transmission pathways for the farm F1. The solid dashed arrows P2 and P3 are the source of pig shipment to F1 and can transmit disease directly to it. The green dashed arrow P1 is the shipment with no risk factor on F1 as both source and destination of the shipment are outside of the defined area.



(a) AUC-ROC score versus the number of features
(b)Stability Versus the number of features

Figure 2

(a) Stability score for different regularization parameters (*C*) in Support Vector Classifier (SVC) for different sizes of feature sets. (b) the area under the ROC curve (AUC-ROC) score for different regularization parameters for different sizes of feature set. The strength of the regularization is inversely proportional to regularization parameter *C*. The regularization *C*=10 yields the best AUC-ROC but the regularization *C*=1 has better stability (equivalent to part of the table 1).



(a)AUC-ROC score versus the number of feature (b) Stability Versus the number of features

Figure 3

Performance assessment of Recursive Feature Elimination (RFE) merged with different classifiers for PRRSV outbreak prediction in terms of: (a) the area under the ROC curve (AUC-ROC), and (b) Stability score for different size features sets. Each graph is labeled according to the classifier used in the RFE feature selection algorithm.

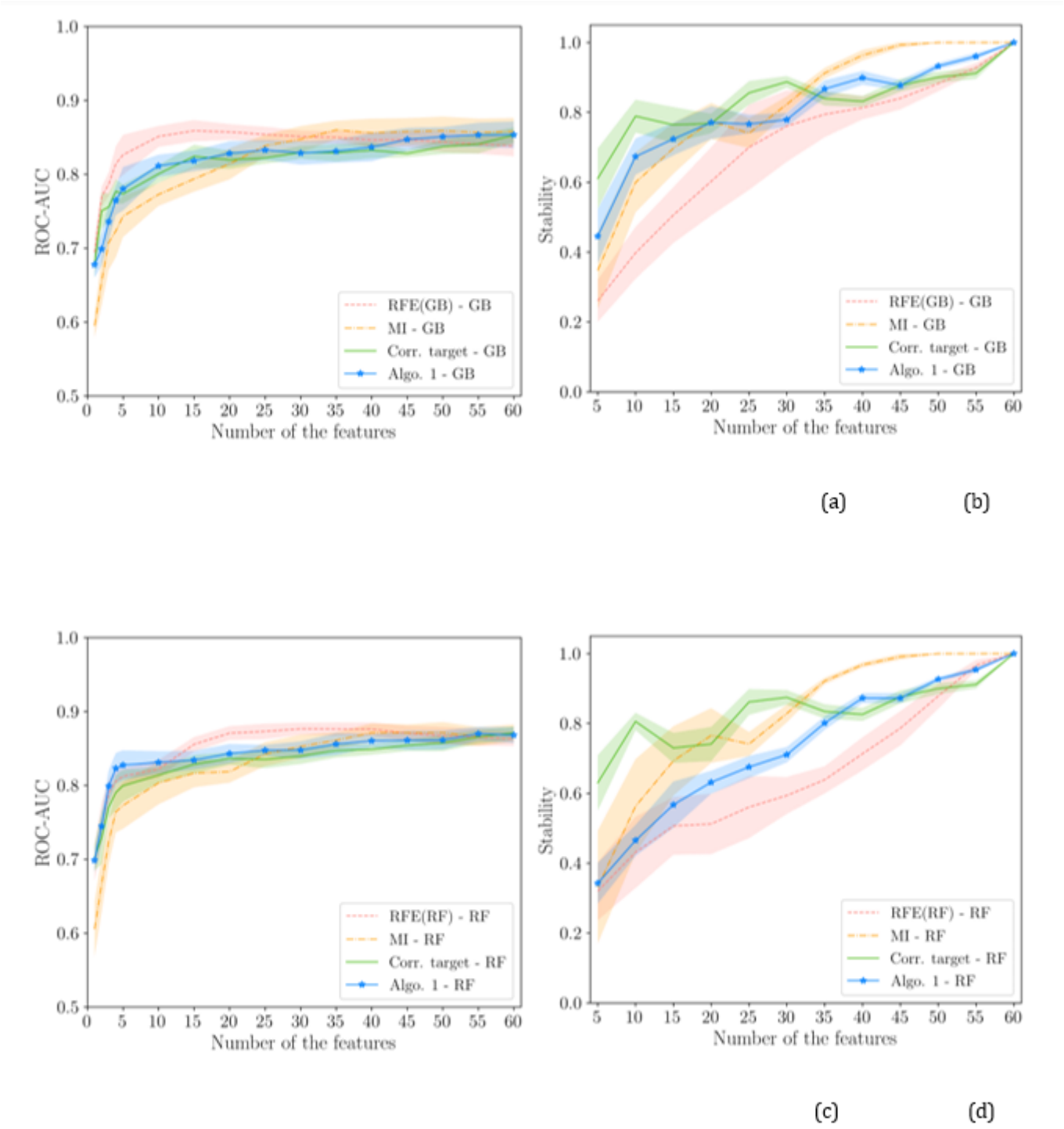


Figure 4

Performance demonstration of different feature selection methods (wrapper and filter methods) merged with the Gradient Boosting (GB) and Random Forest (RF) as classifier for PRRSV outbreak prediction in terms of: (a and c) The area under the ROC curve (AUC-ROC), and (b and d) Stability score for different size features sets. Each graph is labeled according to its feature selection method (on the left of dash) and its classifier method (on the right of dash) in the legend.

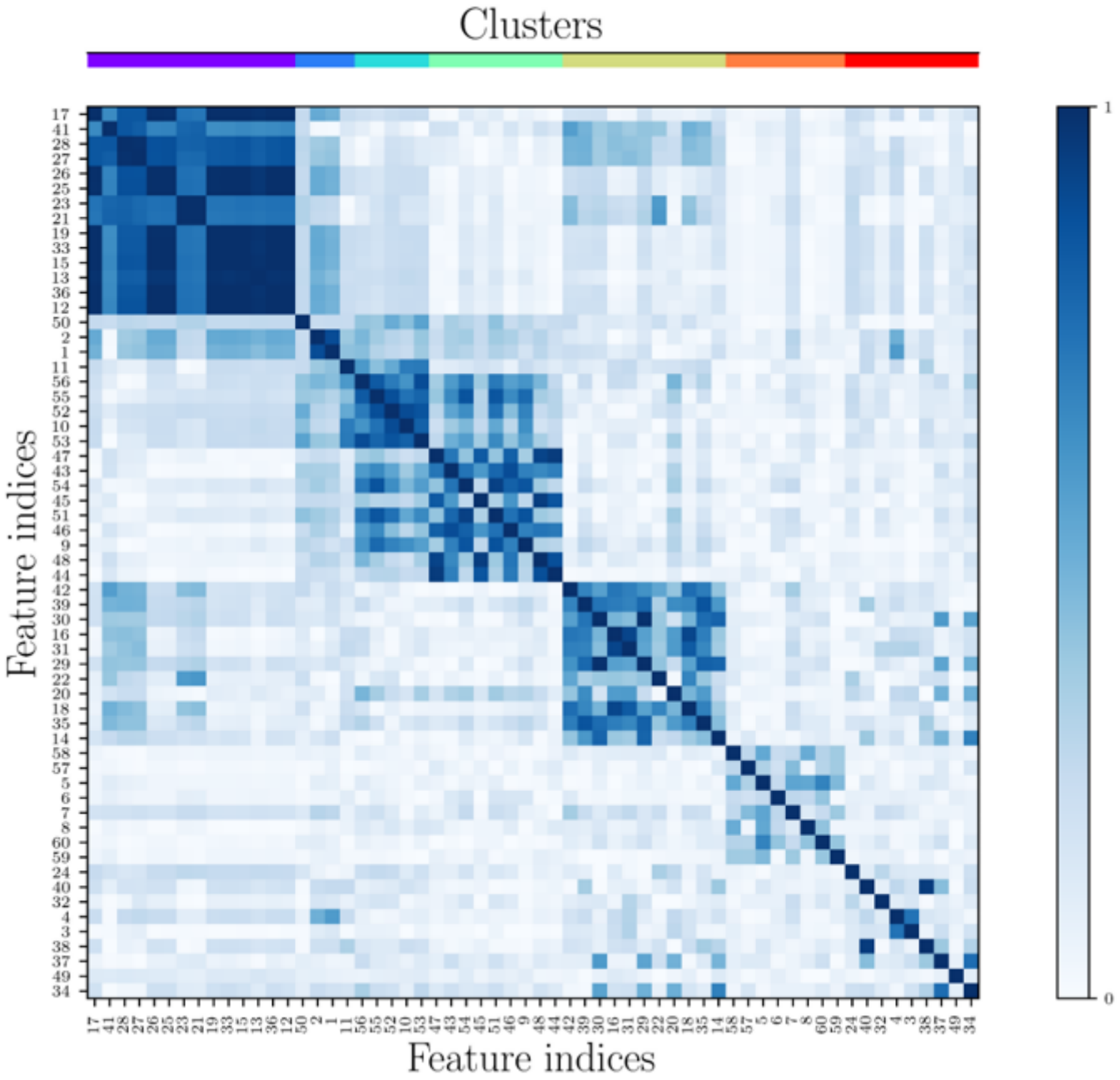


Figure 5

Feature similarity grouped by hierarchical clustering. Hierarchical clustering is used to analyze the similarity between features in terms of their correlation. The darker colors are representative of higher correlations among features. The formed blocks are indicative of the clusters of similar features. The exact margins are shown by the horizontal bar at the top where each color represents a cluster. The description of each feature number can be found in Table 1 of Supplement 1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DataAvailability.docx
- SupplementaryTable.docx