



Prediction of white spot disease susceptibility in shrimps using decision trees based machine learning models

Tran Thi Tuyen¹ · Nadhir Al-Ansari² · Dam Duc Nguyen³ · Hai Minh Le⁴ · Thi Nga Quynh Phan¹ · Indra Prakash⁵ · Romulus Costache^{6,7,8,9} · Binh Thai Pham³

Received: 23 January 2023 / Accepted: 26 October 2023 / Published online: 9 December 2023
© The Author(s) 2023

Abstract

Recently, the spread of white spot disease in shrimps has a major impact on the aquaculture activity worldwide affecting the economy of the countries, especially South-East Asian countries like Vietnam. This deadly disease in shrimps is caused by the White Spot Syndrome Virus (WSSV). Researchers are trying to understand the spread and control of this disease by doing field and laboratory studies considering effect of environmental conditions on shrimps affected by WSSV. Generally, they have not considered spatial factors in their study. Therefore, in the present study, we have used spatial (distances to roads and factories) as well as physio-chemical factors of water: Chemical Oxygen Demand (COD), Dissolved Oxygen (DO), Salinity, NO₃, P₃O₄ and pH, for developing WSSV susceptibility maps of the area using Decision Tree (DT)-based Machine Learning (ML) models namely Random Tree (RT), Extra Tree (ET), and J48. Model's performance was evaluated using standard statistical measures including Area Under the Curve (AUC). The results indicated that ET model has the highest accuracy (AUC: 0.713) in predicting disease susceptibility in comparison to other two models (RT: 0.701 and J48: 0.641). The WSSV susceptibility maps developed by the ML technique, using DT (ET) method, will help decision makers in better planning and control of spatial spread of WSSV disease in shrimps.

Keywords White spot · Random tree · Extra tree · J48 · Disease · Vietnam

✉ Nadhir Al-Ansari
nadhir.alansari@ltu.se

✉ Romulus Costache
romuluscostache2000@yahoo.com

✉ Binh Thai Pham
binhpt@utt.edu.vn

Tran Thi Tuyen
tuyentt@vinhuni.edu.vn

Dam Duc Nguyen
damnd@utt.edu.vn

Hai Minh Le
hailm@vinhuni.edu.vn

Thi Nga Quynh Phan
quynhnga_dhkhtn86@yahoo.com

Indra Prakash
indra52prakash@gmail.com

³ University of Transport Technology, 54 Trieu Khuc, Thanh Xuan, Ha Noi, Vietnam

⁴ Department of Fisheries and Livestock, School of Agriculture and Natural Resources, Vinh University, 182 Le Duan, Vinh City, Vietnam

⁵ DDG (R), Geological Survey of India, Gandhinagar 382010, India

⁶ Research Institute of the University of Bucharest, 90-92 Sos. Panduri, 5th District, 050663 Bucharest, Romania

⁷ National Institute of Hydrology and Water Management, București-Ploiești Road, 97E, 1st District, 013686 Bucharest, Romania

⁸ Department of Civil Engineering, Transilvania University of Brasov, 5, Turnului Str, 500152 Brasov, Romania

⁹ Danube Delta National Institute for Research and Development, 165 Babadag Street, 820112 Tulcea, Romania

¹ Department of Geography, Vinh University, 182 Le Duan, Vinh City, Vietnam

² Department of Civil, Environmental and Natural Resources Engineering, Lulea University of Technology, 971 87 Lulea, Sweden

Introduction

Shrimp is a valuable natural resource that is widely used in the food industry around the world (Li et al. 2016). The recent enhancement of the fatal spread of the White Spot Syndrome Disease Virus (WSSV) in the shrimps all over the world has affected economy of the coastal regions and countries that rely on sea food consumption and export. First time this disease was reported in the year 1992 in the cultured shrimps (*Penaeus japonicus*) in China (Tuyen et al. 2014). There are so many factors responsible for the spread of WSSV disease in shrimps. Quality of the seawater (pH, salinity and other chemical parameters) has a significant effect on shrimp production. For example, the accumulation of nitrogenous waste and ammonia is a very important toxic limiting factor for shrimp because these substances severely pollute the seas, which are the main habitat of shrimp (Chaijarasphong et al. 2019). In other words, environmental parameters and conditions can have a major impact on the physiological characteristics of organisms. The physical parameters like temperature, hypoxia, salinity and ammonia can severely disrupt the physiological processes of shrimp as well as their immune system (Hasan and Haque 2020). A series of statistics provided by global shrimp production indicate that more than 4.5 million tons of shrimp production are minced to the shrimp farming industry (Schleder et al. 2020). The ability of aquaculture to meet many needs and to address the pressures of increasing human population can be very effective. Unfortunately, the aquaculture system is ignored in the community strategy (Caipang et al. 2008). Modified ammonia forms are highly toxic to shrimp due to their ability to diffuse into the cell membranes (Lu et al. 2016). In aqueous surroundings, total ammonia nitrogen is present in two forms: unionized form (ammonia, NH_3) and ionized form (ammonium, NH_4^+). Shrimp fosterage is faced with several diseases around the world including WSSV (Verbruggen et al. 2016). It is worth mentioning that WSSV has affected all great shrimp generating countries, especially south eastern countries like Vietnam. The economic loss caused by this disease is estimated between \$ 8 billion and \$ 15 billion. Worldwide (Millard et al. 2020), the average annual cost used to mitigate this disease is around of \$ 1 billion (Millard et al. 2020). These shrimp diseases are often caused by the interactions between the host environment and the outside environment. Recently, researchers have found that 70% of shrimp are diseased after being caught and harvested (Millard et al. 2020).

The symptoms of WSSV include shrimp natural discoloration, decrease in shrimp nutrition, and white spots on the shrimp (Wang et al. 1999; Nunan and Lightner 1997). However, it should be noted that even in some shrimps,

the occurrence of this disease is asymptomatic and eventually leads to death (Sun et al. 2013; Meng et al. 2010). On the other hand, some recent research has shown that feeding shrimp diets, which should contain a mixture of two seaweed, affects the gut microbiota and can reduce mortality due to the spread of white spot disease, which can also improve the harmful effects of environmental stressors (Zacarias et al. 2021). The study of WSSV in four different groups (AHPND-Vibrio parahaemolyticus infection) showed that there were significant differences in the histopathology of surviving shrimp hepatopancreas among the groups. Shrimps in the co-infection group showed signs of normal histopathology in the hepatopancreas (Han et al. 2019). On the other hand, the toxicity of ammonia that causes disease in shrimp is influenced by the water pH, salinity and temperature as significantly affect immune system of shrimps (Kathyayani et al. 2019). Another study found that other factors (dissolved oxygen concentration, nitrogen, partial pressure of carbon dioxide and pH) related to water quality played a key role in determining the severity of White Spot Disease (Millard et al. 2020).

In this study, we have developed disease susceptibility map considering spatial parameters (distances to roads and factories) and physio-chemical factors (Chemical Oxygen Demand (COD), Dissolved Oxygen (DO), Salinity, NO_3 , P_3O_4 , pH) of water of shrimp aquaculture sites. Three Decision Tree (DT) based ML models namely Random Tree (RT), Extra Tree (ET), and J48 have been used for the data analysis and generation of diseased susceptibility maps. These DT models have been as they have advantage of easy, quick and efficient interpretation of the results by the decision makers (Chien and Chen 2008). The shrimp aquaculture area of Quynh Luu district, Nghe An province, Vietnam, which is severely affected by WSSV has been selected as the study area. First time spatial parameters have been considered in this study for the prediction of White Spot disease susceptibility using ML methods. Standard statistical measures including Area Under the Curve (AUC) were used for the evaluation models performance. Weka and ArcGIS software were used for the data analysis and models development.

Experimental

Decision tree (DT) based methods

Random tree (RT)

RT search model is a ML model, that programs a random track that can effectively find non-convex locations. On the other hand, it should be noted that according to this model,

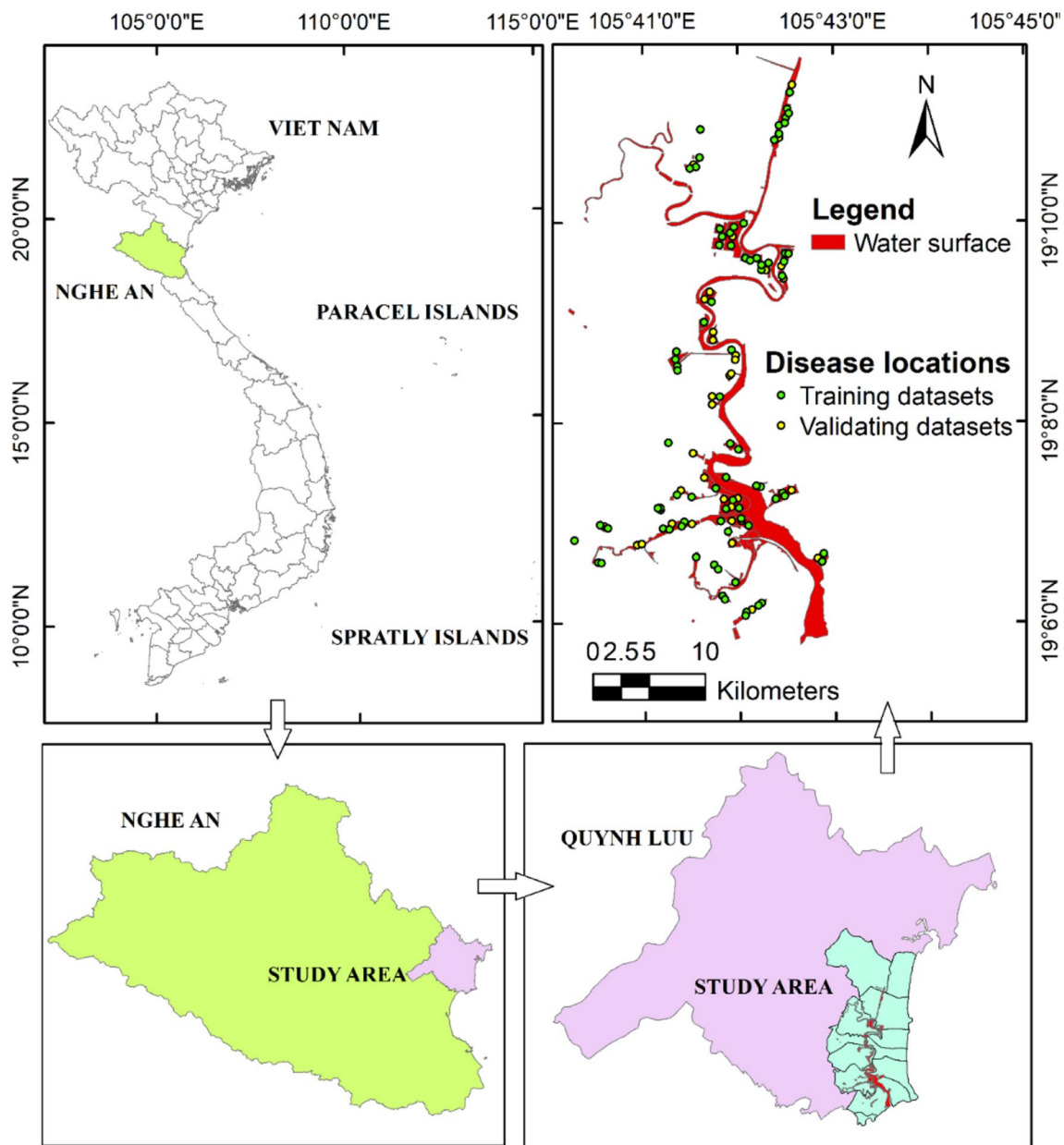


Fig. 1 Location map of the study area showing spread of WSSV disease in Shrimps

continuous computation requires control decisions to navigate the system from the prime x_{init} situation to the x_{goal} target situation. Also, usually this model is characterized by a high variance (Ajayram et al. 2021). The workflow associated with the RT classifier assumes the involvement of input feature vectors and their classification for each tree in the ensemble forest (Díaz et al. 2020). The class label for which the majority votes is assigned will represent the output of the model (Zhang et al. 2020). Different training sets are used

to train the trees having the same parameters. A bootstrap process is followed in order to generate the training sets. In this case, a randomly generated subset for each variable is utilized to find the best split at each node where a new subset will be created with a specific size (Keivani and Sinha 2021). A very important characteristic of RT is given by the fact that the classification error is estimated using the out of bag (OBB) data (Kamiński and Prałat 2019; Rustam et al. 2020; Zito and Cooper 2006).

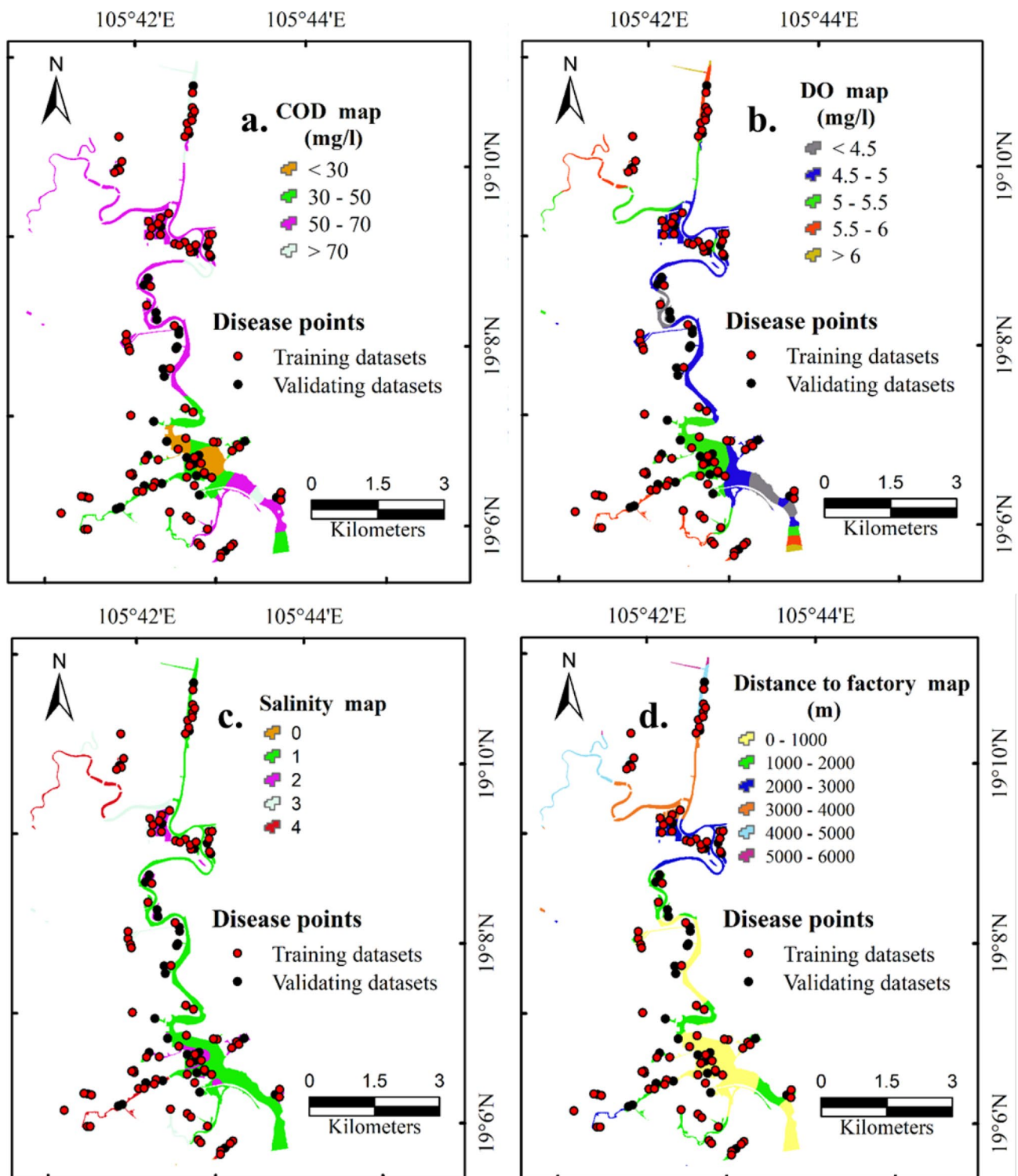


Fig. 2 Thematic maps of the study area (a COD map; b. DO map; c. Salinity map; d. Distance to factory map, e. NO_3 map; f. P_3O_4 map; g. PH map; h. Distance to road map)

Extra tree (ET)

The ET model, or the "Extremely randomized trees" model, is a highly advanced random forest model. This is a very

new method that is a proper subset of ML models. ET is a group of learning methods that use decision tree predictions to increase the accuracy efficiency and decrease computational elaborations (Geurts et al. 2006). In this model, a

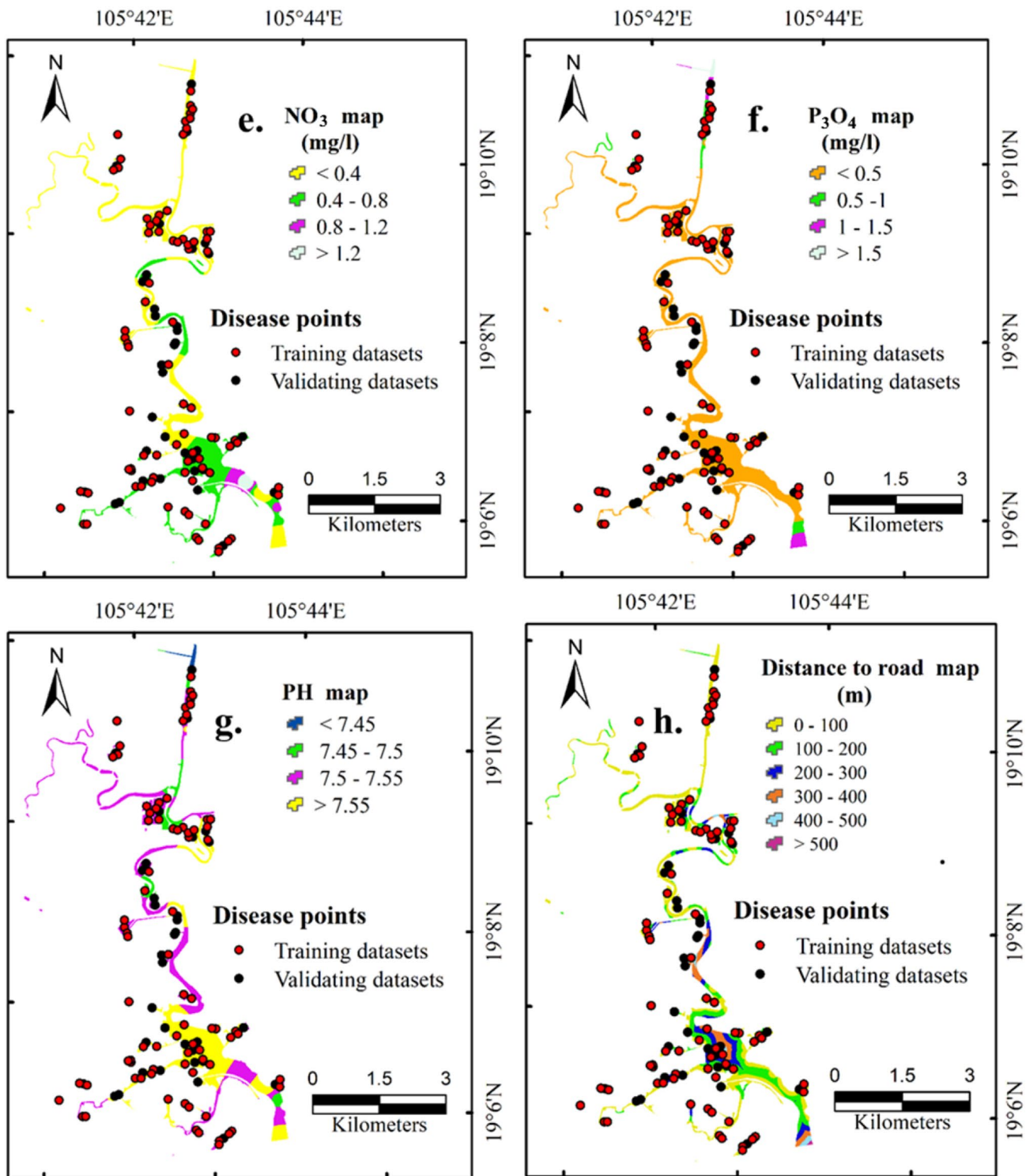


Fig. 2 (continued)

set of trees is generated randomly, and then, the forecasted value of which is added in certain ways. For example, the arithmetic mean is used for the purpose of regression by a maximum judgment in the categorization. This is one of

the principal and basic differences between the ET model and other tree-based grouped models. In the ET model, the splitting procedure in the nodes, is done using a completely random selection of cutting points, and therefore, the trees

will grow using all the training instances against the use of the bootstrap duplicate (John et al. 2016). In other words, the extra tree follows the origin of random forest. It also applies an accidental subset of features to teach every basis assessment (Okoro et al. 2021). ET is divided into two methods to solve regression issues. In the first method, the number of random divisions, i.e., the random selection of both input parameters and cut-off points in each node, is denoted by K . The second method provides the minimum instances size for node division, and is denoted by n_{\min} . The process of tree progression in the extra tree group is continued by determining the value of K in each node until the process of reaching the leaves in which all subsets have a net output, and also the number of learning samples must be measured in n_{\min} . ET has the ability to reduce variance through a clear randomization of input parameters and cut points along with averaging groups (Ahmad et al. 2018). However, the use of all the basic learning examples will help to minimize the bias or the variance. The solution of understanding the number of trees produced in the ET model can control the extent to which the variance of group model association can be reduced (Hou et al. 2020).

J48

J48 model is one of the most popular ML models that have been widely used in recent decades (Pham et al. 2017b). The

method is based on a tree hierarchy which helps to build classification trees that display a naive tree frame. In this simple structure, the non-terminative nodes represent the properties, while the terminative nodes represent the decision results (Sridhar Raj and Nandhini 2018). J48 model represents an example of basic classification because it can support the next class. As a result, it produces a type of hybrid classification that uses a test sample for validation. Predicted results should be compared using training and testing samples. In this section, the results of each classification will be different (Chen et al. 2020). J48 classifier creates a decision tree that consists of data values and instructional samples and also generates a new instance. The construction of the J48 model consists of 5 main steps: (1) Preparation of input layers represented by the independent and dependent variables, (2) Creating the first internal node in the tree structure or the root node, (3) Based on the root node values, the training data should be subdivided into many training samples to create the sub-nodes, (4) Evaluation of the incremental rate under the nodes, and (5) Selecting the most effective agents is an ongoing process (Hong et al. 2018; Madhusudana et al. 2018).

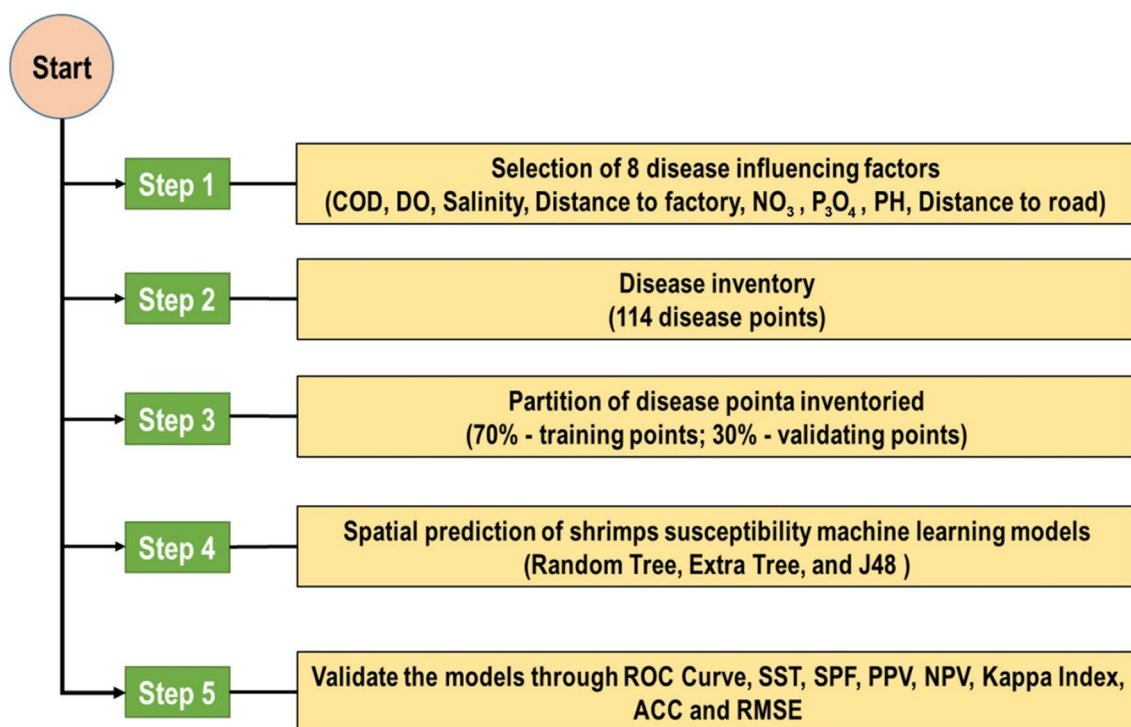


Fig. 3 Methodological flowchart used for Shrimp WSSV disease susceptibility prediction

Table 1 Accuracy analysis of the models

No	Parameters	Training dataset			Testing dataset		
		RT	ET	J48	RT	ET	J48
1	PPV (%)	91.25	91.25	66.25	67.65	67.65	70.59
2	NPV (%)	78.75	78.75	65.00	58.82	58.82	55.88
3	SST (%)	81.11	81.11	65.43	62.16	62.16	61.54
4	SPF (%)	90.00	90.00	65.82	64.52	64.52	65.52
5	ACC (%)	85.00	85.00	65.63	63.24	63.24	63.24

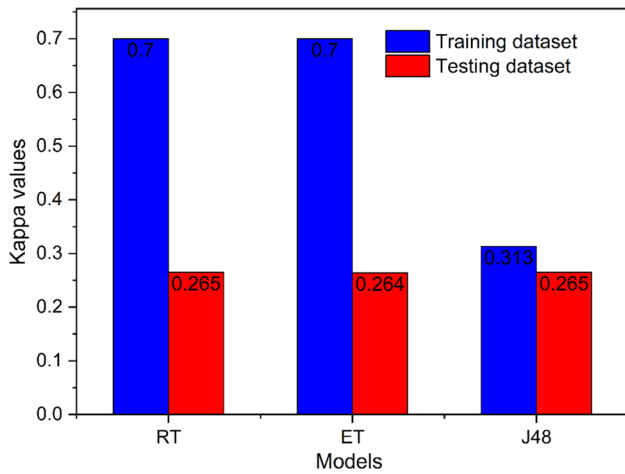


Fig. 4 Kappa analysis of the models using training dataset and testing dataset

Evaluation methods of model’s performance

Statistical indicators

The data is usually split randomly into a 70:30 ratio, where 70% is used for training and 30% is used for validation of the models (Nguyen et al. 2021). The training dataset (70%) evaluates how well the models fit the data, while the testing dataset (30%) evaluates the predictive capability of the models. In the present study, in order to evaluate the accuracy of DT-based ML Models (RT, ET, and J48) the following statistical metrics were used: Positive Predictive Value (PPV), Negative Predictive Value (NPV), Root Mean Square Error (RSME), Accuracy (ACC), Sensitivity (SST), Specificity (SPF), and Kappa (K) were used. Out of these, PPV and NPV are the number of pixels classified as "Shrimp’s White Spot Disease" and "non- Shrimp’s White Spot Disease" susceptibility. The ratio of Shrimp’s WSSV pixels is illustrated by the SST and ratio of non-Shrimp’s WSSV pixels is displays by the SPF. K index is used to check and analyze the accuracy of models. The value of K ranges between 0 and 1 (Tangirala 2020). The closer K is to the number 1, the higher the accuracy of the model. The ACC value of the true prediction rate indicates the whole number of forecasts

for Shrimp’s White Spot Disease. Finally, the RMSE criterion shows the difference between the estimated dataset and the observed dataset. Values close to zero indicate the high accuracy of the model (Tangirala 2020; Pham et al. 2017a; Van Phong et al. 2020). The criterion equations are described below:

$$SST = TP_{\text{true positive}} / TP + FN_{\text{false negative}} \tag{1}$$

$$SPF = TN_{\text{true negative}} / TN_{\text{true negative}} + FP_{\text{false positive}} \tag{2}$$

$$PPV = TP_{\text{true positive}} / FP_{\text{false positive}} + TP_{\text{true positive}} \tag{3}$$

$$NPV = TN_{\text{true negative}} / FN + TN_{\text{true negative}} \tag{4}$$

$$Kappa = P_p - P_{\text{exp}} / 1 - P_{\text{exp}} \tag{5}$$

$$ACC = TP_{\text{true positive}} + TN_{\text{true negative}} / TP_{\text{true positive}} + TN_{\text{true negative}} + FP_{\text{false positive}} + FN_{\text{false negative}} \tag{6}$$

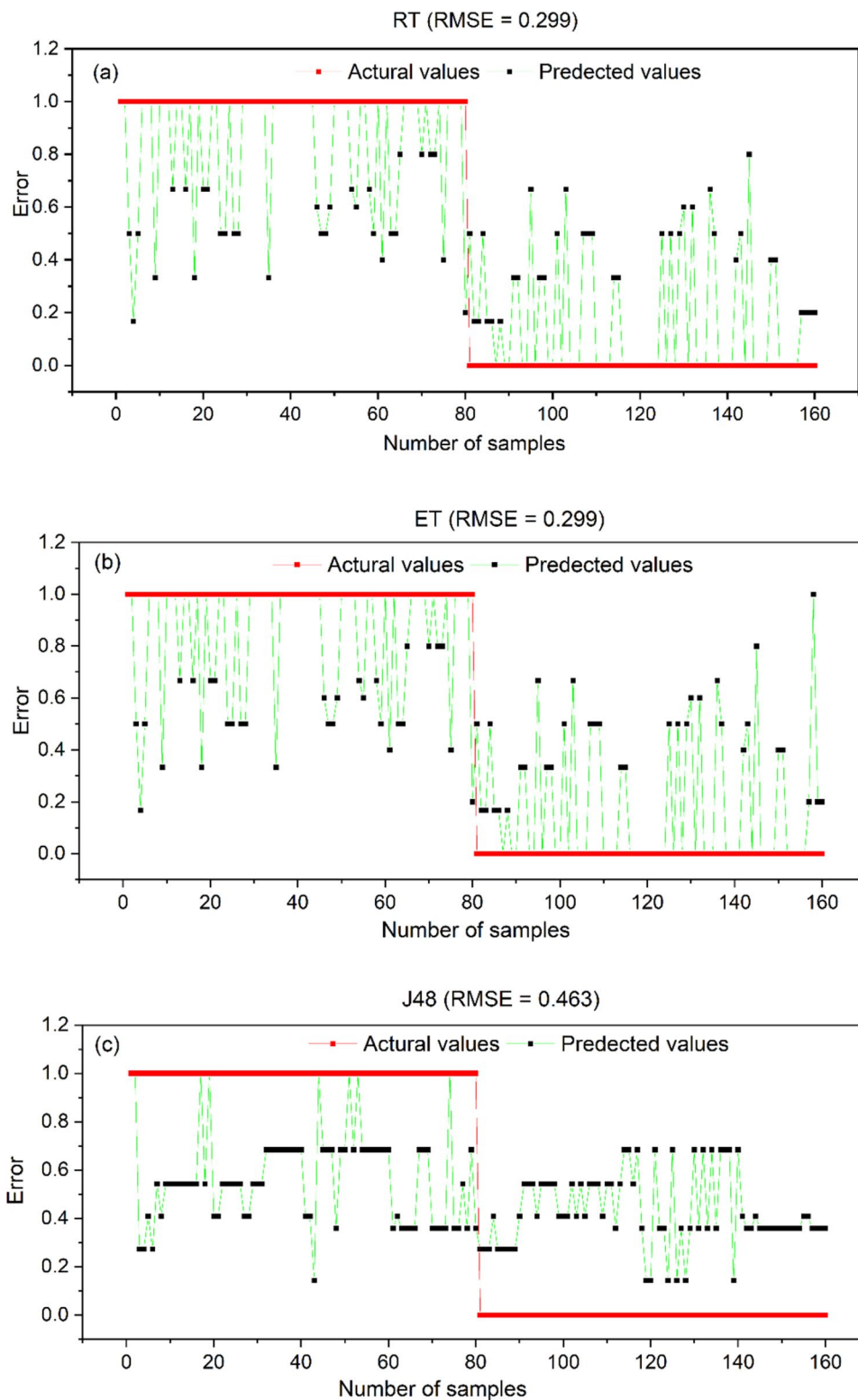
$$RMSE = \sqrt{1/N \sum (X_{\text{predicted}} - X_{\text{actual}})^2} \tag{7}$$

The components of the above formulas are such that FP and FN are the number of pixels that are incorrectly classified as Shrimp’s WSSV and non-Shrimp’s WSSV. P_p is the number of pixels that are correctly classified for Shrimp’s WSSV or non-Shrimp’s WSSV. Predictable adaptations are determined by P_{exp} . Finally, $X_{\text{predicted}}$ and X_{actual} are the forecasted and actual numbers in the training instances or test instances of the models, and N is the whole number of instances in the training instances or test instances.

Receiver operating characteristic (ROC) curve

ROC curve was also used in order to evaluate the model’s performances. The ROC curve is a graph of the balance between the negative and positive error rates for each possible number of slices (Depina et al. 2020). The area under the ROC curve (AUC) indicates the predicted value of the system by describing its ability to accurately estimate the

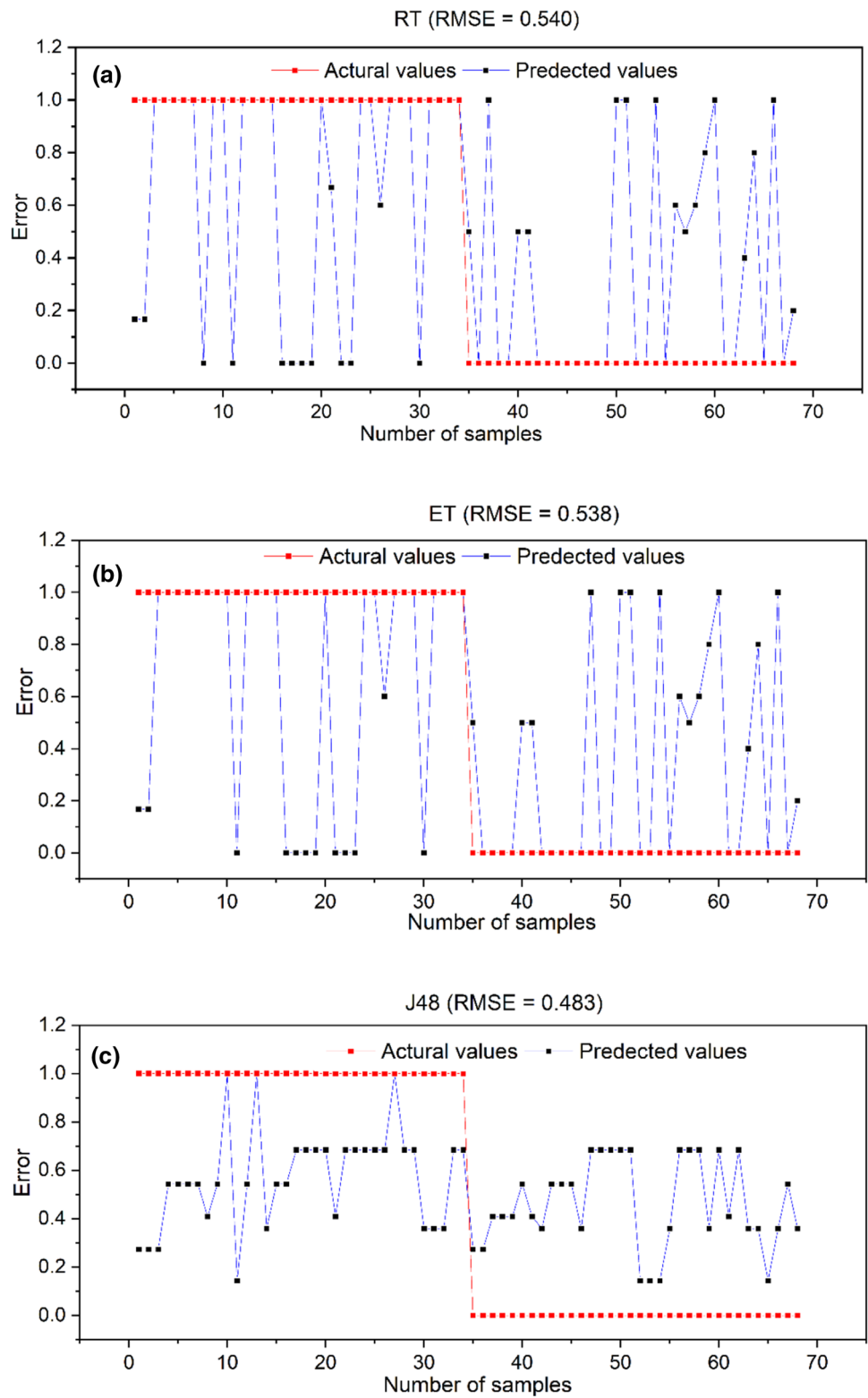
Fig. 5 RMSE analysis of the models using training dataset (a RT; b ET; and c J48)



occurrence and non-occurrence of disease. Its value ranges from 0.5 to 1 (Tien Bui et al. 2019). The closer the sub-curved surface is to 1, the better the accuracy of the zoning map. Classification of the amount of under/sub-curved area is classified as excellent if AUC is between 0.9 and 1, very

good if AUC is between 0.8–0.9, good if AUC is between 0.7 and 0.8, medium if AUC is between 0.6 and 0.7 and poor if AUC is between 0.5 and 0.6 (Tripepi et al. 2009). The equation and numerical value of the area under the curve

Fig. 6 RMSE analysis of the models using validating dataset (a RT; b ET; and c J48)



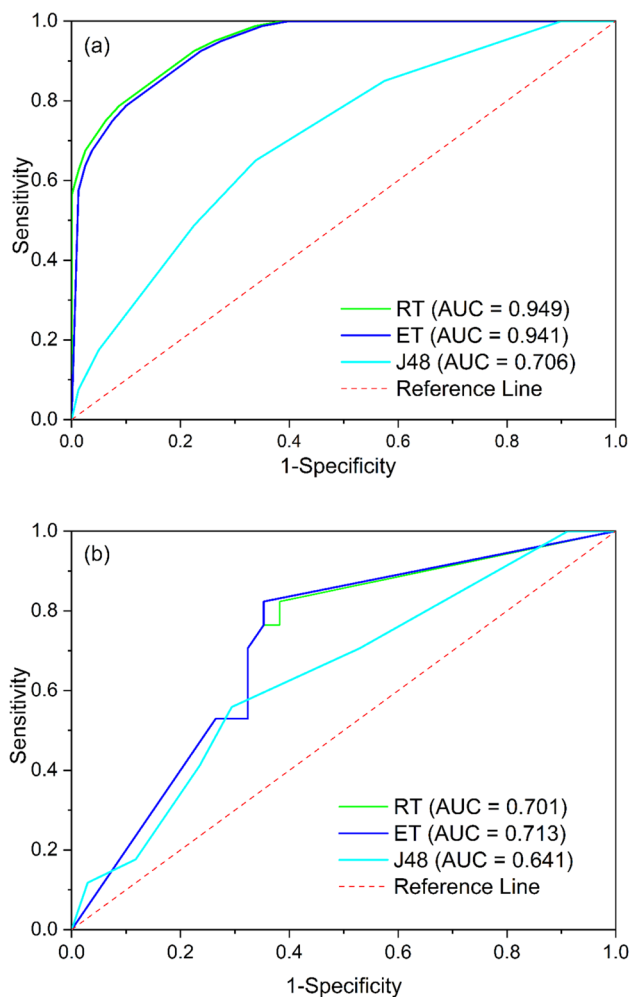


Fig. 7 AUC analysis of the models using training dataset (a) and testing dataset (b)

(AUC) is obtained from the following formula (Dou et al. 2020; Myung et al. 1998):

$$\text{AUC} = \frac{\sum \text{TP}_{\text{true positive}} + \sum \text{TN}_{\text{true negative}}}{P + N} \quad (8)$$

where TP and TN are presented the value of pixels categorized truly as disease and non-disease, P and N are the total number of disease and non-disease, respectively.

Description of study area

The study area of shrimp farming is located in the north-central coast region of Vietnam in the Quynh Luu district of Nghe An province, covering approximately 445.1 km² area and 307,000 population (Fig. 1). Topography of the district is divided from West to East into hills, plains and coastal areas. The territory has many natural lakes and 9.5 km long coastline passing through nine communes and two estuaries (Quen and Thoi canals).

The dominant vegetation in the hilly and mountainous area is represented by planted forests, fruit trees, delta crops, rice and vegetables. The coastal areas are populated with shrimp ponds and mangroves. The climate is a tropical monsoon that is divided into two main seasons: warm summers and cold winters. The average temperature and average annual cumulative precipitation of area are 25°C and 1600 mm, respectively. The warm season starts in May and ends in October, with the average temperature is 30°C. The highest temperature recorded in July was 40°C. The cold season starts from November and ends in April due to the influence of the Northeast Monsoon. Monsoon humid climate and a long coastline with brackish water at the river mouth are considered favorable geographical conditions for the aquaculture farming. The aquaculture in Nghe An province includes shrimp, shellfish, crab, and fish in about 2555 ha surface area. Quynh Luu is the largest area of intensive shrimp farming in this province with 465 ha, raised into 2–3 crops per year, yield 3.5–4 tons per ha. This economic sector accounts for 67% of the workforce and contributes with 72% for the region's GDP.

Data used

Inventory of white spot syndrome virus (WSSV) affected Shrimps

For the development of WSSV infection in shrimp's inventory, disease statistics include secondary and primary data. Shrimp disease statistics data, from 2010 to 2020, was obtained from the Department of Agriculture of Quynh Luu District and the Fisheries Department of Nghe An province. In the study area, the shrimp ponds were most affected by WSSV in the years 2011, 2017, and 2020. Data of 114 disease points were collected by directly interviewing shrimp pond owners about the history of the white spot disease outbreak in their farms. Locations of the shrimp ponds were marked by GPS field surveys (Fig. S2).

Shrimp white spot disease influencing factors

There are many spatial and physio-chemical factors which affect the spread of WSSV in shrimps. The physio-chemical parameters of water affecting shrimp diseases include Chemical Oxygen Demand (COD) (Fig. 2a), Dissolved Oxygen (DO) (Fig. 2b), temperature (Khiem et al. 2020). In addition, salinity (Fig. 2c) has a great impact on the health and growth of the shrimp (Falconer et al. 2016). Besides, spatial factors such as distance to the roads (access and transport networks) (Fig. 2h) and distance to factory (including hatchery and medicine) influence the

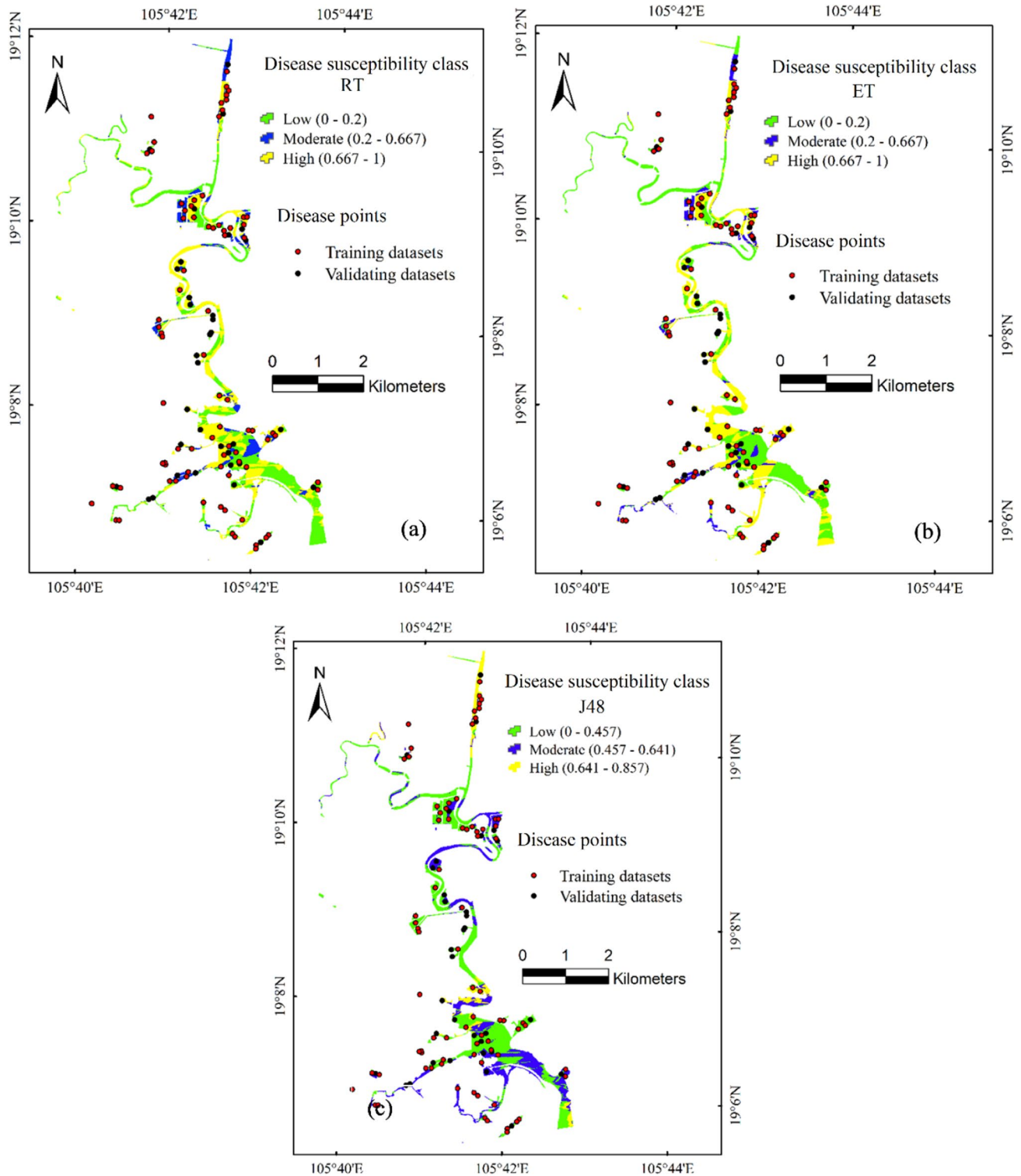


Fig. 8 Disease Susceptibility Maps (DSM) using: **a** RT model; **b** ET model; **c** J48 model

shrimp farming (Fig. 2d) (Giap et al. 2005). Therefore, in the present model’s study, we have selected eight WSSV disease affecting factors namely: pH, Salinity, COD, DO,

NO_3 , P_3O_4 , distance to the factory (shrimp storage) and distance to the road (farming site).

To monitor the parameters of the water environment, 17 water sampling stations were set up on the river, shrimp

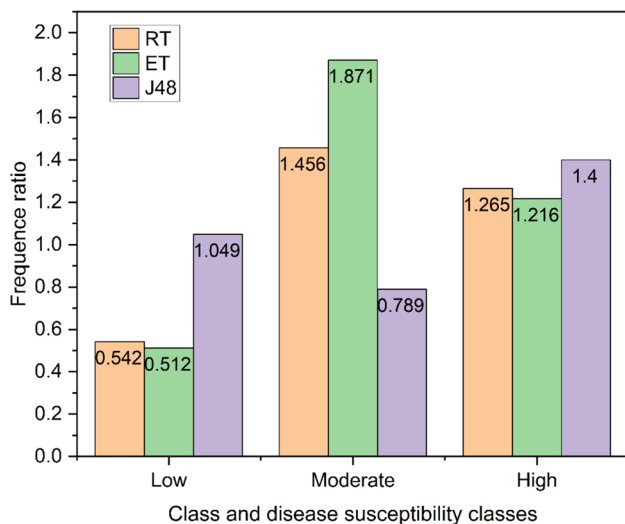


Fig. 9 Analysis of disease density on the susceptibility maps using the models

ponds and estuaries and samples were collected twice in year in the month of January and July, 2020. At those locations, salinity, pH, DO and COD were measured directly by portable instruments such as salinity meter (pH—9909SP), COD meter (YSI 910) and Multi-indicator meter (Hanna HI9829-01042) for NO_3 and P_3O_4 .

Thematic layers of distances to farms and factories were extracted from the cadastral maps of the communes of 1:10,000 scale obtained from the Department of Natural Resources and Environment of Nghe An province in conjunction with Google Earth image.

Methodological flowchart and steps

General methodology of the development of accurate susceptibility map of the Shrimp's WSSV disease based on the spatial and physio-chemical parameters of water using the DT based models is described below (Fig. 3). In the first step, eight factors which affects to the Shrimp's WSSV disease were selected and their spatial database were collected from various source. In the second step, the inventory of Shrimp's WSSV disease which includes 114 disease points and 114 non-disease points were collected from the field investigation. In the third step, the inventory of disease was split in two parts with 70:30 ratios; out of these, 70% of inventory was used for generating training dataset used for building the models and 30% remaining was used for generating testing dataset used for validating the models. In the four step, training dataset was then used to construct the DT-based models namely RT, DT, and J48). In the final step, testing dataset was used to validate the performance of the

models using various indicators namely PPV, NPV, RSME, ACC, SST, SPF, AUC and K.

Results and discussion

Validation of the models

Using training dataset, the accuracy of the three DT models in the prediction of WSSV disease was evaluated. The results show that PPV, NPV, SST, SPF and ACC values of both the RT and ET models have same values that is 91.25, 78.75, 81.11, 90.00 and 85.00, respectively, whereas J48 model show relatively lower values 66.25, 65.00, 65.43, 65.82 and 65.63, respectively (Table 1). The results of Kappa coefficient based on the training data are also same (K: 0.7) for RT and ET models and lower (0.313) for the J48 model (Fig. 4). Similar results were found for RMSE value (0.299) in comparison to J48 model (RMSE: 0.463) (Fig. 5a–c). However, AUC value of model RT (0.949) is the highest, followed by ET (0.941) and J48 models (0.706), respectively.

Using testing dataset, the results showed that RT and ET models have similar results of PPV, NPV, SST, SPF and ACC values: 67.65, 58.82, 62.16, 64.52 and 63.24, respectively but J48 model show different values: 70.59, 55.88, 61.54, 65.52 and 63.24, respectively that is higher values except in case of SST and PPF. Moreover, Kappa Index is the highest for RT and J48 models (0.265) (Fig. 4). The RMSE results also showed that the J48 (0.483) is the most accurate model among the 3 models used (Fig. 6a–c). However, AUC value on testing data set show that ET model (0.713) is the best, followed by RT (0.701) and J48 (0.641) in predicting the susceptibility of Shrimp's WSSV disease in this area (Fig. 7a and b).

In general, all three models (RT, ET and J48) have shown good performance in the development WSSV susceptibility maps. However, ET model has shown the highest accuracy (AUC: 0.713) on testing dataset in comparison to other two models (RT: 0.701 and J48: 0.641) which means that the ET model has a better predictive capability compared with other models, namely RT and J48. One of the advantages of ET method is that it adapts individual trees for global conjecture/ approximation and popularization/ generalization. It has more power due to their smaller error in running this model because the outliers create a limit for predicting Extra Tree. Moreover, ET can estimate with a superior computational efficiency than single trees and can be more accurate than the irrelevant properties of the variables. Further, ET has very little variance compared to single trees, and the accuracy of ET predictions increments with the number of trees in the model, belonging on the type of issue/ problem (Seyyedattar et al. 2020). Finding of this study is also in line with previous published works which stated that ML based

models are good tools for prediction and identification of WSSV disease (Edeh et al. 2022; Ramachandran et al.).

Construction of Shrimp's WSSV disease susceptibility maps

Shrimp's WSSV disease susceptibility was finally constructed using the results of training the DT models. In the first step, all pixels of the study areas were assigned disease susceptibility indexes generated by the training process of the models. Thereafter, the Natural Break classification method was used to classify these indexes into different classes namely low susceptibility, moderate susceptibility, high susceptibility (Fig. 8). Figure 9 shows the validation of the susceptibility maps using the disease frequency ratio analysis. It can be observed that most of the disease points were observed in the high susceptibility classes, which means the susceptibility maps generated are reliable. In practice, these maps can be used to provide valuable information related with the areas with higher susceptibility to WSSV disease for shrimp farmers and aquaculture managers that can help them make informed decisions about the location and conditions for shrimp farming. In addition, the farmers can implement preventive measures, adjust farming practices, and even consider alternative locations to minimize the risk of disease outbreaks.

Conclusions

White spot syndrome is a serious problem for shrimp farming as shrimps affected by WSSV can spread disease in other healthy shrimps within ponds and farms when they come in contact with each other, causing the reduction of effectiveness of shrimp farming. In this study, we have used the three DT models: RT, ET and J48, in the development of WSSV-susceptibility maps, which might help in effective shrimp farming management.

Results showed that the DT-based ML models especially ET (AUC: 0.713) are promising tools in prediction of WSSV and production of reliable WSSV-susceptibility maps which will help decision makers in better planning and control of spatial spread of WSSV disease in shrimps. Limitation of this work is that the limited number of factors affecting the WSSV disease were used for shrimp's WSSV disease susceptibility assessment, therefore, in future study, we proposed to incorporate other disease affecting parameters, such as temperature of pond, farm and sea-water; local chemicals used and present in the water; mixing of other diseased crustaceans from the land and sea, which might improve accuracy of the susceptibility maps of the area. Moreover, other ML models will also be used to check possibility of further improvement in the quality of maps.

Acknowledgements This research is supported by Ministry of Education and Training, Vietnam (Grant No. B2021-TDV-08).

Funding Open access funding provided by Lulea University of Technology.

Availability of data and material The data and materials will be made available at request.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethical approval and consent to participate Not applicable.

Consent for publication When accepted the authors give the journal authority to publish.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ahmad MW, Reynolds J, Rezguy Y (2018) Predictive modelling for solar thermal energy systems: a comparison of support vector regression, random forest, extra trees and regression trees. *J Clean Prod* 203:810–821. <https://doi.org/10.1016/j.jclepro.2018.08.207>
- Ajayaram KA, Jegadeeshwaran R, Sakthivel G, Sivakumar R, Patange AD (2021) Condition monitoring of carbide and non-carbide coated tool insert using decision tree and random tree – A statistical learning. *Mater Today Proceed*. <https://doi.org/10.1016/j.matpr.2021.02.065>
- Caipang CMA, Verjan N, Ooi EL, Kondo H, Hirono I, Aoki T, Kiyono H, Yuki Y (2008) Enhanced survival of shrimp, *Penaeus (Marsupenaeus) japonicus* from white spot syndrome disease after oral administration of recombinant VP28 expressed in *Brevibacillus brevis*. *Fish Shellfish Immunol* 25(3):315–320. <https://doi.org/10.1016/j.fsi.2008.04.012>
- Chaijarasphong T, Thammachai T, Itsathitphaisarn O, Sritunyulucksana K, Suebsing R (2019) Potential application of CRISPR-Cas12a fluorescence assay coupled with rapid nucleic acid amplification for detection of white spot syndrome virus in shrimp. *Aquaculture* 512:734340. <https://doi.org/10.1016/j.aquaculture.2019.734340>
- Chen W, Zhao X, Tsangaratos P, Shahabi H, Ilia I, Xue W, Wang X, Ahmad BB (2020) Evaluating the usage of tree-based ensemble methods in groundwater spring potential mapping. *J Hydrol* 583:124602. <https://doi.org/10.1016/j.jhydrol.2020.124602>
- Chien C-F, Chen L-F (2008) Data mining to improve personnel selection and enhance human capital: a case study in high-technology industry. *Exp Syst Appl* 34(1):280–290
- Depina I, Oguz EA, Thakur V (2020) Novel Bayesian framework for calibration of spatially distributed physical-based landslide

- prediction models. *Comput Geotech* 125:103660. <https://doi.org/10.1016/j.compgeo.2020.103660>
- Díaz JD, Hansen E, Cabrera G (2020) A random walk through the trees: Forecasting copper prices using decision learning methods. *Resour Policy* 69:101859. <https://doi.org/10.1016/j.resourpol.2020.101859>
- Dou J, Yunus AP, Bui DT, Merghadi A, Sahana M, Zhu Z, Chen C-W, Han Z, Pham BT (2020) Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed. *Japan Landslides* 17(3):641–658. <https://doi.org/10.1007/s10346-019-01286-5>
- Edeh MO, Dalal S, Obagbuwa IC, Prasad BS, Ninoria SZ, Wajid MA, Adesina AO (2022) Bootstrapping random forest and CHAID for prediction of white spot disease among shrimp farmers. *Sci Rep* 12(1):20876
- Falconer L, Telfer TC (2016) Ross LGJJoem investigation of a novel approach for aquaculture site selection. *J Environ Manag* 181:791–804
- Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. *Mach Learn* 63(1):3–42. <https://doi.org/10.1007/s10994-006-6226-1>
- Giap DH, Yi Y, Yakupitiyage AJO, Management C (2005) GIS for land evaluation for shrimp farming in Haiphong of Vietnam. *Ocean & Coastal Manag* 48(1):51–63
- Han JE, Kim J-E, Jo H, Eun J-S, Lee C, Kim JH, Lee K-J, Kim J-W (2019) Increased susceptibility of white spot syndrome virus-exposed *Penaeus vannamei* to *Vibrio parahaemolyticus* causing acute hepatopancreatic necrosis disease. *Aquaculture* 512:734333. <https://doi.org/10.1016/j.aquaculture.2019.734333>
- Hasan NA, Haque MM (2020) Dataset of white spot disease affected shrimp farmers disaggregated by the variables of farm site, environment, disease history, operational practices, and saline zones. *Data Brief* 31:105936. <https://doi.org/10.1016/j.dib.2020.105936>
- Hong H, Liu J, Bui DT, Pradhan B, Acharya TD, Pham BT, Zhu AX, Chen W, Ahmad BB (2018) Landslide susceptibility mapping using J48 decision Tree with AdaBoost, bagging and rotation forest ensembles in the Guangchang area (China). *CATENA* 163:399–413. <https://doi.org/10.1016/j.catena.2018.01.005>
- Hou N, Zhang X, Zhang W, Xu J, Feng C, Yang S, Jia K, Yao Y, Cheng J, Jiang B (2020) A new long-term downward surface solar radiation dataset over China from 1958 to 2015. *Sensors (basel)* 20(21):6167. <https://doi.org/10.3390/s20216167>
- John V, Liu Z, Guo C, Mita S, Kidono K Real-Time Lane Estimation Using Deep Features and Extra Trees Regression. In: Bräunl T, McCane B, Rivera M, Yu X (eds) *Image and Video Technology*, Cham, 2016// 2016. Springer International Publishing, pp 721–733
- Kamiński B, Prałat P (2019) Sub-trees of a random tree. *Discret Appl Math* 268:119–129. <https://doi.org/10.1016/j.dam.2019.05.003>
- Kathayani SA, Poornima M, Sukumaran S, Nagavel A, Muralidhar M (2019) Effect of ammonia stress on immune variables of Pacific white shrimp *Penaeus vannamei* under varying levels of pH and susceptibility to white spot syndrome virus. *Ecotoxicol Environ Saf* 184:109626. <https://doi.org/10.1016/j.ecoenv.2019.109626>
- Keivani O, Sinha K (2021) Random projection-based auxiliary information can improve tree-based nearest neighbor search. *Inf Sci* 546:526–542. <https://doi.org/10.1016/j.ins.2020.08.054>
- Khiem NM, Takahashi Y, Oanh DTH, Hai TN, Yasuma H, Kimura NJFS (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. *Fisheries Sci* 86:673–683
- Li C, Gao X-X, Huang J, Liang Y (2016) Studies of the viral binding proteins of shrimp BP53, a receptor of white spot syndrome virus. *J Invertebr Pathol* 134:48–53. <https://doi.org/10.1016/j.jip.2016.01.006>
- Lu X, Kong J, Luan S, Dai P, Meng X, Cao B, Luo K (2016) Transcriptome analysis of the hepatopancreas in the Pacific white shrimp (*Litopenaeus vannamei*) under acute ammonia stress. *PLoS ONE* 11(10):e0164396
- Madhusudana CK, Kumar H, Narendranath S (2018) Fault diagnosis of face milling tool using decision tree and sound signal. *Mater Today: Proceed* 5:12035–12044. <https://doi.org/10.1016/j.matpr.2018.02.178>
- Meng X-H, Jang IK, Seo H-C, Cho Y-R (2010) A TaqMan real-time PCR assay for survey of white spot syndrome virus (WSSV) infections in *Litopenaeus vannamei* postlarvae and shrimp of farms in different grow-out seasons. *Aquaculture* 310(1):32–37. <https://doi.org/10.1016/j.aquaculture.2010.10.010>
- Millard RS, Ellis RP, Bateman KS, Bickley LK, Tyler CR, van Aerle R, Santos EM (2020) How do abiotic environmental conditions influence shrimp susceptibility to disease? A critical analysis focussed on white spot disease. *J Invertebrate Pathol*. <https://doi.org/10.1016/j.jip.2020.107369>
- Myung SJ, Kim YS, Seo DW, Shim KN, Kim HJ, Won SY, Yang SH, Lee SK, Kim MH, Min YI (1998) The new strategy in the diagnosis of pancreatic cancer and cholangiocarcinoma with CA19-9: New cutoff value from ROC (receiver operating characteristic curve. *Gastroenterology* 114:A650–A651. [https://doi.org/10.1016/S0016-5085\(98\)82662-0](https://doi.org/10.1016/S0016-5085(98)82662-0)
- Nguyen QH, Ly H-B, Ho LS, Al-Ansari N, Le HV, Tran VQ, Prakash I, Pham BT (2021) Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering* 2021
- Nunan LM, Lightner DV (1997) Development of a non-radioactive gene probe by PCR for detection of white spot syndrome virus (WSSV). *J Virol Methods* 63(1):193–201. [https://doi.org/10.1016/S0166-0934\(96\)02128-3](https://doi.org/10.1016/S0166-0934(96)02128-3)
- Okoro EE, Obomanu T, Sanni SE, Olatunji DI, Igbinedion P (2021) Application of artificial intelligent in predicting the dynamics of bottom hole pressure for under-balanced drilling: Extra tree compared with feed forward neural network model. *Petroleum*. <https://doi.org/10.1016/j.petlm.2021.03.001>
- Pham BT, Tien Bui D, Pourghasemi HR, Indra P, Dholakia MJT, Climatology A (2017) Landslide susceptibility assessment in the Uttarakhand area India using GIS a comparison study of prediction capability of naïve bayes multilayer perceptron neural networks and functional trees methods. *Theor Appl Climatol* 128:255–273
- Pham BT, Tien Bui D, Prakash I (2017b) Landslide susceptibility assessment using bagging ensemble based alternating decision trees, logistic regression and J48 decision trees methods: a comparative study. *Geotech Geol Eng* 35:2597–2611
- Ramachandran L, Mohan V, Senthilkumar S, Ganesh J Early detection and identification of white spot syndrome in shrimp using an improved deep convolutional neural network. *Journal of Intelligent & Fuzzy Systems (Preprint)*:1–12
- Rustam F, Mehmood A, Ullah S, Ahmad M, Muhammad Khan D, Choi GS, On BW (2020) Predicting pulsar stars using a random tree boosting voting classifier (RTB-VC). *Astronomy and Comput* 32:100404. <https://doi.org/10.1016/j.ascom.2020.100404>
- Schleder DD, Blank M, Peruch LGB, Poli MA, Gonçalves P, Rosa KV, Fracalossi DM, Vieira FdN, Andreatta ER, Hayashi L (2020) Impact of combinations of brown seaweeds on shrimp gut microbiota and response to thermal shock and white spot disease. *Aquaculture* 519:734779. <https://doi.org/10.1016/j.aquaculture.2019.734779>
- Seyyedattar M, Ghiasi MM, Zendejboudi S, Butt S (2020) Determination of bubble point pressure and oil formation volume factor: extra trees compared with LSSVM-CSA hybrid and ANFIS

- models. *Fuel* 269:116834. <https://doi.org/10.1016/j.fuel.2019.116834>
- Sridhar Raj S, Nandhini M (2018) Ensemble human movement sequence prediction model with Apriori based probability tree classifier (APTC) and bagged J48 on machine learning. *J King Saud University – Comput Inform Sci*. <https://doi.org/10.1016/j.jksuci.2018.04.002>
- Sun Y-L, Zhang X-Y, Sun T, He N, Li J-Y, Zhuang Y, Zeng Q, Yu J, Fang Q, Wang R (2013) The anti-inflammatory potential of neuropeptide FF in vitro and in vivo. *Peptides* 47:124–132. <https://doi.org/10.1016/j.peptides.2013.07.003>
- Tangirala S (2020) Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *Int J Adv Comput Sci Appl* 11(2):612–619
- Tien Bui D, Shirzadi A, Shahabi H, Geertsema M, Omidvar E, Clague JJ, Thai Pham B, Dou J, Talebpour Asl D, Bin Ahmad B, Lee S (2019) New ensemble models for shallow landslide susceptibility modeling in a semi-arid watershed. *Forests* 10(9):743
- Tripepi G, Jager KJ, Dekker FW, Zoccali C (2009) Diagnostic methods 2: receiver operating characteristic (ROC) curves. *Kidney Int* 76(3):252–256. <https://doi.org/10.1038/ki.2009.171>
- Tuyen N, Verreth J, Vlaskovic J, De Jong M (2014) Horizontal transmission dynamics of White spot syndrome virus by cohabitation trials in juvenile *Penaeus monodon* and *P. vannamei*. *Preventive Veterinary Med* 117(1):286–294
- Van Phong T, Ly H-B, Trinh PT, Prakash I, Btjvjoes P (2020) Landslide susceptibility mapping using Forest by Penalizing Attributes (FPA) algorithm based machine learning approach. *Vietnam J Earth Sci* 42(3):237–246
- Verbruggen B, Bickley LK, Van Aerle R, Bateman KS, Stentiford GD, Santos EM, Tyler CR (2016) Molecular mechanisms of white spot syndrome virus infection and perspectives on treatments. *Viruses* 8(1):23
- Wang Q, White BL, Redman RM, Lightner DV (1999) Per os challenge of *Litopenaeus vannamei* postlarvae and *Farfantepenaeus duorarum* juveniles with six geographic isolates of white spot syndrome virus. *Aquaculture* 170(3):179–194. [https://doi.org/10.1016/S0044-8486\(98\)00425-6](https://doi.org/10.1016/S0044-8486(98)00425-6)
- Zacarias S, Fegan D, Wangsoontorn S, Yamuen N, Limakom T, Carboni S, Davie A, Metselaar M, Little DC, Shinn AP (2021) Increased robustness of postlarvae and juveniles from non-ablated Pacific whiteleg shrimp, *Penaeus vannamei*, broodstock post-challenged with pathogenic isolates of *Vibrio parahaemolyticus* (VpAHPND) and white spot disease (WSD). *Aquaculture* 532:736033. <https://doi.org/10.1016/j.aquaculture.2020.736033>
- Zhang L, Lin Z, Wang J, He B (2020) Rapidly-exploring Random Trees multi-robot map exploration under optimization framework. *Robot Auton Syst* 131:103565. <https://doi.org/10.1016/j.robot.2020.103565>
- Zito M, Cooper C (2006) Dominating sets of random recursive trees. *Electron Notes in Dis Math* 27:107–108. <https://doi.org/10.1016/j.endm.2006.08.078>

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