# Using the Term Frequency-Inverse Document Frequency for the Problem of Identifying Shrimp Diseases with State Description Text

Luyl-Da Quach[1], Anh Nguyen Quynh[2], Khang Nguyen Quoc[3], An Nguyen Thi Thu[4]
FPT University, Cantho city, Vietnam[1, 2, 3]
RMIT University, Ho Chi Minh city, Vietnam[4]

***Abstract*—With the increasing demand for research on shrimp disease recognition to assist far-off farmers who need the proper assistance for their shrimp farming, shrimp disease prediction research is still in the initial stage. Most current methods utilize vision-based models, which mainly face challenges: symptom detection and image quality. Meanwhile, there are few researches which are language-based to get over the issues. In this study, we will experiment with natural language processing based on recognizing shrimp diseases; based on descriptions of shrimp status. This study provides an efficient solution for classifying multiple diseases in shrimp. We will compare different machine learning models and deep learning models (SVM, Logistic Regression, Multinomial Naive Bayes, (a4) Bernoulli Naive Bayes, Random forest, DNN, LSTM, GRU, BRNN, RCNN) in terms of accuracy and performance. The study also evaluates the TF-IDF technique in feature extraction. Data were collected for 12 types of shrimp diseases with 1,037 descriptions. Firstly, the data is preprocessed with standardised Vietnamese accent typing, tokenized words, converted to lowercase, removed unnecessary characters and stopwords. Then, TF-IDF is utilized to express the text feature weight. Machine learning-based and deep learning-based models are trained. The experimental results show that Random forest (F1-Score micro: 98%) and DNN (Validation accuracy: 84%) are the most efficient models.**

***Keywords*—TF-IDF; machine learning; deep learning; CNN; shrimp disease classification***

## I. INTRODUCTION

Shrimp is the most commonly consumed worldwide, accounting for 15.5% of total global aquaculture production. Production of shrimp-related goods has increased globally over the past 16 years (from 2000 to 2016) by more than 20%. In which, Vietnam accounted for 9% of total export value (ranked 2nd in the world)[1]. Shrimp farming production has increased significantly to meet a market need, but needs to increase more to keep up with the demand[2]. Research [3] shows that the top five shrimp-farming countries are Ecuador, India, Indonesia, Thailand and Vietnam.

Shrimp farming also faces many challenges due to resource use issues in shrimp farming and shrimp diseases. Diseases in shrimp aquaculture have hindered the sector from growing, directly influencing management and production costs. In 2012, an infectious disease, acute hepatopancreatic necrosis syndrome (AHPNS) or early mortality syndrome (EMS), severely damaged the shrimp farming region, with one-sixth of the area in Thailand and Vietnam affected[4]. In 2013, in India, the epidemic caused a loss of INR 10,221 million, 48,717 tons and 2.15 million working days were lost[5]. In 2016, white spot disease caused more than $8 billion in damage and cumulative mortality can reach 90-100% over a period of 3 to 10 days[6].

Most of the current studies have focused on using image processing or machine learning methods to detect shrimp diseases based on the visual features of shrimp. There is research related to disease recognition in shrimp, such as biochemical diagnosis or image-based AI methods. Firstly, diseases can be detected by biochemical techniques performed in the laboratory with measures such as CRISPR to detect white spot virus, acute hepatopancreatic necrosis disease (AHPND)[7], [8], amplification of recombinase polymerase to detect acute hepatopancreas, hemocyte iridescent virus [9], [10], PCR method to detect AHPND[11], etc. There is also an advance in shrimp disease identification methods by applying artificial intelligence algorithms on shrimp disease images, such as: using CNN architecture to create ShrimpNet[12]–[14] to detect yellow head disease in shrimp. This shows that the result of detection of diseases in shrimp is quite impressive, but there are limitations in terms of time and precision. These methods require high-quality images of shrimp and may fail to capture the subtle or complex symptoms that are expressed in natural language. Moreover, these methods may not be easily accessible or affordable for small-scale farmers who lack the necessary equipment or expertise.

Because of the above reason, we propose a novel approach to detect shrimp diseases from text, based on the textual symptom descriptions that can be obtained from various sources. There are also a few papers that have performed text-based classification of shrimp diseases, but they only used a few machine learning models and achieved not high results[32]. Therefore, we conduct this research with different machine learning models and deep learning models (SVM, Logistic Regression, Multinomial Naive Bayes, Bernoulli Naive Bayes, Random forest, DNN, LSTM, GRU, BRNN and RCNN), and improve the accuracy. We hope that our research will contribute to the advancement of NLP applications in the field of aquaculture and provide a valuable tool for shrimp farmers and experts to diagnose shrimp diseases in an efficient and reliable way. We perform in-depth exploratory research on:

- Shrimp disease data is collected from texts describing the status of shrimp on the farm. This is something that farmers often describe in daily farm management through observation. The study collected 1,037 symptom descriptions of 12 common diseases in Vietnamese (attached data set). For example, the description is "Shrimp is a weak, limp, soft shell, intestines without food, swimming sluggishly on the water surface, on the shore, slow to grow; hepatopancreas is more yellow than usual; gills, tail swollen."
- Data preprocessing with standardised Unicode and Vietnamese accent typing, tokenising words, converting to lowercase, removing unnecessary characters and stopwords.
- Then, the study combined TF-IDF technique with machine learning (ML) and deep learning (DL) algorithms.

The remainder of this article is organized as follows: in Section II, we give a brief review of this domain research and re-evaluate the descriptive dataset of related studies. Section III describes the data processing steps, the dataset[1], TF-IDF technique, training process and popular ML/DL algorithms. The experimental method is described in Section IV. The results after implementing the system are covered in Section V. The discussion and conclusion are presented in Section VI and Section VII respectively.

## II. Literature Review

### A. Traditional Methods

Currently, a variety of techniques are utilized to identify diseases. The issue in this field can be broken down into two different classification techniques: biochemistry and AI-based. Many researchers focus their work on computer vision for categorization using AI; moreover, more attempts have been made with text recognition, particularly regarding shrimp diseases.

The conventional method, which is often used, relies on visual inspection, observation, and testing on shrimp samples to identify the disease's presence[15]. High accuracy is used when using this technique. The drawback of this approach is that it necessitates a laboratory, time, and specialist topic knowledge.

Another option is to utilise test kits, such as bacterial test kits, viral test kits, and antigen test kits to detect infections in shrimp. This method has several advantages, including high accuracy, simplicity, speed, and convenience, but it also has disadvantages, including the difficulty of distinguishing many different diseases.

The following technique, which uses PCR and immuno-antibody analysis, is based on genetic analysis to identify disease resistance. RFLP (Restriction Fragment Length Polymorphism)[16], [17], RAPD (Random Amplification of Polymorphic DNA)[18], and SSR (Simple Sequence Repeat)[19], [20] are examples of specific approaches. The benefits of this procedure are the same as those of test kits and conventional methods. But it still has high cost, complexity, and inability to detect all diseases.

To sum up, the biochemical approach also has advantages for swiftly detecting diseases in shrimp. However, it has the drawback of requiring time to assess the severity of the sickness and choose the best approach.

### B. Method of using Artificial Intelligence

Artificial intelligence is an approach that has been widely used recently, especially in research on shrimp disease classification based on images, genetic data, chemical data, etc. This method can be divided into different categories. The first is an image-based shrimp disease classification method using deep learning models such as Convolutional Neural Network (CNN)[12], [13], [21], YOLO model [22], and machine learning[23], [24]. The effectiveness of this strategy depends mainly on the picture size and quality. The environment is the major obstacle affecting the accuracy of the results and data processing. Next, the study uses machine learning algorithms to accurately detect the disease in shrimp by identifying its early symptoms. This shows that it is possible to identify diseases in shrimp using natural language processing techniques and methods.

Some encouraging findings have been made regarding the classifying diseases using natural language processing. In the medical field, more than 20,000 findings use an NLP-based approach to classify diseases. The majority of trials[25]–[28] produced accurate disease diagnosis outcomes. There are about 9,000 studies agriculture employing NLP to enhance agricultural development and productivity. In studies on crops [29], [30], rice[31] have achieved promised accuracy (over 90%) by disease description and support chatbot system. In addition, research [32] has shown the first steps in approaching using NLP to identify diseases in shrimp with basic machine learning techniques with an accuracy of over 80%. It proved that the use of NLP in diagnosing of shrimp diseases is essential.

NLP techniques have been increasingly developed. Among them, there are processing techniques such as tokenization to divide sentences or paragraphs into smaller ones, stop word removal from removing words that have no critical meaning, stemming used to remove suffixes from meaningless words, lemmatization to return words to their infinitive form (lemma), part-of-speech (POS) tagging to classify each word in a sentence into word type parts, named entity recognition (NER) to recognize and classify named objects, sentiment analysis to analyze emotions in text, topic modelling to find the main themes in a corpus, word embeddings to convert words into vectors[33]. Data processing techniques in NLP are commonly used in text classification, automatic translation, chatbots and many other fields.

In conclusion, research into several sectors demonstrates the effectiveness of NLP in text processing. However, no studies currently combine NLP data processing methods with ML and DL analyses to identify shrimp diseases. As a result, data collecting and diagnostic processing related to shrimp

[1]Dataset: https://github.com/nqanh312/shrimp-diseases-dataset.

sickness are crucial, which is considered the study's novelty, which makes an important contribution to the establishment of a system to answer farmers' questions regarding shrimp diseases (Fig. 1).

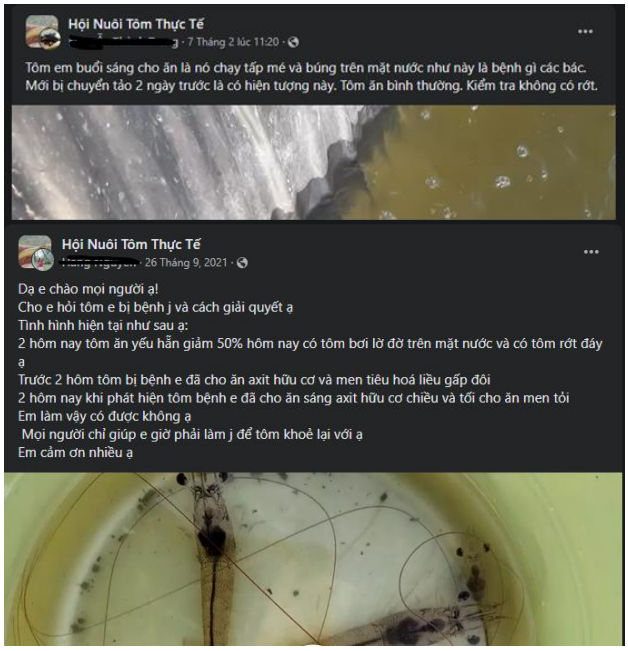



Fig. 1. Some statuses of shrimp farmers describe to seek treatment on the social network Facebook.

## III. Materials and Methods

In this paper, we conduct a research on term frequency/inverse document frequency (TF-IDF) approach to extract characteristic words to construct the sentence embedding. Then, we apply the sentence embedding based machine learning and deep learning to categorize the documents into 1 type of diseases. Our proposed method comprises of following steps as shown in Fig. 2.

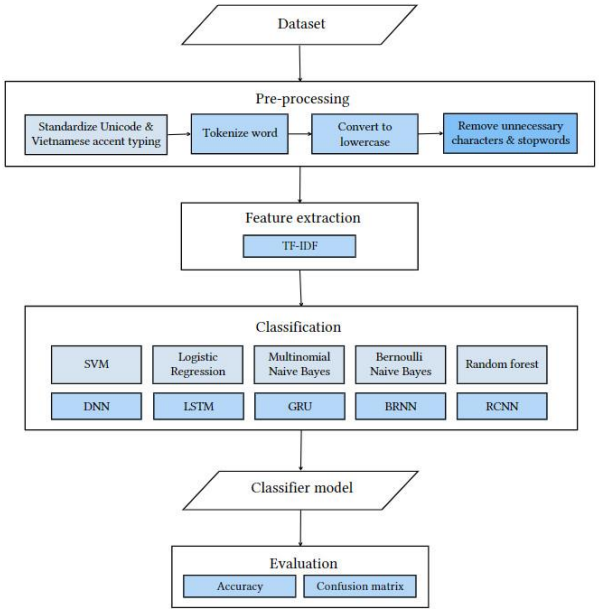



Fig. 2. System components are proposed in this research.

### C. Dataset

We prepared a novel dataset which was collected from the internet, the majority of which came from certain aquaculture pages, as there isn't an available dataset of infected shrimp. We perform statistics on the frequency of occurrence of the words shown in Fig. 3.

The collection of 1,037 documents includes descriptions of diseased shrimps, as seen in Table I. We gathered altogether, which are then utilized to train and test our model.

### D. Data Pre-Processing

Preparing sentences for analysis is a step that transforms an input into understandable data. Steps that sentences pass:

- All of the data is processed using the most straightforward and efficient method of text preprocessing—lowercase.
- Stemming is a heuristic technique that correctly transforms words into their root form by chopping off the ends of words.
- Stopword removal: removes poor information terms from the text so that the work can concentrate on the keywords.
- Turning a text into a standard form is known as normalization. This stage removes distracting text elements, including abbreviations, typos, and words that are not commonly used.
- Remove any letters, numbers, or text fragments that can interfere with text analysis by doing noise reduction. The result is shown in Fig. 4.

TABLE I. Statistics of Data Used in this Research

| No. | Name | Quantity |
|---|---|---|
| 1 | Acute hepatopancreatic necrosis | 74 |
| 2 | Black gill | 77 |
| 3 | Filamentous bacterial | 77 |
| 4 | Infection with vibrio | 91 |
| 5 | Infectious myonecrosis | 101 |
| 6 | Loose shell | 139 |
| 7 | Luminous bacteria disease | 81 |
| 8 | Plaque disease in shrimp | 90 |
| 9 | Taura | 72 |
| 10 | VitaminC deficiency in shrimp | 78 |
| 11 | White feces | 79 |
| 12 | Yellow head | 78 |
| **Total** | | **1.037** |

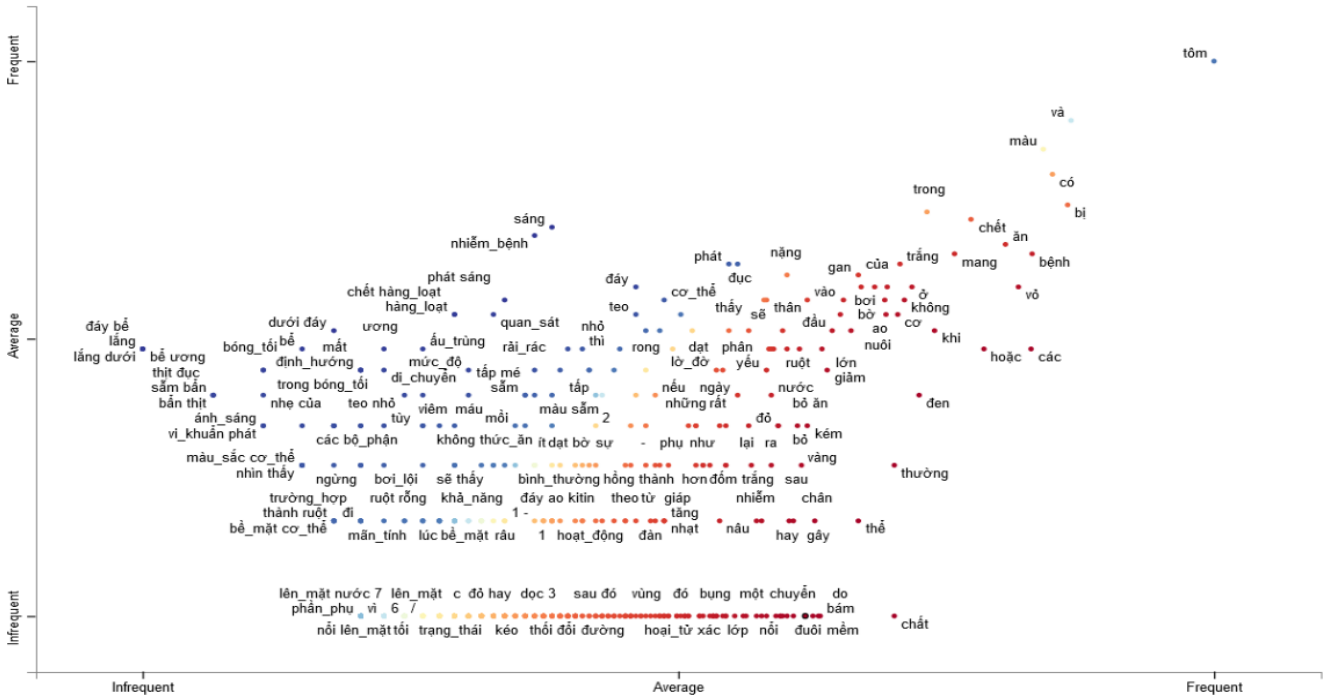



Fig. 3. List of words and its frequency.

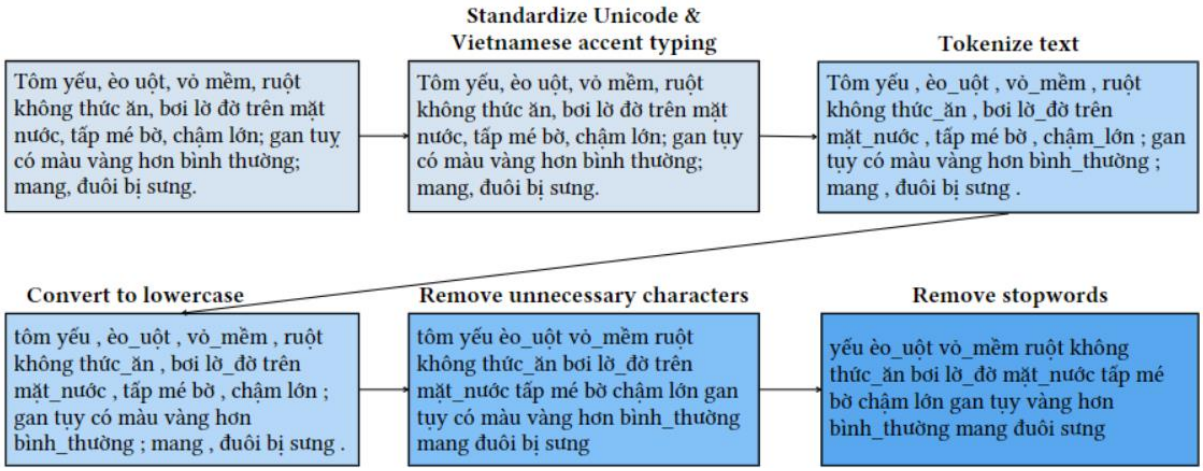



Fig. 4. The result of the preprocessing.

### C. Feature Extraction with TF-IDF

By converting text into vector space (VSM) or sparse vectors, the TF-IDF method is frequently used in text data mining to estimate a word's significance [34]. Sentence embedding:

- The number of words in a set of documents is counted by TF-IDF. We typically assign a score to each word to show how much weight it has in the corpus and document. This approach is every now and again utilized in data recovery and text mining.
- Frequency of Term (TF): Using the heuristic that a term's importance is proportional to how frequently it appears in a document, it shows how important a word is to a document.

IDF:

- Outlines the genuine significance of an expression. It is pointless for terms like stopwords, which frequently show up in certain records, to be critical (the, that, of, and so on). Stopwords ought to have stayed away since they dark the unique circumstance. They are ignored by IDF because of the way it operates.
- It penalizes the word used frequently throughout documents.
- The appropriate term receives a more excellent IDF score, while the stopword receives a lower weight.

In this research, the frequency of words related to shrimp diseases $tf_{Shp}$ is the number of words $t_{Shp}$ compared to $d_{Shrp}$.

$$tf_{Shp}(t_{Shp}, d_{Shrp}) = \frac{f_{Shrp}(t_{Shrp}, d_{Shrp})}{max\{f_{Shrp}(w_{Shp}, d_{Shp}) : w_{Shrimp} \in d_{Shp}\}} (1)$$

In which:

- $tf_{Shp}(t_{Shp}, d_{Shp})$: term frequency of $t_{Shp}$ in $d_{Shp}$.
- $f_{Shp}(t_{Shp}, d_{Shp})$: number of the $t_{Shp}$ appears in $d_{Shp}$.
- $max\{f_{Shp}(w_{Shp}, d_{Shp}) : w_{Shp} \in d_{Shp}\}$: the maximum of number of terms related to shrimp diseases in $d_{Shp}$.

The IDF of a term indicates the percentage of corpus documents that contain the term. Words that are only found in a small number of documents, such as technical jargon terms, are given more excellent relevance ratings than words that are used in all documents, such as a, the, and. $idf_{Shp}(t_{Shp}, D_{Shp})$ is calculated as the following formula:

$$idf_{Shp}(t_{Shp}, D_{Shp}) = log \frac{|D_{Shp}|}{|\{d_{Shp} \in D_{Shp} : t_{Shp} \in d_{Shp}|} \quad (2)$$

In which:

- $idf_{Shp}(t_{Shp}, D_{Shp})$ : inverse document frequency $idf_{Shrimp}$ of term $t_{Shp}$ in $D_{Shp}$.
- $|D_{Shp}|$: number of document in the corpus $|D_{Shp}|$
- $|\{d_{Shp} \in D_{Shp} : t_{Shr} \in d_{Shp}|$ : number of documents $d_{Shp}$ in the corpus $D_{Shp}$ contain the term $t_{Shp}$.

The $tf - idf(t_{Shp}, d_{Shp}, D_{Shp})$ is calculated by multiplying TF and IDF scores:

$$tf - idf(t_{Shp}, d_{Shp}, D_{Shp}) = tf_{Shp}(t_{Shp}, d_{Shp}) \times idf_{Shp}(t_{Shp}, D_{Shp}) \quad (3)$$

### D. Training

The classification model will be trained using machine learning (ML) algorithms and deep learning (DL) models using the training dataset. The model will learn from labeled data, which have been digitized into feature vectors through feature extraction. On this processed data set, parameters will be learned and optimized by machine learning and deep learning algorithms. The classification model will receive data (extracted features) after learning to predict results and return the appropriate label as a result (Fig. 5).

In addition to dividing the models/algorithms used into two types of deep learning/machine learning. They can be divided into three other categories based on structured data, i.e. regression algorithms, binary classifiers and multiclass classification algorithms on structured data.

Types of regression algorithms include Linear Regression (LiR), Random Forest (RF). LiR is used to predict the value of a continuous variable based on different input variables[35]. The RF algorithm builds multiple decision trees and combines their predictions to make the final prediction[36].

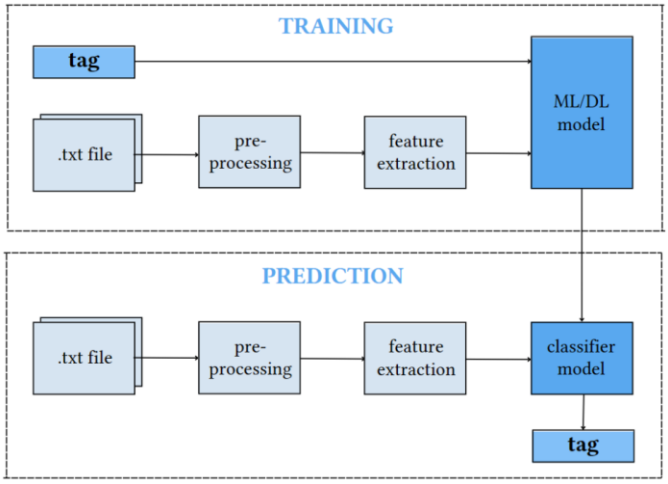



Fig. 5. Illustrate the process of using ML/DL for training and recognition.

Types of binary classification algorithms include Logistic Regression (LoR), and Bernoulli Naive Bayes (BNB). LoR is based on the sigmoid function to predict the probability of a linearly combined data sample of the input feature; the advantage of this algorithm is that it is simple and can explain the results[37]. Similar to LoR, BNB calculates the probability of each input feature, but it will conditionally consider each class based on the probabilities[38].

Types of multi-class classification algorithms on structured data include Support Vector Machine (SVM), Deep Neural Network (DNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional Recurrent Neural Network (BRNN), Recurrent Convolutional Neural Network (RCNN), Multinomial Naive Bayes (MNB). In which SVM classifies data by finding the best boundary[39], DNN uses many hidden layers to learn complex features of data [40]. LSTM uses an artificial neural network to record and store previous information to predict outcomes[41]. GRU has similar characteristics to LSTM. Still, it uses fewer parameters[42], BRNN uses two symmetric neural networks to learn data features [43], RCNN combines recurrent neural network and convolutional neural network to process sequence data[44], MNB is also based on the assumptions of BNB but features independent and differentiated input Multinomial distribution on each class[45].

In general, each predictive model has its advantages and limitations, depending on the specific data set and intended use. However, these models are all predictive and can be applied to the dataset after vectorization using the TF-IDF technique.

## IV. Evaluation

To evaluate the performance of the model, we use a confusion matrix which comprises four building blocks, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

TP and TN allude to situations where the forecasts are exact and negative. Positively false predictions are called FP, while negative false predictions are called FN. We use the confusion matrix to evaluate our model to generate additional

distinct metrics. Accuracy, Precision, Recall, and the F1 score are the precise metrics calculated using the formulas below.

*1) Accuracy*: Accuracy is one of the evaluating metrics and is informally interpreted in Eq. (4). It is the proportion of specifically classified shrimp diseases and the total number of shrimp diseases in the test set. The result shows how good the model works. The higher score of accuracy the more accurate our method is:

$$Accuracy = \frac{TruePositive+TrueNegative}{TotalPredictions} \quad (4)$$

*2) Precision:* The proposition of classified disease (TP) and the ground truth (the sum of TP and FP) defines the precision. It calculates the percentage of accurately classified disease as Eq. ( 5).

*3) Recall:* The percentage of correctly classified disease among all diseases belonging to that class Eq. (6).

*4) F1-score:* The following Eq. (7) is used to calculate the metric: the symphonic average of precision and recall. F1-score will be in (0,1], as F1 score higher as better model is.

$$Precision = \frac{TruePositive}{TruePositive+Falsepositive} \quad (5)$$

$$Recall = \frac{TruePositive}{TruePositive+Falsenegative} \quad (6)$$

$$F1-Score = 2\frac{Precision*Recall}{Precision+Recall} \quad (7)$$

In this study, we used Micro avg: F1-score is calculated by considering all classes' total number of true positives, false negatives and false positives. This method is often used to measure the correct prediction ratio of the model over the entire dataset. This research belongs to the case of a multi-class classification problem with the condition that each sample belongs to only one class; test accuracy will be equal to the F1-score micro.

The confusion matrix shows and summarizes a better view of the performance of a classification algorithm. It evaluates the results of the classification problem by considering both the accuracy and generality of the prediction for each class. The accurate/false prediction is shown as a percentage of each class.

## V. Result

This experiment is based on Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz, RAM 8 Gbytes. In this research, we adopted machine learning and deep learning method in natural language processing to classify deceases of shrimps. The model's performance was evaluated using accuracy and F1-score micrometric on the validation and test dataset. The result is shown in Table II.

According to the Table II, it is found that:

- Machine learning method: SVM has the highest accuracy score on the validation data (0.83), while Random forest obtained the highest F1-score micro on the test data (0.96). This shows that SVM works for matching better while Random forest has better generalization
- Deep learning: DNN has the highest accuracy score on both validation and test set. It is proved in Table II that DNN outperforms other models in shrimp classification. Other models, such as LSTM, GRU, BRNN, and RCNN, all have close results, with the micro F1-score ranging from 0.93 to 0.97.

Moreover, we used a confusion matrix for each algorithm and model to evaluate the confusion among shrimp diseases. Based on Fig. 6, the research found that the confusion of the models on acute hepatopancreatic necrosis, Filamentous bacteria, Infection with vibrio, Taura, and Vitamin C deficiency in shrimp, White feces, Yellow heads are quite low, while Plaque disease in shrimp confusion prediction obtained a high rate. The results prove that DNN is the most suitable model for shrimp disease classification.

TABLE II. Accuracy and F1-Score Results of ML Algorithms and DL Models on Shrimp Disease Dataset

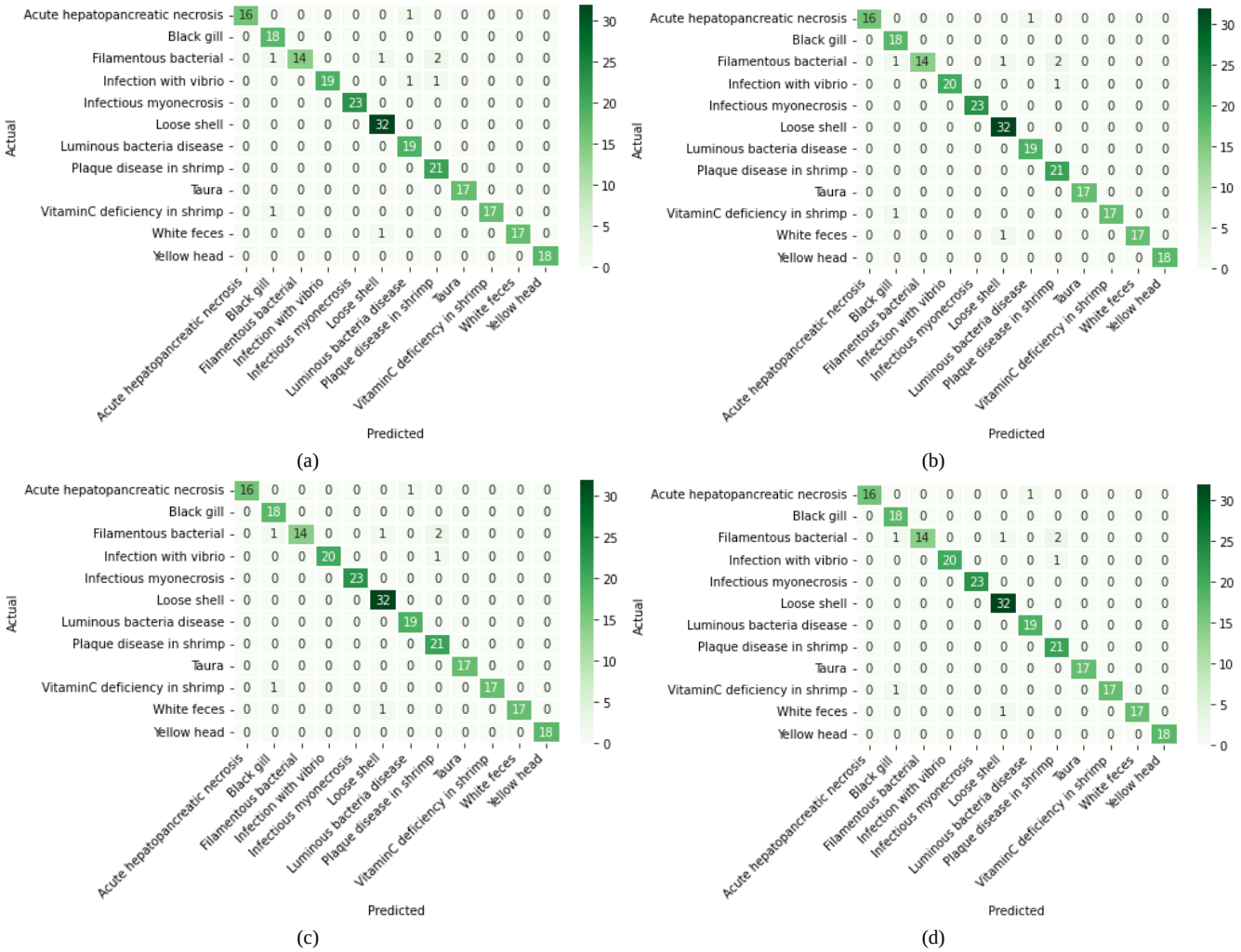
| Methods | SVM | LoR | MNB | BNB | RF | DNN | LSTM | GRU | BRNN | RCNN |
|---|---|---|---|---|---|---|---|---|---|---|
| Validation accuracy (without TF-IDF) | 0.725 | 0.727 | 0.724 | 0.646 | **0.779** | 0.738 | 0.542 | 0.479 | 0.665 | 0.671 |
| Validation accuracy (using TF-IDF) | 0.825 | 0.800 | 0.725 | 0.763 | 0.788 | **0.838** | 0.575 | 0.575 | 0.788 | 0.750 |
| F1-score micro (without TF-IDF) | 0.896 | 0.858 | 0.808 | 0.704 | 0.917 | **0.971** | 0.779 | 0.742 | 0.854 | 0.842 |
| F1-score micro (using TF-IDF) | 0.963 | 0.908 | 0.821 | 0.767 | **0.975** | 0.971 | 0.929 | 0.950 | 0.967 | 0.971 |

## VI. Discussion

In this section, we will discuss the results obtained and the remaining limitations of the study. Based on the accuracy and validation accuracy results, the research finds that the test data set has a different distribution from the validation dataset and is more suitable for the model, leading to a difference in the error in the test data model. However, when comparing ML algorithms and DL models, the performance of DL is higher in this problem. This can be explained by the ability of DL models to learn complex and semantic features of expressions. DL models also have the advantage of dealing with unbalanced data so that minority classes can be correctly classified.

In machine learning algorithms, the algorithm with the highest results on the test set is RF. The algorithm can minimise overfitting and increase the diversity of decision trees. The SVM, LoR and NB algorithms are all based on linear classifier architecture. Although these algorithms are easy to implement, they are not suitable due to the nonlinearity of shrimp disease data.

Deep Neural Network gives the best results among deep learning models because of their ability to represent non-linear features of the data. However, this model is challenging to train and adjust parameters and does not use sequence information of disease expression description in shrimp. Thus may need to understand the significance of this model regarding contextual meaning. The models based on Recurrent Neural Network architecture (LSTM, GRU, BRNN and RCNN), although capable of using information about the sequence of expression, have problems of vanishing gradient or exploding gradient when training because of long-term dependence between time steps in the data series (if the link weights between steps are too small or too large, the gradient will disappear or explode when propagating back through the long sequence).



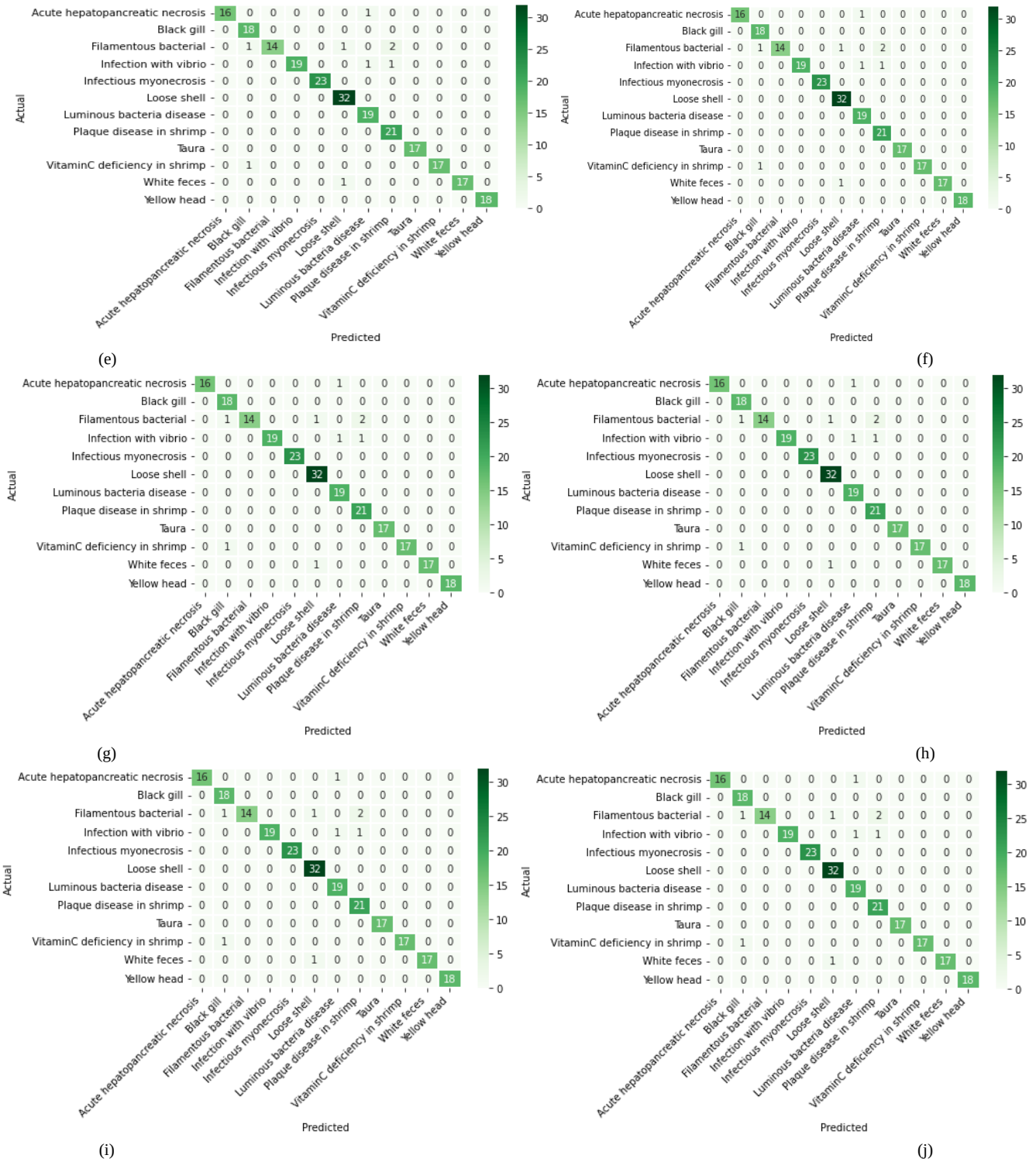



Fig. 6. Confusion matrix of algorithms: (a)SVM, (b)LoR, (c)MNB, (d)BNB, (e)RF, (f)DNN, (g)LSTM, (h)GRU, (i)BRNN, (j)RCNN.

## VII. Conclusion

This research classified shrimp diseases based on deep learning and machine learning methods. We preprocessed 1.037 samples of 12 prevalent shrimp diseases and divided them into three groups: training data, test sets, and validation sets. After training, the model is put through its paces on the validation and test sets. Compared to other models, the outcome demonstrates that DNN achieves the highest performance on this data.

This study applies the theoretical results of natural language processing (NLP) to analyze shrimp description and classify shrimp disease to enhance the productivity and

sustainability of aquaculture by providing timely and accurate diagnosis of shrimp diseases. Another motivation is to reduce the economic losses caused by shrimp diseases and increase the competitiveness of the shrimp industry. Furthermore, NLP can facilitate the prevention and treatment of shrimp diseases by offering instant services recommendation and early interventions. Additionally, NLP can advance the scientific knowledge and innovation in the field of NLP and its applications for aquaculture. However, this study still has some limitations that need to be overcome. First, the data size is relatively small and uneven across disease classes. This can affect the generalization ability of algorithms and models. Second, we only use TF-IDF as a feature extraction method for machine learning algorithms. TF-IDF is a simple and effective method, but it cannot represent the semantic meaning of the expression. Therefore, in the future, we will continue to collect more data to assess the methods on different datasets, and use other feature extraction methods, such as word2vec or BERT, to compare with TF-IDF.

## REFERENCES

[1] N. M. Khiem, Y. Takahashi, K. T. P. Dong, H. Yasuma, and N. Kimura, "Predicting the price of Vietnamese shrimp products exported to the US market using machine learning," Fish Sci, vol. 87, no. 3, pp. 411–423, May 2021, doi: 10.1007/s12562-021-01498-6.

[2] F. Asche et al., "The economics of shrimp disease," Journal of Invertebrate Pathology, vol. 186, p. 107397, Nov. 2021, doi: 10.1016/j.jip.2020.107397.

[3] C. E. Boyd, R. P. Davis, and A. A. McNevin, "Comparison of resource use for farmed shrimp in Ecuador, India, Indonesia, Thailand, and Vietnam," Aquaculture Fish & Fisheries, vol. 1, no. 1, pp. 3–15, Dec. 2021, doi: 10.1002/aff2.23.

[4] T. Pongthanapanich, K. A. T. Nguyen, and C. M. Jolly, "Risk management practices of small intensive shrimp farmers in the Mekong Delta of Viet Nam," FAO Fisheries and Aquaculture Circular, vol. C1194, pp. 1–20.

[5] M. Salunke, A. Kalyankar, C. D. Khedkar, M. Shingare, and G. D. Khedkar, "A Review on Shrimp Aquaculture in India: Historical Perspective, Constraints, Status and Future Implications for Impacts on Aquatic Ecosystem and Biodiversity," Reviews in Fisheries Science & Aquaculture, vol. 28, no. 3, pp. 283–302, Jul. 2020, doi: 10.1080/23308249.2020.1723058.

[6] S. Yaemkasem, V. Boonyawiwat, M. Sukmak, S. Thongratsakul, and C. Poolkhet, "Spatial and temporal patterns of white spot disease in Rayong Province, Thailand, from october 2015 to september 2018," Preventive Veterinary Medicine, vol. 199, p. 105560, Feb. 2022, doi: 10.1016/j.prevetmed.2021.105560.

[7] T. J. Sullivan, A. K. Dhar, R. Cruz-Flores, and A. G. Bodnar, "Rapid, CRISPR-Based, Field-Deployable Detection Of White Spot Syndrome Virus In Shrimp," Sci Rep, vol. 9, no. 1, p. 19702, Dec. 2019, doi: 10.1038/s41598-019-56170-y.

[8] P. Naranitus, P. Aiamsa‐at, T. Sukonta, P. Hannanta‐anan, and T. Chaijarasphong, "Smartphone‐compatible, CRISPR ‐based platforms for sensitive detection of acute hepatopancreatic necrosis disease in shrimp," Journal of Fish Diseases, vol. 45, no. 12, pp. 1805–1816, Dec. 2022, doi: 10.1111/jfd.13702.

[9] H. N. Mai, L. F. Aranguren Caro, R. Cruz-Flores, and A. K. Dhar, "Development of a Recombinase Polymerase Amplification (RPA) assay for acute hepatopancreatic necrosis disease (AHPND) detection in Pacific white shrimp (Penaeus vannamei)," Molecular and Cellular Probes, vol. 57, p. 101710, Jun. 2021, doi: 10.1016/j.mcp.2021.101710.

[10] Z. Chen, J. Huang, F. Zhang, Y. Zhou, and H. Huang, "Detection of shrimp hemocyte iridescent virus by recombinase polymerase amplification assay," Molecular and Cellular Probes, vol. 49, p. 101475, Feb. 2020, doi: 10.1016/j.mcp.2019.101475.

[11] T.-D. Mai-Hoang et al., "A novel PCR method for simultaneously detecting Acute hepatopancreatic Necrosis Disease (AHPND) and mutant-AHPND in shrimp," Aquaculture, vol. 534, p. 736336, Mar. 2021, doi: 10.1016/j.aquaculture.2020.736336.

[12] W.-C. Hu, H.-T. Wu, Y.-F. Zhang, S.-H. Zhang, and C.-H. Lo, "Shrimp recognition using ShrimpNet based on convolutional neural network," J Ambient Intell Human Comput, Jan. 2020, doi: 10.1007/s12652-020-01727-3.

[13] N. Duong-Trung, L.-D. Quach, and C.-N. Nguyen, "Towards Classification of Shrimp Diseases Using Transferred Convolutional Neural Networks," Adv. sci. technol. eng. syst. j., vol. 5, no. 4, pp. 724–732, 2020, doi: 10.25046/aj050486.

[14] T. Q. Bao, T. C. Cuong, N. D. Tu, L. H. Dang, and L. T. Hieu, "Designing the Yellow Head Virus Syndrome Recognition Application for Shrimp on an Embedded System," EIRJ, vol. 6, no. 2, pp. 48–63, Apr. 2019, doi: 10.31273/eirj.v6i2.309.

[15] T. H. O. Dang, T. N. T. Nguyen, and N. U. Vu, "Investigation of parasites in the digestive tract of white leg shrimp (Litopenaeus vannamei) cultured at coastal farms in the Mekong Delta," CTUJS, vol. 13, no. Aquaculture, pp. 79–85, Jun. 2021, doi: 10.22144/ctu.jen.2021.020.

[16] T. Kobayashi et al., "Microbiological properties of Myanmar traditional shrimp sauce, hmyin-ngan-pya-ye," Fish Sci, vol. 86, no. 3, pp. 551–560, May 2020, doi: 10.1007/s12562-020-01415-3.

[17] P. Pérez-Barros, N. V. Guzmán, V. A. Confalonieri, and G. A. Lovrich, "Molecular identification by polymerase chain reaction-restriction fragment length polymorphism of commercially important lithodid species (Crustacea: Anomura) from southern South America," Regional Studies in Marine Science, vol. 34, p. 101027, Feb. 2020, doi: 10.1016/j.rsma.2019.101027.

[18] S. Thiyagarajan, B. Chrisolite, and S. V. Alavandi, "Degenerate primed randomly amplified polymorphic DNA (DP-RAPD) fingerprinting of bacteriophages of Vibrio harveyi from shrimp hatcheries in Southern India," Microbiology, preprint, Aug. 2021. doi: 10.1101/2021.08.10.455891.

[19] J. Zhao et al., "Transcriptome Analysis Provides New Insights into Host Response to Hepatopancreatic Necrosis Disease in the Black Tiger Shrimp Penaeus monodon," J. Ocean Univ. China, vol. 20, no. 5, pp. 1183–1194, Oct. 2021, doi: 10.1007/s11802-021-4744-x.

[20] J. Yuan et al., "Simple sequence repeats drive genome plasticity and promote adaptive evolution in penaeid shrimp," Commun Biol, vol. 4, no. 1, p. 186, Feb. 2021, doi: 10.1038/s42003-021-01716-y.

[21] A. Ashraf and A. Atia, "Comparative Study Between Transfer Learning Models to Detect Shrimp Diseases," in 2021 16th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, Egypt: IEEE, Dec. 2021, pp. 1–6. doi: 10.1109/ICCES54031.2021.9686116.

[22] D. J. A. Amora, D. P. M. Alulod, K. A. B. Debolgado, J. R. M. Magcale, C. R. A. Tobias, and S. U. Arenas, "Design of a P. Vannamei White Spot Syndrome Virus (WSSV) Detection System Utilizing YOLOv5n," in 2022 IET International Conference on Engineering Technologies and Applications (IET-ICETA), Changhua, Taiwan: IEEE, Oct. 2022, pp. 1–2. doi: 10.1109/IET-ICETA56553.2022.9971656.

[23] M. O. Edeh et al., "Bootstrapping random forest and CHAID for prediction of white spot disease among shrimp farmers," Sci Rep, vol. 12, no. 1, p. 20876, Dec. 2022, doi: 10.1038/s41598-022-25109-1.

[24] [24] L. Đ. Quách, T. N. Phan, T. T. Hùng, and N. C. Ngôn, "Kiểm thử giải thuật AI trong nhận diện bệnh tôm qua hình ảnh," CTUJSVN, vol. 57, no. CĐ Thủy Sản, pp. 192–201, Jun. 2021, doi: 10.22144/ctu.jvn.2021.078.

[25] F. B. Putra et al., "Identification of Symptoms Based on Natural Language Processing (NLP) for Disease Diagnosis Based on International Classification of Diseases and Related Health Problems (ICD-11)," in 2019 International Electronics Symposium (IES), Surabaya, Indonesia: IEEE, Sep. 2019, pp. 1–5. doi: 10.1109/ELECSYM.2019.8901644.

[26] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani, "Natural Language Processing of Clinical Notes on Chronic

Diseases: Systematic Review," JMIR Med Inform, vol. 7, no. 2, p. e12239, Apr. 2019, doi: 10.2196/12239.

[27] R. Garg, E. Oh, A. Naidech, K. Kording, and S. Prabhakaran, "Automating Ischemic Stroke Subtype Classification Using Machine Learning and Natural Language Processing," Journal of Stroke and Cerebrovascular Diseases, vol. 28, no. 7, pp. 2045–2051, Jul. 2019, doi: 10.1016/j.jstrokecerebrovasdis.2019.02.004.

[28] [28] T. A. Koleck, C. Dreisbach, P. E. Bourne, and S. Bakken, "Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review," Journal of the American Medical Informatics Association, vol. 26, no. 4, pp. 364–379, Apr. 2019, doi: 10.1093/jamia/ocy173.

[29] [29] L. Li, S. Zhang, and B. Wang, "Plant Disease Detection and Classification by Deep Learning—A Review," IEEE Access, vol. 9, pp. 56683–56698, 2021, doi: 10.1109/ACCESS.2021.3069646.

[30] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, "Machine Learning Applications for Precision Agriculture: A Comprehensive Review," IEEE Access, vol. 9, pp. 4843–4873, 2021, doi: 10.1109/ACCESS.2020.3048415.

[31] V. K. Shrivastava, M. K. Pradhan, and M. P. Thakur, "Application of Pre-Trained Deep Convolutional Neural Networks for Rice Plant Disease Classification," in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India: IEEE, Mar. 2021, pp. 1023–1030. doi: 10.1109/ICAIS50930.2021.9395813.

[32] L.-D. Quach, L. Q. Hoang, N. D. Trung, and C. N. Nguyen, "TOWARDS MACHINE LEARNING APPROACHES TO IDENTIFY SHRIMP DISEASES BASED ON DESCRIPTION," in KỶ YẾU HỘI NGHỊ KHOA HỌC CÔNG NGHỆ QUỐC GIA LẦN THỨ XII NGHIÊN CỨU CƠ BẢN VÀ ỨNG DỤNG CÔNG NGHỆ THÔNG TIN, Hanoi, Vietnam: Publishing House for Science and Technology, Oct. 2019. doi: 10.15625/vap.2019.00063.

[33] S. S. Aljameel et al., "A Sentiment Analysis Approach to Predict an Individual's Awareness of the Precautionary Procedures to Prevent COVID-19 Outbreaks in Saudi Arabia," IJERPH, vol. 18, no. 1, p. 218, Dec. 2020, doi: 10.3390/ijerph18010218.

[34] L. Havrlant and V. Kreinovich, "A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation)," International Journal of General Systems, vol. 46, no. 1, pp. 27–36, Jan. 2017, doi: 10.1080/03081079.2017.1291635.

[35] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," JASTT, vol. 1, no. 4, pp. 140–147, Dec. 2020, doi: 10.38094/jastt1457.

[36] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," The Stata Journal, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: 10.1177/1536867X20909688.

[37] E. Y. Boateng and D. A. Abaye, "A Review of the Logistic Regression Model with Emphasis on Medical Research," JDAIP, vol. 07, no. 04, pp. 190–207, 2019, doi: 10.4236/jdaip.2019.74012.

[38] M. Artur, "Review the performance of the Bernoulli Naïve Bayes Classifier in Intrusion Detection Systems using Recursive Feature Elimination with Cross-validated selection of the best number of features," Procedia Computer Science, vol. 190, pp. 564–570, 2021, doi: 10.1016/j.procs.2021.06.066.

[39] D. A. Pisner and D. M. Schnyer, "Support vector machine," in Machine Learning, Elsevier, 2020, pp. 101–121. doi: 10.1016/B978-0-12-815739-8.00006-7.

[40] J. Gawlikowski et al., "A Survey of Uncertainty in Deep Neural Networks," 2021, doi: 10.48550/ARXIV.2107.03342.

[41] Y. Liu et al., "A long short-term memory-based model for greenhouse climate prediction," Int J Intell Syst, vol. 37, no. 1, pp. 135–151, Jan. 2022, doi: 10.1002/int.22620.

[42] Q. Ni, J. C. Ji, and K. Feng, "Data-Driven Prognostic Scheme for Bearings Based on a Novel Health Indicator and Gated Recurrent Unit Network," IEEE Trans. Ind. Inf., vol. 19, no. 2, pp. 1301–1311, Feb. 2023, doi: 10.1109/TII.2022.3169465.

[43] S. S. Tng, N. Q. K. Le, H.-Y. Yeh, and M. C. H. Chua, "Improved Prediction Model of Protein Lysine Crotonylation Sites Using Bidirectional Recurrent Neural Networks," J. Proteome Res., vol. 21, no. 1, pp. 265–273, Jan. 2022, doi: 10.1021/acs.jproteome.1c00848.

[44] B. Wang, Y. Lei, T. Yan, N. Li, and L. Guo, "Recurrent convolutional neural network: A new framework for remaining useful life prediction of machinery," Neurocomputing, vol. 379, pp. 117–129, Feb. 2020, doi: 10.1016/j.neucom.2019.10.064.

[45] E. Hossain, O. Sharif, and M. Moshiul Hoque, "Sentiment Polarity Detection on Bengali Book Reviews Using Multinomial Naïve Bayes," in Progress in Advanced Computing and Intelligent Engineering, C. R. Panigrahi, B. Pati, B. K. Pattanayak, S. Amic, and K.-C. Li, Eds., in Advances in Intelligent Systems and Computing, vol. 1299. Singapore: Springer Singapore, 2021, pp. 281–292. doi: 10.1007/978-981-33-4299-6_23.