Fisheries Science (2020) 86:673–683
https://doi.org/10.1007/s12562-020-01427-z

ORIGINAL ARTICLE

Aquaculture

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

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

Received: 13 March 2020 / Accepted: 16 April 2020 / Published online: 6 May 2020

Abstract

Predicting the outbreak of disease is essential when managing shrimp farms. Acute hepatopancreatic necrosis disease (AHPND) caused by *Vibrio parahaemolyticus* is a serious disease in shrimp. It is essential that shrimp farmers on the east coast of the Mekong Delta detect the disease as early as possible, because the mortality rate can reach 100%. Here, we used machine learning to predict AHPND development based on data collected since 2010 from shrimp farms in Tra Vinh, Ben Tre, Bac Lieu, and Ca Mau provinces. We initially hypothesized that the dependent variable, AHPND, was affected by 31 independent variables, but ultimately used 15 key variables to train the models. Logistic regression, artificial neural network, decision tree, and K-nearest neighbor analyses were performed, and the accuracy of the predictions was evaluated using hold-out and cross-validation tests. Logistic regression, as the most stable algorithm, was thus used to predict AHPND outbreaks in shrimp farms.

Keywords Acute hepatopancreatic necrosis disease (AHPND) · Machine learning · Shrimp disease

Introduction

Aquaculture is a key economic driver in Vietnam. Long-term sustainable shrimp-farming contributes 4–5% of the gross domestic product (GDP) of Vietnam and currently employs more than 4 million people. By exploiting the available resources, the fishery industry has grown very rapidly, producing 6.1 and 6.7 million tonnes of product in 2015 and 2016, respectively (COFI 2019). The Vietnamese coastline is 3260 km long, and is amenable to both fish and shrimp farming. Many shrimp farms have been developed in southern Vietnam, where the Mekong Delta (area of 40,000 km^2) is one of the most productive fishing and shrimping zones, contributing about 80% of all farmed shrimp exports. Shrimp farming is a key component of Vietnamese aquaculture. However, about 80% of all shrimp farms report regular disease outbreaks (ADB-NACA 1998) that are initially difficult to detect. Acute hepatopancreatic necrosis disease (AHPND), caused by *Vibrio parahaemolyticus*, first appeared in 2010 and has caused severe shrimp losses not only in Vietnam, but worldwide. Overt disease is evident within 20–30 days of infection (10 days for shrimp in ponds receiving early casein supplements); AHPND is thus also termed early mortality syndrome (EMS). Environmental factors, shrimp seedstocks, and poor nutritional management may contribute to infection (Crane 2019). The causative agent can be detected in the hepatopancreas; the characteristic disease signs include a pale and tough hepatopancreas, and fluid in the gut. AHPND has caused massive shrimp losses (Zheng et al. 2018). From 2010 to 2017, AHPND caused major economic losses in Thailand; the value of shrimp traded in the Mahachai Market fell by about $US7.4 billion, with

✉ 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, CanTho University, Can Tho, Vietnam

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

a further $US4.2 billion loss in exports. In the Mekong Delta of Vietnam, the losses attributed to AHPND in 2015 exceeded $US26 million (Shinn et al. 2018). Thus, AHPND is a serious problem in the Mekong Delta. Most shrimp farms are very large; the farmers pay minimal attention to shrimp nutrition, water temperature, or pH. The risk of AHPND is very high, and enormous areas may be affected. In 2012, the total area of affected farms was 46,093 ha [Soc Trang province, 23,371 ha (56.6% of the total area); Bac Lieu, 16,919 ha (41.9%); Tra Vinh, 12,224 ha (49.5%); and Ben Tre, 2237 ha (29.0%)] (Dang et al. 2018).

Much effort has been devoted to the detection and prevention of AHPND. A Thai case study evaluated farm characteristics and management, pond and water preparation, feed management, and post-larval shrimp and stock management (Boonyawiwat 2017). Chlorine treatment and reservoir availability were assessed, and predatory fish counted. Dhar et al. (2019) used the polymerase chain reaction technique to show that the toxin-encoding *pirA* and *pirB* genes played key roles in AHPND development in white leg shrimp *Litopenaeus vannamei* on a shrimp farm in Texas. In the Philippines, AHPND has caused major losses of cultured *Litopenaeus vannamei* and *Penaeus monodon* (Leobert et al. 2015). AHPND was evident in both shrimps of marketable size and late-stage juveniles.

The Vietnamese government has established a national task force to prevent and manage AHPND. Of the factors affecting AHPND development in the Mekong Delta, environmental conditions are of particular concern, because these are closely associated with disease development (Glenn 1976). Pond and water quality management is very poor, and pollution is widespread; pond water is taken from canals into which effluent is discharged, and cross-contamination among farms is rife (Claude et al. 2019). Environmental factors increasing the risk of AHPND at ponds and farms include larger ponds, sun-drying of pond bottoms, and proximity to already infected farms (Boonyawiwat 2018). Some practical solutions (an appropriate stocking density, monitoring of environmental parameters, and nutrition monitoring in the first month of life) have been suggested (Dang et al. 2018). However, such physical changes work only temporarily; they are not long-term solutions.

In recent years, machine learning has been applied for disease prediction in aquaculture (Rahman and Tasnim 2014). However, only a few studies have sought to predict shrimp diseases. Data collection is difficult, and technical barriers may be encountered. Disease timing is a particularly important issue. Some studies have used computers to process digital images, to facilitate the diagnosis of diseases of aquaculture (Rao 2017). Artificial neural networks have been used to diagnose protozoan and bacterial infections of fish (Lopes et al. 2011), and shrimp diseases have been reliably diagnosed using logistic regression (Leung and Tran 2000).

AHPND symptoms are easily confused with those of other diseases ("white feces" and gut conditions). Water parameters (dissolved oxygen level, pH, and temperature) and shrimp density and nutrition affect AHPND development. An understanding of the relationships among symptoms, external conditions, and AHPND status would be highly useful. However, no computer-based model predicting AHPND is yet available. Here, to understand the relationship between symptoms, external conditions, and acute hepatopancreatic necrosis disease (AHPND), we used machine learning to evaluate historical shrimp farm data from the Mekong Delta. We used an artificial neural network, logistic regression, the K-nearest neighbor approach, and a decision tree to analyze the data, and built a predictive algorithm using the best approach.

Materials and methods

Dataset

Data from 2010 to 2019 were collected from shrimp farms on the east coast of the Mekong Delta. This region has been heavily affected by AHPND since 2010 (Fig. 1); the research area includes four provinces: Soc Trang, Ca Mau, Tra Vinh, and Bac Lieu. The white leg shrimp *L. vannamei* and the tiger prawn *P. monodon* are commonly cultured in the delta; 763 samples were collected from 80 shrimp ponds of 50 farms and then analyzed (Table 1). Both continuous and categorical parameters were evaluated. If a variable was associated with disease symptoms, it was coded as 1, and otherwise as 0. The dataset was divided into symptoms, visceral status, environmental factors, and general management practices. Symptoms included a curved body, opaque muscles, poor growth, black spots, a lack of appetite, a soft shell, and dirty gills. Visceral status included gut and hepatopancreatic swelling, atrophy, toughness, pallor, emptiness, fluid in the gut, and a discontinuous gut. Environmental factors included pond water pH and temperature, dissolved oxygen, NO_2, NH_4, NH_3, salinity, and alkalinity levels, and chemical oxygen demand. Management factors included shrimp age, the time of symptom detection, fresh smear test results, mortality rates, shrimp density and seedstock origin, province, and pond area. The data were collected at different times from different provinces; any missing categorical data were scored as 0 (no symptoms). The mortality rate, timing of symptom detection, and seedstock origin were predicted by logistic regression. Mean values were entered into the analysis for shrimp age and density, water pH, temperature, and salinity, NO_2, dissolved oxygen, NH_4, and NH_3 levels, and pond area.

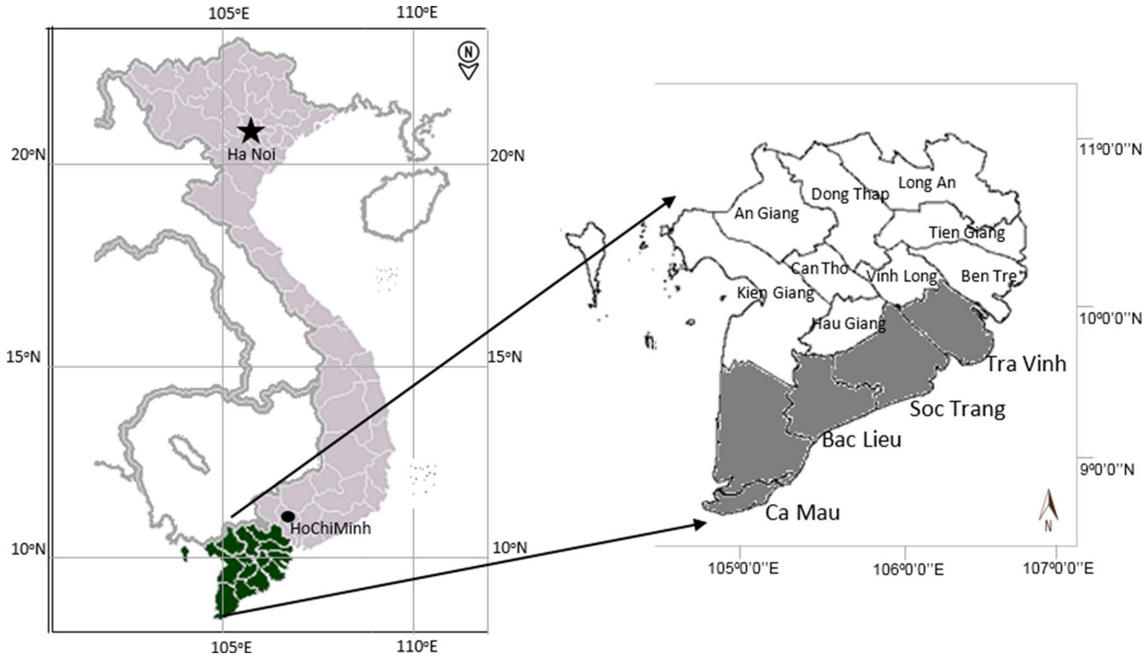


Fig. 1 The east coast of the Mekong Delta

Machine learning

We included variables considered informative in the literature (Crane 2019), and according to the results of machine learning; these included poor seedstock quality, excessive water temperature and salinity, and shrimp mortality. Machine learning can be accomplished using backward elimination or forward selection. The former approach was preferred here, because it retains potentially useful variables in the analysis. Independent variables that correlated strongly with dependent variables were selected as follows. First, probit regression was used to identify potentially informative variables, and backward elimination was then employed to exclude certain of these variables to improve model accuracy.

A probit model is a regression procedure for binary classification; the dependent variable takes one of only two values. Here, as mentioned above, a variable that affected AHPND status was coded as 1, and otherwise as 0. All independent variables with p values ≥ 0.05 were excluded from the model. We then proceeded to delete additional, less informative variables, performing logistic regression after omitting each variable to assess model accuracy. The first model included all 16 variables; each variable was then omitted on an individual basis to determine whether it was informative. The five least significant variables were omitted according to this process.

For machine learning, the dataset was divided into training and test datasets (3:1 ratio). Hold-out and cross-validation tests (of 572 and 191 samples, respectively) were performed to determine model accuracy. The accuracy of models generated during training was assessed using the validation subset. The cross-validation tests used algorithm-dependent, iterative -fold values (k values, where k is the number of tests); each fold included 673/k samples. Each sample had to appear once in the testing subset and could be in both the training and testing subsets during any one test. We evaluated four machine learning approaches: logistic regression, the KNN algorithm, a decision tree, and an ANN.

Logistic regression

Logistic regression models the binary probability and is often used for probabilistic prediction. The following equation (where the x values are independent variables) is used:

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

where p is the probability of the outcome, and coefficients β_1 to β_n are assigned to variables x_1 to x_n. The outcome is the logarithm of the ratio of two probabilities, i.e., that AHPND will or will not develop. The algorithm can also be represented by the sigmoid (or cost) function, which is a continuous approximation of the step function yielding an output between 0 and 1:

$$\sigma(x) = \frac{1}{1 - e^{-x}}, \tag{2}$$

Table 1 List of variables

	Variable	Variable Description
General management		
1	Shrimp age	Continuous variable
2	Day detect symptom	Continuous variable, value is smaller than shrimp age
3	Fresh smear test	1: Yes (already tested); 0: No (not tested yet)
4	Mortality	Continuous variable, value from 0 to 100
5	Density	Continuous variable
6	Seed origin	1–14 ID of hatchery 1: Hoang Gia-CP 2: Tran Hau Dien 3: Costal Seed Production 1 4: Le Tuan Phat 5: Mien Trung 6: Ninh Thuan 7: Xuan Bay 8: Dong Khoi 9: Kim Sa 10: Duong Hung 11: CP 12: Long Phu 13: Tung Bach 14: Thong Thuan
7	Province	1–4: ID of province on east coast of Mekong Delta 1: Ca Mau 2: Tra Vinh 3: Soc Trang 4: Bac Lieu
8	Area	Continuous value
Viscera status		
10	Hepatopancreas swelling	1: Yes; 0: No
11	Empty gut	1: Yes; 0: No
12	Fluid in gut	1: Yes; 0: No
13	Hepatopancreas atrophy	1: Yes; 0: No
14	Hepatopancreas tough	1: Yes; 0: No
15	Hepatopancreas pale	1: Yes; 0: No
16	Discontinuous gut	1: Yes; 0: No
Symptom appearance		
17	Curved body	1: Yes; 0: No
18	Opaque muscle	1: Yes; 0: No
19	Low growth	1: Yes; 0: No
20	Black spot	1: Yes; 0: No
21	Low eating	1: Yes; 0: No
22	Soft shell	1: Yes; 0: No
23	Dirty gills	1: Yes; 0: No
Environmental factors		
24	pH	Continuous value, > 7
25	Temperature	Continuous value, > 30
26	DO	Continuous value
27	NO_2	Continuous value
28	NH_4	Continuous value
29	NH_3	Continuous value
30	Salinity	Continuous value

Table 1 (continued)

	Variable	Variable Description
31	kH	Continuous value
32	COD	Continuous value

where x includes independent variables x_1–x_n. This is more useful than a linear function that can yield values > 1 or < 0. Unlike linear regression, the logit procedure uses the maximum-likelihood method; the mean and variance are used to determine specific parametric values. We used the sklearn Python package in this study (Cournapeau 2007).

K-nearest neighbor algorithm

The KNN algorithm is used for both classification and regression, and assumes that similar data are in close proximity (Zhang 2016). The KNN algorithm is based on the concept of similarity (i.e., the distance between, or closeness of, the points on a graph). We transformed data points mathematically and calculated the Euclidean distances between them. In particular, when $K = 1$, the object is simply assigned to the single nearest neighbor. In machine learning terms, this is considered "lazy learning," because all approximations are local. Nearest neighbors affect classification more so than distant neighbors. The algorithm run time depends on the K-values; if K is large, the algorithm runs rapidly. We set K to 5 and used the KNN procedure of the scikit Python package (Cournapeau 2007).

Decision tree

A decision tree uses the shape of a tree to predict target values from input variables. One root node and multiple internal nodes are the inputs; each leaf is an output. The dataset is classified into specific classes. Here, we used a decision tree to predict AHPND status (yes or no). The algorithm is depth-dependent; a deeper tree is better trained and more accurate. We used the decision tree of the scikit Python package (Cournapeau 2007).

ANN

An ANN is a complex algorithm inspired by the human brain (Harston et al. 1990). ANNs have many connected nodes that process data and yield outputs based on simple mathematical operations. Several parameters must be established at every node. These nodes, where computations occur, are organized into layers. Every node combines input data with coefficients or weights to both learn and yield an output for the next node. For complex calculations, the ANN has multiple hidden layers between the input and output layers. The nodes on the hidden layers allow many relationships (termed mapping functions) between the input and output layers to be tested. Each hidden layer node must learn by minimizing a cost function, which is a measure of how effectively the ANN detects the relationship between a given input and the expected output. We used the ReLu activation function implemented in the ANN of the scikit Python package (Cournapeau 2007). ANN complexity depends on the number of layers; we used one input layer, one hidden layer, and one output layer.

Results

Probit regression showed that 16 of 27 independent variables were useful predictors of AHPND status (Table 2). The results of backward elimination are shown in Table 3. Of the 16 originally included variables, 5 were omitted (province, timing of symptom detection, shrimp age, fresh smear test results, and poor growth), such that 11 variables remained in the model. The model accuracy increased after each removal step. We then input 15 variables into the machine learning algorithms: 4 were manually selected, and the remaining 11 were those mentioned above.

Logistic regression

The logistic regression results are shown in Table 4. Hold-out tests using 572 training samples and 191 test samples yielded accuracy rates of 90.33% and 85.50%, respectively. For the cross-validation test, the data were divided into three folds, each with 254 samples, and tested three times (training subset, 509 samples; testing subset, 254 samples). The accuracy was 83.04%.

ANN

The ANN results are shown in Table 5. The predictive accuracy of the hold-out test was 86.43% for the test subset and 89.35% for the training subset. Regarding the cross-validation test, the highest predictive accuracy, of 73.05%, was obtained using nine folds (i.e., six more than for the logistic regression). The difference in accuracy was marked; the ANN could not determine any trend in the data.

Table 2 The probit statistics

Variable	Coefficient	Standard error	z	*p* value
Opaque muscle	−1.3375	0.535	−2.501	0.012
Poor growth	−1.9242	0.716	−2.688	0.007
Poor appetite	−2.1870	0.941	−2.323	0.020
Dirty gills	3.8308	0.930	4.120	0.000
Empty gut	3.2240	0.413	7.814	0.000
Hepatopancreatic atrophy	5.2931	0.793	6.671	0.000
Tough hepatopancreas	3.7129	0.942	3.941	0.000
Discontinuous gut	4.6700	0.544	8.579	0.000
Soft shell	0.5026	0.320	1.572	0.016
Shrimp age	−0.0557	0.027	−2.050	0.040
Timing of symptom detection	−0.2188	0.035	−6.183	0.000
Fresh smear test result	4.7398	0.552	8.594	0.000
NH_4 level	−2.2574	0.387	−5.840	0.000
Pond area	0.0048	0.001	5.807	0.000
Water pH	−0.0229	0.006	−4.017	0.000
Province	−1.1319	0.240	−4.707	0.000

Table 3 Predictive accuracy of probit regression after eliminating variables

Step	Eliminated variable	Accuracy after elimination (%)
1	Province	72.82
2	Timing of symptom detection	74.40
3	Shrimp age	75.59
4	Fresh smear test result	83.03
5	Poor growth	83.04

Decision tree

The decision tree results are shown in Table 6. The predictive accuracy rates of the hold-out test were 97.57% and 99.10% for the validation and training subsets, respectively. The cross-validation accuracy, of 73.42%, was significantly lower than that of the hold-out test.

Table 4 Classification accuracy of the logistic regression model according to the hold-out test

		Training subset			Validation subset		
		Predicted		Percent correct	Predicted		Percent correct
		0	1		0	1	
Observed	0	201	29	87.39	61	14	81.33
	1	23	319	93.27	12	104	89.66
Overall				90.33			85.50

Table 5 Classification accuracy of the neural network model according to the hold-out test

		Training subset			Validation subset		
		Predicted		Percent correct	Predicted		Percent correct
		0	1		0	1	
Observed	0	204	20	91.07	65	10	86.67
	1	43	305	87.64	16	100	86.20
Overall				89.35			86.43

Table 6 Classification accuracy of the decision tree model according to the hold-out test

		Training subset			Validation subset		
		Predicted		Percent correct	Predicted		Percent correct
		0	1		0	1	
Observed	0	220	4	98.21	72	3	96.00
	1	0	348	100.00	1	115	99.14
Overall				99.10			97.57

K-nearest neighbor

The results of the KNN tests are shown in Table 7. For the hold-out test, the accuracy rates were 91.82% and 87.26% for the training and validation subsets, respectively. The cross-validation test yielded the worst result of all algorithms: 57.80% accuracy using seven folds. As with the decision tree, the accuracy of the KNN algorithm differed notably between the hold-out and cross-validation tests.

Discussion

Logistic regression was more stable than the other algorithms based on the results of both the validation and cross-validation tests; the accuracy rates according to the hold-out test were 90.30% and 85.50% for the training and validation subsets, respectively, and the cross-validation accuracy was similar, at 83.04%.

Notably, the hold-out test showed that all four algorithms were remarkably accurate (≥ 85.00% correct for both the training and testing subsets). The decision tree accuracy rates were 97.57% and 99.10%, respectively. Compared with previous studies using machine learning to predict shrimp disease (Leung and Tran 2000), the accuracy rates of our hold-out tests were higher. However, the hold-out test is somewhat simple, calculating predictive accuracy rapidly even for large datasets. Most algorithms employ 75% of the data for training, and then use the model to predict the outcomes of the validation subset (the remaining 25% of the data). Our dataset was not particularly large (763 individuals with or without AHPND), so it may have been susceptible to bias, i.e., to the tendency to “learn incorrectly” as a result of not considering all available information during dichotomous classification. For the KNN and decision tree methods, the respective accuracy rates were 91.82% and 99.10% (training subset) and 87.26 and 97.57% (validation subset). However, both algorithms performed poorly according to the results of cross-validation tests (accuracy rates of 57.80% and 73.42%, respectively). The results for the KNN and decision tree methods were markedly different. When 15 variables were used, the difference in predictive performance was about 33% and 25% according to the hold-out and cross-validation tests, respectively. The decision tree may have been generally unsuitable, working well only in small parts (i.e., folds) of the test subset. The KNN algorithm focused on the details, rather than trends, of the trained data during the hold-out test. Also, some of the dataset was unsuitable, which reduced its predictive power. Over-fitting may have been a factor: while the hold-out test accuracy rate was the highest among all algorithms, the cross-validation test accuracy rate was the lowest. Neither the KNN nor the decision tree was reliable in the present study.

The ANN algorithm was moderately better, with hold-out test accuracy rates of 89.35% and 86.43% for the training and validation subsets, respectively, and 73.05% in the cross-validation test. Thus, the difference in accuracy between the two tests was approximately 14%. Although this difference was less than that for the KNN and decision tree, it nevertheless showed that the model was poorly trained and failed to recognize trends in the data. Also, an ANN can be considered a “black box” (Leung and Tran 2000); although the weightings of the variables are available, they do not clearly explain the relative contributions of the variables to the prediction. Moreover, the weights (and model accuracy rates) were unstable when the number of hidden layers was

Table 7 Classification accuracy of the K-nearest neighbor model according to the hold-out test

		Training subset			Validation subset		
		Predicted		Percent correct	Predicted		Percent correct
		0	1		0	1	
Observed	0	197	27	87.94	63	12	84.00
	1	15	333	95.69	11	105	90.51
Overall				91.82			87.26

varied. The input weights varied considerably as the models changed; it was impossible to determine the extent to which each variable affected the overall prediction. Furthermore, ANNs are more time-consuming than other algorithms, and require more computational resources because of the need to generate a weight for each variable input into every neuron in every layer.

In contrast, logistic regression performed well according to both the hold-out and cross-validation tests. Hold-out tests using 572 training samples and 191 test samples yielded accuracy rates of 90.33% and 85.50%; the cross-validation test accuracy was 83.04%. The dichotomous nature of the prediction did not require extensive computational resources; the results could be interpreted without the need to scale the inputs. The small difference in accuracy rate between the two tests, of about 4%, showed that the logistic regression model performed well. Moreover, in logistic regression models, the contribution of each independent variable is clear (Leung and Tran 2000).

Many researchers have used ANNs to make dichotomous predictions. However, our ANN was unstable, where the coefficients of the input variable changed when the number of hidden layers was varied. Meanwhile, the KNN algorithm and decision tree showed over-fitting, where the results of the hold-out and cross-validation tests differed greatly.

All input variables in the model were removed one-by-one, and then restored, to evaluate their importance. The model including 15 selected independent variables was faster and more reliable than the original 31-variable model. The literature since 2009 (e.g., Crane 2019) suggests that water temperature and salinity, shrimp mortality rate, and seedstock source affect AHPND development. In this study, we identified additional parameters via backward selection, including other disease signs (discontinuous gut, a soft shell, and hepatopancreas status). We backward-selected 15 variables and evaluated the utility of logistic regression, KNN, ANN, and decision tree methods for predicting AHPND. Figure 2 shows the contributions of those 15 variables to the prediction of acute hepatopancreatic necrosis disease. Logistic regression was more stable than the other methods according to both the hold-out and cross-validation tests, and was superior in identifying important predictive variables such as hepatopancreatic atrophy, toughness, pallor, and a high temperature. On the contrary, the NH_3 level, dirty gills, poor appetite, and pond area were less important.

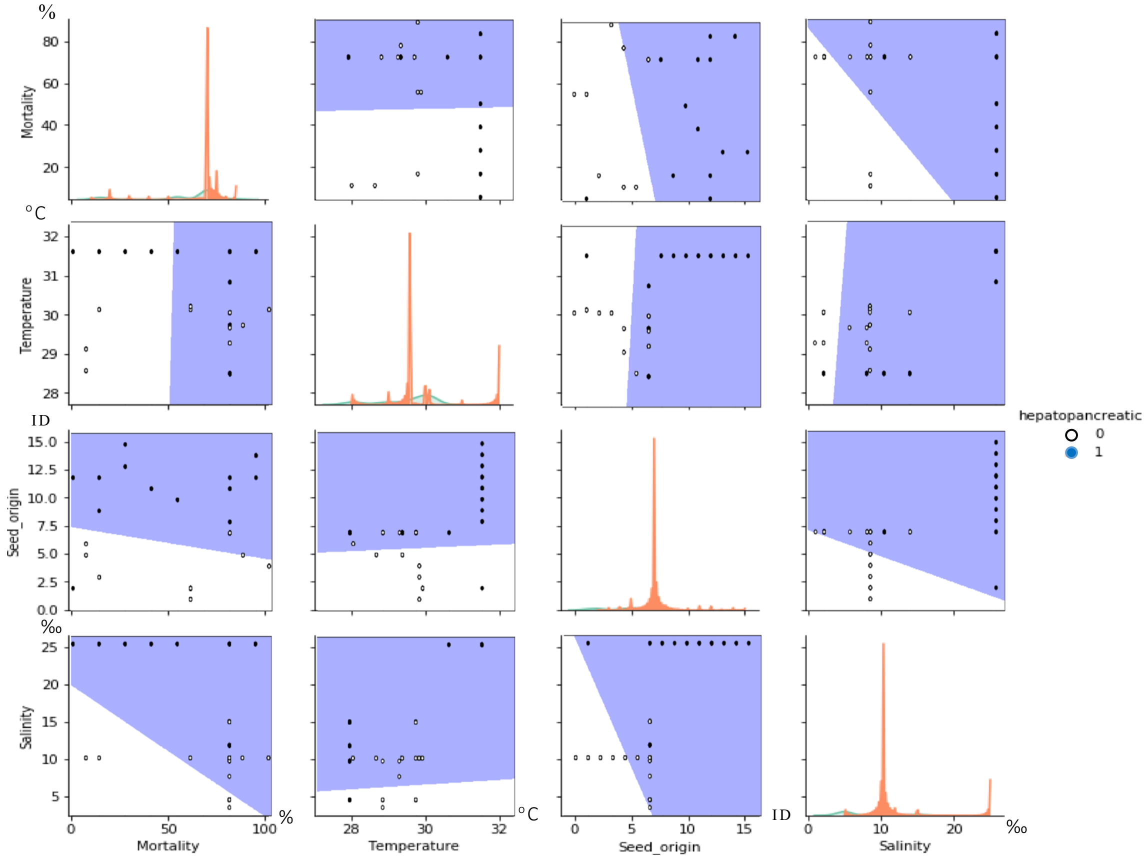


Fig. 2 Relationships among daily shrimp mortality, water temperature and salinity, and seedstock origin. The values of several samples from the same ponds were identical

The appearance of symptoms is a useful early predictor of AHPND mortality. As mentioned above, AHPND is also known as EMS. A warmer water temperature and higher salinity increased the risk of AHPND. Seedstock origin was also important; some shrimp farms were infected because their seedstocks came from AHPND-affected hatcheries. Figure 3 shows the correlations among the computer-predicted and manually selected factors, including seedstock origin, water temperature and salinity, and shrimp mortality: the manually selected variables showed good reliability for predicting AHPND development. The panels in the figure show the correlations between pairs of variables. Although all variables were important, the logistic procedure showed that some were more influential through the pairing of variables. For example, a high mortality rate was more important than a high water temperature. However, seedstock origin and salinity were more important in the mortality/seedstock origin and mortality/salinity pairs. In the temperature/seedstock origin and temperature/salinity pairs, temperature was not prioritized, while in the seedstock origin/salinity pair, seedstock origin was prioritized. Thus, high-quality seedstock reduced AHPND development regardless of water temperature. Hepatopancreatic atrophy had a remarkable effect on the predictions, as expected: shrimp with AHPND characteristically exhibit atrophy exceeding 50% (Leobert et al. 2015). Atrophy was evident in 53% of the AHPND-affected shrimp in our dataset (Table 8). We used machine learning to predict disease outbreaks on shrimp farms based on symptoms, visceral status, environmental factors, and general management practices. The predictive application was built for fish farmers, as shown in Fig. 4. Low (compared to high) salinity reduced the disease risk. Our dataset for generating predictive models is valid only up to the time of the present study. The interrelationships among variables were explored during the data collection period. For example, seedstock origin was not independently predictive of AHPND, instead exerting an effect in combination with 30 other environmental, symptomatic, and visceral parameters. AHPND status was previously shown to be affected by the poor-quality seedstocks used in some hatcheries, but high-quality seedstocks are now available.

Table 8 Numbers of samples exhibiting hepatopancreatic atrophy

	Atrophy	No atrophy
AHPND	248	216
No AHPND	3	296

AHPND acute hepatopancreatic necrosis disease

Hepatopancreatic disease has many signs, including hepatopancreas swelling, atrophy, pallor, and a discontinuous gut. These symptoms are also appeared in many other disease conditions, such as white feces, low growth, and intestinal parasites. The model proposed here can predict AHPND with high accuracy without using an empirical method. Therefore, the threat of AHPND can be considered before making an actual fishpond using the proposed application and the proposed model helps shrimp farmers.

Application of this technology to the fishery industry in the Mekong Delta has been considered. Many large shrimp farms have recently been converted into intensive or semi-intensive farms to improve aquaculture farm management by improving product quality and saving costs. In intensive and semi-intensive farms, data on the water environment can be easily collected using Internet of Things technology. The proposed model can be applied not only in relation to AHPND but also to other diseases, if we have enough data. Therefore, the proposed model is considered especially

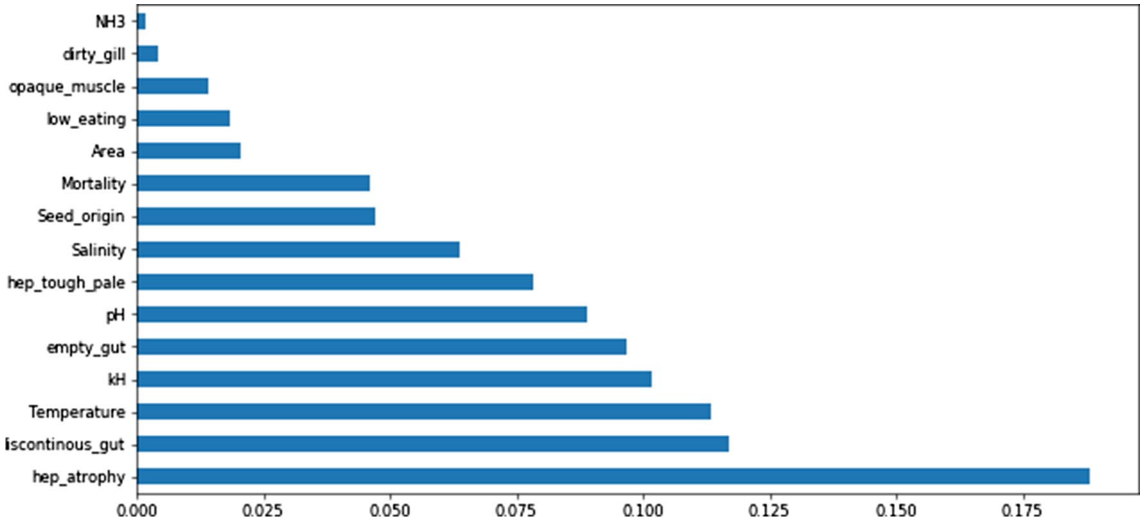
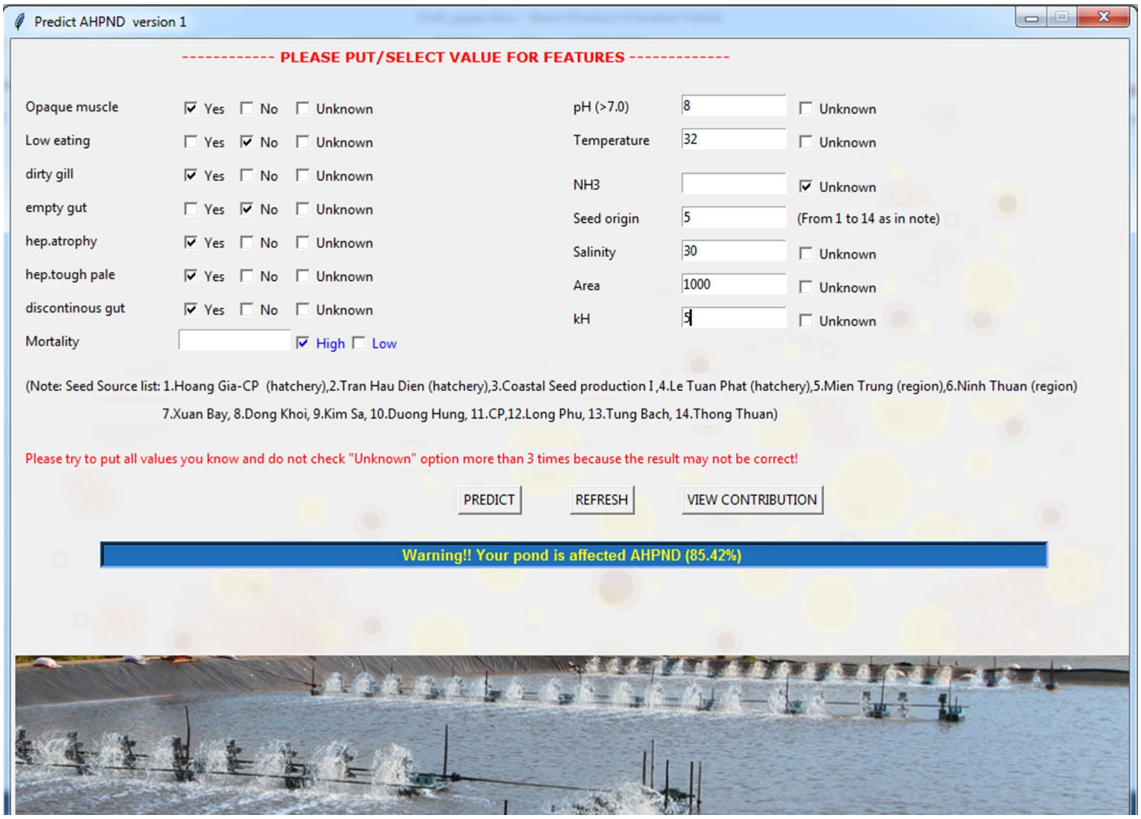


Fig. 3 The contributions of the variables to the prediction of acute hepatopancreatic necrosis disease



------------ PLEASE PUT/SELECT VALUE FOR FEATURES -------------

Opaque muscle ☑ Yes □ No □ Unknown
Low eating □ Yes ☑ No □ Unknown
dirty gill ☑ Yes □ No □ Unknown
empty gut □ Yes ☑ No □ Unknown
hep.atrophy ☑ Yes □ No □ Unknown
hep.tough pale ☑ Yes □ No □ Unknown
discontinous gut ☑ Yes □ No □ Unknown
Mortality ☑ High □ Low

pH (>7.0) 8 □ Unknown
Temperature 32 □ Unknown
NH3 ☑ Unknown
Seed origin 5 (From 1 to 14 as in note)
Salinity 30 □ Unknown
Area 1000 □ Unknown
kH 5 □ Unknown

(Note: Seed Source list: 1.Hoang Gia-CP (hatchery),2.Tran Hau Dien (hatchery),3.Coastal Seed production I ,4.Le Tuan Phat (hatchery),5.Mien Trung (region),6.Ninh Thuan (region) 7.Xuan Bay, 8.Dong Khoi, 9.Kim Sa, 10.Duong Hung, 11.CP,12.Long Phu, 13.Tung Bach, 14.Thong Thuan)

Please try to put all values you know and do not check "Unknown" option more than 3 times because the result may not be correct!

Warning!! Your pond is affected AHPND (85.42%)

Fig. 4 The graphic user interface of the predictive application

suitable for intensive farms, and the risk of all potential diseases, not just AHPND, can be predicted.

In this study, we targeted AHPND on the east coast of the Mekong Delta. AHPND also occurs in other regions in Vietnam; however, the proposed model is not considered suitable for other regions because of the different environmental conditions. In the future, we should collect data for all Vietnamese aquaculture farms to develop a model to predict AHPND in various regions. Finally, the proposed model can help to improve the efficiency of Vietnamese aquaculture and achieve smart aquaculture fisheries.

Acknowledgements This study was funded in part by the Can Tho University Improvement Project VN14-P6, supported by a Japanese Official Development Assistance (ODA) loan. Data collection was partially funded by the technical cooperation project "Building capacity for Can Tho University to be an excellent institution of education, scientific research and technology transfer" of the Japan International Cooperation Agency (JICA). We are grateful to the anonymous reviewers who made helpful comments.

References

ADB–NACA (1998) Aquaculture sustainability and the environment, a report on a regional study and workshop. Asian Development Bank and Network of Aquaculture Centers in the Asia-Pacific, Bangkok

Boonyawiwat V, Patanasatienkul T, Kasornchandra J, Poolkhet C, Yaemkasem S, Hammell L, Davidson J (2017) Impact of farm management on expression of early mortality syndrome/acute hepatopancreatic necrosis disease (EMS/AHPND) in penaeid shrimp farms in Thailand. J Fish Dis. https://doi.org/10.1111/jfd.12545

Boonyawiwat V, Nga NTV, Bondadreantaso MG (2018) Risk factors associated with acute hepatopancreatic necrosis disease (AHPND) outbreak in the Mekong Delta, Viet Nam. Asian Fish Sci 31:226–241

Boyd C, Truong P (2019) Environmental factors and acute hepatopancreatic necrosis disease (AHPND) in shrimp ponds in Viet Nam: practices for reducing risks. Asian Fish Sci 31:121–136

COFI (2019) Fishery and aquaculture country profiles: the Socialist Republic of Viet Nam. FAO, Rome

Cournapeau D (2007) Scikit-learn: machine learning in Python. JMLR 12:2825–2830

Crane M (2019) Hepatopancreatic necrosis disease. In: OIE - manual of diagnostic tests for aquatic animals. World organisation for Animal Heath, Paris

Dang TL, Pham AT, Phan TV (2018) Acute Hepatopancreatic Necrosis Disease (AHPND) in Vietnam. Asian Fish Sci 31:274–282

Dhar AK, Piamsomboon P, Aranguren Caro LF, Kanrar S, Adami R Jr, Juan YS (2019) First report of acute hepatopancreatic necrosis disease (AHPND) occurring in the USA. Dis Aquat Organ. https://doi.org/10.3354/dao03330

Harston CT (1990) The neurological basis for neural computations. In: Maren AJ, Harston C, Pap RZ (eds) Handbook of neural computing applications. Academic Press, San Diego, pp 29–44

Hoffman GL (1976) Fish diseases and parasites in relation to the environment. Fish Pathol 10(2):123–128

Lopes JNS, Gonçalves ANA, Fujimoto RY, Carvalho JCC (2011) Diagnosis of fish diseases using artificial neural networks. Int J Comput Sci 8(6):68–74

Molnar C (2019) Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Lulu, Germany

Peña LD, Cabillon NAR, Catedral DD, Amar EC, Usero RC, Monotilla WD, Saloma CP (2015) Acute hepatopancreatic necrosis disease (AHPND) outbreaks in Penaeus vannamei and *P. monodon* cultured in the Philippines. Dis Aquat Organ 116(3):251–254

Ping SL, Liem TT (2000) Predicting shrimp disease occurrence: artificial neural networks vs logistic regression. Aquaculture 187(49):35–49. https://doi.org/10.1016/S0044-8486(00)00300-8

Rahman A, Tasnim S (2014) Application of machine learning techniques in aquaculture. Int J Comput Trends Technol. https://doi.org/10.14445/22312803/IJCTT-V10P137

Shinn AP, Pratoomyot J, Griffiths D, Trong TQ, Vu NT, Jiravanichpaisal P, Briggs M (2018) Asian shrimp production and the economic costs of disease. Asian Fish Sci 31:29–58

Tu JV (1996) Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol 49(11):1225–1231

Venkateswara Rao P (2017) Computer aided shrimp disease diagnosis in aquaculture. IJRASET. https://doi.org/10.22214/ijraset.2017.2079

Zhang Z (2016) Introduction to machine learning: k-nearest neighbors. Ann Transl Med 4(11):218. https://doi.org/10.21037/atm.2016.03.37

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 Genom. https://doi.org/10.1186/s12864-018-4728-4

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.