Journal of Computer Science and Cybernetics, V.42, N.1 (2026), 41-56
DOI no. 10.15625/1813-9663/22392

VNSED: VIETNAMESE SPAM EMAIL DETECTION USING MULTI DEEP LEARNING MODELS

NGUYEN TAN CAM[1,2,*], NGO THANH BINH[1,2]

[1] *University of Information Technology, Quarter 34, Linh Xuan Ward, Ho Chi Minh City, Viet Nam*

[2] *Vietnam National University Ho Chi Minh City, Quarter 33, Linh Xuan Ward, Ho Chi Minh City, Viet Nam*

Abstract. This study introduces VNSED, a Vietnamese spam email detection system built upon a newly constructed dataset comprising 6,008 labeled emails. The system evaluates the performance of both traditional machine learning models such as Naïve Bayes, SVM (Support Vector Machine), and deep learning architectures such as CNN (Convolutional Neural Network), BiLSTM (Bidirectional Long Short-Term Memory), LSTM, PhoBERT, and a hybrid PhoBERT+CNN model. Experimental results indicate that SVM achieves the highest accuracy (89.77%), while PhoBERT obtains the highest recall (93.56%) and an impressive F1-score (87.43%). To improve training efficiency, model pruning was applied by reducing the number of encoder layers in PhoBERT, effectively shortening training time with minimal performance loss. PhoBERT with pruning still achieved an F1-score of 87.36%. Furthermore, the study employs explainable artificial intelligence (XAI) techniques to enhance the interpretability and transparency of the model's predictions.

Keyword. Spam email detection, convolutional neural network, bidirectional long short-term memory, PhoBERT.

1. INTRODUCTION

Email spam has become a significant issue for individuals and organizations worldwide. According to statistics from Statista [1], nearly 45.6% of all emails in 2023 were spam. Besides advertising emails, anonymous and phishing emails are also common types of spam, although they are sent less frequently than advertising emails. These types of emails are particularly dangerous and can cause physical and mental harm to the recipients. Spam emails often waste the recipient's time and effort when receiving these emails because the recipient has to skim through the received email before determining that it is spam [2].

The content of spam emails usually has the following characteristics: (1) The general, unspecific greetings to the recipients. (2) Promises like winning prizes, receiving gifts, etc. (3) Special promotions and advertising programs. (4) The emails have urgent content to lead the user to take a certain action. (5) Asking the recipients to provide personal or account information. (6) Guiding users into clicking on malicious links or attachments.

Although many technical solutions have been applied to prevent spam, the most important thing to determine whether an email is spam or not still lies in the content of the email sent

*Corresponding author.
E-mail addresses: camnt@uit.edu.vn (N. T. Cam); binhngt.16@grad.uit.edu.vn (N. T. Binh).

to the users. Nowadays, machine learning models as well as deep learning models have been applied in detecting spam based on the content of sent emails with high accuracy. The biggest challenge is the lack of a dataset to train the models.

Unlike widely studied languages such as English, Vietnamese presents unique linguistic challenges, including the use of diacritics, specific grammatical structures, and contextual nuances that complicate the accurate detection and classification of spam content. Current spam detection models are primarily optimized for English, resulting in less effective protection for Vietnamese-speaking users. By focusing on Vietnamese spam detection, this study aims to enhance the security infrastructure, protect sensitive information, and ensure a safer and more reliable email communication environment for individuals and businesses in Vietnam. This paper presents the construction of a Vietnamese spam email dataset and the application of deep learning models, including CNN (Convolutional Neural Network) [3], BiLSTM (Bidirectional Long Short-Term Memory) [4], and PhoBert [5] to train models for detecting Vietnamese spam emails.

The main contributions of this study are as follows: (1) Constructing a Vietnamese spam email dataset for training models. (2) Developing a Vietnamese spam email detection system, named VNSED, using deep learning models including CNN, BiLSTM, LSTM, PhoBERT, and PhoBERT+CNN.

The remainder of the study consists of Section 2, 3, 4, and 5. Sections 2 presents related works. The proposed system model and experimental evaluation are described in Sections 3 and 4, respectively. Finally, Section 5 is the conclusion.

2. RELATED WORKS

Many related studies have used machine learning models as well as deep learning models to detect spam email [6, 7, 8, 9, 10]. These studies are mainly conducted in the English language due to the popularity of English. AbdulNabi and Yaseen [11] proposed to fine-tune the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model to detect spam email. Their proposed model is compared with a deep learning model, BiLSTM, and two other machine learning models, K-Nearest Neighbors, and Naïve Bayes. Due to the limitation of computational resources, the length of the email content fed into the model is 300, and the accuracy of their proposed model achieves 98.67% and the *F1_score* achieves 98.66%.

Siddique et al. [12] proposed a model for detecting spam email in Urdu, a language of South Asian countries, especially Pakistan. Their proposed method uses Naïve Bayes, Support Vector Machine, LSTM, and CNN. Their results show that the LSTM deep learning model achieves the highest accuracy of 98.4% in detecting Urdu spam email, but it also takes longer to train. Meanwhile, Tida et al. [13] proposed a model based on the pre-trained BERT model to classify spam email. The dataset used in their model consists of four English spam email datasets, including Enron, Spamassassin, Lingspam, and Spamtext combined. The model using this combined dataset is fine-tuned the hyperparameters based on the hyperparameters of the models using the separate datasets and achieved better results than the models trained on the separate datasets. The accuracy of their model achieved 97% and the F1_score achieved 96%.

To contribute to fighting Vietnamese spam SMS messages, Tuan et al. [14] combined a DNN (deep neural network) model and pre-trained PhoBERT to detect Vietnamese spam SMS messages. Their proposed model achieved 99.53% accuracy in detecting Vietnamese spam SMS messages. To evaluate the effectiveness of the model, they tested this dataset with machine learning models (Support Vector Machine, Naïve Bayes, and K-Nearest Neighbors) and CNN.

The results show that their proposed model achieves high accuracy compared to other models. The model also needs to be further optimized by optimizing different parameters of the model.

Guo et al. [15] proposed an effective spam email detection method using a pre-trained BERT model combined with four well-known classification machine learning algorithms, namely Support Vector Machine, K-Nearest Neighbors, Random Forest, and Logistic Regression. Experimental results demonstrate that the Logistic Regression algorithm achieves the best classification on both datasets. However, this model needs further evaluation in incorporating more complex layers inside BERT.

From the above studies, deep learning models are highly effective in detecting spam email but are mainly trained on English datasets or translated from English datasets to other languages through GoogleTrans API. Tuan et al. [14] have collected a spam dataset, but it is a Vietnamese spam SMS dataset. Providing training data for machine learning models as well as deep learning models plays an important role in evaluating the effectiveness of the proposed model. Conducting a Vietnamese spam email dataset for model training will contribute to detecting Vietnamese spam email more effectively and preventing the spam problem.

3. THE PROPOSED SYSTEM

Figure 1 presents the proposed system, named VNSED, for detecting Vietnamese spam email using deep learning models, including CNN, PhoBERT, PhoBERT+CNN, BiLSTM, and LSTM. This proposed system includes the steps of collecting, cross-checking, and classifying Vietnamese emails; extracting the subject and body of the emails; assigning SPAM and HAM labels to the emails; performing data preprocessing; and representing the data in the form of a numerical vector before feeding it into deep learning models for training. The output is a model capable of detecting Vietnamese spam email.

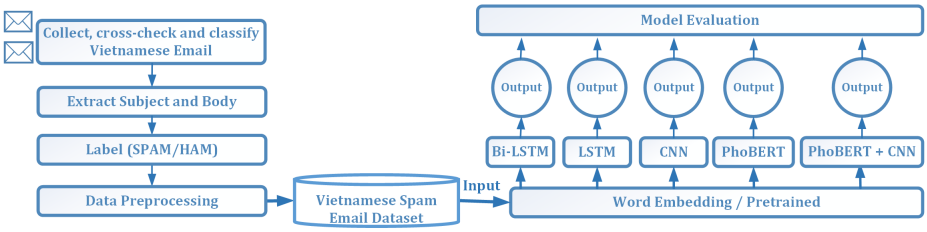

Figure 1: VNSED: Vietnamese spam email detection system

3.1. Vietnamese spam email dataset

The effective spam detection is a critical component in maintaining the security and efficiency of email communication systems. In the context of Vietnamese-language emails, the development of a specialized dataset is essential for training and evaluating robust spam detection deep learning based models. The existing datasets primarily focus on English or other widely spoken languages, leaving a significant gap in resources tailored to Vietnamese. To bridge this gap, this study propose a systematic approach to constructing a comprehensive Vietnamese spam email dataset.

Emails are labeled as spam based on spam identification criteria [16], including: urgent content, requests for recipients to provide personal information, emails containing malicious links or attachments, generic greetings, and content with enticing promises (such as winning prizes or receiving gifts). Subsequently, spam emails are classified into common categories, including advertising emails, phishing emails, anonymous emails, and fundraising emails. In

the future, as new types of Vietnamese spam emails emerge, the dataset will be updated accordingly.

The proposed algorithm (Algorithm 1) outlines a step-by-step process for building this dataset from raw email files. It begins with the collection of email file names and proceeds through the extraction and preprocessing of email content. By using this algorithm, each email is categorized based on its type and labeled accordingly as "HAM" (legitimate) or "SPAM". The algorithm ensures consistency and accuracy in data labeling, which is paramount for the effectiveness of subsequent deep learning models. Finally, the processed data is formatted into a structured *CSV* file, ready for use in training and evaluation tasks.

Algorithm 1 Building a Vietnamese spam email dataset

Require: A list containing the names of email files: `msg_files`; A list storing the processed content of emails: `msg_texts[]`; A list storing the corresponding email categories: `msg_types[]`; A list storing the corresponding email labels: `msg_labels[]`
Ensure: `dataset.csv` // Dataset

Initialize $i \leftarrow 0$
for each `item` in `msg_files` **do**
  if $i <$ length(msg_files) **then**
    Extract `msgsub`, `msgbody` **of** `item` using `BytesParser`
    `msgtext` ← `msgsub` + '. ' + `msgbody`
    `msgtext` ← `Preprocess_text(msgtext)`
    `type` ← **Extract Type category from** `item name`
    if `item` in `Ham` **then**
      `label` ← 'ham'
    else if `item` in `Spam` **then**
      `label` ← 'spam'
    else
      Print('Folder does not exist.')
    end if
    `list_msgtexts.append(msgtext)`
    `list_msgtypes.append(type)`
    `list_msglabels.append(label)`
    $i \leftarrow i + 1$
  else
    Print('.eml file does not exist')
  end if
end for
`df_eml` ← **Format** `list_msgtexts`, `list_msgtypes`, `list_msglabels` using Pandas DataFrame
`df_eml.to_csv('dataset.csv')`
Return `dataset.csv`

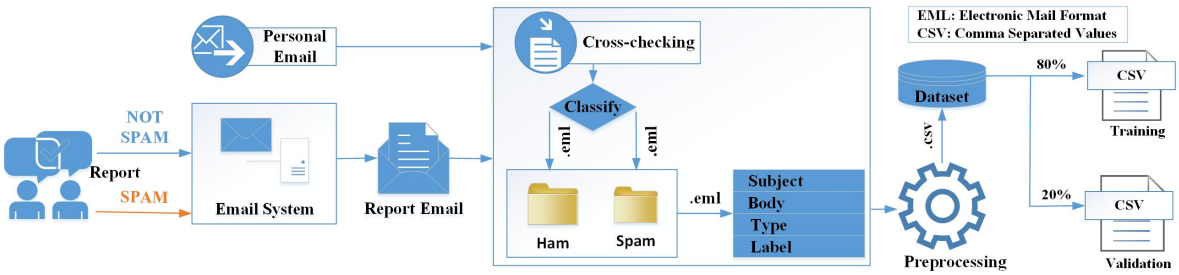

Figure 2: The process of constructing the Vietnamese spam email dataset

Figure 2 describes the process of conducting the Vietnamese spam email dataset. The detailed steps are as follows:

- Step 1: (Email collection): The Vietnamese spam email dataset is sourced from (i) the author's personal emails and (ii) user spam reports (Report/Mark as Spam, Not Spam).

Emails reported as spam go to the spam account; those reported as not spam go to the ham account.

- Step 2: (Labeling and *.eml* storage): Emails are cross-checked, classified, and saved as *.eml* using the pattern *email type_subject.eml*. Verified "not spam" reports are filed in the Ham folder by domain (labeled HAM). Verified "spam" reports are filled in the Spam folder by type (labeled SPAM). During preprocessing, *BytesParser* reads each *.eml*, extracts subject/body, derives email type from the filename, and infers the label from the parent folder.
- Step 3: (Text preprocessing): Clean subject/body by removing stopwords, numbers, punctuation, symbols, links, email addresses, sender signatures, etc., to reduce noise and training time.
- Step 4: (Export to *.csv*): Each row contains label, email type, and content. The final dataset has 6,008 emails: 3,368 HAM and 2,640 SPAM (details in Table 1).
- Step 5: (Dataset splitting): Divide the dataset for training (80%) and testing (20%), respectively.

Table 1: Detailed Information about the Vietnamese spam email dataset

Dataset characteristics	Value	Dataset characteristics	Value
Total number of emails	6,008	Total number of regular emails	3,368
Total number of spam emails	2,640	Shortest regular email (words)	6
Longest regular email (words)	1,142	Average regular email (words)	214
Shortest spam email (words)	5	Longest spam email (words)	1,432
		Average spam email (words)	255

3.2. Deep learning models for Vietnamese email spam detection

3.2.1. Numeric vector representation

There are many ways to represent text content such as emails in the form of numeric vectors for model training. Word2Vec is a natural language processing model developed by Google, capable of representing the context of words, helping to understand the language better. Word2Vec will be trained on a corpus of words in the Vietnamese spam email dataset and create numeric vector representations of corresponding Vietnamese words. Semantically close words will have high similarity. The numerical vectors are then fed into the CNN and BiLSTM models to train the Vietnamese spam email detection model. In addition to using the Word2Vec model in combination with CNN, LSTM, and BiLSTM, the proposed system also uses PhoBERT and a combination of PhoBERT and CNN for model comparison and evaluation.

3.2.2. CNN

The CNN model is widely used in image processing, but now CNN is also applied in natural language processing, typically text classification, which can be applied in detecting Vietnamese spam email [17]. In the text classification problem, such as Vietnamese spam email detection, Conv1D and MaxPooling1D layers are used to extract features from the input data instead of Conv2D and MaxPooling2D, which are commonly used for image input data (see Figure 3). The input data consists of words in emails that have been represented as numerical vectors using a pre-trained Word2Vec model. The GlobalMaxPooling1D layer aggregates important

features across the entire input text sequence before passing them to a fully connected layer for classification. The system then determines whether an email is a regular email (corresponding to a value of 0) or a spam email (corresponding to a value of 1). Based on the classification results, the system can effectively detect whether a Vietnamese email is spam.

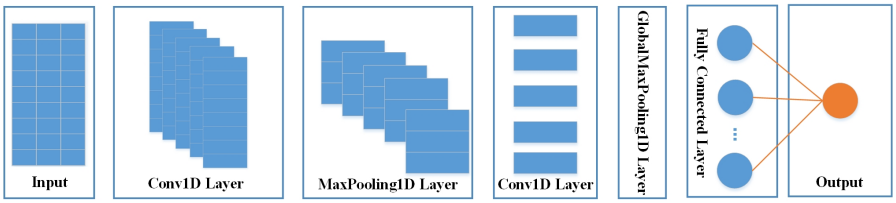

Figure 3: CNN model for Vietnamese email spam detection

3.2.3. LSTM and BiLSTM

The LSTM model [18] is used in the field of natural language processing due to its ability to process data in the form of sequences, such as text. This model is improved to overcome the problems of the RNN (recurrent neural network) model. BiLSTM is an extension of the LSTM model, allowing data processing in both directions, forward and backward, to understand data in the form of sequences better than the one-way data processing of the LSTM model. To process data in both directions, BiLSTM uses two LSTM layers, one forward LSTM layer and one backward LSTM layer (see Figure 4). Because of its ability to process sequence data, it can be used for text classification, thereby applying it to Vietnamese spam email detection. The input is the email content represented by a numeric vector through a pretrained Word2Vec model. The input data is then fed into the network in both forward and reverse directions. In the LSTM forward layer, the input is processed sequentially from left to right. At the same time, in the LSTM backward layer, the input is processed sequentially from right to left. The results are then aggregated and fed into the fully connected layer to perform classification. Based on the classification results, the system can detect Vietnamese spam email.

3.2.4. PhoBERT

The BERT model [19] has two main models, BERT-base and BERT-large. The BERT-base model consists of 12 stacked representation layers (encoders) (see Figure 5). Input data is fed into BERT-base to learn the context representation of each word through the multi-head attention mechanism. This mechanism allows us to associate each word in a sentence with all the remaining words in the sentence to learn the relationship between words and the context of the words. The FeedForward neural network in each representation layer of BERT-base consists of 768 hidden nodes, so the size of the context representation of each word in the output is 768. A special token (CLS token) is added to the beginning of the input data and will contain a representation of the entire input data (denoted as R_{CLS}) after calculation. This token will be used for various natural language processing tasks, like text classification.

PhoBERT [20] is a model based on Google's BERT model, pre-trained on the Vietnamese language to perform various natural language processing tasks in Vietnamese, such as language inference, question and answer, text classification, etc. in which text classification task can be applied in detecting Vietnamese email spam. The PhoBERT-base model (135 million parameters) is chosen to fine-tune on the Vietnamese spam email dataset for classifying and detecting Vietnamese spam email. The result of this model is compared with three other deep

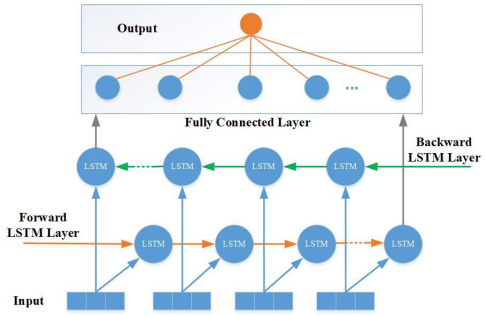

Figure 4: BiLSTM model for Vietnamese email spam detection.

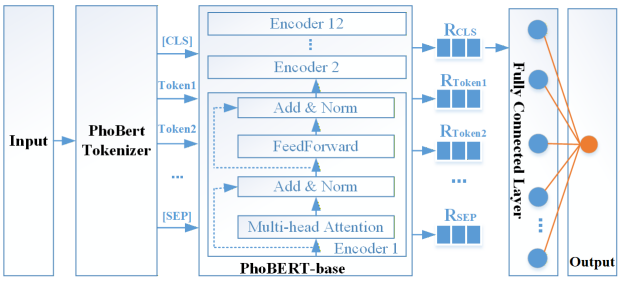

Figure 5: PhoBERT model for Vietnamese email spam detection.

learning models, such as CNN, BiLSTM, and LSTM, to evaluate the accuracy in detecting Vietnamese spam email.

3.2.5. PhoBERT+CNN fusion

To leverage both semantic understanding and structural pattern recognition for Vietnamese spam email detection, this study proposes a hybrid model that combines the strengths of PhoBERT and Convolutional Neural Networks (CNN). PhoBERT is employed to extract contextualized word embeddings from the input email text. These embeddings capture rich semantic and syntactic information based on the surrounding context of each token. Subsequently, CNN is applied as a feature extractor that scans the sequence of embeddings using multiple filters of varying sizes. This enables the model to detect important local patterns such as common phrases, token co-occurrences, or structural cues that are indicative of spam. By integrating the global contextual representation from PhoBERT with the local pattern learning capabilities of CNN, the hybrid model effectively captures both high-level semantics and discriminative features typical for spam emails. This fusion allows for more robust and accurate classification, especially in the presence of noisy or obfuscated content often found in real-world spam email.

To implement this approach, the output of PhoBERT (*last_hidden_state*) has the shape [*sequence_length* × *hidden_size*], where *sequence_length* is the number of tokens corresponding to each word in the email, and *hidden_size* is the dimensionality of each token embedding. For the PhoBERT-Base model used in this study, the maximum sequence length is 256 and the hidden size is 768. This output is then passed into a CNN model, where a *Conv1D* layer is used to extract features, followed by a *GlobalMaxPooling1D* layer to summarize the most important features. Finally, the output is fed into a fully connected *Dense* layer to perform classification and detect spam emails.

3.2.6. Used activation function

Vietnamese spam email detection is a type of text classification task, so the sigmoid activation function is used to normalize the output to 0 and 1 corresponding to HAM and SPAM (Formula (1)). The value will be normalized to 0, corresponding to HAM, or 1, corresponding to SPAM.

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{1}$$

3.2.7. Used loss function

The loss function measures the difference between the model's predicted value and the true value. If the model's prediction is correct, the error will be small and if the model's prediction is incorrect, the error will be large. During training, the models will learn to make better predictions by adjusting the parameters to minimize the loss. For binary classification problems such as the Vietnamese spam email detection, the loss function has the form of a binary cross-entropy function as Formula (2).

$$\mathcal{L}(\theta, x, y) = -\frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)}\log(\hat{y}^{(i)}) + (1-y^{(i)})\log(1-\hat{y}^{(i)})\right], \tag{2}$$

where $\mathcal{L}$ is the loss function, θ is model parameter, N is the total number of email samples, $y^{(i)}$ is the true value of the label of the i^{th} email sample (taking the value 0 or 1), $\hat{y}^{(i)}$ is the predicted probability of the label of the i^{th} email sample at the output, $\hat{y}$ will be in the form of a sigmoid activation function, as Formula (1), to normalize the output to a value in the range [0, 1].

3.3. Model evaluation metrics

To evaluate the effectiveness of deep learning models in detecting spam email based on the Vietnamese spam email dataset, accuracy parameters, including *Accuracy, Precision, Recall*, and *F1_score*, are recorded to evaluate the accuracy along with evaluating the model training time. The accuracy parameters are calculated using the following formulas.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{3}$$

$$Precision = \frac{TP}{TP + FP}, \tag{4}$$

$$Recall = \frac{TP}{TP + FN}, \tag{5}$$

$$F1\_Score = 2 * \frac{Recall * Precision}{Recall + Precision}. \tag{6}$$

In this work, the parameters used to compute accuracy for the deep learning models are defined succinctly as follows: TP (True Positive): the model predicts an email as SPAM and it is actually SPAM; FP (False Positive): the model predicts SPAM but the email is HAM; FN (False Negative): the model predicts HAM while the email is SPAM; TN (True Negative): the model predicts HAM and the email is HAM.

4. EXPERIMENT RESULTS

4.1. Experimental configuration

The experiments were carried out on a 64-bit Windows 11 operating system configured with a 32-core CPU and 32 GB of RAM, providing sufficient computational resources to train and evaluate deep learning models efficiently. The development and experimentation environment was built using Jupyter Notebook [21], an open-source web-based interactive platform that facilitates the combination of executable code, rich text, visualizations, and data outputs

within a single document. To implement and train the deep learning models, the TensorFlow framework [22] was used. TensorFlow offers comprehensive tools and libraries for building and deploying machine learning and deep learning applications, and it supports GPU acceleration, model serialization, and a wide range of APIs for model training and evaluation.

4.2. Vietnamese spam email dataset

This study builds the Vietnamese email spam dataset. The collected Vietnamese email spam dataset includes 6008 emails, of which 3368 are labeled HAM and 2640 are labeled SPAM. The length of the emails is mainly at the threshold of 400 words. For HAM labeled emails, the emails are classified according to different fields, including education and training, business, finance, securities, technology, sports and tourism, healthcare, real estate, public services, and others, as in Figure 6. For SPAM-labeled emails, the emails are classified according to common types of SPAM email, including advertising emails, anonymous emails, phishing emails, and other spam emails, as in Figure 7.

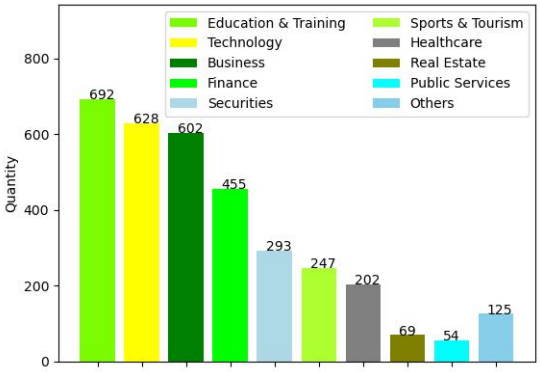

Figure 6: Statistics of emails classified according to different fields.

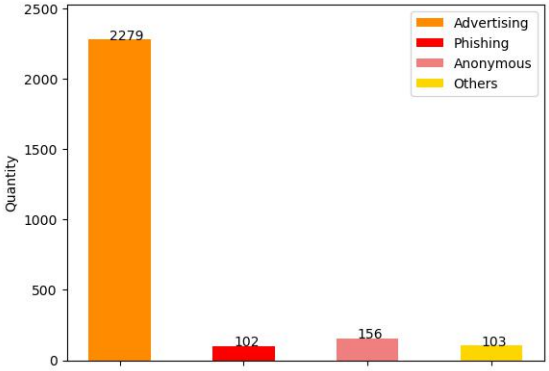

Figure 7: Statistics of common spam email types.

4.3. Experiment results

The Vietnamese spam email dataset was divided in an 80:20 ratio, where 80% of the data was used to train the models and the remaining 20% was reserved for testing and evaluation. This common practice ensures a sufficient amount of training data while preserving a separate validation set for unbiased performance assessment.

Model's performance analysis

Figures 8 to 17 illustrate the training behavior and detection performance of five deep learning models (CNN, LSTM, BiLSTM, PhoBERT, and PhoBERT+CNN) on the Vietnamese spam email detection task. The training and validation accuracy curves reflect the learning efficiency and generalization capability, while the confusion matrices provide insight into how each model handles class-wise predictions.

The CNN model (Figures 8 and 9) achieved a strong performance with an F1-score of 86.02%. The model demonstrates rapid convergence within the first few epochs, and the confusion matrix reveals relatively balanced performance across both spam and ham classes, with 623 true positives and only 51 false negatives.

In contrast, the LSTM model (Figures 10 and 11) achieved the lowest F1-score of 80.04%. Although the training and validation accuracy curves show consistent improvement, the confusion matrix indicates a higher number of false positives (103), which affects precision.

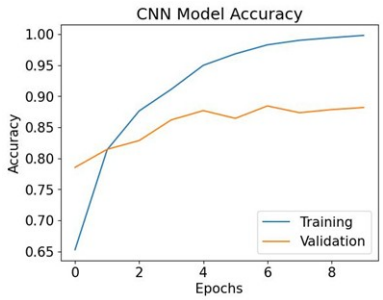

Figure 8: Training and validation accuracy trends of the CNN model.

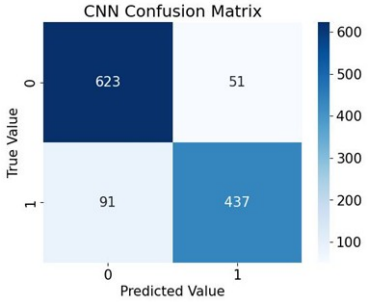

Figure 9: Confusion matrix illustrating CNN model detection performance.

The BiLSTM model (Figures 12 and 13) slightly improves recall (85.98%) but suffers from lower precision (76.69%), yielding an F1-score of 81.07%. Although the model generalizes better than LSTM, it misclassifies more non-spam emails as spam, as reflected by 138 false positives.

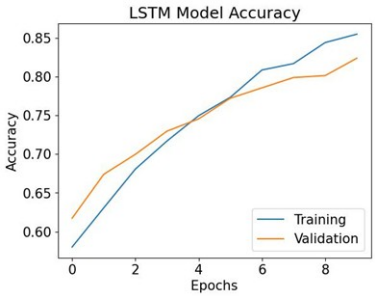

Figure 10: Training and validation accuracy trends of the LSTM model.

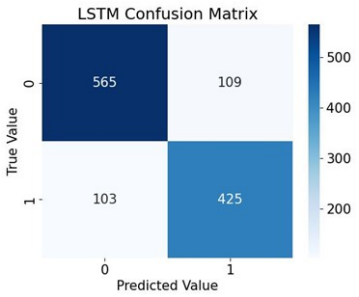

Figure 11: Confusion matrix illustrating LSTM model detection performance.

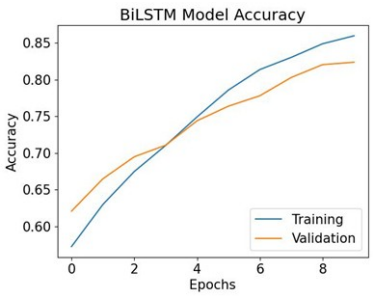

Figure 12: Training and validation accuracy trends of the BiLSTM model.

PhoBERT (Figures 14 and 15) shows remarkable recall (93.56%), which indicates superior sensitivity to spam emails. The learning curve shows rapid early learning with some fluctuations around epoch 6. This model achieves an overall F1-score of 87.43%, proving its strong contextual understanding, although with some false positives (69).

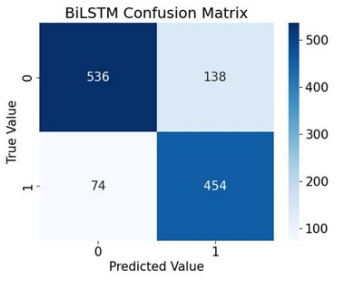

Figure 13: Confusion matrix illustrating BiLSTM model detection performance.

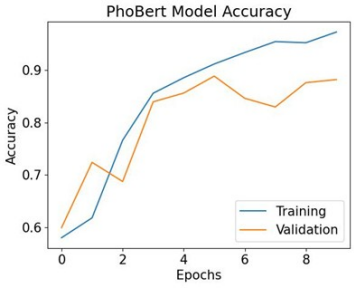

Figure 14: Training and validation accuracy trends of the PhoBERT model.

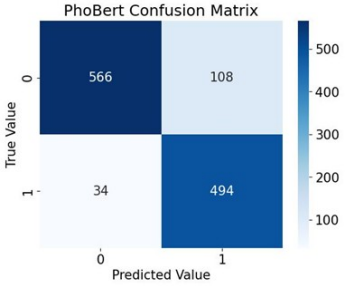

Figure 15: Confusion matrix illustrating PhoBERT model detection performance.

The hybrid PhoBERT+CNN model (Figures 16 and 17) yields the best balance between all evaluation metrics, with an F1-score of 86.29%. The hybrid model demonstrates enhanced stability and consistency in both training and validation, suggesting that combining semantic (PhoBERT) and spatial (CNN) features enhances robustness. The confusion matrix confirms improved class separation with reduced misclassification compared to PhoBERT alone.

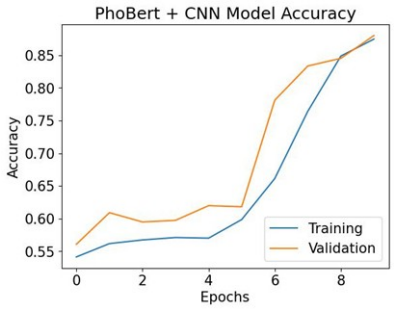

Figure 16: Training and validation accuracy trends of the PhoBERT+CNN model.

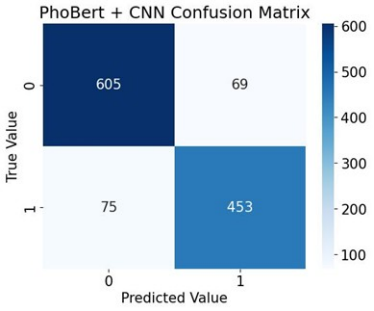

Figure 17: Confusion matrix illustrating PhoBERT+CNN model detection performance.

To comprehensively assess model effectiveness, we extended our experiments by evaluating two traditional machine learning classifiers: Naïve Bayes and Support Vector Machine (SVM). These models serve as strong baselines due to their widespread use in spam detection.

As shown in Table 2, SVM outperforms Naïve Bayes across all evaluation metrics, achieving the highest accuracy (89.77%) among all models tested. It also maintains high precision (89.78%) and F1-score (88.14%), reflecting its robust classification capability. In contrast, Naïve Bayes achieves an accuracy of 84.53%, with comparatively lower recall (77.84%) and F1-score (81.55%), indicating a tendency to misclassify some spam samples as HAM.

The confusion matrices in Figures 18 and 19 further support these findings. The Naïve Bayes model results in a higher false-negative count (spam predicted as ham), which limits its recall. Meanwhile, SVM demonstrates a more balanced prediction capability across both classes. Compared to deep learning models, SVM surpasses BiLSTM and LSTM in all metrics and closely rivals CNN and PhoBERT-based models. Although deep learning models such as PhoBERT achieve a higher recall (93.56%) and competitive F1-scores, they require significantly more computational resources and training time.

Those results show that while deep learning models exhibit superior recall and generalization, SVM provides an efficient and competitive alternative for spam classification with significantly lower complexity. This highlights the trade-off between computational cost and performance, and suggests that SVM remains a viable option for lightweight deployment scenarios. While SVM demonstrates impressive performance with lower computational overhead, deep learning models such as PhoBERT are more suitable for future deployment in Federated Learning (FL) environments. FL frameworks benefit from models capable of capturing semantic and contextual nuances across distributed nodes, a strength inherent in deep architectures. Moreover, the adaptability of deep learning models to non-IID data distributions and their ability to scale with increased training data make them a more robust and future-proof solution. Therefore, this study focuses on evaluating deep learning approaches, laying the groundwork for secure, privacy-preserving, and decentralized email spam detection systems in the future.

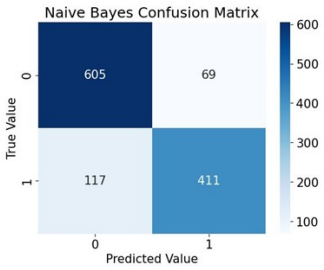

Figure 18: Confusion matrix illustrating Naïve Bayes model detection performance.

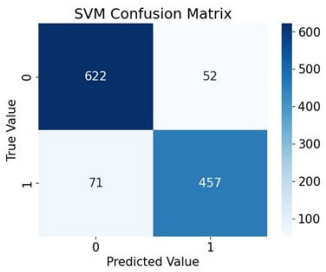

Figure 19: Confusion matrix illustrating SVM model detection performance.

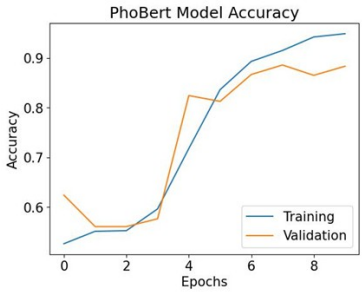

Figure 20: Training and validation accuracy trends of the PhoBERT model with pruning

Table 2: Evaluation metrics for various models

Metric	Naïve Bayes	SVM	CNN	LSTM	BiLSTM	PhoBERT	PhoBERT+CNN
Accuracy	84.53%	89.77%	88.19%	82.36%	82.36%	88.19%	88.02%
Precision	85.62%	89.78%	89.55%	79.59%	76.69%	82.06%	86.78%
Recall	77.84%	86.55%	82.77%	80.49%	85.98%	93.56%	85.80%
F1_Score	81.55%	88.14%	86.02%	80.04%	81.07%	87.43%	86.29%

Training time analysis

To evaluate the computational efficiency of each model, we measured the training time on a computing system with a hardware configuration described in Section 4.1. The results, presented in Table 3, show a clear trade-off between model complexity and training duration.

Table 3: Training time of models

Model	Training time (Minutes)	Model	Training time (Minutes)
Naïve Bayes	0.05	LSTM	21.17
SVM	0.43	BiLSTM	42.80
CNN	1.13	PhoBERT	130.87
		PhoBERT+CNN	144.43

Traditional machine learning models such as Naïve Bayes and SVM exhibit extremely low training times of 0.05 and 0.43 minutes, respectively, making them highly efficient and suitable for resource-constrained environments. Among deep learning models, CNN achieves a good balance between performance and efficiency, with a moderate training time of 1.13 minutes. In contrast, sequential models like LSTM and BiLSTM require significantly longer training times (21.17 and 42.80 minutes), due to their recurrent architecture and the need to process sequential dependencies. Transformer-based models show the highest training cost: PhoBERT and PhoBERT+CNN required 130.87 and 144.43 minutes, respectively. This is expected due to the large number of parameters and complex contextual representations inherent in transformer architectures. Given the considerable training overhead of PhoBERT-based models, optimization techniques such as pruning, particularly reducing the number of encoder layers or using lightweight transformer variants, can be employed to mitigate computational burden. To evaluate the impact of pruning on model performance and efficiency, this study conducted experiments comparing PhoBERT and PhoBERT+CNN models before and after applying pruning. Specifically, the number of encoder layers in PhoBERT was reduced from 12 to 4.

As shown in Table 4, the pruned PhoBERT model (see Figure 20 and Figure 21) maintained competitive performance with a slight improvement in accuracy (from 88.19% to 88.35%) and precision (from 82.06% to 83.45%), while recall slightly decreased (from 93.56% to 91.67%), resulting in a nearly equivalent F1-score (87.43% vs. 87.36%). These results demonstrate that the pruning process did not significantly affect classification performance.

Table 4: Performance comparison of models with and without pruning

Model	Accuracy	Precision	Recall	F1_Score
Naïve Bayes	84.53%	85.62%	77.84%	81.55%
SVM	89.77%	89.78%	86.55%	88.14%
CNN	88.19%	89.55%	82.77%	86.02%
LSTM	82.36%	79.59%	80.49%	80.04%
BiLSTM	82.36%	76.69%	85.98%	81.07%
PhoBERT	88.19%	82.06%	93.56%	87.43%
PhoBERT + Pruning	88.35%	83.45%	91.67%	87.36%
PhoBERT + CNN	88.02%	86.78%	85.80%	86.29%
PhoBERT + CNN + Pruning	86.77%	84.23%	85.98%	85.10%

In the case of PhoBERT+CNN, pruning led to a moderate drop in accuracy (from 88.02% to 86.77%) and F1-score (from 86.29% to 85.10%), yet the model still preserved strong performance in recall (85.80% to 85.98%), as observed in Figure 22 and Figure 23. More importantly, pruning considerably reduced training time for both models, as indicated in Table 5. For PhoBERT, training time dropped from 130.87 minutes to 110.32 minutes (a reduction of 15.7%). For PhoBERT+CNN, training time decreased from 144.43 minutes to 113.53 minutes (a reduction of 21.4%). This significant reduction highlights the effectiveness of pruning in enhancing training efficiency while preserving model accuracy.

Table 5: Training time (in Minutes) of models with and without pruning

Model	Time (Minutes)	Model	Time (Minutes)
Naïve Bayes	0.05	LSTM	21.17
SVM	0.43	BiLSTM	42.80
CNN	1.13	PhoBERT	130.87
PhoBERT + Pruning	110.32	PhoBERT + CNN	144.43
		PhoBERT + CNN + Pruning	113.53

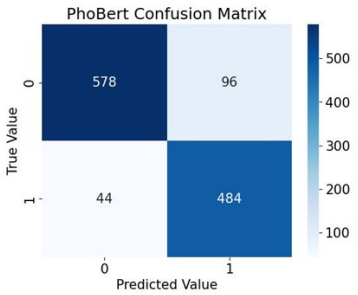

Figure 21: Confusion matrix illustrating PhoBERT model performance with pruning

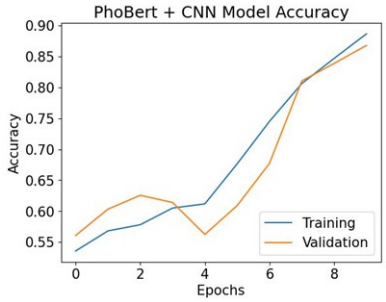

Figure 22: Training and validation accuracy trends of the PhoBERT+CNN with pruning

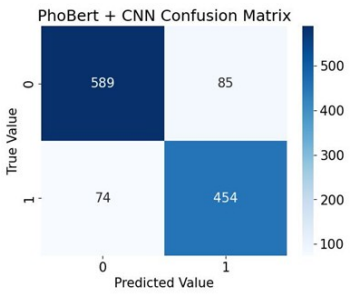

Figure 23: Confusion matrix illustrating PhoBERT+CNN performance with pruning

4.4. Prediction result explaining with explainable artificial intelligence

To gain insight into the prediction result based on the model's inputs and outputs, the LIME (Local Interpretable Model-Agnostic Explanations) technique [23] is used to explain the model's prediction by building a model that can interpret the prediction locally. LIME works by slightly perturbing the input data and observing the changes in the output. From there, LIME identifies important features that affect the prediction result. LIME constructs a

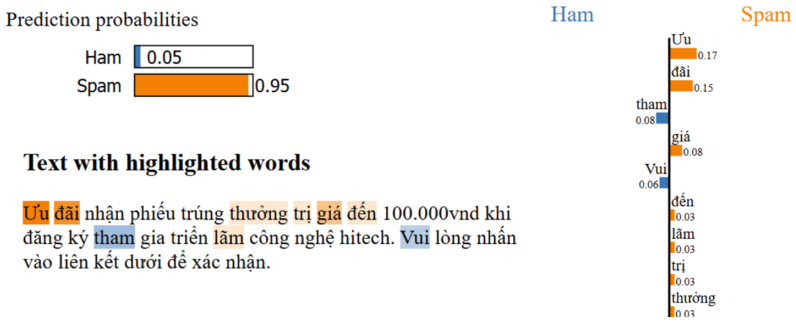

Figure 24: Spam email prediction explanation with LIME

locally interpretable model $\hat{f}(x)$ [24] that approximates the behavior of the black-box model f around a specific instance x.

$$\hat{f}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g), \tag{7}$$

where $\hat{f}(x)$ is the locally interpretable model; G is the set of possible interpretable models; $L(f, g, \pi_x)$ is the loss function that measures how well g approximates f in the neighborhood of x; π_x is the proximity function. It assigns higher weights to perturbed instances closer to x; $\Omega(g)$ is the complexity term to ensure the interpretability by limiting the number of features.

Figure 24 shows the results of using LIME to explain why the Vietnamese spam email detection model classifies an email as spam. This figure illustrates the analysis results of the model's prediction using LIME for a Vietnamese email. (In English: "Special offer: Receive a prize voucher worth up to 100,000 VND when registering to attend the Hitech technology exhibition. Please click the link below to confirm".) Based on the predicted probability of this email being spam, as Figure 24 (the predicted probability is 0.95 which is close to 1), words like "Special offer", "worth" (in English), ... are important words or features to predict that email is spam.

In summary, this study constructed a Vietnamese spam email dataset and conducted extensive experiments on five deep learning models and two traditional machine learning models with explainable AI. Compared with the related works discussed in Section 2, this study contributes significantly to enhancing the detection of Vietnamese spam emails by providing a new dataset, evaluating a diverse set of models, and incorporating explainable AI techniques.

5. CONCLUSION

This study introduced a newly constructed Vietnamese spam email dataset comprising 6,008 emails and proposed the VNSED system for detecting Vietnamese spam emails. To evaluate the effectiveness of spam classification, we implemented and compared five deep learning models (CNN, LSTM, BiLSTM, PhoBERT, and PhoBERT+CNN) as well as two traditional

machine learning models (*Naïve Bayes* and SVM). The results show that deep learning models generally achieved high performance, particularly PhoBERT and its hybrid versions, with PhoBERT+Pruning attaining an F1-score of 87.36% and PhoBERT+CNN achieving 86.29%. To reduce training time and computational overhead, this study applied model pruning by reducing the number of encoder layers in PhoBERT-based architectures. The pruned models maintained competitive accuracy while significantly decreasing training time. These findings demonstrate that pruning is a promising strategy for improving the deployment feasibility of large language models in resource-constrained or distributed environments. Moreover, this study used explainable artificial intelligence (XAI) techniques, specifically LIME, to interpret the predictions of the deep learning models. This enhances model transparency and helps end-users better understand which textual features influence classification outcomes. In future work, additional spam samples from various sources will be collected and annotated to capture more diverse linguistic patterns and spam techniques. Other future directions include applying attention mechanisms to better capture key content within long emails and evaluating the model adaptability in real-time spam filtering systems.

ACKNOWLEDGEMENT

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

REFERENCES

[1] Statista, "Global spam volume as percentage of total e-mail traffic from 2011 to 2023," 2024. [Online]. Available: https://www.statista.com/statistics/420400/spam-email-traffic-share-annual/

[2] Kaspersky, "Employees can lose around two business days per year sorting out spam emails," 2022. [Online]. Available: https://www.kaspersky.com/about/press-releases/2022_employees-can-lose-around-two-business-days-per-year-sorting-out-spam-emails

[3] E. H. Tusher, M. A. Ismail, M. A. Rahman, A. H. Alenezi, and M. Uddin, "Email spam: A comprehensive review of optimize detection methods, challenges, and open research problems," *IEEE Access*, vol. 12, pp. 143 627–143 657, 2024.

[4] A. Poobalan, K. Ganapriya, K. Kalaivani, and K. Parthiban, "A novel and secured email classification using deep neural network with bidirectional long short-term memory," *Computer Speech & Language*, vol. 89, p. 101667, 2025.

[5] H. Q. Anh, P. T. Anh, P. S. Nguyen, and P. D. Hung, "Federated learning for Vietnamese SMS spam detection using pre-trained PhoBERT," in *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 2024, pp. 254–264.

[6] M. T. Hadi and S. S. Baawi, "Email spam detection by machine learning approaches: A review," in *International Conference on Forthcoming Networks and Sustainability in the AIoT Era*. Springer, 2024, pp. 186–204.

[7] T. Amutha and S. Geetha, "Automated spam detection using sandpiper optimization algorithm-based feature selection with the machine learning model," *IETE Journal of Research*, vol. 70, no. 2, pp. 1472–1479, 2024.

[8] G. Nasreen, M. M. Khan, M. Younus, B. Zafar, and M. K. Hanif, "Email spam detection by deep learning models using novel feature selection technique and BERT," *Egyptian Informatics Journal*, vol. 26, p. 100473, 2024.

[9] C. N. Mohammed and A. M. Ahmed, "A semantic-based model with a hybrid feature engineering process for accurate spam detection," *Journal of Electrical Systems and Information Technology*, vol. 11, no. 1, p. 26, 2024.

[10] J. Chen, L. Zhang, and M. Jiang, "A 1D convolutional neural network for spam classification," in *Intelligent Management of Data and Information in Decision Making: Proceedings of the 16th FLINS Conference on Computational Intelligence in Decision and Control & the 19th ISKE Conference on Intelligence Systems and Knowledge Engineering (FLINS-ISKE 2024)*. World Scientific, 2024, pp. 291–298.

[11] A. Isra'a and Y. Qussai, "Spam email detection using deep learning techniques," *Procedia Computer Science*, vol. 184, pp. 853–858, 2021.

[12] Z. B. Siddique, M. A. Khan, I. U. Din, A. Almogren, I. Mohiuddin, and S. Nazir, "Machine learning-based detection of spam emails," *Scientific Programming*, vol. 2021, no. 1, p. 6508784, 2021.

[13] V. S. Tida and S. Hsu, "Universal spam detection using transfer learning of BERT model," *arXiv preprint arXiv:2202.03480*, 2022.

[14] V. M. Tuan, "Vietnamese SMS spam detection with deep learning and pre-trained language model: Array," *Journal of Science and Technology on Information and Communications*, vol. 1, no. 2, pp. 71–75, 2022.

[15] Y. Guo, Z. Mustafaoglu, and D. Koundal, "Spam detection using bidirectional transformers and machine learning classifier algorithms," *Journal of Computational and Cognitive Engineering*, vol. 2, no. 1, pp. 5–9, 2023.

[16] The Messaging Company, "7 tips to identify spam emails," 2024. [Online]. Available: https://themessaging.co/blog/is-this-a-spam-email/

[17] B. Guo, C. Zhang, J. Liu, and X. Ma, "Improving text classification with weighted word embeddings via a multi-channel TextCNN model," *Neurocomputing*, vol. 363, pp. 366–374, 2019.

[18] C. Olah, "Understanding LSTM networks," 2015. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[19] S. Ravichandiran, *"Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT"*. Packt Publishing Ltd, 2021.

[20] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," *arXiv preprint arXiv:2003.00744*, 2020.

[21] Jupyter Council Members, "Project Jupyter," 2024. [Online]. Available: https://jupyter.org/

[22] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and M. Isard, "TensorFlow: a system for large-scale machine learning," in *12th USENIX symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016, pp. 265–283.

[23] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, and G. Menegaz, "A perspective on explainable artificial intelligence methods: SHAP and LIME," *Advanced Intelligent Systems*, vol. 7, no. 1, p. 2400304, 2025.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144.

Received on February 10, 2025
Revised on September 26, 2025