An Ensemble Learning Technique for Predicting Mortality Rate in Red Tilapia (*Oreochromis niloticus* Linn.) Fingerlings

Roongparit Jongjaraunsuk[1], Wara Taparhudee[1*] and Putra Ali Syahbana Matondang[2]

ABSTRACT

Aquaculture has witnessed a gradual transformation owing to advancement in automatic and intelligent technology. Coupled with the power of high-performance computers, these innovations have given rise to machine learning technologies capable of extracting valuable insights from data. Consequently, these technologies are poised to usher smart aquaculture into a new era of efficiency and productivity. This particular study focused on enhancing the predictive accuracy of mortality rates in red tilapia (*Oreochromis niloticus* Linn.) fingerlings raised in outdoor earthen ponds with a recirculating aquaculture system. To achieve this, the study leveraged a voting-based ensemble learning technique based on the combination of three single predictive algorithms: decision tree, deep learning, and naïve bayes (EL–$V_{(DT–DL–NB)}$). The initial phase of the research involved the compilation of a comprehensive dataset encompassing parameters were water temperature (ºC), dissolved oxygen (mg·L^{-1}), pH, total ammonia nitrogen (mg·L^{-1}), nitrite–nitrogen (mg·L^{-1}), transparency (cm), alkalinity (mg·L^{-1}), date, month and mortality rate (fish·day^{-1}). Following the collection and cleaning of the dataset, 173 samples with 12 attributes were used in this study. The outcomes of this investigation revealed that the performance of the individual predictive models was eclipsed by the proposed EL–$V_{(DT–DL–NB)}$ model, boasting an impressive accuracy rate of 90.85%, precision of 84.00%, recall of 77.50%, and area under curve of 0.896. These results affirm the potential utility of the proposed model for accurately forecasting the mortality rate of red tilapia fingerling in aquaculture settings, thereby contributing significantly to the optimization of aquaculture practices.

Keywords: Data mining, Ensemble learning, Red tilapia, Mortality rate

INTRODUCTION

Tilapia is one of the most valuable global inland aquaculture commodities with a production contribution of 4.4 million tonnes in 2020, accounting for approximately 9 percent of the world's production of major aquaculture species (FAO, 2022). One of the most popular species of tilapia is the red tilapia (*Oreochromis niloticus* Linn.). This fish has gained popularity in freshwater aquaculture throughout the world, especially in Thailand (Jongjaraunsuk and Taparhudee, 2022; Taparhudee *et al*., 2023).

The increase global consumption of tilapia highlights a growing trend in fish consumption within communities, necessitating a corresponding boost in production to meet rising demand. However, there are several factors that limit the expansion of tilapia production to satisfy consumer needs. One such factor is the deterioration of water quality due to nitrogen (N) waste originated from feces, uneaten feed residues, and gill excretions, resulting in decreased fish productivity (Putra *et al*., 2019; Hasibuan *et al*., 2023). On the other hand, red tilapia during the nursing period is very vulnerable to environmental factors. Therefore,

[1]Department of Aquaculture, Faculty of Fisheries, Kasetsart University, Bangkok, Thailand
[2]Fishery Science and Technology, Faculty of Fisheries, Kasetsart University, Bangkok, Thailand
[*]Corresponding author. E-mail address: ffiswrt@ku.ac.th
Received 1 September 2023 / Accepted 5 February 2024

determining the optimal environmental conditions to achieve adequate fingerling growth performance is important in optimizing production and maintaining cultivation profitability (Azaza *et al*., 2008; Ekasari *et al*., 2015; García-Ríos *et al*., 2019). Furthermore, water quality management has emerged as a critical component of aquaculture practices. It is crucial to comprehend the parameters of water quality and their interrelationships, which can affect fish health and growth and also can determine the failure or success of aquaculture practices as a whole (Putra *et al*., 2019; El-Sayed, 2020).

Automation and intelligent technological advancements are rapidly reshaping the landscape of aquaculture. Combined with high-performance computing, machine learning technologies can sift through vas datasets, unravel cause-and-effect associations, anticipate issues, and provide intelligent solutions in aquaculture (Liakos *et al*., 2018; Sharma and Gupta, 2021; Zhao *et al*., 2021; Gladju *et al*., 2022). Numerous studies have investigated creation of prediction models utilizing traditional learning algorithms, such as naïve bayes (NB), deep learning (DL), and decision trees (DT) (LeCun *et al*., 2015; Schmidhuber, 2015; Huan *et al*., 2020). Traditionally, a single predictive model has been crafted to forecast the class level of each classification problem. Specifically, in predictive modeling and data analysis in general, a single predictive algorithm is used for a given sample of data. However, the applied single predictive model sometimes faces problems, such as low predictive performance on a limited dataset. This occurs because each of the several algorithms can contribute more effectively using its own strength. Hence, this study proposed an ensemble learning (EL) technique that combines the predicted results of several algorithms to improve predictive performance.

The EL technique is a type of learning model in which instead of a single learning algorithm, multiple learning schemes are used to solve the same problem; therefore, it can be called a multilevel predictive classification model (Sagi and Rokach, 2018; Marcello *et al*., 2020).

Compared to single predictive algorithms, EL techniques can yield more accurate and reliable predictions with improved generalization performance and wider applications (Knudby *et al*., 2010; Raza, 2019; Lin *et al*., 2022; Mienye and Sun, 2022). In this study, a voting-based ensemble learning technique ($EL–V_{DT–DL–NB}$) was proposed, combining three algorithms: DT (Quinlan, 1986), DL, and NB (Lewis, 1998). These algorithms are widely recognized in supervised machine learning for aquaculture (Zhao *et al*., 2021).

By implementing an EL-V technique using water quality parameters data, we aimed to enhance predictive performance by integrating multiple machine learning algorithms in a unified process. A successful outcome is expected to provide the aquaculture management system with improved recommendations and insights, enabling fish farmers to make informed decisions and actions to prevent mortality, produce higher quality fingerlings, and increase the productivity of red tilapia for sustainable aquaculture.

MATERIALS AND METHODS

Study area

The study area for this research was the Patthamarach Farm, situated in the Lam Plai Mat district of Buriram province, Thailand. The farm, located at coordinates (15°04'01.9" N 102°47'20.3" E), comprises various components, including a reservoir pond spanning 9,200 m^2, 2 treatment ponds with sizes of 1,600 m^2 and 4,600 m^2, respectively, 4 nursing ponds each covering an area of 1,600 m^2, 18 grow-out ponds, each with a size of 1,600 m^2, 1 sedimentation pond spanning 2,000 m^2, 2 biological treatment pond over an area of 8,200 m^2 and 7,200 m^2, 1 ready-to-use pond covering 3,600 m^2, and 1 treatment pond for water outlet 800 m^2.

Data sources

The fish were cultivated in earthen ponds lined with polyethylene, employing a recirculating aquaculture system featuring mono-sex fish in 4 ponds, each spanning an area of 1,600 m^2. The stocking density was set at 25 fish·m^{-2}, with an average initial weight of the fish was 50 g·$fish^{-1}$.

Each pond was equipped with four paddle wheel, operated 24 h a day. The fish were fed floating pellets comprising 35% crude protein three times a day (8:00 a.m., 11:30 a.m., and 4:30 p.m.) until satiation, using an automatic feeder.

The water quality parameters measured daily during maintenance included water temperature (Temp) dissolved oxygen (DO), pH, total ammonia nitrogen (TAN), nitrite-nitrogen (NO_2–N), transparency (Trans), and alkalinity (ALK). Temp, DO, and pH were measured twice daily (morning and evening) using a YSI Professional Plus instrument (Yellow Springs, OH, USA). These values were divided into two parts: average and difference. Temp average, DO average, and pH average represented the central or typical value from both morning and evening measurements. On the other hand, Temp difference, DO difference, and pH difference indicated the changes between morning and evening measurements. TAN, NO_2–N, and ALK were measured once a day following the procedures outlined in APHA (2005). In addition, Trans was measured daily using a Secchi disk.

Fish were reared until they reached an average weight of 200 g·fish^{-1}. The assessment of red tilapia mortality during the rearing process involved daily counting of the dead. Fish were categorized as low mortality if the death rate was less than 10 fish·day^{-1} and as high mortality if the death rate was more than 10 fish·day^{-1}.

Data processing

RapidMiner version 9.10 software was applied for all data processing in this study.

Data cleaning

The dataset obtained during a study typically contains incomplete entries, such as missing or unfilled data. In the present study, missing or unfilled data were replaced with values derived from the dataset (minimum, maximum, or average, depending on the characteristics of the attribute). Irrelevant data are discarded because its existence can reduce the quality or accuracy of data mining later (Kotu and Deshpande, 2019).

Data normalization

After the cleaning process, all data were standardized in RapidMiner using the 'Normalize' operator, utilizing z-transformation method. This helps to normalize the scale of numeric features (Figure 1). Because the applied parameters were in difference scales, a normalization technique was used to scale and transform features within a dataset to a similar scale, preventing certain features from dominating the learning process due to their larger values. The equation for z-transformation, represented as Equation 1, is as follows.

$$z = \chi - \mu / \sigma \quad (1)$$

Where: z is the resulting value from the z-transformation, χ is the individual data value, μ is the mean (average) of the data, and σ is the standard deviation of the data.

Data modelling

Once the dataset had been cleaned and normalized, it was imported into the RapidMiner. The ‘date’ attribute was changed from ‘regular’ to ‘id’, indicating that this column was used solely for identification purposes and not for classification or prediction. Furthermore, the ‘mortality rate’ attribute was used as a ‘label’ for analysis. The dataset was stored in a local or temporary repository in the RapidMiner.

To apply the EL model for predicting the mortality rate of red tilapia, a voting method was used. Initially, the ‘retrieve dataset’ and ‘optimize parameters (Grid)’ operators were introduced into the process. The Grid operator is a nested operator designed to execute the subprocess on all combinations of selected parameter values, providing the optimal parameter values that yield the best predictive performance. The optimization parameters chosen for this study align with the specifications detailed in Table 1.

Within the ‘optimize parameters (Grid)’ operator, the 'cross validation' operator is employed to gauge the practical accuracy of model performance. This technique was selected toidentify overfitting or underfitting and to provide a more reliable estimation of a model's performance, as opposed to relying solely on a single train-test split. This operator has two subprocesses namely training (learning) and testing (validation), employing a 10-fold configuration. The processing and optimizing parameters structures can be seen in Figure 1.

Figure 2(a) illustrates the validation structure. In the training subprocess, the ‘Vote’ operator is used to build a classification model. The advantage of the voting operator lies in the freedom to choose more single algorithms and thus combine them to prove their effectiveness in predicting the mortality rate of the red tilapia. The maximum vote or average received for a certain class is then predicted. The ‘Vote’ operator is a nested operator with a subprocess that requires a minimum of two base learners. In this study, we introduced three predictive models as base

Table 1. Optimization parameters.

Operator	Parameter	Range
DT	Criterion	Gain ratio, Gini index, Information gain, Accuracy
	Minimal leaf size	2–6
	Maximal depth	1–10
DL	Activation	Tanh, Rectifier, Maxout, ExpRectifier
NB	Laplace correction	False, True

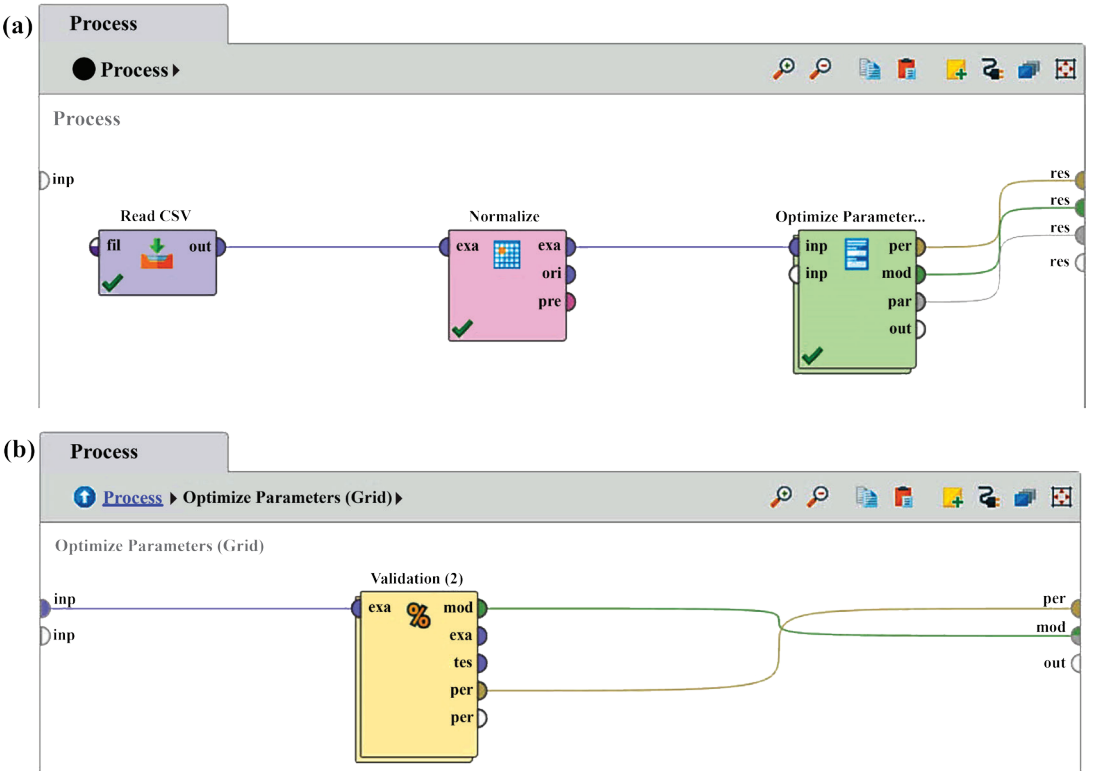

Figure 1. Structures for processing (a) and under optimizing parameters (grid) (b).

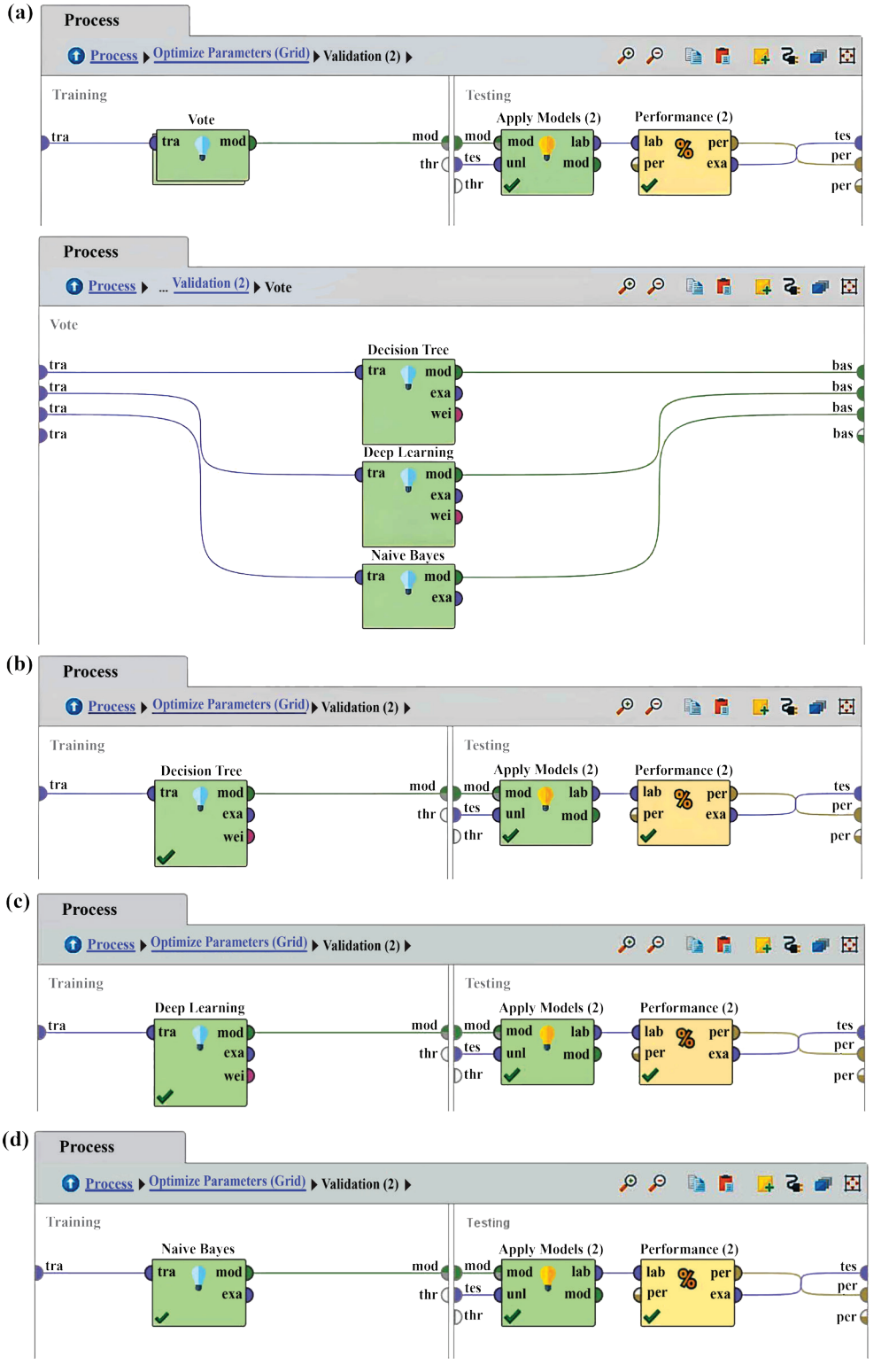

Figure 2. Validation structure under optimize parameter (Grid) with (a) EL–$V_{(DT–DL–NB)}$, (b) DT, (c) DL, and (d) NB.

learners: DT, DL, and NB (Figure 2b-2d). Together, they constituted the EL–$V_{(DT–DL–NB)}$ model. In the testing subprocess, the 'apply model' and 'performance' operators are used. The 'apply model' operator is used to obtain predictions on unseen data or to transform data by applying a preprocessing model. The 'performance' operator determines the learning task type, which also generates the most typical criterion for that type. The main page allows processing and editing of the parameter settings in the 'optimize parameters (Grid)' operator (Table 2). Finally, the EL–$V_{(DT–DL–NB)}$ model is formed and loaded to view the result. Moreover, the DT, DL, and NB individual models are evaluated against the EL–V model, employing a comparable procedure for execution. Nevertheless, the "vote" operator was replaced with the models under examination.

Description of the single algorithms

Decision tree

The DT algorithm, proposed by Quinlan (1986), is a method that partitions the input space into multiple regions, each having its distinct parameters. It operates by extracting a tree-based classification model from a random training sample. This process involves segmenting the feature space and constructing decision trees. Within the decision tree, each non-leaf node encapsulates a category rating characteristic, while each leaf node signifies a final scoring category (Liakos *et al*., 2018; Zhao *et al*., 2021). Mathematically, the entropy and information gain of DT can be expressed as shown in equations (2) and (3).

$$H(S) = -\sum_{i=1}^{n} p_i \log_2(p_i) \quad (2)$$

Where H(S) = is the entropy of set S before the split, n is the number of classes, and p(i) is the proportion of examples in set S that belong to class i.

$$IG(S,A) = H(S) - \sum_{t\in T} p(t)H(t) = H(S) - H(S|A) \quad (3)$$

Where IG(S,A) is the information gain of attribute A on the set S, t iterates over each individual value in the set T, T is the set of possible values of attribute A, p(t) is the probability of occurrence of value t of attribute A in set S, H(t) is the entropy of the subset of S for which attribute A has the value t, and H(S|A) is the conditional entropy of set S given attribute A.

Deep learning

DL is a particular kind of algorithm that learns the fundamental principles and layers of sample data representation. By using functions to integrate lower-level features, DL creates more abstract high-level features that reflect attribute categories. Finding the data's distributed feature representation makes nonlinear relationship modeling a great deal simpler. It is an artificial network that is utilized to address more challenging and complicated data mining tasks. (Zhao *et al*., 2021; Bilal *et al*., 2022).

Naïve bayes

NB (Lewis, 1998) is a type of algorithm that is based on the Bayesian principle, which states that sample data sets can be classified using statistical knowledge of probability. This method is known to perform better than even extremely advanced methods in a number of practical issues. It is simple to construct, extremely scalable, adaptable to large datasets, and needs a number of parameters (features) (Raza, 2019). Mathematically, the NB algorithm can be expressed as shown in equation (4).

$$Q(c|x) = Q(x|c) \times Q(c) / Q(x) \quad (4)$$

Where Q(c) and Q(x) are the prior probabilities of class c and feature x, respectively and Q(x|c) is the probability of feature x, given class c, which is called the probability.

All the experimental models utilized the default configuration provided by the RapidMiner Studio program, as outlined in the Table 2.

Table 2. Model architectures of DT, DL and NB models.

Model	Parameter
DT	Criterion = gain ratio, Maximum depth = 10, Apply pruning, Confidence level = 0.1, Minimal gain = 0.01, Minimal leaf size = 2
DL	Activation = rectifier, Hidden layer = 2 layers with 50 neurons of each layer, Epochs = 10, Loss function = automatic
NB	Laplace correction

Performance indicators

Accuracy, precision, and recall

During the classification process, performance indicators are very important in differentiating and obtaining the optimal classifier. Therefore, 3 important classification metrics were considered in this study: accuracy, precision, and recall. In the following equations, TP is the true positive, TN is the true negative, FP is a false positive, and FN is a false negative. Accuracy is the number of correct predictions (positive and negative) of all observed data generated by the model. Accuracy can be mathematically expressed using Equation (5):

$$\text{Accuracy} = (TP + TN / TP + TN + FP + FN) \times 100\% \quad (5)$$

Precision is the number of correct positive data categories divided by the total data classified as positive. Precision can be written as equation (6):

$$\text{Precision} = (TP / TP + FP) \times 100\% \quad (6)$$

Recall is the ratio of correctly positive predictions compared to the overall actual positive data. Recall is calculated with the formula shown in Equation (7):

$$\text{Recall} = (TP / TP + FN) \times 100\% \quad (7)$$

Confusion matrix

The confusion matrix in a classification task is typically utilized to investigate the performance of a suggested model, where the value of the confusion matrix is usually shown as a percentage. Figure 3 shows the confusion matrix to predict the mortality rate of red tilapia.

Additionally, the calculation of the area under curve (AUC) was conducted. This metric serves to evaluate the performance of a binary classification model, with higher AUC values signifying improved ability in distinguishing between classes.

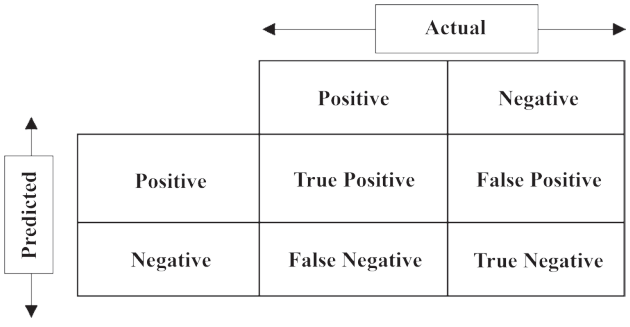

Figure 3. Confusion matrix structure.

RESULTS AND DISCUSSION

The dataset used in this study consisted of 173 examples and 12 attributes. The attributes used were Temp average, Temp difference, DO average, DO difference, pH average, pH difference, TAN, NO_2–N, ALK, Trans, date, and month. The mortality rate of red tilapia fingerlings was used as a label, with 134 cases of low mortality and 39 cases of high mortality. A detailed description of the water quality parameters is provided in Table 3. Table 4 displays the optimal parameters generated for the EL–$V_{(DT–DL–NB)}$, DT, DL and NB models.

The result for the performance indicators of EL–$V_{(DT–DL–NB)}$ in predicting the mortality rate of red tilapia fingerlings had accuracy of 90.85%, precision of 84.00%, and recall of 77.50%. Therefore, the proposed model produced better results compared to the single predictive models based on 10-fold cross-validation on the same dataset, as shown in Figure 4. The second-best predictive model was DT with accuracy of 89.58%, precision of 80.38%, and recall of 79.17%. However, both DL and NB exhibited comparatively lower performance levels.

Meanwhile, the AUC for the EL–$V_{(DT–DL–NB)}$ reached 0.896, whereas individual models like DT, DL, and NB achieved AUC values of 0.874, 0.916, and 0.887, respectively, as depicted in Figure 5–8.

Table 3. Description of water quality parameters.

Water quality parameter	Unit	Min	Max	Mean	Standard deviation
Temp average	ºC	18.50	27.00	22.67	1.92
Temp difference	ºC	0.00	4.00	1.79	0.84
DO average	$mg{\cdot}L^{-1}$	3.60	7.85	5.77	0.87
DO difference	$mg{\cdot}L^{-1}$	0.00	3.90	1.54	0.94
pH average	-	5.15	7.60	7.29	0.23
pH difference	-	0.00	0.90	0.19	0.19
TAN	$mg{\cdot}L^{-1}$	0.00	3.00	1.11	0.73
NO_2–N	$mg{\cdot}L^{-1}$	0.00	1.40	0.71	0.43
ALK	$mg{\cdot}L^{-1}$	60.00	110.00	79.11	9.41
Trans	cm	15.00	45.00	33.78	8.22

Table 4. Optimal values for EL–$V_{(DT–DL–NB)}$, DT, DL and NB models.

Operator	Sub-operator	Parameter	Optimal value
EL–$V_{(DT–DL–NB)}$	DT	Criterion	Gini_index
	DT	Minimal leaf size	2
	DT	Maximal depth	10
	DL	Activation	Rectifier
	NB	Laplace correction	True
DT	-	Criterion	Information_gain
		Minimal leaf size	8
		Maximal depth	4
DL	-	Activation	Rectifier
NB	-	Laplace correction	False

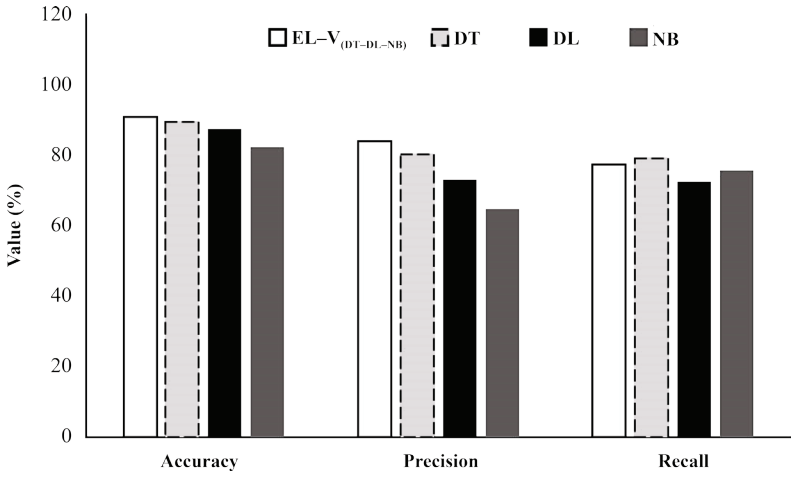

Figure 4. Performance comparison of EL–$V_{(DT–DL–NB)}$ and three single predictive models.

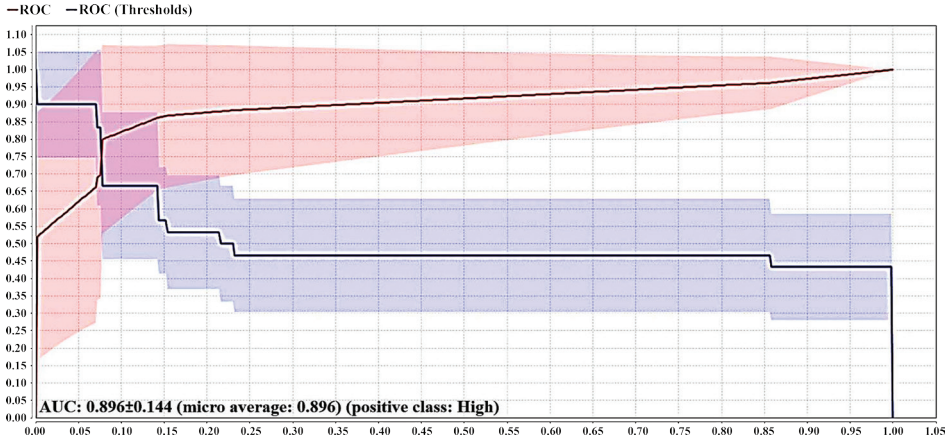

Figure 5. AUC of EL–$V_{(DT–DL–NB)}$.

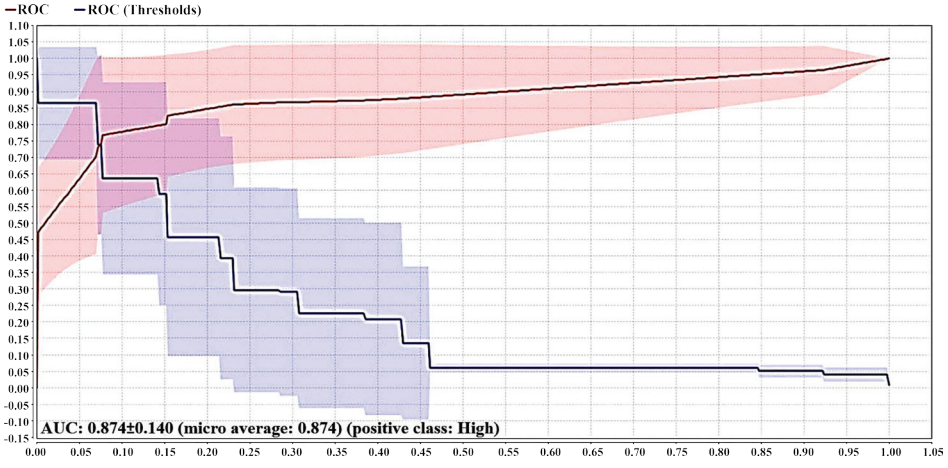

Figure 6. AUC of DT.

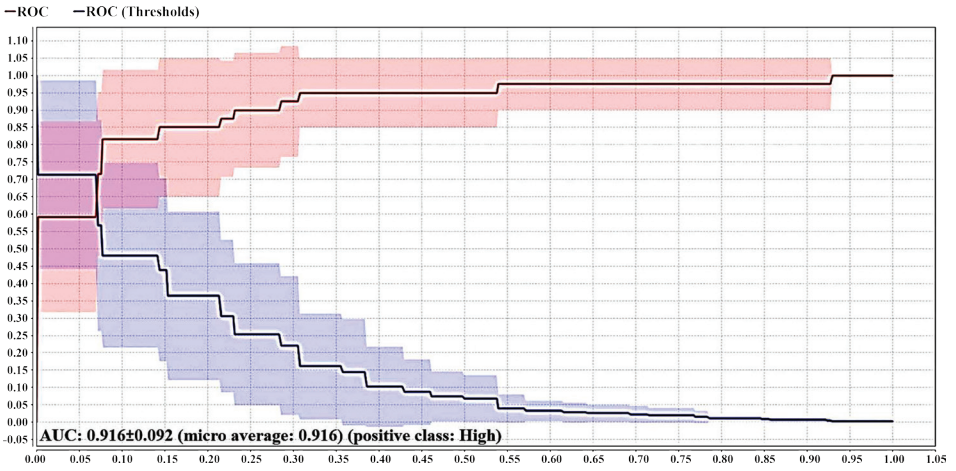

Figure 7. AUC of DL.

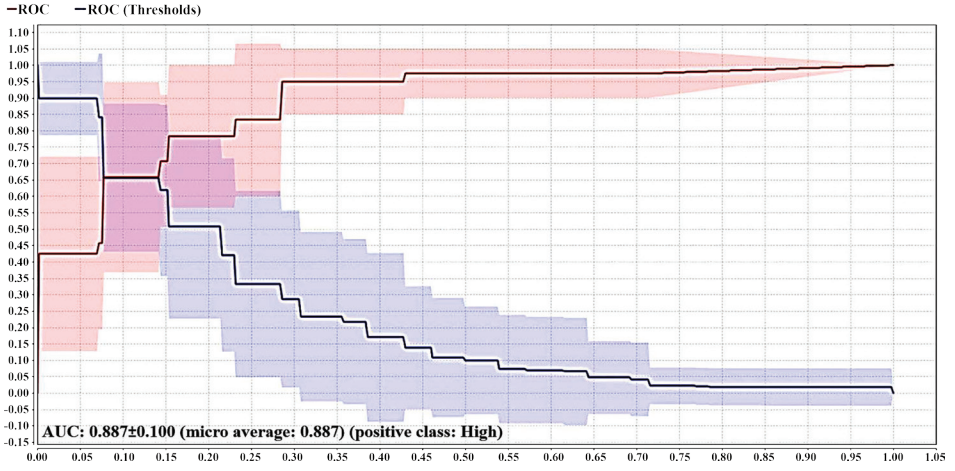

Figure 8. AUC of NB.

The research revealed that ensemble methods (EL–$V_{(DT–DL–NB)}$) outperformed individual models such as DT, DL, and NB, in terms of overall performance. Each model has its strengths and weaknesses. For instance, DT excel in easy interpretation of results, handling missing data well, being less impacted by outliers compared to other models, and handling both linear and non-linear relationships between features and target variables. However, they tend to overfit complex or high-dimensional data and struggle with imbalanced data compared to other methods (Quinlan, 1986; Podgorelec *et al*., 2002; Zhao *et al*., 2021).

DL, on the other hand, handles large-scale data efficiently, learns complex and deep patterns well, and offers a flexible, customizable structure. However, it demands significant computational resources, a substantial amount of high-quality data for effective model training, and sometimes operates as a 'black box', making it challenging to interpret (LeCun *et al*., 2015; Schmidhuber, 2015; Goodfellow *et al*., 2016).

NB demonstrates speed in training and good prediction capabilities, has an easily understandable algorithm with simple code suitable for basic use,

and performs well in predicting and grouping large datasets. However, it often struggles when data is incomplete or when crucial variables have missing values, and it might not handle diverse feature values effectively, leading to lower accuracy (Rish, 2001; Zhang, 2004).

EL–$V_{(DT–DL–NB)}$ leverages the strengths of individual models while mitigating their weaknesses. They enhance overall model accuracy, reduce errors from individual models, mitigate overfitting, and effectively handle complex data (Kuncheva, 2004; Zhou, 2012). Although ensembles aim to reduce overfitting, in some cases, they can still overfit if individual models within the ensemble are highly correlated or if the ensemble is too complex relative to the dataset size (Breiman, 2001; Kuncheva, 2004; Caruana *et al*., 2006).

Additionally, from the data in this study, it was observed that the data used for prediction (low-high mortality rates) was not equal. Dealing with imbalanced data can be approached through various methods, such as employing resampling techniques to balance the dataset, using bagging and boosting to create accurate models adept at handling imbalanced data, importing models from different algorithms into an ensemble, and adjusting class importance values (He and Garcia, 2009; Sun *et al*., 2009; Lemaitre *et al*., 2017).

For this study, we opted for ensemble methods utilizing DT, DL, and NB models to enhance learning from imbalanced data. This study's conclusions were supported by various research works, such as Jardim *et al*. (2021) advising on fisheries management; Alfatinah *et al*. (2023) predicting skipjack area for each time slice; Chen *et al*. (2023) projecting fish distribution in response to climate changes; and Khiem *et al*. (2023) predicting the growth of abalone.

However, to increase accuracy, there should be additional data collection with comparable quantities. The next study step should involve integrating real-time internet of things sensory systems for future predictive applications.

CONCLUSION

Machine learning provides technical support in data mining and has been widely applied in aquaculture. The use of single predictive algorithms sometimes encounters challenges in mining, particularly with a limited dataset. Here, we proposed a voting-based ensemble learning technique, combining three algorithms (EL–$V_{(DT–DL–NB)}$), to enhance the predictive performance of the mortality rate of red tilapia fingerlings based on a water quality dataset. The performance indicators of the proposed model outperform the compared models in all aspects, achieving an accuracy of 90.85%, precision of 84.00%, and recall of 77.50%. Overall, the results of this study demonstrate that the reliability and implementation of the proposed technique could be useful in predicting the mortality rate of red tilapia fingerlings raised in outdoor earthen ponds using a recirculating aquaculture system. Therefore, the EL–$V_{(DT–DL–NB)}$ model may serve as a valuable component for developing future decision support systems for the aquaculture industry.

In the future, this study could be expanded by utilizing a larger dataset to predict the mortality rate of red tilapia during their grow-out period. In addition, for classification purposes, other classification algorithms could be tested to obtain the best performance results.

ACKNOWLEDGEMENTS

This study was supported by the "A decision support system using machine learning techniques for finding best practices of red tilapia (*Oreochromis niloticus* Linn.) reared in an outdoor recirculating aquaculture system" project, Faculty of Fisheries, Kasetsart University, Bangkok, Thailand. All procedure were approved by the ethic and animal welfare committee of Kasetsart University, Bangkok, Thailand (ethical protocol number: ACKU 66-FIS-004), in accordance with the university's guidelines for good scientific practice.

LITERATURE CITED

Alfatinah, A., H.J. Chu, Tatas and S.R. Patra. 2023. Fishing Area Prediction Using Scene-Based Ensemble Models. **Journal of Marine Science and Engineering** 11: 1398. DOI: 10.3390/jmse11071398.

American Public Health Association (APHA). 2005. **Standard Methods of the Examination of Water and Wastewater,** 21st ed. American Public Health Association, Washington, D.C., USA. 541 pp.

Azaza, M.S., M.N. Dhraïef and M.M. Kraïem. 2008. Effects of water temperature on growth and sex ratio of juvenile Nile tilapia *Oreochromis niloticus* (Linnaeus) reared in geothermal waters in southern Tunisia. **Journal of Thermal Biology** 33(2): 98–105. DOI: 10.1016/j.jtherbio.2007.05.007.

Bilal, S.F., A.A. Almazroi, S. Bashir, F.H. Khan and A.A. Almazroi. 2022. An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry. **PeerJ Computer Science** 8: 8:e854. DOI: 10.7717/peerj-cs.854.

Breiman, L. 2001. Random forests. **Machine Learning** 45(1): 5–32. DOI: 10.1023/A:1010933404324.

Caruana, R., A. Munson and A. Niculescu-Mizil. 2006. **Getting the most out of ensemble Selection.** Proceeding of the 6th International Conference on Data Mining (ICDM'06) 2006: 828–833.

Chen, Y., X. Shan, H. Gorfine, F. Dai, Q. Wu, T. Yang, Y. Shi and X. Jin. 2023. Ensemble projections of fish distribution in response to climate changes in the Yellow and Bohai Seas, China. **Ecological Indicators** 146: 109759. DOI: 10.1016/j.ecolind.2022.109759.

Ekasari, J., D.R. Rivandi, A.P. Firdausi, E.H. Surawidjaja, M. Zairin, P. Bossier and P. De Schryver. 2015. Biofloc technology positively affects Nile tilapia (*Oreochromis niloticus*) larvae performance. **Aquaculture** 441: 72–77. DOI: 10.1016/j.aquaculture.2015.02.019.

El-Sayed, A.F.M. 2019. **Environmental requirement.** In: Tilapia Culture, 2nd ed. (ed. A.F.M. El-Sayed), pp. 47–67. Academic Press, Massachusetts, USA.

Food and Agriculture Organization of the United Nations (FAO). 2022. **The state of world fisheries and aquaculture.** https://www.fao.org/3/cc0461en/cc0461en.pdf. Cited 15 Nov 2023.

García-Ríos, L., A. Miranda-Baeza, M.G. Coelho-Emerenciano, J.A. Huerta-Rábago and P. Osuna-Amarillas. 2019. Biofloc technology (BFT) applied to tilapia fingerlings production using different carbon sources: Emphasis on commercial applications. **Aquaculture** 502: 26–31. DOI: 10.1016/j.aquaculture.2018.11.057.

Gladju, J., B.S. Kamalam and A. Kanagaraj. 2022. Applications of data mining and machine learning framework in aquaculture and fisheries: A review. **Smart Agricultural Technology** 2: 100061. DOI: 10.1016/j.atech.2022.100061.

Goodfellow, I., Y. Bengio, A. Courville and Y. Bengio. 2016. **Deep Learning,** Volume 1. The MIT Press, Massachusetts, USA. 800 pp.

Hasibuan, S., S. Syafriadiman, N. Aryani, M. Fadhli and M. Hasibuan. 2023. The age and quality of pond bottom soil affect water quality and production of *Pangasius hypophthalmus* in the tropical environment. **Aquaculture and Fisheries** 8(3): 296–304. DOI: 10.1016/j.aaf.2021.11.006.

He, H. and E.A. Garcia. 2009. Learning from imbalanced data. **IEEE Transactions on Knowledge and Data Engineering** 21(9): 1263–1284. DOI: 10.1109/TKDE.2008.239.

Huan, J., H. Li, M. Li and B. Chen. 2020. Prediction of dissolved oxygen in aquaculture based on gradient boosting decision tree and long short-term memory network: A study of Chang Zhou fishery demonstration base, China. **Computers and Electronics in Agriculture** 175: 105530. DOI: 10.1016/j.compag.2020.105530.

Jardim, E., M. Azevedo, J. Brodziak, E.N. Brooks, K.F. Johnson, N. Klibansky, C.P Millar, C. Minto, I. Mosqueira, R.D.M. Nash, P. Vasilakopoulos and B.K. Wells. 2021. Operationalizing ensemble models for scientific advice to fisheries management. **ICES Journal of Marine Science** 78(4): 1209–1216. DOI: 10.1093/icesjms/fsab 010.

Jongjaraunsuk, R. and W. Taparhudee. 2022. Weight estimation model for red tilapia (*Oreochromis niloticus* Linn.) from images. **Agriculture and Natural Resources** 56(1): 215–224. DOI: 10.34044/j.anres. 2021.56.1.20.

Khiem, N.M., Y. Takahashi, T. Masumura, G. Kotake, H. Yasuma and N. Kimura. 2023. A machine learning ensemble approach for predicting growth of abalone reared in land-based aquaculture in Hokkaido, Japan. **Aquacultural Engineering** 103: 102372. DOI: 10.1016/j.aquaeng.2023. 102372.

Knudby, A., A. Brenning and E. LeDrew. 2010. New approaches to modelling fish–habitat relationships. **Ecological Modelling** 221: 503–511. DOI: 10.1016/j.ecolmodel.2009. 11.008.

Kotu, V. and B. Deshpande. 2019. **Getting started with RapidMiner.** In: Data Science Concepts and Practice, 2nd ed. (eds. V. Kotu and B. Deshpande), pp. 491–521. Morgan Kaufmann, Massachusetts, USA.

Kuncheva, L.I. 2004. **Combining Pattern Classifiers: Methods and Algorithms.** John Wiley and Sons, New Jersey, USA. 350 pp.

LeCun, Y., Y. Bengio and G. Hinton. 2015. Deep learning. **Nature** 521: 436–444. DOI: 10. 1038/nature14539.

Lemaitre, G., F. Nogueira and C.K. Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. **Journal of Machine Learning Research** 18(17): 1–5.

Lewis, D.D. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. Machine Learning: **ECML** 98: 4–15. DOI: 10.1007/BFb0026666.

Liakos, K.G., P. Busato, D. Moshou, S. Pearson and D. Bochtis. 2018. Machine learning in agriculture: A review. **Sensors (Basel)** 18(8): 2674. DOI: 10.3390/s18082674.

Lin ,S., H. Zheng, B. Han, Y. Li, C. Han and W. Li. 2022. Comparative performance of eight ensemble learning approaches for the development of models of slope stability prediction. **Acta Geotechica** 17(4): 1477–1502. DOI: 10.1007/s11440-021-01440-1.

Marcello, B., C. Davide, F. Marco, G. Roberto, M. Leonardo and P. Luca. 2020. An ensemble-learning model for failure rate prediction. **Procedia Manufacturing** 42: 41–48. DOI: 10.1016/j.promfg.2020.02. 022.

Mienye, I.D. and Y. Sun. 2022. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. **IEEE Access** 10: 99129–99149. DOI: 10.1109/access.2022.3207287.

Podgorelec, V., P. Kokol, B. Stiglic and I. Rozman. 2002. Decision trees: an overview and their use in medicine. **Journal of Medical systems** 26(5): 445–463. DOI: 1016409 317640.

Putra, I., I. Effendi, I. Lukistyowati and U.M. Tang. 2019. Growth and survival rate of red tilapia (*Oreochromis* sp.) cultivated in the brackish water tank under biofloc system. **Advance in Engineering Research** 190: 96–99. DOI: 10.2991/iccelst-st-19.2019. 19.

Quinlan, J.R. 1986. Induction of decision trees. **Machine Learning** 1: 81–106. DOI: 10. 1007/BF00116251.

Raza, K. 2019. **Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule.** In: U-Healthcare Monitoring Systems (eds. N. Dey, A.S. Ashour, S.J. Fong and S. Borra), pp. 179–196. Academic Press, Massachusetts, USA.

Rish, I. 2001. An empirical study of the naive Bayes classifier. **IJCAI 2001 workshop on empirical methods in artificial intelligence** 3(22): 41–46.

Sagi, O. and L. Rokach. 2018. Ensemble learning: a survey. **Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery** 8(4): e1249. DOI: 10.1002/widm.1249.

Schmidhuber, J. 2015. Deep learning in neural networks: An overview. **Neural Networks** 61: 85–117. DOI: 10.1016/j.neunet.2014.09.003.

Sharma, S. and Y.K. Gupta. 2021. Predictive analysis and survey of COVID-19 using machine learning and big data. **Journal of Interdisciplinary Mathematics** 24(1): 175–195. DOI: 10.1080/09720502.2020.1833445.

Sun, Y., A.K. Wong and M.S. Kamel. 2009. Classification of imbalanced data: A review. **International Journal of Pattern Recognition and Artificial Intelligence** 23(4): 687–719. DOI: 10.1142/S0218001409007326.

Taparhudee, W., R. Jongjaraunsuk, S. Nimitkul and W. Mathurossuwan. 2023. Application of unmanned aerial vehicle (UAV) with area image analysis of red tilapia weight estimation in river-based cage culture. **Journal of Fisheries and Environment** 47(1): 119–131.

Zhang, H. 2004. **The optimality of naive bayes.** Proceeding of the 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004) 2004: 12–14.

Zhao, S., S. Zhang, J. Liu, H. Wang, J. Zhu, D. Li and R. Zhao. 2021.Application of machine learning in intelligent fish aquaculture: A review. **Aquaculture** 540: 736724. DOI: 10.1016/j.aquaculture.2021.736724.

Zhou, Z.H. 2012. **Ensemble Methods: Foundations and Algorithms.** CRC Press, Florida, USA. 236 pp.