Fisheries Science (2022) 88:1–13
https://doi.org/10.1007/s12562-021-01577-8

ORIGINAL ARTICLE

Fisheries

Use of GIS and machine learning to predict disease in shrimp farmed on the east coast of the Mekong Delta, Vietnam

Nguyen Minh Khiem[1,4] · Yuki Takahashi[2] · Hiroki Yasuma[2] · Dang Thi Hoang Oanh[3] · Tran Ngoc Hai[3] · Vu Ngoc Ut[3] · Nobuo Kimura[2]

Received: 19 August 2021 / Accepted: 24 November 2021 / Published online: 13 January 2022

Abstract
Diseases in shrimp farms in the Mekong Delta of Vietnam cause significant crop losses and are therefore of great concern to producers. Once a pond becomes infected, it is difficult to prevent spread of the disease to nearby shrimp farming areas. Thus, predicting the occurrence of disease is an essential part of reducing the risk for shrimp farmers. In this study, we applied an integrated geographic information system and machine learning system to predict three serious diseases of shrimp, namely, acute hepatopancreatic necrosis, white spot syndrome disease, and *Enterocytozoon hepatopenaei* infection, based on data collected from shrimp farms in the Tra Vinh, Bac Lieu, Soc Trang, and Ca Mau provinces of Vietnam. We first constructed a map showing the distribution of these diseases using the locations of affected farms, and then we conducted spatial analysis to acquire the geographical features of the affected locations. This latter information was combined with environmental factors and clinical signs to form the set of independent variables affecting the outbreak of diseases. The neural network model outperformed the logistic regression, random forest, and gradient boosting methods in terms of predicting infection to estimate the probability of disease occurrence in farmed areas. Acute hepatopancreatic necrosis disease infected farms downstream of the Co Chien and Hau Rivers of Tra Vinh and west of Ca Mau. *Enterocytozoon hepatopenaei* infection is distributed in Soc Trang Province, while white spot syndrome virus has spread to the coastal districts of Soc Trang and Bac Lieu Provinces, where it is highly associated to water from a complex canal system.

Keywords GIS · Machine learning · Acute hepatopancreatic necrosis · White spot syndrome · *Enterocytozoon hepatopenaei*

Introduction

Fisheries in Vietnam contribute to the development of sustainable livelihoods and to the general economy, especially in the Mekong Delta. Shrimp farming is the most significant fisheries activity in the country and has reduced poverty remarkably, provided employment opportunities, and increased exports (Ha et al. 2013). The Mekong Delta of Vietnam has the greatest potential for shrimp aquaculture, an activity that plays a vital role in rural development and helps approximately one million fish farmers to achieve a sustainable livelihood (Duc 2009; Phuong and Oanh 2010).

However, shrimp farms in the Mekong Delta are being affected by various diseases which seriously constrain sustainable shrimp farming due to high shrimp mortality among the affected shrimp farms, thereby affecting shrimp farmers' incomes. A number of these diseases, such as acute hepatopancreatic necrosis disease (AHPND), diseases caused by white spot syndrome virus (WSSV), and the disease caused by *Enterocytozoon hepatopenaei* (EHP disease), have disasterous effects on shrimp farms.

✉ Nguyen Minh Khiem
nmkhiem@cit.ctu.edu.vn

1 Graduate School of Fisheries Sciences, Hokkaido University, Hakodate, Hokkaido 041-8611, Japan

2 Faculty of Fisheries Sciences, Hokkaido University, Hakodate, Hokkaido 041-8611, Japan

3 College of Aquaculture and Fisheries, Can Tho University, Can Tho, Vietnam

4 College of Information and Communication Technology, Can Tho University, Can Tho, Vietnam

AHPND, caused by *Vibrio parahaemolyticus,* was first detected in 2010 and has resulted in huge losses for the shrimp industry (Zheng et al. 2018). In the years following 2010, AHPND has damaged approximately 59,000 ha of farming ponds and has spread rapidly to 294 communes of 86 districts in 25 provinces in Vietnam (Dang et al. 2018). Boonyawiwat et al. (2016) described the relationships among AHPND and farm characteristics, management, water preparation, and post-larval shrimp in Thailand. Khiem et al. (2020) used machine learning to predict the occurrence of AHPND in the Mekong Delta.

A disease in shrimp caused by the microsporidian parasite *E. hepatopenaei*, subsequently called EHP disease, was first detected in the tiger prawn in Thailand in 2003 (Chayaburakul 2004) and first recorded in Vietnam in 2010 (Ha et al. 2010). In shrimp ponds affected by EHP, shrimp growth is normal for the first month after stocking; however, growth slows thereafter, with the direct consequence of a decrease in shrimp farmers' incomes. The clinical signs of EHP disease are indistinct; as such, it is difficult to recognize the infection. Aranguren et al. (2017) identified the association between EHP and AHPND in the shrimp *Penaeus vannamei* in Asian countries. To date, EHP disease has been very difficult to recognize and treat.

White spot disease, caused by infection by WSSV, the sole member of the virus family *Nimaviridae*, is a viral disease that causes high mortality in shrimp within a short time. In 2001, more than 20,854 ha of shrimp farms in the Mekong Delta were infected with this virus, with farms in seven coastal provinces (Dong Nai, Ben Tre, Tra Vinh, Soc Trang, Bac Lieu, Ca Mau, and Kien Giang) being were seriously impacted (Oanh and Phuong 2005). The disease has been identified as the most serious disease affecting shrimp in coastal farms. The WSSV has been the focus of much research which has resulted in the identification of the relationship between WSSV and salinity (Ramos et al. 2014) and determination of the viability of WSSV in pond sediment (Satheesh et al. 2019).

In New Zealand, geographic information system (GIS) technology was implemented to identify suitable and sustainable locations for aquaculture-management areas (Peter et al. 2008). Moreover, the Food and Agriculture Organization (FAO) has used GIS applications for fisheries, as described by Meaden and Aguilar-Manjarrez (2013), Kapetsky et al. (2013), and Aguilar-Manjarrez and Crespi (2013). In Vietnam, GIS technology has been applied to forecasting tuna fishing grounds (Thuoc and Thanh 2013), and a GIS-based approach was used to identify suitable areas for shrimp farming in Ben Tre Province in the Mekong Delta (Thanh et al. 2008). Giap et al. (2003) used a GIS-based approach in Thai Nguyen Province of Vietnam to detect land-use changes and estimate potential areas for aquaculture development in watershed ponds. GIS-based methodologies are promising tools for providing comprehensive evaluations of disease outbreaks and for evaluating disease symptoms, seed source, and living environments. However, the application of a GIS-based system to research and/or prevent shrimp disease in the Mekong Delta remains limited, partially due to the economic instability that characterizes shrimp-farming households.

Machine learning is an advanced computer technique that can provide strong support for fisheries, and one of its applications is the prediction of disease outbreak. Examples of this application include processing digital imagery to diagnose disease (Rao et al. 2017), using an artificial neural network and logistic regression to predict the occurrence of shrimp disease (Leung and Tran 2000), and applying machine learning to predict AHPND (Khiem et al. 2020). Machine learning has also been applied to forecasting the distribution of fishing activities (Soykan et al. 2014) and supporting decision-making pertaining to aquaculture shellfish farm closure (Shahriar and McCulluch 2014). However, research involving integrated systems of GIS and machine learning that focus on the prediction of diseases in shrimp has not yet been performed for shrimp farming in the Mekong Delta.

To reduce the risk of diseases that threaten shrimp farming, we first attempted to assess the status of disease infection in shrimp farms through visualization of the distribution of three serious diseases, namely, WSSV, EHP, and AHPND, on a map of farms located on the east coast of the Mekong Delta. We then extracted geographical information from this map, which was examined as a feature related to disease outbreak. Then, various factors, including the clinical signs of infected shrimp, environmental impact, and geographical features influencing disease, were investigated. The machine-learning technique was applied to these factors to predict the occurrence of each disease based on classification algorithms.

Materials and methods

Dataset

The data for this study have been collected since 2013 at shrimp farms in four provinces located on the east coast of the Mekong Delta: Tra Vinh, Soc Trang, Bac Lieu, and Ca Mau (Fig. 1). The two main shrimp species cultured in these farms are the tiger prawn *Penaeus monodon* and the whiteleg shrimp *Litopenaeus vannamei*. The data were collected from 182 farms (Ca Mau, 84 farms; Bac Lieu, 24 farms; Soc Trang, 16 farms; and Tra Vinh, 58 farms). Due to the focus of the data collection on disease outbreak, most of the farms were infected with WSSV, AHPND, or EHP, although some were affected by multiple diseases. Thus, 125

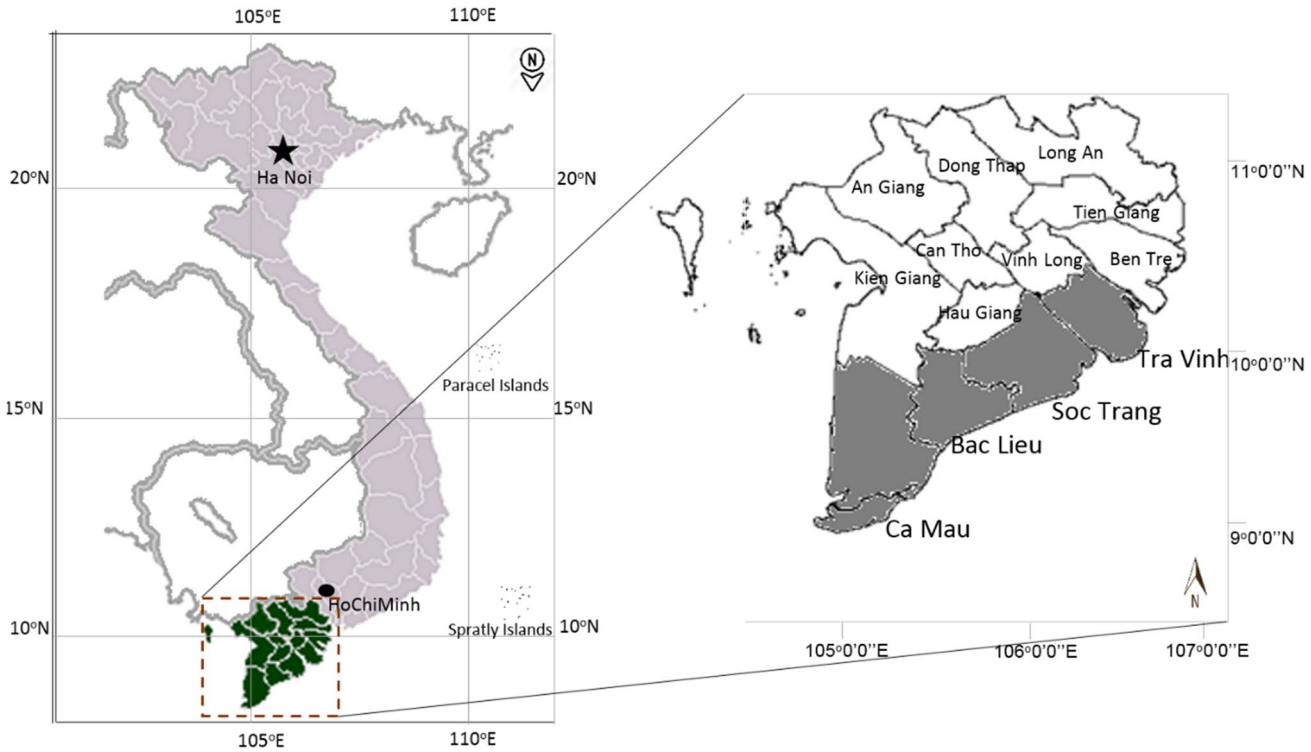


Fig. 1 East coast of the Mekong Delta of Vietnam

farms were affected by AHPND, 57 by WSSV, and 42 farms by EHP; 35 farms had no recorded disease.

The collected dataset consisted of two main parts: clinical signs and environmental factors. The categorical variables were the clinical signs, and the continuous variables were the environmental factors. If a categorical variable was associated with the symptoms of disease, its value was set as 1; otherwise, it was set as 0. Here, the clinical signs included: (1) gut status, differentiated as discontinuous gut, yellow liquid in gut, little food in gut, or empty gut; (2) hepatopancreas status, defined as hepatopancreatic paleness and atrophy; (3) slow growth; (4) soft shell; (5) white feces; (6) white spots; (7) vermiform structure; and (8) gregarine infection. The environmental factors consisted of temperature, salinity, pH, NO_2, and NH_4. However, the environmental data for all farms were not always complete because the data were collected at different times in different provinces. To solve this problem, interpolation based on GIS was used to predict the values of missing data. The tool used is part of the Spatial Analyst extension component of ArcGIS (Esri Inc., Redlands, CA, USA), which uses an inverse distance weighted (IDW) technique to interpolate a raster surface from points. The weight in the IDW technique is a function of inverse distance. The process, based on geographical point data, utilizes available values to estimate missing values of neighboring spatially correlated points. Output values from this process are limited to within the range of the highest and lowest values of the input. Missing values are filled by averaging the values of the sample data points in the neighborhood of each processing point. The numbers of points used to calculate the interpolated value can be selected in two different ways: by directly specifying the number of points or by specifying a fixed radius including the points to be used in the interpolation. A variable search radius flexibly determines the radius distance from each interpolated point that is less than a given maximum distance, while the fixed radius requires a constant value for the radius. This radius is used to create a circle that covers all of the points specified for the calculation. The principle of the task is the concept that more proximal locations tend to have more similar characteristics. When all farms were considered as points on the map, they were suitable for application of this tool to predict missing values. For instance, if the temperature calculated for a farm was high, other farms in the same district could be predicted with a high level of confidence also to have hot weather, but there would be lower confidence for temperatures at farms in distant provinces. The interpolation was applied to predict missing salinity, pH, NO_2, and temperature values of farms in the dataset. Here, we used the variable search radius option to estimate missing values; that is, the number of farms was set to ten and the maximum search radius was set to 10 km. Hence, interpolation was limited to ten farms or farms within a radius of 10 km. As the topography across

the Mekong Delta is similar, the environments among farms inside the 10-km radius were considered to be similar. Thus, predicted values were used for missing data.

ArcGIS version 10.5 was used to process geographical information for this study. The longitude and latitude of the shrimp farms were first determined. A hatchery on the east coast of the Mekong Delta was considered to be one of the pathways to disease outbreak because it provided the seed source for shrimp farms in the same area. Disease can easily break out and spread to many farms when farms are stocked with an infected seed source from the same hatchery or if the hatchery has been affected because it is near a diseased farm. The distance between any one shrimp farm and the hatchery was calculated as a variable that contributed to the estimation of disease occurrence. Then, the distance from the hatchery closest to each affected farm was selected based on the premise that the closest hatchery would have high confidence in the transmission of disease between locations.

The water source is the epidemic pathway of most concern for shrimp farmers. Pathogens may exist in the water environment and be transmitted from infected farms to healthy farms via a river or drainage canal, thus spreading disease throughout a large area. In the Mekong Delta, tributaries of the Mekong River provide most of the water for shrimp farms and also play an important role in the spread of disease. By applying spatial analysis in ArcGIS software, the distance between each affected farm and the closest river was calculated and used as a variable to predict disease outbreak.

Cross-contamination is another pathway for the transference of viruses or bacteria among the shrimp farms in an area. The probability of disease occurrence at a farm is very high when neighboring farms are affected as pathogens are transmitted easily. We considered the distance between each farm and its nearest neighboring farm as a variable leading to disease occurrence.

The three above-mentioned variables from the geographical analysis—distance between a farm and its closest neighboring farm, closest river, and closest hatchery—were added to the dataset. All dataset variables are shown in Table 1. Using the machine-learning technique, this dataset was used to predict outbreaks of EHP, WSSV, and AHPND on the shrimp farms.

Machine learning

To predict the occurrence of disease, the dataset was divided into training and testing datasets (4:1 ratio) comprising 145 and 37 farms, respectively. The training dataset was used to generate the prediction model, and the testing dataset was used to determine the model's accuracy. Here, the dataset consisted of the three dependent variables WSSV, EHP, and AHPND, and multiple disease labels were assigned to each farm. Each variable was a binary output coded as 1 if the farm was affected by the corresponding disease and 0 if not. Multi-label classification was determined for our dataset to predict the occurrence of each disease. This technique is a generalization of multiclass classification of machine learning. Many strategies are used to solve multi-label classification problems, such as the classifier chain (Jesse et al. 2009), binary relevance (Montañes et al. 2014), and one-versus-rest (Xu 2011) methods. The one-versus-rest approach involves splitting the multi-class dataset into multiple binary classification problems. A binary classifier is then trained on each binary classification problem, and prediction is performed using the model with the highest confidence. In other words, this method involves training a single classifier per class, which means that samples belonging to that class are positive samples and all other samples are negative ones. In our study, we applied the one-versus-rest method because it can explicitly provide the probability of occurrence for each disease. This method was implemented in the scikit-learn Python module for machine learning (Pedregosa 2011). The estimator parameter in one-versus-rest, which is an equation for picking the "best model" based on real observations, was evaluated by the logistic regression, neural network, gradient-boosting, and random forest algorithms.

Logistic regression

The logistic regression model is often applied to probabilistic prediction. Its equation can be written as follows:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

where p is the probability of the outcome and β is the coefficient of each independent variable. Here, each disease occurrence was the outcome, i.e., whether WSSV, EHP, or AHPND will or will not occur. In this study, the scikit-learn package for Python was used to implement the prediction (Pedregosa 2011).

Neural network

The neural network method evolved from the concept of simulating the human brain (Zou et al. 2008). The structure of a neural network consists of many nodes (neurons) located in layers. There are three main layers: the input, hidden, and output layers. The input layer consists of neurons that receive information from the outside world, the hidden layer maps internal information patterns, and the output layer plays the role of relaying outcomes to the outside world. Data are processed in each node based on mathematical operations to yield the outcome. Every node in the hidden layer receives the output from previous nodes and

Table 1 Definitions and descriptions of the variables in the dataset

Variable number	Variables	Variable descriptions
Farm information		
1	Address/name	Text
2	Longitude	Long decimal
3	Latittude	Long decimal
Geographical factor		
4	Distance between farm and nearest hatchery	Continuous value
5	Distance between farm and nearest river	Continuous value
6	Distance between farm and nearest affected farm	Continuous value
Environmental factor		
7	pH	Continuous value, > 7
8	Temperature	Continuous value, > 30
9	NO_2	Continuous value
10	NH_4	Continuous value
11	Salinity	Continuous value
Clinical sign		
12	Discontinuous gut	1: Yes; 0: No
13	Empty gut	1: Yes; 0: No
14	Food in gut	1: Yes; 0: No
15	Yellow liquid in gut	1: Yes; 0: No
16	Soft shell	1: Yes; 0: No
17	Hepatopancreas atrophy	1: Yes; 0: No
18	Hepatopancreas pale	1: Yes; 0: No
19	Slow growth	1: Yes; 0: No
20	White feces	1: Yes; 0: No
21	White spot	1: Yes; 0: No
22	Vermiform	1: Yes; 0: No
23	Gregarine	1: Yes; 0: No
Disease		
24	WSSV	1: Yes; 0: No
25	EHP	1: Yes; 0: No
26	AHPND	1: Yes; 0: No

AHPND Acute hepatopancreatic necrosis disease, *EHP* disease caused by *Enterocytozoon hepatopenaei*, *WSSV* diseases caused by white spot syndrome virus

then combines it with coefficients, or weights, to learn and compute the result for the next nodes. The weights in each node are a negative or positive value, which influences the received and output data. The intelligence of this algorithm occurs through the connection and weight of nodes. In our study, we used a solver for weight named "*lbfgs*", which is an optimizer in the quasi-Newton method family, implemented in the scikit Python package (Pedregosa 2011).

Random forest

Random forest is based on the decision tree principle, by which decision tree root and internal nodes are the input and leaf nodes are the output. Random forest makes a prediction model by selecting samples randomly and uses features to build multiple decision trees. In each tree, a random vector value is determined (Breiman 2001). The result is obtained by majority voting of decision trees; therefore, the random forest is more suitable and powerful than a single decision tree. The random forest method belongs among the bagging techniques, which train many individual models in parallel.

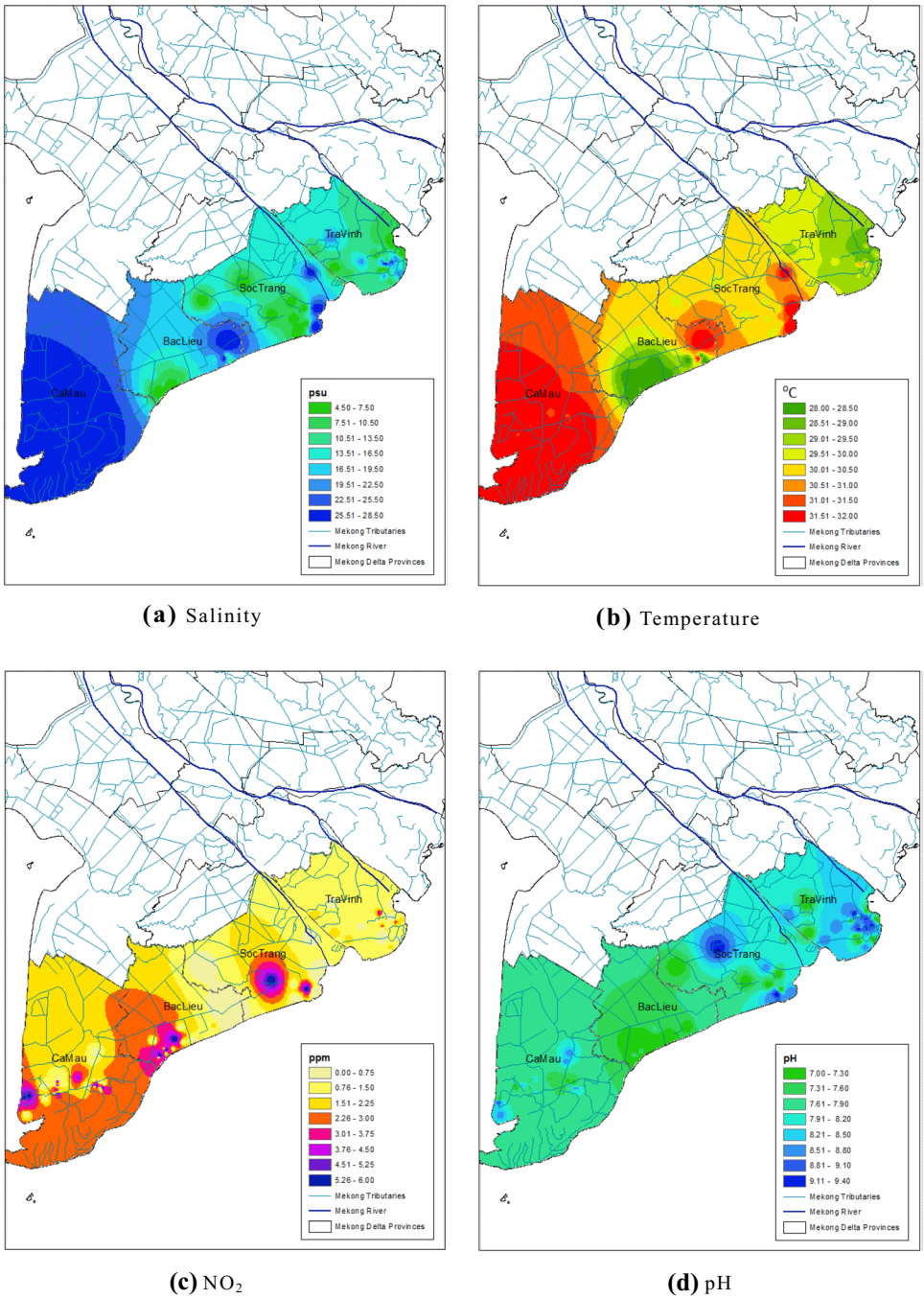
Gradient boosting

Gradient boosting is also a powerful machine learning technique based on a decision tree. This algorithm belongs to the group of boosting techniques that train a group of individual models in a sequential way, and each individual model learns from the mistakes made by the previous model. The aim of this algorithm is to make a weak learner into a strong learner, and it is developed through many applications (Natekin 2013). Here, we used gradient boosting and

random forest implemented in scikit-learn of the Python package (Pedregosa 2011).

Results

Missing values for the four environmental factors of temperature, salinity, pH, and NO_2 were obtained by interpolation, as shown in Fig. 2, and missing data on approximately 38 farms were obtained based on available data on the other 144 farms. The interpolated value of each variable for each farm fell in the range between the highest and lowest values of the farm. Once all data were complete, mapping was undertaken to visualize the locations of all included shrimp farms and hatcheries, as illustrated in Fig. 3. This mapping provided an overview of the distribution of the included farms and hatcheries on the east coast of the Mekong Delta. The distribution of the included farms was mainly in the coastal regions of Tra Vinh, Soc Trang, and Bac Lieu Province, but Ca Mau Province had a high concentration of farms in the



(a) Salinity **(b)** Temperature

(c) NO_2 **(d)** pH

Fig. 2 Interpolation of environmental factors

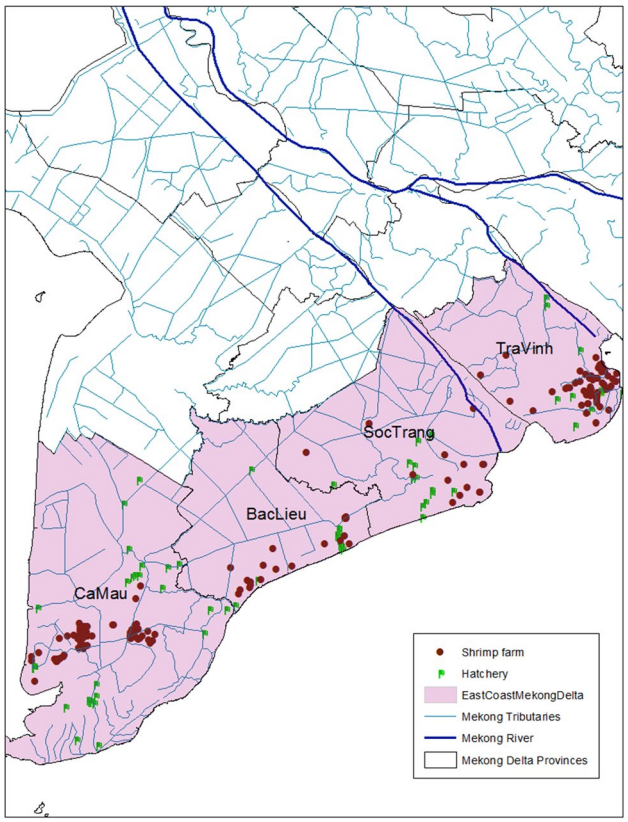


Fig. 3 Locations of shrimp farms and hatcheries on the east coast of the Mekong Delta

south. The geographic correlations of farms and hatcheries were also shown on the map.

To establish the distributions of the farms infected by each disease, we separately mapped the locations of farms with AHPND, EHP, and WSSV in the four provinces on the east coast of the Mekong Delta, as shown in Fig. 4. This visualization revealed the concentrations of infected farms as well as the areas that were heavily affected or less affected. Specifically, the density of farms infected by AHPND was high in Ca Mau and Tra Vinh Provinces, while EHP had less effect on farms in Bac Lieu Province, and WSSV was sparsely distributed throughout the entire study area. Subsequently, machine learning using these algorithms was used for prediction. The prediction accuracy of the logistic regression is shown in Table 2. The model attained values of 88.96% for WSSV, 86.89% for EHP, and 97.93% for AHPND; however, this algorithm showed low accuracy in the testing dataset: 72.97% (WSSV), 72.97% (EHP), and 91.89% (AHPND). The neural network model functioned better than the logistic regression model in our dataset, as shown in Table 3. For the training dataset, the accuracies of the neural network model was 97.24% for WSSV, 95.86% for EHP, and 96.55% for AHPND. Notably, the model was stable in predicting the testing dataset: 83.78%, 75.67%, and 91.89% for WSSV, EHP, and AHPND, respectively.

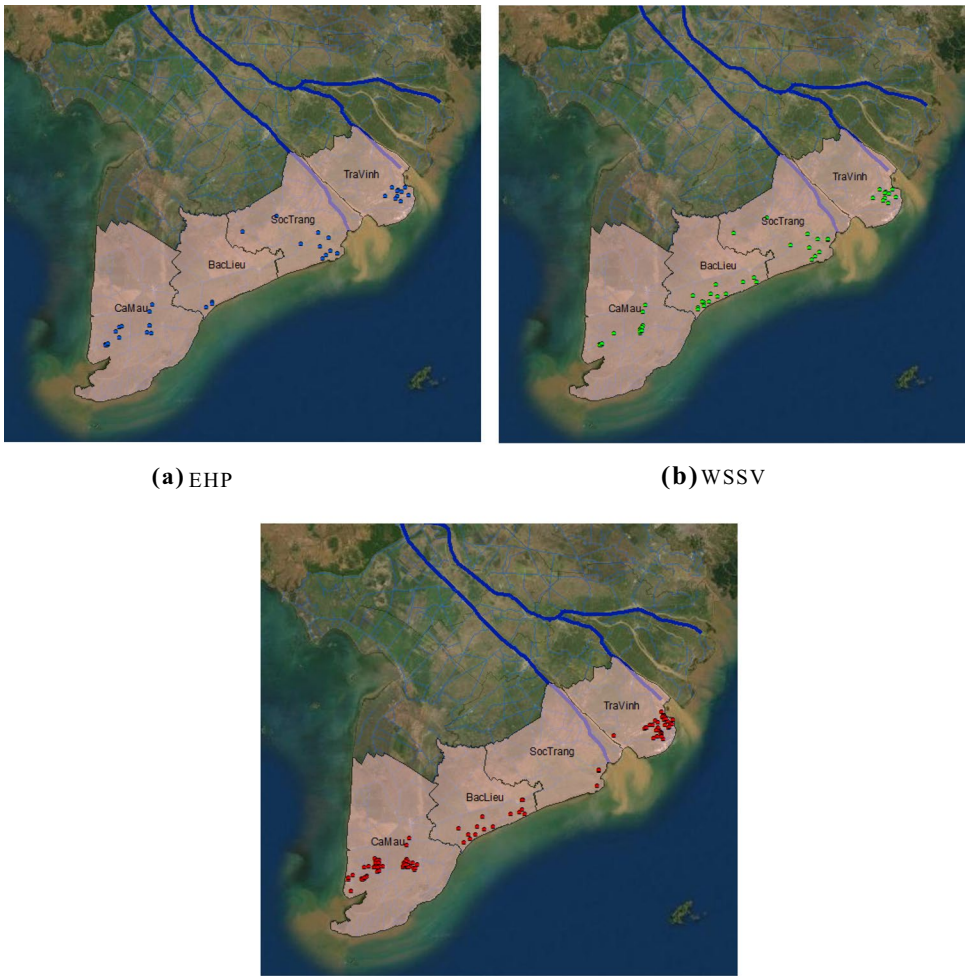
The random forest and gradient boosting methods provided over-fit models for our dataset, as indicated in Tables 4 and 5, respectively. Because these models learned details, they performed well with the training data. However, they could not ascertain the main trends of the dataset, which resulted in worse performance. For the training dataset, accuracy was 100% for all-disease prediction; nonetheless, these models yielded low accuracies in the testing set in comparison with those in the training set. Specifically, the random forest model predicted with accuracies of 83.78% for WSSV, 78.37% for EHP, and 83.78% for AHPND, and the gradient boosting method obtained accuracies of 78.37% for WSSV, 78.37% for EHP, and 81.08% for AHPND. The large difference in accuracy between the training dataset and the testing dataset showed that these two algorithms were not suitable for our analysis.

The results for the testing set (37 farms) of the neural network were mapped to reveal the accuracy of the method, which was obtained by comparing both affected and non-affected farms between the predicted number and actual cases of each disease. The results are shown in Fig. 5. The prediction accuracy was highest for AHPND farms (34 correctly identified farms), lowest for EHP farms (only 28y identified farms), and relatively good for WSSV farms (31 correctly identified and 6 incorrectly identified farms).

Next, the prediction for each disease was interpolated to obtain an estimation of infected areas on the east coast of the Mekong Delta. The estimated affected areas for each disease are visualized in Fig. 6, which provides a means to predict the areas infected by each disease. Areas heavily infected with EHP disease were mainly in Soc Trang Province; WSSV disease occurred in Bac Lieu Province and a portion of Soc Trang Province, and AHPND occurred most strongly in Tra Vinh and Ca Mau Provinces. These predictions will be useful to shrimp farmers for determining suitable locations to develop new farms.

Discussion

In this study, machine learning was used to predict the occurrence of AHPND, WSSV, and EHP in shrimp farms on the east coast of the Mekong Delta. GIS technology was used to map the distribution of these diseases in the target area. The extracted geographical information was then used for prediction. The one-versus-rest method of multi-label classification with machine learning played an important role in our predictions. Accurate predictions were achieved by the neural network method for both the training dataset and the testing dataset, and this method outperformed the logistic regression, random forest, and gradient boosting



(a) EHP **(b)** WSSV **(c)** AHPND

Fig. 4 Distribution of each shrimp disease under study. **a** Disease caused by *Enterocytozoon hepatopenaei* (*EHP*), **b** diseases caused by white spot syndrome virus (*WSSV*), **c** acute hepatopancreatic necrosis disease (*AHPND*)

Table 2 Results of predictions based on the logistic regression model

Disease	Training dataset		Testing dataset	
	Number of farms correct	Percentage correct	Number of farms correct	Percentage correct
WSSV	129	88.96	27	72.97
EHP	126	86.89	27	72.97
AHPND	142	97.93	34	91.89

Table 3 Results of predictions based on the neural network model

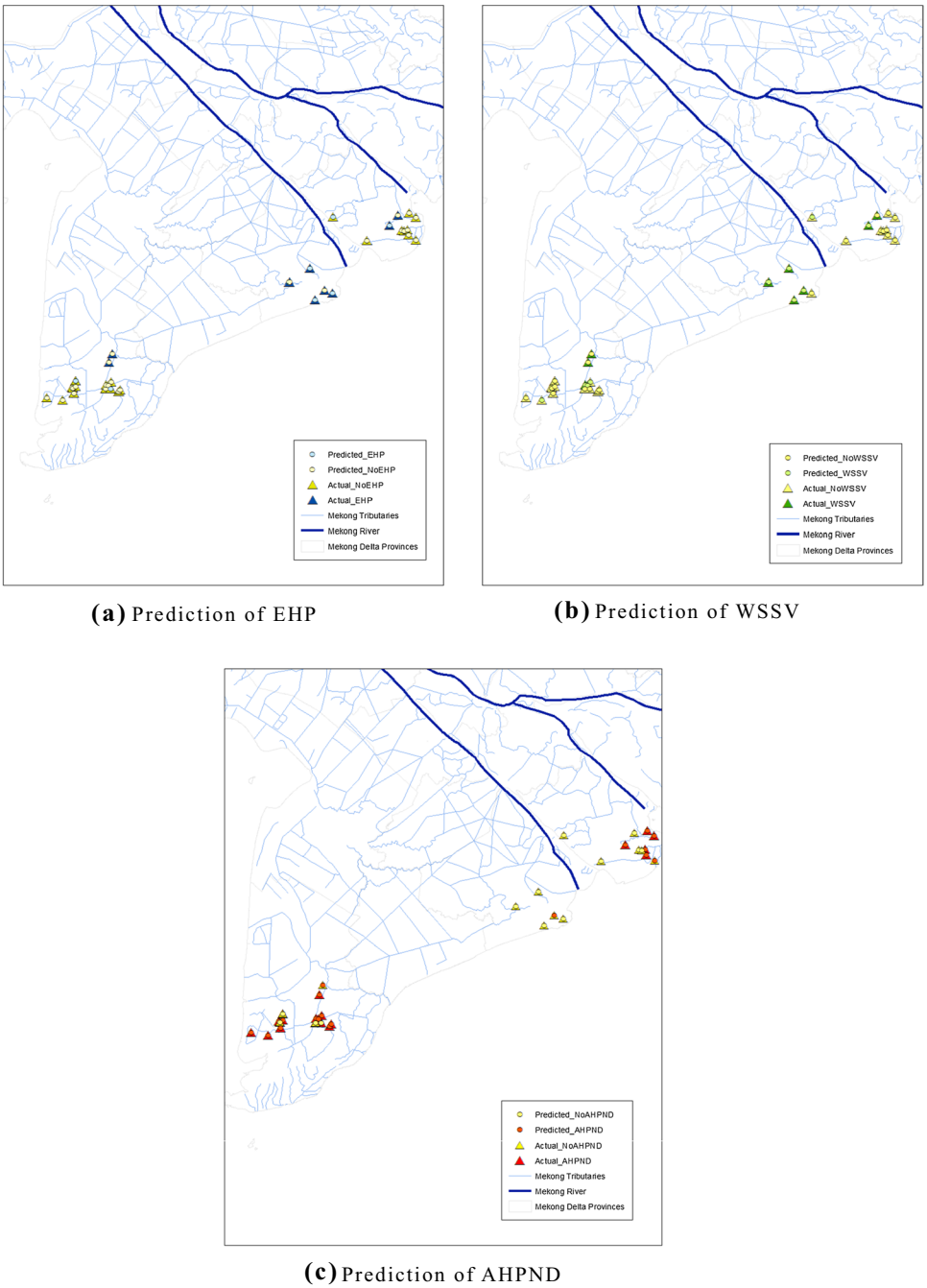
Disease	Training dataset		Testing dataset	
	Number of farms correct	Percentage correct	Number of farms correct	Percentage correct
WSSV	141	97.24	31	83.78
EHP	139	95.86	28	75.67
AHPND	140	96.55	34	91.89

Table 4 Results of predictions based on the random forest model

Disease	Training dataset		Testing dataset	
	Number of farms correct	Percentage correct	Number of farms correct	Percentage correct
WSSV	145	100	31	83.78
EHP	145	100	29	78.37
AHPND	145	100	31	83.78

Table 5 Results of predictions based on the gradient boosting model

Disease	Training dataset		Testing dataset	
	Number of farms correct	Percentage correct	Number of farms correct	Percentage correct
WSSV	145	100	29	78.37
EHP	145	100	29	78.37
AHPND	145	100	30	81.08



(a) Prediction of EHP

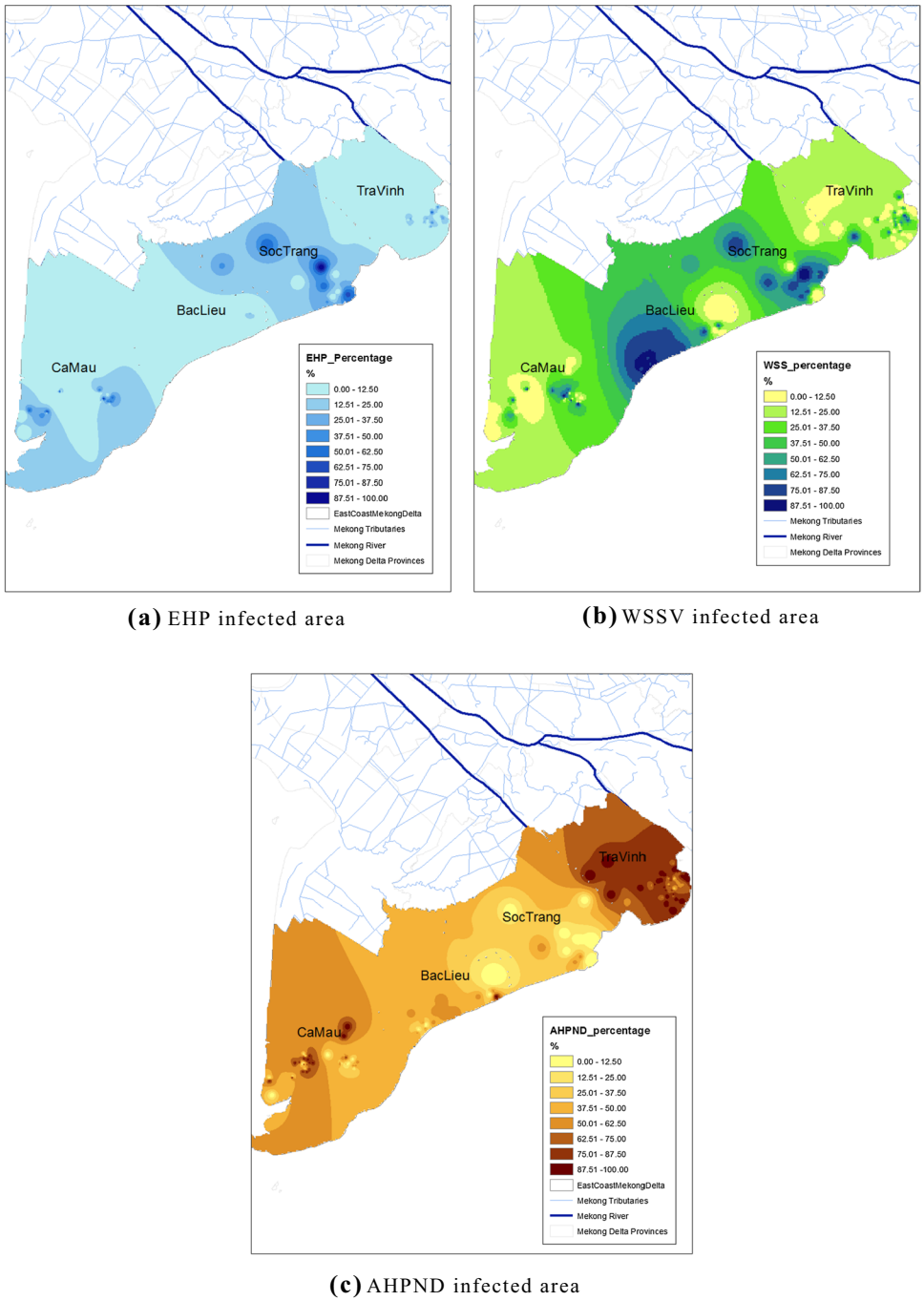
(b) Prediction of WSSV

(c) Prediction of AHPND

Fig. 5 Prediction results of the neural network model for the testing dataset

methods. The performance of the logistic regression method was stable, but its accuracy was low, while with the random forest and gradient boosting methods there were considerable differences between the estimates based on the training dataset and those based on the testing dataset.

This study contributes to disease management by helping shrimp farmers to understand how GIS-based technology can be used to visualize disease outbreaks and to determine strategies for reducing the risk of disease. The combination of GIS and machine learning provided comprehensive prediction and an intuitive map which provided visualization of the distribution of disease; this visualization is meaningful in terms of evaluating the status of infection in the coastal area of the Mekong Delta. Knowledge of disease status at the local levels also allows assessment of the effectiveness of disease management activities. Heavily infected areas may be related to weak farm management, with the latter contributing to cross-contamination between farms or an infected seed source, whereas areas with low levels of infection suggest good disease management on farms. Based on such information, shrimp farmers can easily determine suitable



(a) EHP infected area

(b) WSSV infected area

(c) AHPND infected area

Fig. 6 Distribution of each disease in the study area

locations for new farms or prepare appropriate solutions to avoid infection.

The use of GIS in this study contributed to the clarification of the outbreak and spread of disease that was analyzed based on the locations of farms, hatcheries, and river tributaries. This information was used to provide an initial diagnosis of the source of infection. For example, a new outbreak of disease at a farm could arise from its proximity to a previously infected farm or an infected hatchery seed source. Therefore, the GIS reinforced the prediction probability. Moreover, each disease caused damage to shrimp farms in different areas, resulting in different levels of local infection. The visualization of predicted affected areas (shown in Fig. 6) indicated that disease affected proximal areas more heavily than distant areas. EHP disease spread locally and was more destructive in the Tran De District of Soc Trang Province than elsewhere in this province because of the high local density of EHP-infected farms. AHPND widely infected the Duyen Hai District and spread to adjacent areas, such as the Cau Ngang and Tra Cu districts of Tra Vinh

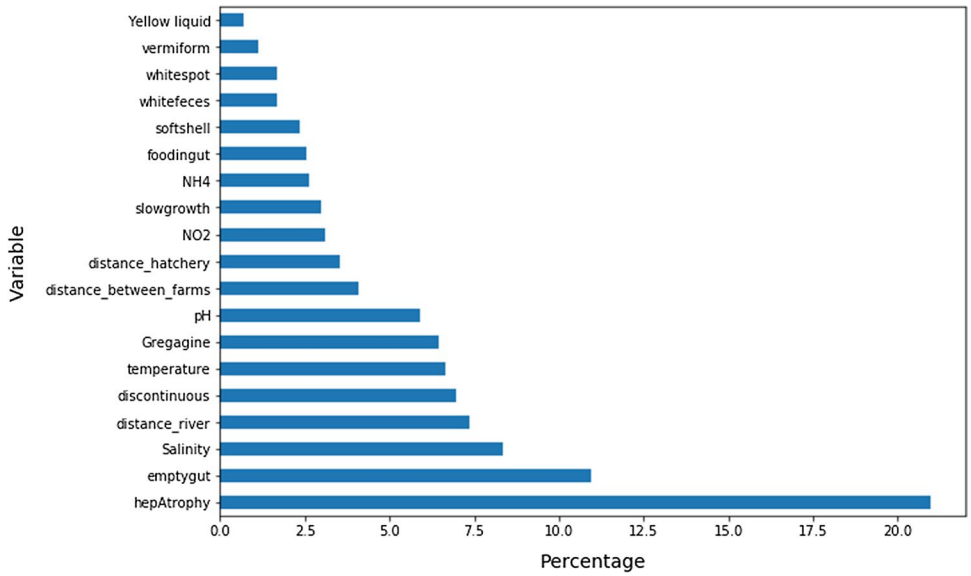


Fig. 7 Contribution of each variable to disease prediction

Province, both of which are downstream of the Co Chien and Hau Rivers. In Ca Mau Province, Phu Tan District was heavily infected with AHPND, which then spread to neighboring districts, including Cai Nuoc, the western part of Tran Van Thoi, and Nam Can. Notably, the infected areas in Ca Mau Province were within mangrove forests where there were canals. WSSV widely infected the coastal region of Tran De in the Vinh Chau District of Soc Trang Province and the Dong Hai District of Bac Lieu Province. Thus, adjacent infected areas strongly contributed to expansion of disease in this region. This expansion of disease was considered to have resulted from between-farm spread due to the transmission of pathogens on clothing, farm equipment, and/or birds. The prediction map showed that many proximal farms in the downstream region of the Co Chien River of Tra Vinh Province and the Cai Doi canal in Ca Mau Province were affected by AHPND, which spread to surrounding areas, whereas farms affected by WSSV were found around the Ganh Hao canal in Bac Lieu Province.

investigation of the closest distance between farms and the river revealed that some farms that shared the same river water source. Downstream farms are at high risk when infection occurs upstream because the pathogen can spread by the river water. To avoid the spread of infection to neighboring farms, shrimp farmers must manage the quality of their water sources and isolate any infected farms. Once an infected farm is detected, other farms using the same water source should be notified. Sources of water for shrimp farms should be checked frequently for pathogens, a step which requires a rigorous farm management policy.

Furthermore, to increase the comprehensiveness of prediction, we examined environmental factors related to conditions suitable for strong activation of pathogens. Temperature and salinity strongly affect disease, which tends to break out in hot weather and under conditions of high salinity, but other factors, such as pH, NH_4, and NO_2 levels, also influence infection rates. For example, temperatures > 28 °C and salinity > 20 ppt are suitable for the rapid spread of AHPND, and the probability of WSSV increases with changes in salinity, temperature, and pH (World Organization for Animal Health [OIE] 2019). These environmental factors are particularly noticeable in the Mekong Delta, where hot and dry weather result in conditions favorable for higher risk of disease. Among the environmental factors, salinity contributed the most to disease prediction, followed by temperature, pH, NO_2, and NH_4 (Fig. 7). Although the environment affects the estimate of the area of disease spread, this process is mainly based on evidence of whether infected farms are present. Accordingly, Fig. 6c shows the estimates of the area inland of Soc Trang Province where there was less AHPND infection due to conditions of low temperature and low salinity. Additionally, no AHPND-infected farms were found.

This study has a number of limitations. Most importantly, it included a relatively low amount of observed data for the target area. The neural network had a relatively low prediction accuracy for EHP, but a high accuracy for AHPND and WSSV (75.67, 91.89, and 83.78%, respectively). This difference in prediction accuracies was due to it being easier to distinguish farms affected by AHPND and WSSV because these diseases present typical signs, such as hepatopancreatic paleness and atrophy for AHPND and white spots for WSSV. In contrast, EHP is not easily recognized because it is evidenced only by the slow growth of shrimp and can be confused with white feces disease or symptoms of poor nutrition. To improve EHP prediction accuracy, more data are required, such as the density of shrimp in ponds and details of feeding and care regimes. Although the Mekong

Delta has many shrimp farms and disease is highly prevalent, as evidenced by huge economic losses, disease data are difficult to collect. Also, because disease outbreaks constitute a sensitive situation, shrimp farmers usually do not share information on the status of their infected farm. Furthermore, farmers usually find treatments themselves. Additionally, disease research requires long periods of sufficient data, especially for extensive farms. However, if data could be collected from all shrimp farms in the region, including both healthy and infected farms, the visualization of disease distribution would be clearer and prediction would be more accurate. Full mapping of all farms would provide a foundation for future research, such as detection of affected populations, the effects of industrial pollutants on disease, analysis of the most suitable areas for farm development, and assessments of annual changes in shrimp farm distribution.

In the future, this research can be extended to other diseases of shrimp that occur in cultured shrimp farms, such as white feces disease, yellow head disease, Taura syndrome, opaque muscle disease, among others. Additional data on clinical signs and additional environmental and geographical factors will be collected. The geographical data can be enriched by analyzing more variables, including pond density, factory locations, and rainfall. In machine learning, the more data that are investigated, the more comprehensive will be the prediction. Infection prediction will become more accurate when the dataset is updated with additional data and, accordingly, the estimated area of the disease will be reliably visualized in the infected area. Additionally, given suitable data this research can be applied to protect shrimp farms in regions other than those located on the east coast of the Mekong Delta. It is anticipated that this study will be implemented into an application supported by modern techniques, including sensors and GPS-enabled smart phones, to collect data and also analyze data in real time. Our research will contribute to the development of sustainable shrimp farming in the Mekong Delta region.

Acknowledgements This study is funded in part by the Can Tho University Improvement Project VN14-P6, supported by a loan of Japan's Official Development Assistance (ODA). The English in this document has been checked by at least two professional editors, both native speakers of English. For a certificate, please see: http://www.textcheck.com/certificate/pYPXsc

References

Aguilar-Manjarrez J, Crespi V (2013) National Aquaculture Sector Overview map collection. User manual. FAO, Rome. https://www.fao.org/3/i3103b/i3103b00.htm. Accessed 20 Jun 2021

Aranguren LF, Han JE, Tang KFJ (2017) Enterocytozoon hepatopenaei (EHP) is a risk factor for acute hepatopancreatic necrosis disease (AHPND) and septic hepatopancreatic necrosis (SHPN) in the Pacific white shrimp *Penaeus vannamei*. Aquaculture. https://doi.org/10.1016/j.aquaculture.2016.12.038

Boonyawiwat V, Patanasatienkul V, Kasornchandra J, Poolkhet C, Yaemkasem S, Hammel L, Davidson J (2016) Impact of farm management on expression of early mortality syndrome/acute hepatopancreatic necrosis disease (EMS/AHPND) on penaeid shrimp farms in Thailand. Fish Dis. https://doi.org/10.1111/jfd.12545

Breiman L (2001) Random Forests. Mach Learn. https://doi.org/10.1023/A:1010933404324

Chayaburakul K, Nash G, Pratanpipat P, Sriurairatana S, Withyachumnarnkul B (2004) Multiple pathogens found in growth-retarded black tiger shrimp *Penaeus monodon* cultivated in Thailand. Dis Aquat Organ. https://doi.org/10.3354/dao060089

Dang TL, Pham AT, Phan TV (2018) Acute hepatopancreatic necrosis disease (AHPND) in Vietnam. Asian Fish Sci. https://doi.org/10.33997/j.afs.2018.31.S1.020

Duc NM (2009) Application of econometric models for price impact assessment of antidumping measures and labelling laws on global markets: a case study of Vietnamese striped catfish. Rev Aquac. https://doi.org/10.1111/j.1753-5131.2010.01024.x

Giap DH, Yi Y, Cuong NX, Luu LT, Diana JS, Lin CK (2003) Application of GIS and remote sensing for assessing watershed ponds for aquaculture development in Thai Nguyen, Vietnam. Map Asia Conference, Thailand, GIS Development Network, 8 pp

Ha N, Ha D, Thuy N, Lien V (2010) Enterocytozoon hepatopenaei has been detected parasitizing tiger shrimp (*Penaeus monodon*) cultured in Vietnam and showing white feces syndrome (in Vietnamese with English abstract). Sci Technol J Agri Rural Devel 12:45–50

Ha TTP, Dijk HV, Bosma R, Sinh LX (2013) Livelihood capabilities and pathways of shrimp farmers in the Mekong Delta, Vietnam. Aquac Econ Manag. https://doi.org/10.1007/s12562-020-01427-z

Jesse R, Pfahringer B, Holmes G, Frank E (2009) Classifier chains for multi-label classification. In: Buntine W, Grobelnik M, Mladenić D, Shawe-Taylor J (eds) Machine learning and knowledge discovery in databases. ECML PKDD 2009. Lecture Notes in Computer Science, vol 5782. Springer, Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04174-7_17

Kapetsky JM, Aguilar-Manjarrez J, Jenness J (2013) A global assessment of potential for offshore mariculture development from a spatial perspective. FAO Fisheries and Aquaculture Technical Paper. No. 549. FAO, Rome. https://www.fao.org/3/i3100e/i3100e.pdf. Accessed 20 Jun 2021

Khiem NM, Takahashi Y, Oanh DTH, Hai TN, Yasuma H, Kimura N (2020) The use of machine learning to predict acute hepatopancreatic necrosis disease (AHPND) in shrimp farmed on the east coast of the Mekong Delta of Vietnam. Fish Sci. https://doi.org/10.1007/s12562-020-01427-z

Leung PS, Tran LT (2000) Predicting shrimp disease occurrence: artificial neural networks vs. logistic regression. Aquaculture. https://doi.org/10.1016/S0044-8486(00)00300-8

Meaden GJ, Aguilar-Manjarrez J (2013) Advances in geographic information systems and remote sensing for fisheries and aquaculture. Summary version. FAO Fisheries and Aquaculture Technical Paper No. 552. FAO, Rome. https://www.fao.org/3/i3102e/i3102e00.htm. Accessed 20 Jun 2021

Montañes E, Senge R, Barranquero J, Quevedo JR, Coz JJ, Hüllermeier E (2014) Dependent binary relevance models for multi-label classification. Pattern Recogn. https://doi.org/10.1016/j.patcog.2013.09.029

Natekin A, Knoll A (2013) Gradient boosting machines. Front Neurorobot 7:21. https://doi.org/10.3389/fnbot.2013.00021

Oanh DTH, Phuong NT (2005) Prevalence of white spot syndrome virus (WSSV) and monodon aculovirus (MBV) infection in *Penaeus monodon* postlarvae in Vietnam. In: Walker P, Lester R, Bondad Reantaso MG (eds) Diseases in Asian aquaculture V. Fish Health Section, Asian Fisheries Society, Manila, pp 395–404

Pedregosa F (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830. https://scikit-learn.org/stable/. Accessed 10 May 2021

Peter CL, Terry RH, Black KP (2008) An integrated GIS approach for sustainable aquaculture management area site selection. Ocean Coast Manag. https://doi.org/10.1016/j.ocecoaman.2008.06.010

Phuong NT, Oanh DTH (2010) Stripped catfish aquaculture in Vietnam: a decade of unprecedented development. In: De Silva SS, Davy FB (eds) Success stories in Asian aquaculture. Springer, Dordrecht, pp 131–147

Ramos-Carreño S, Valencia-Yáñez R, Correa-Sandoval F, Ruíz-García N, Díaz-Herrera F, Giffard-Mena I (2014) White spot syndrome virus (WSSV) infection in shrimp (*Litopenaeus vannamei*) exposed to low and high salinity. Adv Virol. https://doi.org/10.1007/s00705-014-2052-0

Rao PV, Reddy AR, Sucharita V (2017) Computer aided shrimp disease diagnosis in aquaculture. IJRASET. https://doi.org/10.22214/ijraset.2017.2079

Satheesh KS, Bharathi RA, Rajan JJS, Chitra V, Muralidhar M, Alavandi SV (2019) Viability of white spot syndrome virus (WSSV) in shrimp pond sediments with reference to physicochemical properties. Aquacult Int. https://doi.org/10.1007/s10499-019-00394-2

Shahriar MS, McCulluch J (2014) A dynamic data-driven decision support for aquaculture farm closure. Procedia Comput Sci. https://doi.org/10.1016/j.procs.2014.05.111

Soykan CU, Eguchi T, Kohin S, Dewar H (2014) Prediction of fishing effort distributions using boosted regression trees. Ecol Appl. https://doi.org/10.1890/12-0826.1

Thanh NK, Tripathi NK, Duan HD, Gallardo WG (2008) GIS-based planning for sustainable shrimp farming In Thanh Phu District, Ben Tre Province, Viet Nam. In: The International Conference on GeoInformatics for Spatial-Infrastructure Development in Earth & Allied Sciences (GIS-IDEAS). Hanoi, Vietnam. https://gisws.media.osaka-cu.ac.jp/gisideas08/viewabstract.php?id=273. Accessed 15 Jul 2021

Thuoc P, Thanh ND (2013) Application of GIS technology for fishing ground forecasting of Tuna fish in Vietnam sea waters. Vietnam J Mar Sci Technol. https://doi.org/10.15625/1859-3097/12/4/2580

World Organization for Animal Health (OIE) (2019) Manual of diagnostic tests for aquatic animals. Paris, World Organization for Animal Health

Xu J (2011) An extended one-versus-rest support vector machine for multi-label classification. Neurocomputing. https://doi.org/10.1016/j.neucom.2011.04.024

Zheng Z, Aweya JJ, Wang F, Yao D, Lun J, Li S, Ma H, Zhang Y (2018) Acute Hepatopancreatic Necrosis Disease (AHPND)-related microRNAs in *Litopenaeus vannamei* infected with an AHPND-causing strain of *Vibrio parahemolyticus*. BMC Genomics. https://doi.org/10.1186/s12864-018-4728-4

Zou J, Han Y, So SS (2008) Overview of artificial neural networks. Methods Mol Biol 458:15–23. https://doi.org/10.1007/978-1-60327-101-1_2