Proceedings of the 10th International Conference on Fisheries and Aquaculture, Vol. 10, Issue 1, 2023, pp. 35-44

ISSN 2386-1282 online
DOI: https://doi.org/10.17501/23861282.2023.10104

COMBINING GENERATIVE MODEL AND RANDOM FOREST TO PREDICT SHRIMP DISEASE OCCURRENCE

Lukman H*, Syauqy NA and Liris M

JALA TECH Pte Ltd, Indonesia

Abstract: Penaeus vannamei is one of the most cultured species. The global production of Penaeus (*Litopenaeus*) vannamei reached 5.8 million tonnes in 2020, contributing to 51.7% of total shrimp production. However, despite its high production, there are still many issues in this industry. One of those is the disease. The disease threatens shrimp farming, such as slowing shrimp growth rate and even mortality. To help the farmers in mitigating the impact of disease we tried to develop a predictive model that is able to give early warning of disease occurrence. We focused on predicting acute hepatopancreatic necrosis disease (AHPND), white feces disease (WFD), infectious myonecrosis virus (IMNV), and white spot disease (WS). We used data from 1839 cultivation cycles. The cycles are managed by 383 Farms. The data covered 4 physical parameters measured twice daily (in the morning and evening). Those parameters are water temperature, dissolved oxygen, salinity, and pH. The data also cover disease tests. We conducted several processes to develop the predictive model. First, we improve the data quality using the Generative Adversarial Network model (GAN). The improved data is then used for feature engineering and model training. We used the Random Forest Model as the predictor to the data we managed to achieve an average F1 score of 0.91 for the four diseases. The model achieved an F1 score of 0.91 for AHPND, 0.89 accuracy for WFD, 0.93 accuracy for IMNV, and 0.9 accuracy for WS. Those results indicate a good possibility to predict the disease occurrence based on water quality data. Hence the method can be used as an early warning system to help the farmer in mitigating disease occurrence.

Keywords: Shrimp Disease, Machine Learning and Shrimp Farming

Introduction

Aquaculture is one of the food sectors with the fastest growth rate. Amongst the various branches of aquaculture, shrimp culture has expanded rapidly across the world because of faster growth rate of shrimps, short culture period, high export value and demand in the market (Rahman et al., 2015). Shrimp farming is a significant source of livelihood for people in some countries (Dastidar et al., 2013). Shrimp is a nutritious food that is high in protein (20%) and contains essential vitamins, minerals, omega-3 fatty acids, antioxidants, and selenium [2]. It is also a delicious seafood option that can be prepared in various ways (Lifestyle Lounge: Health & Fitness 2012).

Indonesia is one of the largest shrimp producers in Southeast Asia, according to the Food and Agriculture Organization (FAO) (FAO, 2020). The shrimp farming industry in Indonesia began in the late 1980s in East Java and has since spread throughout the country. However, like other major shrimp farming countries, bacterial and viral diseases pose a threat to the sustainable development of shrimp farming in Indonesia, which can lead to severe economic losses affecting yield and survival rate (Sunarto et al., 2004; Walker et al., 2009; Ali et al., 2018).

The practice of high-density aquaculture, particularly shrimp farming, has led to an increase in the occurrence of diseases in shrimp. This situation has highlighted the necessity for regular laboratory

*Corresponding Author's Email: *lukman@jala.tech*

testing to analyze disease infection. However, routine checking poses challenges for small farms, especially when they lack access to laboratory facilities and face additional operational costs. Therefore, it is crucial to develop effective and quantitative measures to prevent and predict these diseases (Xiong, J et al., 2016).

Table 1: Variables in the Dataset

No.	Parameters	Description	Roles
1.	Pond area	Measured in square meter	Independent Variable
2.	Total seed	Total seed of shrimp (tails)	Independent Variable
3.	Daily feed	Total feed used for each pond in one day (kg)	Independent Variable
4.	Day of cultivation	The age of cultivation in days	Independent Variable
5.	Temperature	Measured in Celsius in the morning (3 to 9 am) and evening (17 to 21 pm)	Independent Variable
6.	Dissolved Oxygen	Measured in ppm in the morning (3 to 9 am) and evening (17 to 21 pm)	Independent Variable
7.	Salinity	Measured in ppm in the morning (3 to 9 am) and evening (17 to 21 pm)	Independent Variable
8.	pH	Measured in the morning (3 to 9 am) and evening (17 to 21 pm)	Dependent Variable
9.	Disease Occurrence	Type of diseases: ● AHPND ● WFD ● IMNV ● WSSV	Dependent Variable

MATERIALS AND METHODS

Dataset

The study utilized a dataset gathered from various sites across Indonesia. This dataset comprises 11040 sample measurements taken from 1839 cultivation cycles. These cycles were conducted in 1255 shrimp ponds from 2021 to 2022. The timing and duration of these cycles varied, with some ponds experiencing two or more cultivation cycles. The dataset includes several significant parameters, which are detailed in TABLE 1. Disease occurrence variables such as AHPND, WFD, IMNV, and WSSV were treated as dependent variables, while the remaining variables served as predictors.

Data Cleaning

Data cleaning was done to make sure that the data has good quality. In this work the dataset was cleaned through two steps, outlier handling and data imputation.

Outlier handling

Outlier was done to make sure that all of the data are valid and to exclude anomaly conditions. This research uses univariate gaussian distribution to detect outliers on each variable. Equation (1) shows probability distribution function (PDF) of gaussian distribution using mean (μ) and standard deviation (π) Zhang, X. (2011). For each variable, the two parameters were calculated using maximum likelihood estimation (MLE).

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad (1)$$

The obtained mean and standard deviation then used to estimate quantiles 5% and 95% of every variable. The obtained quantile values are then used to filter the data. In this research we only use the data if the value is within the range of quantile 5% and 95%.

Data imputation

This research used a generative adversarial network (GAN) to impute the missing data. GAN is a type of machine learning framework that can learn from a set of training data and generate new data with the same characteristics as the training data. For example, a GAN trained on images of human faces can create realistic-looking faces that do not exist in reality. A GAN consists of two neural networks: a generator and a discriminator. The generator takes a random seed as input and produces fake data, such as synthetic images or audio. The discriminator takes either real data from the training set or fake data from the generator as input and tries to classify them as real or fake (Goodfellow et al., 2014). The generator and the discriminator are trained in an adversarial manner, meaning that they compete against each other. The generator tries to fool the discriminator by generating more realistic data, while the discriminator tries to improve its accuracy by rejecting the fake data. The training process stops when the generator and the discriminator reach an equilibrium, where the discriminator cannot distinguish the real data from the fake data. A GAN is a powerful generative model that can capture complex patterns and distributions in the data and create novel and diverse samples. Figure 1 shows architecture of the Generative Adversarial Network used in this research. In this research both Generator and Discriminator used multilayer perceptron architecture with 2 Fully Connected Layer.

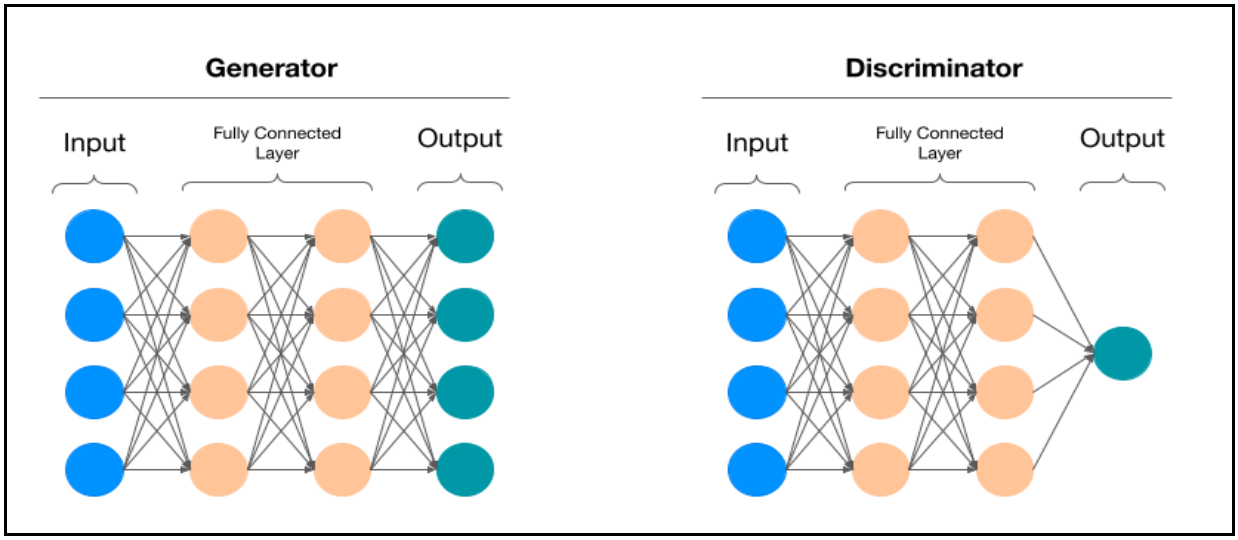

Figure 1: *Generative Adversarial Network Architecture*

Feature engineering

During feature engineering the data was transformed into new features that can be used in building predictive models. During this phase, seed density and 7-days windowed moving average was calculated.

Stocking density

This study utilized stocking density as a method to determine the density of shrimp in a pond. The stocking density is influential in the growth and survival rate of the shrimp (Marlina, e., et al, 2020), This parameter calculated with equation (2):

$$Stocking\ density = \frac{Total\ Seed}{Pond\ area} \tag{2}$$

With:

$$Total\ Seed = Number\ of\ stocked\ seed\ (Tails)$$

$$Pond\ area = Area\ of\ shrimp\ pond\ (m^2)$$

$$Stocking\ density = Number\ of\ stocked\ shrimp\ per\ squared\ area\ (Tails/m^2)$$

7-days windowing

We utilized a technique known as data windowing, a concept borrowed from time series analysis. This method involves the creation of a sliding window, either of fixed or variable size, that traverses the data, extracting segments of observations to be used as input variables. The input consists of a sequence of current and preceding time steps. The application of data windowing can aid in identifying temporal dependencies and patterns within time series data, while also reducing data dimensionality and noise. Furthermore, data windowing can be employed to resample data at varying frequencies - hourly, daily, or weekly - contingent on the objective of the analysis and the data available.

Kernelized Principal Component Analysis (KPCA)

Classical PCA algorithm aims at finding a linear subspace of lower dimension than the original space. KPCA is an extension of Principal Component Analysis (PCA). Unlike normal PCA, KPCA achieves non-linear dimensional reduction of data through kernel function. In this research, a polynomial kernel as explained in Shaft-Taylor (2011) was used. Steps of PCA can be found at Ezuwokwe (2019).

Z-score Normalization

In some datasets there are different ranges of values for each attribute. The difference in the range of the value might cause the malfunction of the attribute which has a much smaller value compared to other attributes (Henderi., et al, 2021). Hence transformation toward the dataset such as normalization is needed. Normalization is a way to adjust values measured in different scales to a notationally

common scale. Z-score normalization normalize values by using mean (μ) and standard deviation (σ). It can be calculated with equation (3) (Aldhyani et al., 2020):

$$Z\ Score = \frac{(x - \mu)}{\sigma} \tag{3}$$

Random Forest Classification

Random forest is a group of un-pruned classification or regression trees made from the random selections of samples of the training data. Random features are selected in the induction process based on the selected samples. Prediction is made by averaging the prediction of the ensemble from all of the trees (Ali and Ahmad 2012). The basic steps of random forest algorithm are follows (Xu., et al, 2021; Natekin and Knoll, 2013):

1. K sets of data are created from the training set data through bootstrap sampling with replacement. Each dataset is then split into two parts: sampled and un-sampled data. The sampled data is utilized during the training phase, while the un-sampled data is used during the testing phase. A decision tree is generated from each dataset during the training phase.
2. Every decision tree is trained using the training data. At each node, a random selection of *m* features is made. The best features are chosen based on the Gini metric.
3. Each decision tree that has been created is evaluated using the un-sampled data. The prediction error observed during this phase is utilized to identify the most accurate decision tree.
4. The decision tree models determined from each dataset are utilized for making predictions. The predicted value is derived by calculating the average of the prediction results produced by these determined decision tree models.

Results and Discussion

This section provides a foundation for understanding the effectiveness of developed machine learning models. The following discussion will also delve deeper into the implications of these outcomes, their significance in the field of aquaculture, and how they can guide future research. The discussion will start from results of data imputation using GAN and then proceed to accuracy of shrimp disease prediction model.

Data imputation with generative network

In the preceding sections of this research paper, we have discussed the methodology and implementation of data imputation using a Generative Adversarial Network (GAN). The results of this process are visually represented in the accompanying graphs, which compare the data distribution before and after imputation.

Figure 2 depicts temperature, dissolved oxygen (DO) levels, salinity, and pH for 2 measurement times (morning and evening). Each graph contains two lines: one representing the actual data (blue line) and the other representing the synthetic data generated by the GAN (orange line). These graphs provide a clear visual representation of how closely the synthetic data aligns with the actual data, demonstrating the effectiveness of our GAN-based imputation method.

We also try to evaluate the imputation results quantitatively using KSKomplement. The KComplement of data imputation shown by *TABLE 2*. Evaluation with KSComplement showed that the imputation with synthetic data managed to represent the natural distribution of the original data. It means that the imputation data can be used for further process.

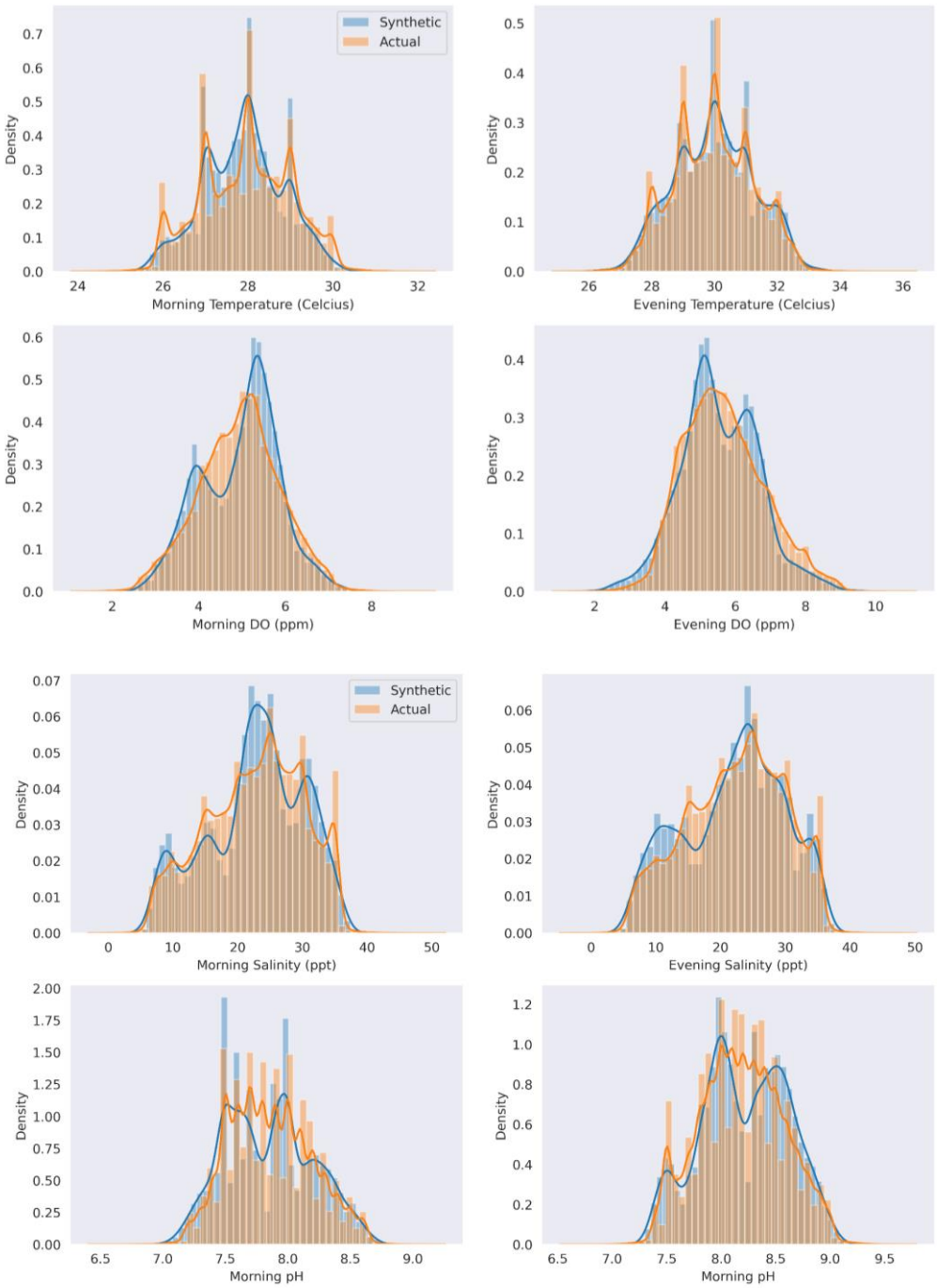

Figure 2: Results of Data Imputation with GAN

Table 2: KSComplement of Data Imputation Results

Parameter	KSComplement	
	Morning Measurement	**Evening Measurement**
Temperature	0.912	0.96
DO	0.933	0.942
Salinity	0.935	0.950
pH	0.911	0.915

Shrimp disease prediction model

This part will discuss the accuracy of prediction models. First we will discuss the performance of the model over several trials with different data splits. Figure 3 shows F1 scores of prediction models on 4 trials. In all four trials the model managed to achieve F1 scores higher than 0.85 both in the training set and test set with lowlight there is a noticeable gap between performance on the training set and the test set.

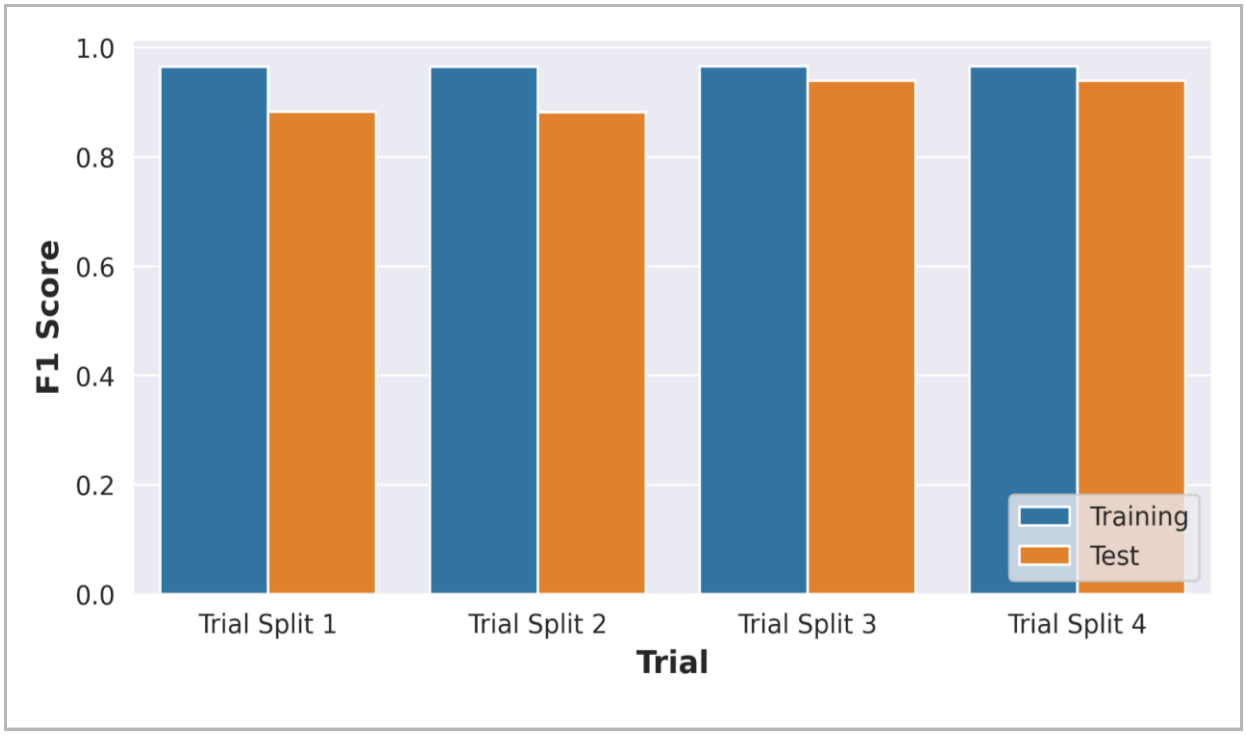

Figure 3: F1 of best model on different data split

Table 3: F1 score of AHPND Prediction on Test Set

	Metrics		
	Precision	**Recall**	**F1-Score**
Infected	0.94	0.92	0.865
Not Infected	0.95	0.93	0.883
Avg. Accuracy	**0.945**	**0.925**	**0.87**

AHPND Prediction

TABLE 3. Depict precision, recall, and F1-score metrics for the prediction of Acute Hepatopancreatic Necrosis Disease (AHPND). The metrics are calculated for two classes: “Infected” and “Not Infected”. The precision, recall, and F1-score for the “Infected” class are all above 0.86, indicating a high accuracy in predicting the infected cases. Similarly, for the “Not Infected” class, these metrics are also above 0.88, suggesting a high accuracy in predicting the non-infected cases as well. The table also provides the average accuracy of the model, which is calculated by averaging the precision, recall, and F1-score of both classes. The average accuracy is also above 0.93, demonstrating that the model performs well in predicting both infected and non-infected cases. This high level of accuracy suggests that the model could be a reliable tool for predicting AHPND infection.

IMNV

The performance of the model in prediction IMNV occurrence is evaluated using several metrics, including Precision, Recall, and F1-Score. The metrics are shown in *TABLE 4*. .For the "Infected" class, our model achieved a Precision of 0.94, a Recall of 0.82, and an F1-Score of 0.89. This indicates that our model is highly accurate when predicting infected instances and is able to correctly identify a significant proportion of all actual infected instances. For the "Not Infected" class, the model achieved even higher scores with a Precision of 0.93, a Recall of 0.97, and an F1-Score of 0.90. These results highlight the model's effectiveness in correctly predicting not infected instances and its ability to identify the majority of actual not infected instances.

On average, across both classes, our model's predictions were correct 89% of the time, as indicated by the average accuracy score of 0.89. These results demonstrate the potential of using machine learning techniques in disease prediction in aquaculture. The combination of kernel PCA and Random Forest not only provides accurate predictions but also offers insights into the important features contributing to shrimp health. This research contributes to the broader goal of improving health management practices in shrimp farming.

Table 4: F1 score of IMNV Prediction on Test Set

	Metrics		
	Precision	**Recall**	**F1-Score**
Infected	0.94	0.82	0.89
Not Infected	0.93	0.97	0.90
Avg. Accuracy	**0.93**	**0.89**	**0.89**

WFD Prediction

TABLE 5 presents the performance metrics of a machine learning model that predicts shrimp disease. The model, which combines Kernel PCA and Random Forest (RF), was evaluated on a test set using three metrics: Precision, Recall, and F1-Score. For the 'Infected' class, the model achieved a Precision of 0.93, a Recall of 0.67, and an F1-Score of 0.62. For the 'Not Infected' class, the model achieved a Precision of 0.93, a Recall of 0.96, and an F1-Score of 0.89. The average accuracy across both classes was found to be quite high with a Precision of 0.93, a Recall of 0.81, and an F1-Score of 0.75. This suggests that the model performs well in predicting both infected and non-infected cases, with a slightly better performance in predicting non-infected cases.

Table 5: F1 score of WFD Prediction on Test Set

	Metrics		
	Precision	**Recall**	**F1-Score**
Infected	0.93	0.67	0.62
Not Infected	0.93	0.96	0.89
Avg. Accuracy	**0.93**	**0.81**	**0.75**

From the precision there are indications that the model managed to avoid false positives nicely. However, the recall score of infected cases indicates that the model still has problems in detecting positive cases. In other words, when the model predicts that a shrimp is infected, it's usually correct, but it also misses a lot of infected shrimps that it labels as not infected. This could be problematic in this case because failing to identify an infected shrimp could lead to the spread of the disease. Adjustment of the model or consider using a different one that can improve recall without sacrificing too much precision.

WSSV Prediction

The last disease that we tried to predict using this method is WSSV. *TABLE 6.* The F1 score of WSSV Prediction on Test Set" presents the performance metrics of our machine learning model for predicting White Spot Syndrome Virus (WSSV) in shrimps. The metrics include Precision, Recall, and F1-Score for both "Infected" and "Not Infected" classes, as well as the average accuracy. For the "Infected" class, the Precision is 0.94, indicating that when the model predicts an instance is infected, it is correct 94% of the time. The Recall is 0.82, meaning the model correctly identifies 82% of all actual infected instances. The F1-Score is 0.89, which is the harmonic mean of precision and recall, providing a single metric that balances both considerations.

Table 6: F1 score of WSSV Prediction on Test Set

	Metrics		
	Precision	**Recall**	**F1-Score**
Infected	0.92	0.88	0.80
Not Infected	0.94	0.96	0.90
Avg. Accuracy	**0.93**	**0.92**	**0.87**

For the "Not Infected" class, the Precision is 0.93, which means when the model predicts an instance is not infected, it is correct 93% of the time. The Recall is 0.97, meaning the model correctly identifies 97% of all actual not infected instances. The F1-Score is 0.90, balancing precision and recall for this class. The last row shows the average accuracy across both classes, with all metrics being 0.89. This suggests that on average, the model's predictions are correct 89% of the time. These results indicate that our combination of kernel PCA and random forest techniques has performed well in predicting WSSV in shrimps. It has particularly high accuracy in identifying not infected instances, while still performing reasonably well for infected instances. This research contributes to improving health management practices in shrimp farming by providing a reliable tool for early detection of WSSV.

Conclusion

The practice of high-density aquaculture, especially shrimp farming, has resulted in a rise in the incidence of diseases in shrimp. The issue raises the importance of shrimp disease monitoring. However, frequent monitoring increases the operational cost of the cultivation.Therefore, it is crucial to develop effective and quantitative measures to prevent and predict these diseases. This research tried to build a disease occurrence prediction model based on physical water quality measurement data. The prediction functions as probability estimation of shrimp disease occurrence. The results showed that the model was able to predict shrimp disease occurrence with an F1 score higher than 0.85 for 4 types of shrimp disease (AHPND, WFD, WSSV, IMNV). It means that the model can be used as an alternative way to monitor shrimp disease occurrence besides laboratory tests.

References

Aldhyani .H.H., Al-Yaari M., Alkahtani H., Maashi M. 2020. Water Quality Prediction Using Artificial Intelligence Algorithms. Water Quality Prediction Using Artificial Intelligence Algorithms. 2020.

Ali, J., Ahmad, N.(2012). Random Forest and Decision Trees. International Journal of Computer Science Issues. 9.(3).

Ali, H., Rahman, M. M., Rico, A., Jaman, A., Basak, S. K., Islam, M. M., Khan, N., Keus, H. J., & Mohan, C. V. (2018). An assessment of health management practices and occupational health hazards in tiger shrimp (*Panaeus monodon*) and freshwater prawn (*Macrobrachium rosenbergii*) aqualculture in Bangladesh. *Veterinary and Animal Science.* 5. 10-19.

Arnaud de Myttenaere, Boris Golden, Bénédicte Le Grand, Fabrice Rossi.(2016).Mean Absolute Percentage Error for regression models. Neurocomputing. 192. 38-48.

Ezuwokwe K., Zareian S.J. Kernel Methods For Principal Component Analysis (PCA).

John Shawt-Taylor, Nello Cristianini. 2011. Kernel Methods for Pattern Analysis. Cambridge University Press,ISBN:9780511809682, 47-83.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. http://arxiv.org/abs/1406.2661.

Marlina Eulis, Hartono Puji Dwi, Panjaitan Imelda. 2020. Optimal Stocking Density of Vannamei Shrimp Litopenaeus Vannamei at Low Salinity Using Spherical Tarpaulin Pond. Advances in Social Science Education and Humanities Research. 298.

Natekin, A., Knoll, A.(2013). Gradient Boosting Machines, a tutorial. Frotiers in Neurotobitcs. 7 (21).

Ostasevicius V, Paleviciute I, Paulauskaite-Taraseviciene A, Jurenas V, Eidukynas D, Kizauskiene L. Comparative Analysis of Machine Learning Methods for Predicting Robotized Incremental Metal Sheet Forming Force. Sensors (Basel). 2021 Dec 21;22(1):18. doi: 10.3390/s22010018. PMID: 35009560; PMCID: PMC8747513.

Vinod Kothari, Suman Vij, SuneshKumar Sharma, Neha Gupta. Correlation of various water quality parameters and water quality index of districts of Uttarakhand. Environmental and Sustainability Indicators. 9. 100093.

Walker, P. J., & Mohan, C. V. (2009). Viral disease emergence in shrimp aquaculture: origins, impact and the effectiveness of health management strategies. *Reviews in Aquaculture*, 1, 125-154.

Xu, J.; Xu, Z.; Kuang, J.; Lin, C.; Xiao, L.; Huang, X.; Zhang, Y. 2021. An Alternative to Laboratory Testing: Random Forest-Based Water Quality Prediction Framework for Inland and Nearshore Water Bodies. Water 2021. 13. 3262. https://doi.org/10.3390/ w13223262