Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 4 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i04.53383

PAPER

High Performance of LSTM on Dengue Shock Syndrome Detection Using DNA Sequence Encoding Methods

Lailil Muflikhah[1](✉), Agustin Iskandar[2], Novanto Yudistira[1], Bambang Nurdewanto[3]

[1]Faculty of Computer Science, Brawijaya University, Malang, Indonesia

[2]Faculty of Medicine, Brawijaya University, Malang, Indonesia

[3]Faculty of Technology Information, Universitas Merdeka, Malang, Indonesia

lailil@ub.ac.id

ABSTRACT

Dengue fever (DF) is a significant global health challenge, affecting approximately 390 million people annually and imposing substantial public health and economic burdens. Accurate DNA sequence classification is crucial for identifying genetic factors in diseases such as DF. However, many machine learning (ML) models for disease detection rely on basic encoding methods such as one-hot encoding, which fail to fully exploit the sequential and contextual nature of DNA data. To address this limitation, this study applies long short-term memory (LSTM) networks, a neural architecture adept at handling sequential data, to classify DNA sequences for detecting dengue and dengue shock syndrome (DSS). The study evaluates three encoding techniques—one-hot encoding, term frequency-inverse document frequency (TF-IDF), and Word2Vec—using datasets of 3,458 DNA sequences sourced from genomics repositories. Preprocessing included the removal of non-ACGT sequences and duplicates to ensure data integrity, followed by under-sampling to address class imbalance. Experimental results demonstrate that the LSTM model with Word2Vec encoding achieved the highest accuracy (0.98), significantly outperforming other encoding techniques. Word2Vec captures contextual and semantic relationships within DNA sequences, enabling superior classification performance. These findings highlight the potential of combining advanced encoding techniques with LSTM networks to improve the accuracy of disease detection models. The study's approach offers promising implications for genomic diagnostics, particularly in resource-limited settings, and lays the foundation for future research into applying similar methodologies to other diseases or datasets.

KEYWORDS

long short-term memory (LSTM), dengue shock syndrome (DSS), DNA sequence, encoding method

1 INTRODUCTION

Dengue fever (DF) is a critical global health challenge, particularly in tropical and subtropical regions where it is endemic. The World Health Organization (WHO)

Muflikhah, L., Iskandar, A., Yudistira, N., Nurdewanto, B. (2025). High Performance of LSTM on Dengue Shock Syndrome Detection Using DNA Sequence Encoding Methods. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(4), pp. 79–98. https://doi.org/10.3991/ijoe.v21i04.53383

Article submitted 2024-11-18. Revision uploaded 2025-01-15. Final acceptance 2025-01-15.

estimates that approximately 390 million dengue infections occur annually, with about 96 million presenting clinically across over 100 countries [1]. This widespread prevalence imposes substantial strain on healthcare systems, particularly in low- and middle-income countries, where resources are often insufficient to manage the disease burden effectively [2]. The severe manifestation of dengue, known as dengue shock syndrome (DSS), presents life-threatening complications, including vascular leakage, severe bleeding, and organ failure. Without timely diagnosis and intervention, DSS can lead to high morbidity and mortality rates, further exacerbating its impact on public health [3].

Dengue shock syndrome represents the most severe form of dengue virus infection, often characterized by heightened vascular permeability, thrombolytic, and potential organ failure, which can lead to fatal outcomes if not diagnosed and managed swiftly [4], [5], [6]. The dengue virus, a member of the Flaviviridae family and primarily transmitted by Aedes mosquitoes, causes a clinical spectrum of illnesses ranging from mild fever to severe forms, including dengue hemorrhagic fever (DHF) and DSS [7], [8]. The WHO has categorized dengue manifestations as DF, DHF, and DSS, each requiring specific diagnostic and management strategies [9].

The burden of DF extends beyond immediate health consequences, creating significant economic and social challenges. Lost productivity, increased healthcare costs, and long-term impacts on affected communities are among the far-reaching consequences of the disease [10]. Urbanization, global travel, and climate change have further intensified the frequency and severity of dengue outbreaks, emphasizing the urgent need for comprehensive and effective diagnostic and management strategies [11]. Environmental factors, coupled with population dynamics, contribute to the propagation of outbreaks, necessitating an integrated approach to disease control [12].

The underlying pathophysiology of DSS involves a complex interplay between the virus and the host's immune system, with secondary infections often exacerbated by a phenomenon known as antibody-dependent enhancement (ADE). This process intensifies disease severity and makes certain demographics, particularly children, more susceptible to severe outcomes [7], [8]. While early detection is crucial for effective treatment, current diagnostic methods such as serological testing and PCR for viral RNA are not always accessible, especially in resource-limited settings [4], [13]. This emphasizes the need for innovative, accessible, and accurate diagnostic approaches. Despite advancements in diagnostic tools, accurately identifying severe cases such as DSS remains a significant challenge, particularly in resource-constrained settings where laboratory infrastructure is often lacking. Traditional methods, such as serological testing and polymerase chain reaction (PCR) for detecting viral RNA, are effective but frequently inaccessible or cost-prohibitive in underserved regions [14]. These limitations underscore the need for innovative, cost-effective, and scalable diagnostic methods that facilitate early detection and intervention.

Advancements in technology have significantly improved healthcare quality and diagnostics. For example, mobile clinical decision support systems enhance real-time decision-making during clinical practice, improving healthcare delivery in diverse settings [15]. Additionally, intelligent knowledge systems, such as those for monitoring depression, demonstrate the potential of artificial intelligent in early detection and intervention [16]. Furthermore, interactive technologies in healthcare education have enhanced training, preparing professionals to adopt innovative diagnostic tools [17]. These developments align with this study's focus on leveraging machine learning (ML) models and encoding methods for effective DSS detection, addressing critical healthcare challenges.

Machine learning approaches, especially neural network-based models, offer promising solutions for improving diagnostic accuracy in dengue cases. Studies have demonstrated the efficacy of ML models in predicting dengue incidence and severity, showcasing their potential for real-world applications in clinical settings [18]. Among these, long short-term memory (LSTM) networks—a specialized form of recurrent neural networks (RNNs)—stand out for their ability to process sequential data effectively. LSTM networks have been shown to excel in sequence prediction tasks, making them well-suited for analyzing genetic data related to DSS [19]. By integrating genomic data from dengue virus strains, these models can refine predictions, yielding valuable insights into viral behavior and its clinical implications [20].

Recent advancements in ML, especially LSTM networks, offer promising solutions for analyzing DNA sequences related to dengue. LSTM networks, a type of RNN, have shown impressive results in sequence prediction tasks, suggesting their potential to identify and predict DSS cases based on genetic data [21]. These models can capture temporal patterns in epidemiological and genomic data, making them valuable tools in early-warning systems for dengue outbreaks and facilitating timely interventions that can improve patient outcomes [22], [23]. By incorporating genomic data from dengue virus strains, LSTM networks can refine predictions further, yielding insights into viral behavior and clinical implications.

Long short-term memory networks are particularly suited for handling sequential data, which is essential for analyzing DNA sequences. Prior studies, such as those by Gunasekaran et al. (2021), have demonstrated that LSTM and bidirectional LSTM models perform well in DNA sequence classification when paired with feature extraction techniques such as k-mer encoding. This technique divides DNA sequences into overlapping sub-sequences of a specified length k, helping models identify important patterns within the sequences [24]. The choice of k significantly impacts the model's accuracy and efficiency, as observed by Zhang and Shen in 2019, who highlighted the need to balance model complexity and computational feasibility in achieving optimal performance [25]. This balance is critical for DSS detection, where rapid and precise identification of the dengue virus enables timely interventions.

Encoding DNA sequences meaningfully is key to leveraging LSTM models effectively. Techniques such as k-mer encoding, which quantifies nucleotide sequence frequencies, enhance the ability to detect viral signatures within DNA sequences [26]. Additional encoding strategies, including one-hot encoding and Word2Vec embedding, further improve the representation of DNA sequences, making them more compatible with ML algorithms and boosting model performance in classification tasks [22]. These encoding methods transform raw DNA sequence data into more informative formats, ultimately increasing the predictive power of ML models.

Another valuable approach is TF-IDF (term frequency-inverse document frequency), which helps prioritize the most informative k-mer in the datasets, allowing the model to focus on essential features for DSS detection [27]. This method aligns with existing research underscoring the role of effective feature selection in optimizing model performance [25].

In recent years, encoding methods such as TF-IDF and Word2Vec, originally developed for natural language processing, have been adapted to biological sequence data, allowing more contextually aware representations [28]. Word2Vec, for instance, embeds DNA sequences into continuous vector spaces that retain contextual relationships, making it particularly valuable for sequence classification [29], [30]. While these methods have advanced the field, few studies have systematically compared multiple encoding techniques within a single framework. Therefore, this study

addresses this gap by evaluating the effectiveness of one-hot encoding, TF-IDF, and Word2Vec in conjunction with LSTM networks for DNA sequence classification, providing a comprehensive analysis of encoding techniques in DSS detection.

2 THE PROPOSED METHOD

The overall block diagram for DSS detection based on the patient's DNA sequence begins with the input of the DNA sequence datasets into the LSTM model architecture. Before the sequence data is entered, the datasets undergo preprocessing. The next step involves constructing the architecture using the LSTM algorithm. Following this, the model is trained, and the validation process is conducted on the developed model. After the model has been trained using the training data, it is tested on the prepared test data to evaluate its performance in classifying diabetes. A flowchart of the general proposed method is shown in Figure 1.

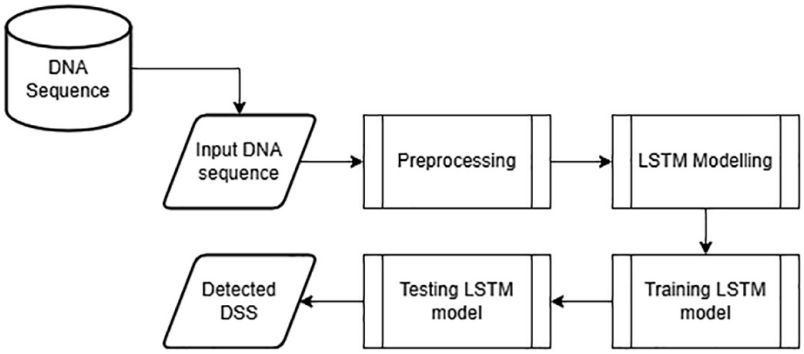

Fig. 1. General proposed research method

2.1 Dataset

The data used in this study is secondary data obtained from a genetic and protein database and health-related literature, specifically from the National Center for Biotechnology Information (NCBI). The data was gathered using keywords related to DSS to ensure relevance to the research focus. Subsequently, the collected sequence data was aligned with the existing NCBI database using the bioinformatics tool BLAST, with the aim of obtaining additional sequences showing a homogeneity level of over 90%. This approach allowed for more comprehensive and representative datasets aligned closely with the study's objectives. The datasets were sourced from NCBI, comprising 3458 DNA sequences with two categories: dengue (59.2%) and shock (40.8%). After removing non-"A-T-G-C" characters and eliminating duplicates, 1961 sequences remained, with a balanced distribution across both classes after under-sampling.

2.2 Preprocessing data

Several preprocessing steps were applied to the dataset to ensure quality and balance. First, non-ACGT sequences were removed, leaving 3,326 sequences that passed the initial filter. Following this, duplicate sequences were identified and removed,

which further reduced the dataset to 1,961 sequences. Lastly, under-sampling was performed to ensure equal representation from both classes, creating a balanced dataset for subsequent analysis. These steps were crucial in preparing the data for accurate and reliable results, as shown in Figure 2.

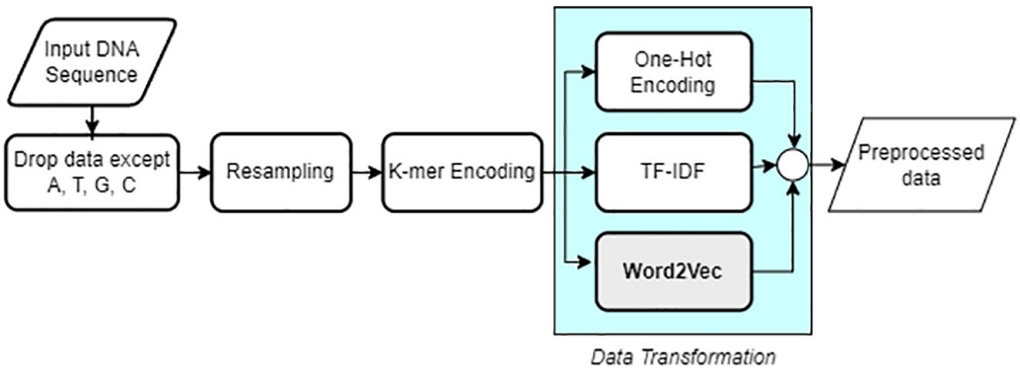

Fig. 2. Preprocessing data steps

K-mer encoding. K-mer encoding is a fundamental technique in bioinformatics, used to analyze DNA or protein sequences by breaking them into smaller, overlapping fragments of a fixed length. Each fragment has a specified length k-mer by fragment of length *k*. It consists of *k* nucleotides or amino acids. For example, a DNA sequence such as "ACGTACG" can be divided into 3 k-mers of length, resulting in fragments such as "ACG," "CGT," "GTA," and "TAC." This k-mer captures the local sequential patterns within the data and serves as the basic unit for computational analysis. This encoding technique has been applied in various bioinformatics applications, such as genome classification, species identification, and protein structure prediction [31], [32].

In our study, k-mer encoding is used to preprocess DNA sequences for input into an LSTM model. By breaking sequences into k-mers, the study leverages this method to capture fine-grained patterns in the DNA data, which are crucial for identifying genetic markers associated with dengue shock syndrome.

Data transformation using encoding methods. In this study, three encoding techniques were employed to represent DNA sequences for classification. One-hot encoding converts each DNA nucleotide (A, C, G, T) into a unique binary vector, ensuring a straightforward numerical representation of all possible nucleotide triplets, though it results in high-dimensional and sparse data. In contrast, TF-IDF assigns weights to these triplets based on their frequency in the datasets, highlighting the importance of less common sequences while diminishing the influence of frequent ones. However, TF-IDF cannot capture positional relationships within the sequences. Lastly, Word2Vec offers a more advanced approach by generating dense, low-dimensional vector embedding that encode the contextual relationships between DNA triplets, treating them such as words in a sentence, as illustrated in Figure 3. This method captures both the semantic and positional nuances in the sequences, making it highly effective for complex genomic data. Together, these techniques provide varied approaches to feature extraction, each with its strengths in terms of dimensional reduction, information retention, and sequence. Furthermore, the encoding stage involves transforming the DNA sequence from a string of letters into a numerical format. A combination of k-mer encoding and word embedding will be used to encode the DNA sequence data. In the k-mer encoding technique,

each DNA sequence is transformed into a k-mer of a specified size. Following this transformation, each DNA sequence represented as a k-mer will undergo an embedding process, converting it into a matrix of real-valued vectors.

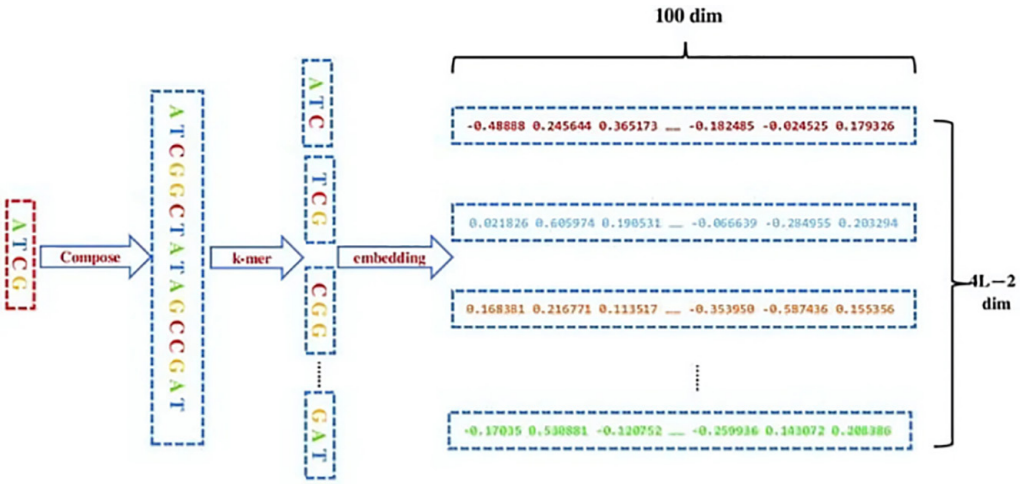

Fig. 3. Word2Vec DNA sequence encoding

2.3 Long short-term memory

Long short-term memory is a modification of the RNN. This method emerged with the addition of memory cells and addresses the problem of vanishing gradients when processing long sequential data in RNNs [33]. LSTM has memory cells and an architecture consisting of an input gate, recurrent connections, a forget gate, and an output gate. LSTM can also retain long-term information. The input gate functions to decide which input values will be forwarded for updating. The output gate is the result of a sigmoid layer that determines which cell will be generated. The forget gate allows memories to be forgotten [34]. Figure 4 presents the LSTM architecture.

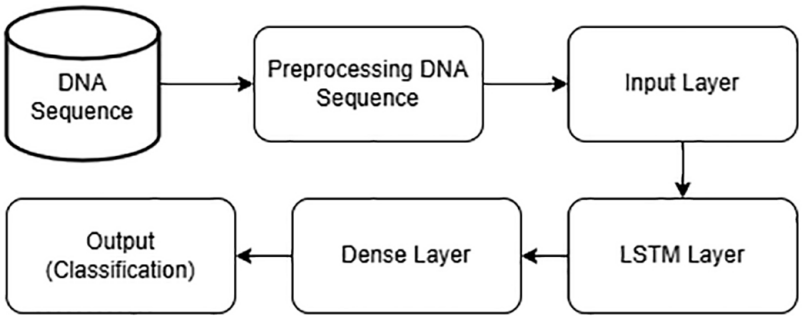

Fig. 4. Architecture of the proposed model

The architecture development process is carried out based on predefined configurations and parameters, which are subsequently tested. If the constructed architecture performs adequately, the configuration will be used in the next steps. Conversely, if the architecture fails to achieve satisfactory classification results, improvements will be made. These improvements involve adding fully connected layers (dense layers) based on the evaluation of the current architecture.

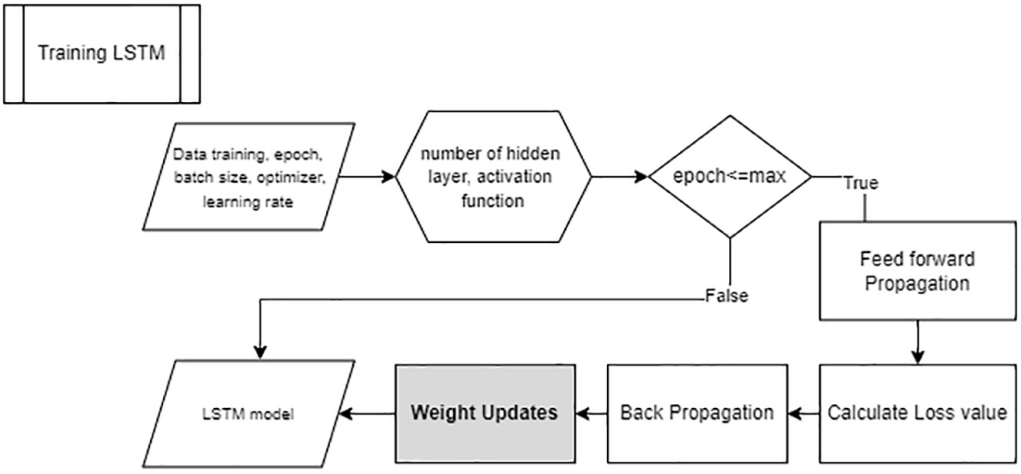

Fig. 5. Flowchart of training model

The model training process is conducted iteratively over a predefined number of epochs, incorporating various adjustments throughout as shown in Figure 5. These adjustments include altering the number of neurons in the hidden layer and changing the activation functions. The iterative process consists of several key steps: forward propagation, loss calculation, backward propagation, and weight updates. Once the iterations reach the maximum epoch, the training process is concluded, yielding an LSTM model that will be utilized in the testing phase, as shown in Figure 6.

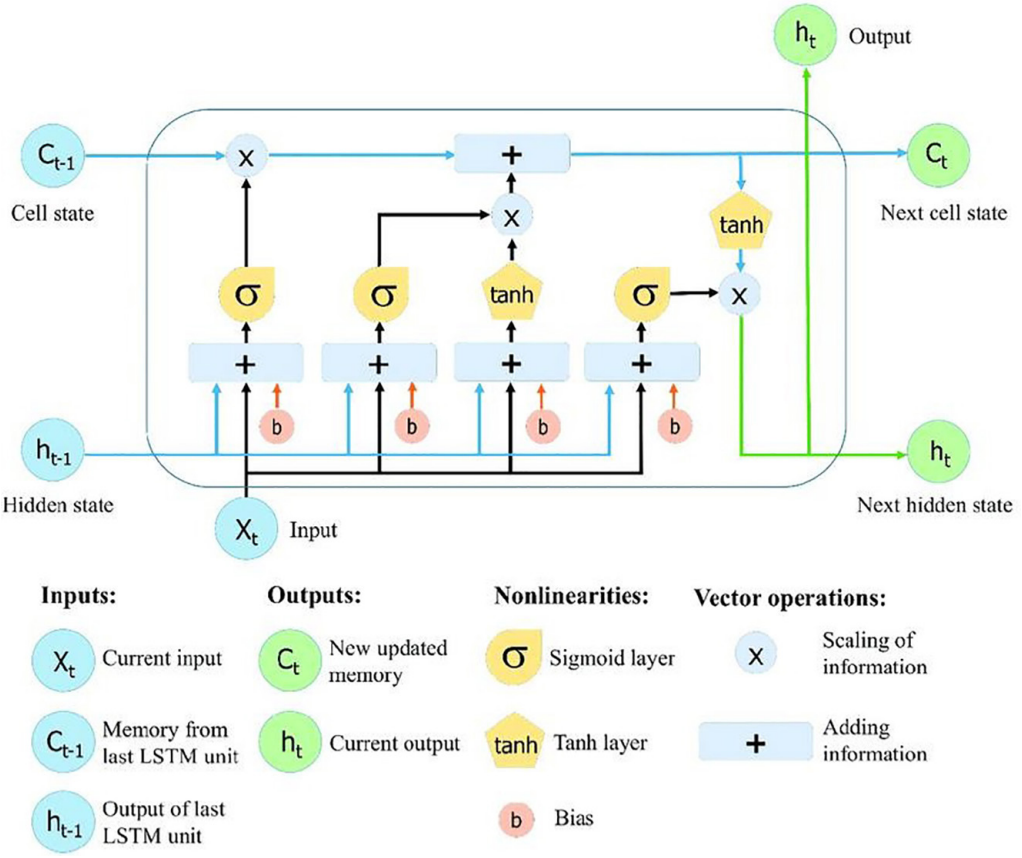

Fig. 6. LSTM architecture

The first step in LSTM is responsible for determining the information to be discarded from C_{t-1} using a sigmoid function applied to the forget gate. This gate receives the concatenated values of s_{t-1} and x_t and produces a value between 0 and 1. When the value is 0, it indicates that the information will be discarded, whereas a value of 1 means the information can be passed forward. The manual calculation of the value from the forget gate can be seen in Equation (1) as follows.

$$f_t = \sigma(W_{hf}.\, x_t + W_{hf}.\, s_{t-1} + b_f) \tag{1}$$

f_t: forget gate
σ: function of sigmoid activation
W_{hf}: weight on each gate (waktu ke-t)
s_{t-1}: the previous state
x_t: current input
b_f: bias on forget gate

The second step determines what information will be added and stored in the cell state. This step results from combining s_{t-1} and x_t using two functions: a sigmoid function as the input gate and a tanh function as the intermediate gate. The outputs of these two functions are multiplied to obtain the information that will be added to the cell state. The value of the input gate is shown in Equation (2), while the candidate value and new cell state can be seen in Equations (3) and (4).

$$i_t = \sigma(W_{hi}.\, x_t + W_{hi}.\, h_{t-1} + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(W_{hc}.x_t + W_{hc}.h_{t-1} + b_c) \tag{3}$$

i_t: input gate
$\tilde{C}_t$: cell state
σ: sigmoid tanh function
W_{hc}: weight of cell state
W_{hi}: weight of input gate
x_t: current input
b_i: bias of input gate
b_c: bias of cell state

$$c_t = (f_t * c_{t-1} + i_t * \tilde{C}_t) \tag{4}$$

c_t: the new cell state
f_t: the result of forget gate
c_{t-1}: cell state before to t
i_t: the result of input gate
$\tilde{C}_t$: candidate value

Next, add the output from the forget gate in the first step. The final step is responsible for determining the output of the LSTM. To generate the output, perform a sigmoid calculation on the combination of s_{t-1} and x_t, applied to the output gate. The output gate determines how much of the cell state value will be produced at s_{t-1}. Then, calculate the tanh function value of e and multiply it by the output gate value. The result of this multiplication will be the output of the LSTM unit, as shown in Equations (5) and (6).

$$o_t = \sigma(W_{ho}.\, x_t + W_{ho}.\, h_{t-1} + b_o) \tag{5}$$

$$h_1 = o_t * \tanh(c_t) \tag{6}$$

3 RESULTS

In this study, testing was conducted on the proposed method involving various encoding representations of DNA sequence data within an LSTM model for detecting DSS. Three encoding methods were applied: one-hot encoding, word2vec, and TF-IDF, to transform the sequence data as input layers for the LSTM model. During the testing phase, the shortest DNA sequence identified during preprocessing was 145 nucleotides, which was set as the standardized length for all datasets used in modeling. DNA sequences longer than 145 nucleotides were excluded from the analysis to maintain uniformity. The training process employed a total of 70 epochs with a learning rate of 0.01 and utilized a single LSTM layer for the model architecture. Predictions were made using the best-performing model weights, selected based on the highest validation accuracy achieved during the training phase. This approach ensured that the model was optimized for accuracy while maintaining computational efficiency.

3.1 Long short-term memory model with one-hot encoding

In the method of one-hot encoding, all combinations of the nucleotide sequence ACGT were divided into groups of three, resulting in 64 unique combinations. The dataset used for this process was balanced to ensure fair model training and evaluation. The best epoch during training, determined based on validation accuracy, was achieved at the 14th epoch, and the model weights from this epoch were utilized for predictions. The training process, including the validation loss and accuracy trends, is illustrated in Figures 7 and 8, demonstrating the model's performance improvement over successive epochs.

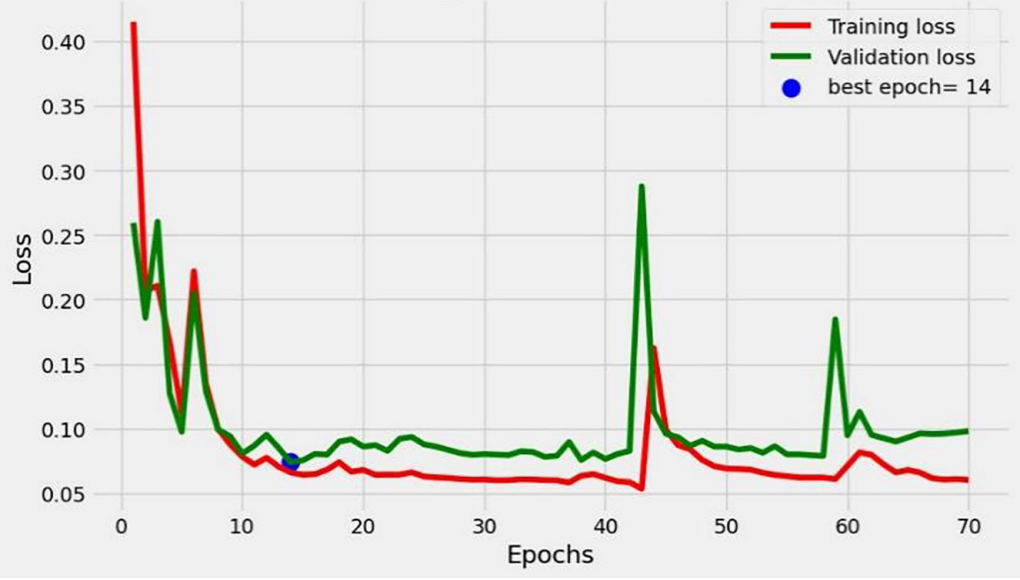

Fig. 7. Training and validation loss result of LSTM with one-hot encoding

The graph, as in Figure 7, presents the training and validation loss over 70 epochs during the model's training process. The red curve represents the training loss, which decreases consistently with some fluctuations, reflecting the model's gradual optimization on the training dataset. The green curve denotes the validation loss, which follows a similar decreasing trend initially but exhibits more variability, particularly

after epoch 40. A blue marker highlights the best epoch (epoch 14), corresponding to the point where the validation loss is at its optimal value. The x-axis represents the number of epochs, ranging from 0 to 70, while the y-axis indicates the loss values, ranging from 0.0 to 0.45. Overall, the graph shows effective learning in the early stages, but the increased variability in validation loss after epoch 40 suggests potential overfitting or instability, emphasizing the importance of early stopping or additional regularization methods.

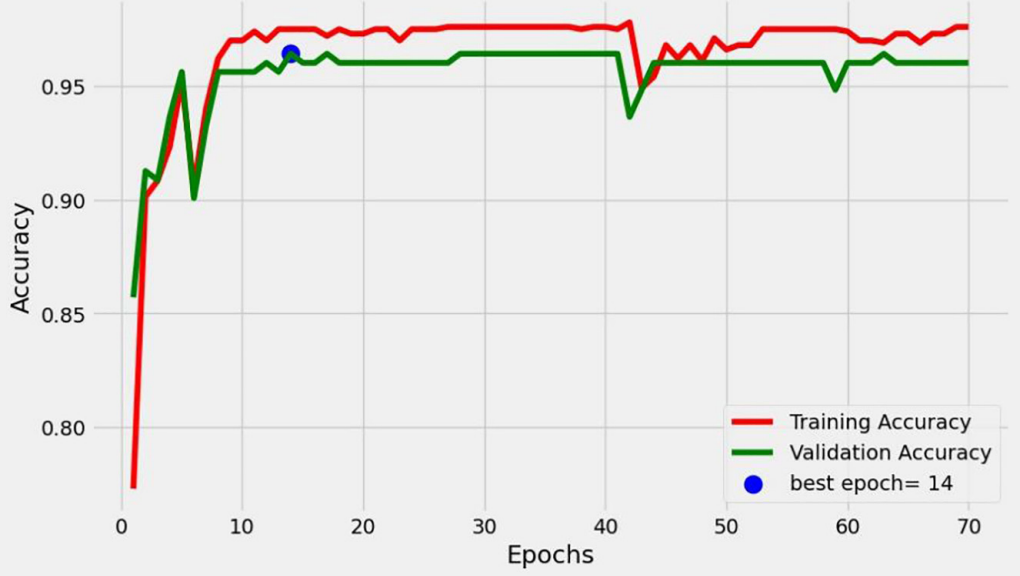

Fig. 8. Training and validation accuracy result of LSTM with one-hot encoding

The graph, as in Figure 8, illustrates the training and validation accuracy of over 70 epochs during the training process. The red line represents training accuracy, which increases rapidly in the initial epochs and stabilizes at a value above 0.95, demonstrating the model's improved performance on the training dataset. The green line indicates the validation accuracy, which also improves early on, stabilizing near 0.95 but with occasional drops in performance, particularly around epoch 40 and epoch 60, suggesting minor fluctuations in generalization performance. A blue marker identifies the best epoch (epoch 14), corresponding to the highest validation accuracy achieved during training. The x-axis displays the number of epochs (0–70), while the y-axis represents accuracy, ranging from 0.80 to 1.00. The graph indicates successful learning, though the fluctuations in validation accuracy suggest potential areas for further refinement, such as tuning hyperparameters or employing regularization techniques.

3.2 LSTM model with TF-IDF

Another encoding method, TF-IDF, DNA sequences were divided into three-part combinations of ACGT and embedded using the TF-IDF technique. The hyperparameters utilized in this process were consistent with those applied in the previous encoding methods. The best epoch during training, based on validation accuracy, was achieved at the 67th epoch. The validation accuracy trends during training are illustrated in Figures 9 and 10, highlighting the model's optimization and performance over the training process.

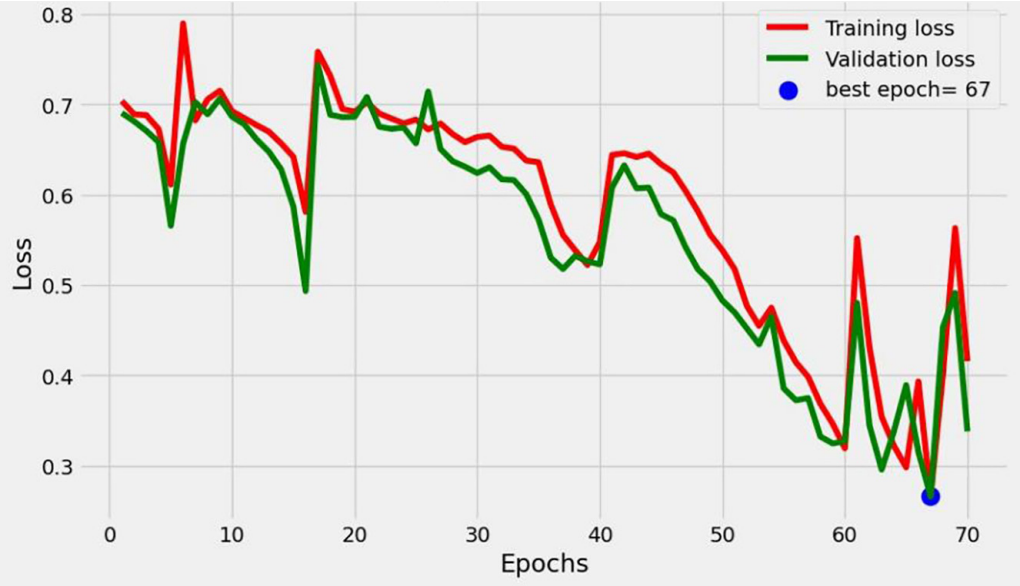

Fig. 9. Training and validation loss result of LSTM with TF-IDF encoding

The graph, as in Figure 9, illustrates the training and validation loss results of the LSTM model with TF-IDF encoding for detecting DSS across 70 epochs. The red curve represents the training loss, while the green curve denotes the validation loss. The loss values demonstrate a downward trend, indicating that the model is learning effectively. Initially, both training and validation loss exhibit fluctuations, with validation loss slightly trailing the training loss. As training progresses, the losses converge, showing stability and suggesting that the model generalizes well without overfitting. Around epoch 67, the lowest validation loss is achieved, marked by a blue dot, which signifies the best epoch for the model. At this point, the model achieves optimal performance with minimal loss. Overall, the graph highlights the effectiveness of combining LSTM with TF-IDF encoding in identifying DSS, showing consistent improvement and a convergence of losses over training epochs.

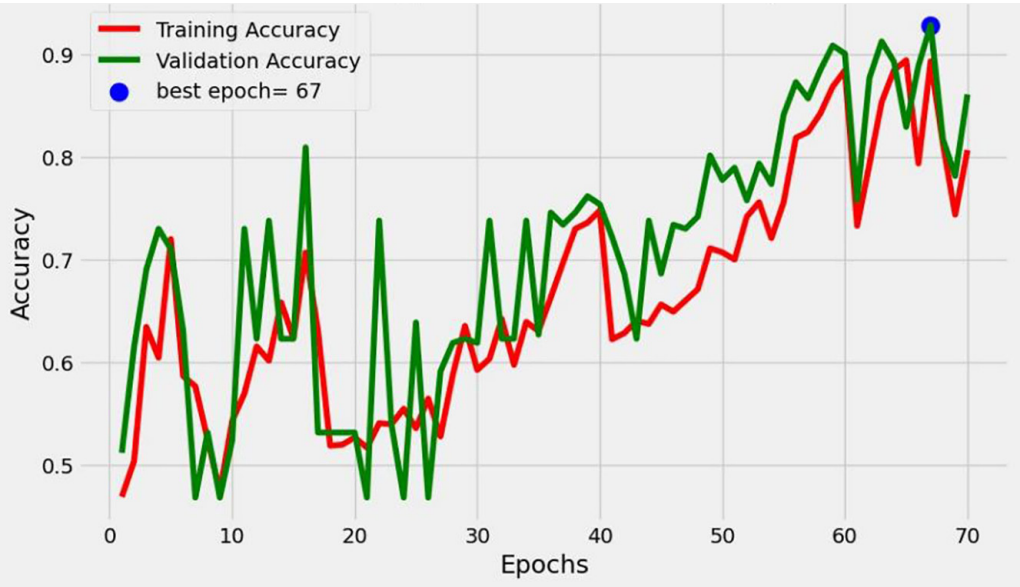

Fig. 10. Training and validation accuracy result of LSTM with TF-IDF encoding

The graph in Figure 10 depicts the training and validation accuracy results of the LSTM model with TF-IDF encoding for detecting DSS across 70 epochs. The red curve represents the training accuracy, while the green curve shows the validation accuracy. Initially, the accuracy for both training and validation fluctuates significantly, reflecting the model's adjustment phase. As training progresses, both accuracies exhibit a steady upward trend, indicating the model's improved performance and learning. Around Epoch 67, the highest validation accuracy is achieved, marked by a blue dot, denoting the best epoch where the model achieves optimal generalization.

The consistent increase and convergence of training and validation accuracies, particularly after epoch 40, suggest that the model performs well on both the training and validation datasets. This highlights the effectiveness of the LSTM model combined with TF-IDF encoding in achieving high accuracy for DSS detection.

3.3 LSTM Model with Word2Vec

The last representation method using Word2Vec encoding, the DNA sequences were split into three-part combinations of ACGT and embedded using the Word2Vec technique. The hyperparameters used in this process were consistent with those employed in the previous encoding method. The best epoch during training, determined based on validation accuracy, was achieved at the 14th epoch. The trends in validation accuracy during the training process are presented in Figures 11 and 12, showcasing the model's performance and optimization over time.

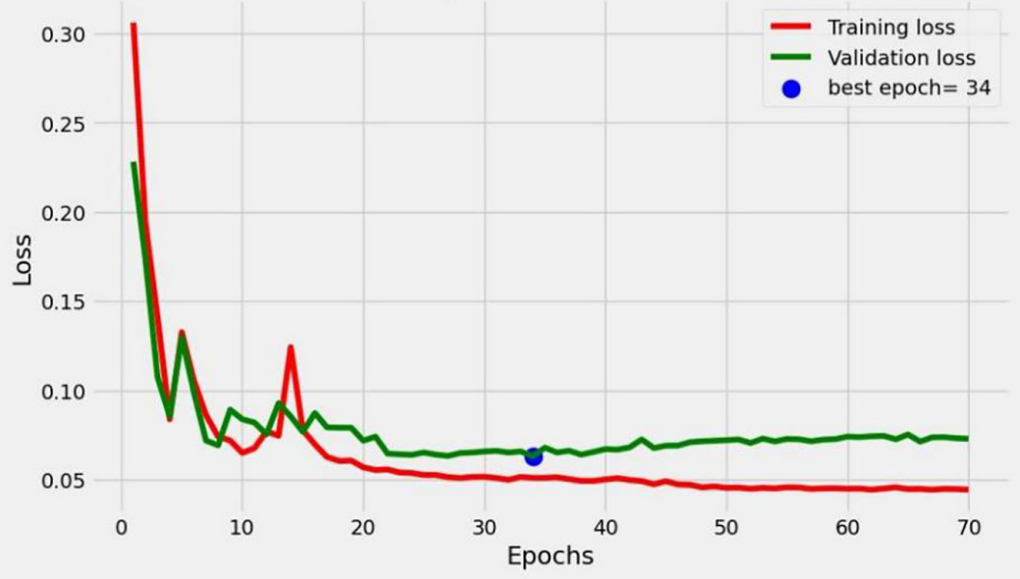

Fig. 11. Training and validation loss result of LSTM with Word2Vec encoding

The graph in Figure 11 illustrates the training and validation loss results of the LSTM model with Word2Vec encoding for detecting DSS across 70 epochs. The red curve represents the training loss, while the green curve denotes the validation loss. At the start, the loss for both training and validation is relatively high but drops sharply within the first 10 epochs, indicating rapid learning by the model during the initial phase. After this, the loss stabilizes and decreases gradually over the

remaining epochs, with the training loss consistently lower than the validation loss, reflecting effective learning without significant overfitting. The best epoch is marked at epoch 34, where the validation loss reaches its minimum, denoted by a blue dot. This signifies the point at which the model achieves optimal performance on the validation datasets. Overall, the graph highlights the effectiveness of combining LSTM with Word2Vec encoding, showcasing a stable and low loss for both training and validation, which indicates high performance and generalization.

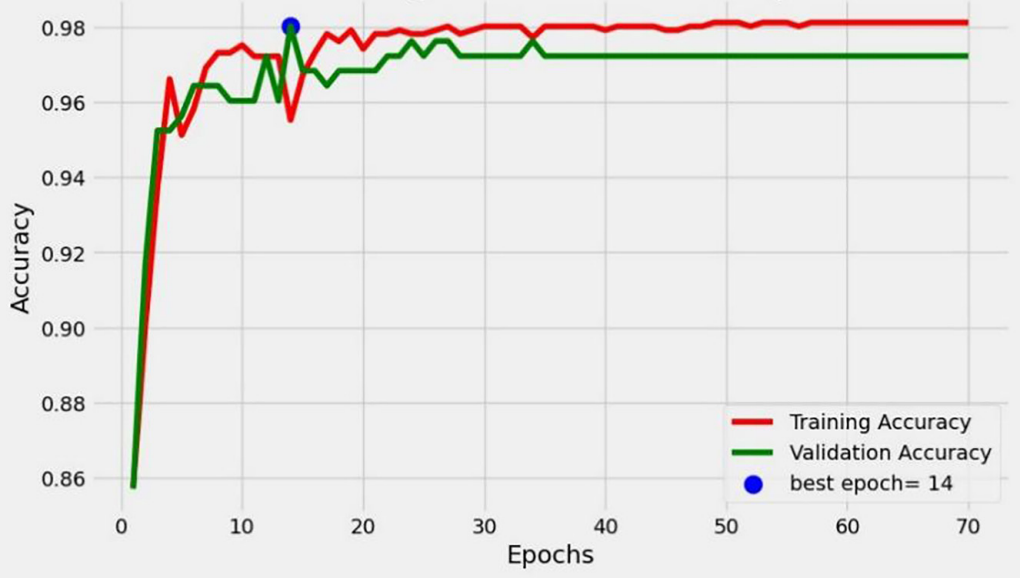

Fig. 12. Training and validation accuracy of LSTM with Word2Vec encoding

Last, the graph in Figure 12 illustrates the training and validation accuracy results of the LSTM model with Word2Vec encoding for detecting DSS over 70 epochs. The red curve represents the training accuracy, while the green curve denotes the validation accuracy. Initially, both training and validation accuracies rise sharply, indicating rapid improvement in the model's ability to learn from the data. By epoch 14, the highest validation accuracy is achieved, marked by a blue dot, which signifies the best epoch with optimal generalization. After this point, both training and validation accuracies stabilize at approximately 98%, showing minimal fluctuations. The close alignment of training and validation accuracies throughout the training process demonstrates the model's robustness and lack of overfitting. The graph highlights the high effectiveness of the LSTM model paired with Word2Vec encoding in achieving exceptional accuracy for DSS detection.

4 DISCUSSION

The study demonstrated that the choice of encoding method significantly impacts the performance of LSTM models for DNA sequence classification. The Word2Vec method consistently outperformed one-hot encoding and TF-IDF, likely due to its ability to capture contextual relationships between DNA bases. The performance of the LSTM models for DSS detection using different encoding methods—one-hot, TF-IDF, and Word2Vec—shows distinct variations across several key metrics, including accuracy, precision, recall, and F1-score, as outlined in Table 1.

Table 1. Comparison of performance for transformation encoding model

Model	Best Epoch	Accuracy	Precision	Recall	F1-Score
LSTM + One-Hot	14	0.96	0.93	1.00	0.96
LSTM + TF-IDF	67	0.93	0.91	0.94	0.93
LSTM + Word2Vec	14	0.98	0.97	0.99	0.98

Table 1 illustrates the performance of LSTM models trained with different encoding techniques, highlighting metrics across accuracy, precision, recall, and F1-score. The LSTM model with one-hot encoding, achieving optimal results at epoch 14, has an accuracy of 0.96, high recall (1.00), and a balanced F1 score (0.96). This configuration shows that one-hot encoding effectively supports high recall, meaning it successfully captures true positives while maintaining a strong overall performance. In contrast, the LSTM with TF-IDF encoding, reaching its best at epoch 67, has lower metrics across the board, with an accuracy of 0.93 and precision, recall, and F1-score all around 0.91–0.94. This indicates that TF-IDF may be less effective in this setting, with a slightly reduced ability to capture relevant patterns compared to other encoders. Lastly, the LSTM with Word2Vec, achieving its best performance at epoch 14, displays the highest scores: an accuracy of 0.98, precision of 0.97, recall of 0.99, and F1-score of 0.98. This shows that Word2Vec provides the most robust representation, supporting high accuracy and balanced recall and precision, making it the most effective encoding method among the three for this LSTM model.

Furthermore, the comparative analysis of the three encoding methods demonstrates a clear distinction in their ability to represent DNA sequences for DSS detection using LSTM models. Word2Vec consistently delivered superior results, achieving the highest accuracy (0.98), precision (0.97), recall (0.99), and F1-score (0.98). This performance underscores Word2Vec's strength in capturing the contextual relationships among nucleotide sequences, effectively translating biological nuances into meaningful computational features. The training and validation accuracy graph (see Figure 12) further supports these findings, showing a rapid increase in accuracy during the early epochs and stable performance thereafter, with minimal fluctuations. Notably, the close alignment between training and validation accuracies indicates the model's robustness and lack of overfitting, a key factor in its success.

In contrast, the LSTM model with one-hot encoding achieved 0.96 accuracy, but its limitations were evident in the higher dimensionality and sparsity of the encoded data. This was reflected in the training and validation loss graph (see Figure 7), which exhibited a more pronounced variability in validation loss after Epoch 14. Similarly, the accuracy graph (see Figure 8) showed occasional drops in validation accuracy despite the overall strong performance. These fluctuations suggest that while one-hot encoding effectively captures true positives (evident from its perfect recall of 1.00), it struggles with finer contextual relationships, leading to occasional misclassifications.

The TF-IDF encoding method, with an accuracy of 0.93, precision of 0.91, recall of 0.94, and F1-score of 0.93, further highlights the challenges of frequency-based feature extraction. The corresponding loss graph (see Figure 9) revealed a stable downward trend, indicating effective learning, but with a higher validation loss compared to Word2Vec. The accuracy graph (see Figure 10) showed consistent improvement but lacked the stability observed in Word2Vec, suggesting a reduced capacity to generalize across the dataset.

Overall, the graphs of loss and accuracy collectively emphasize the critical role of the encoding method in determining model performance. Word2Vec's ability to generate dense, low-dimensional vector embedding that encode semantic and positional nuances of DNA sequences significantly enhanced the LSTM model's performance and stability. These results demonstrate the importance of integrating advanced embedding techniques into sequence classification workflows, paving the way for improved diagnostic tools in bioinformatics and medical applications.

4.1 Model performance and metrics

The performance of each encoding technique is summarized in Table 1. Among the methods tested, the LSTM model with Word2Vec encoding achieved the highest accuracy of 99%, outperforming one-hot encoding and TF-IDF. However, given the high accuracy levels, overfitting concerns were evaluated. To effectively discuss the confusion matrix results, we would describe the performance of the LSTM models using each encoding technique based on the key metrics derived from the matrix, including true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as is shown in Figures 13, 14, and 15.

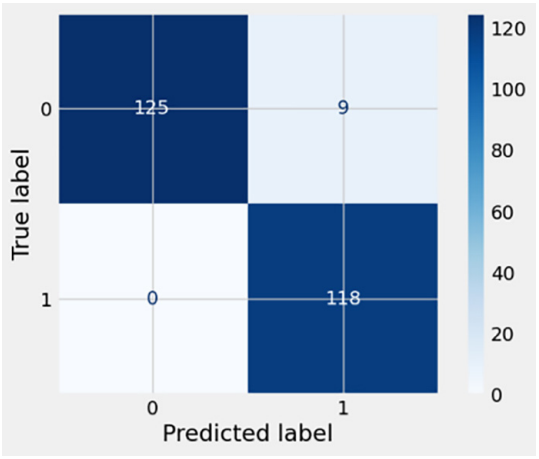

Fig. 13. Confusion matrix of result using LSTM with one-hot encoding

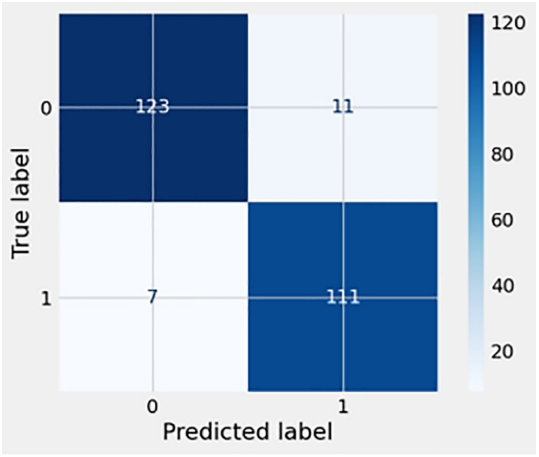

Fig. 14. Confusion matrix of result using LSTM with TF-IDF encoding

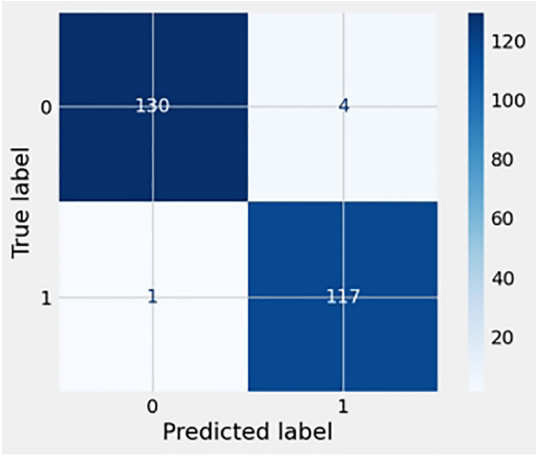

Fig. 15. Confusion matrix of result using LSTM Word 2Vec encoding

The analysis of the three confusion matrices highlights the varied performance of the LSTM model using different encoding techniques for DSS detection. The one-hot encoding method demonstrated high recall, correctly identifying all DSS cases with no false negatives (FN = 0). However, it showed moderate precision due to the presence of 9 false positives (FP), indicating some misclassifications of non-DSS cases as DSS. TF-IDF Encoding offered a balanced performance, with 111 true positives (TP) and 123 true negatives (TN), but the model's ability was slightly hindered by 7 false negatives (FN) and 11 false positives (FP), leading to a reduction in both recall and precision compared to One-Hot Encoding. On the other hand, the Word2Vec encoding method significantly outperformed the others, achieving the highest accuracy with 130 true negatives (TN) and 117 true positives (TP). It exhibited minimal misclassifications, with only 4 false positives (FP) and 1 false negative (FN), showcasing exceptional precision and recall. These results emphasize that Word2Vec encoding, by effectively capturing contextual relationships within DNA sequences, is the most robust and accurate approach for DSS detection among the three methods evaluated.

5 CONCLUSION

This study highlights the effectiveness of integrating LSTM neural networks with advanced encoding techniques, particularly Word2Vec and TF-IDF, for the detection of DSS. Among the encoding strategies, the LSTM model paired with Word2Vec encoding demonstrated superior accuracy and stability, making it a promising tool for real-world diagnostic applications. This approach is particularly valuable in resource-limited settings where rapid and accurate detection is essential to improve patient outcomes. The proposed methodology offers a scalable framework that could be extended to other diseases with similar diagnostic challenges, enhancing medical diagnostics across various domains.

The results underline the potential of ML in addressing critical healthcare needs, especially in genomic diagnostics. The integration of sophisticated neural network models with robust encoding strategies provides a versatile tool for improving diagnostic precision and accessibility. However, further research is necessary to validate these findings using larger and more diverse datasets to ensure the generalizability and robustness of the proposed approach.

6 ACKNOWLEDGMENT

This study was supported by DIKTI under the 2024 Fundamental Research Grant Program with contract number:00309.2/UN10.A0501/B/PT.01.03.2/2024

7 REFERENCES

[1] G. Iqbal, H. Javed, F. A. Raza, U. F. Gohar, W. Fatima, and M. Khurshid, "Diagnosis of acute dengue virus infection using enzyme-linked immunosorbent assay and real-time PCR," *Canadian Journal of Infectious Diseases and Medical Microbiology*, vol. 2023, no. 1, p. 3995366, 2023. https://doi.org/10.1155/2023/3995366

[2] R. P. Khetan *et al.*, "Profile of the 2016 dengue outbreak in Nepal," *BMC Research Notes*, vol. 11, p. 423, 2018. https://doi.org/10.1186/s13104-018-3514-3

[3] H. Harapan, A. Michie, R. T. Sasmono, and A. Imrie, "Dengue: A minireview," *Viruses*, vol. 12, no. 8, p. 829, 2020. https://doi.org/10.3390/v12080829

[4] S. D. Lakshmi, P. N. Devi, and C. Saikumar, "The seroprevalence of dengue in a tertiary care hospital," *Int. J. Curr. Microbiol. App. Sci.*, vol. 7, no. 9, pp. 43–51, 2018. https://doi.org/10.20546/ijcmas.2018.709.006

[5] A. Ahmed *et al.*, "Dengue fever in the Darfur area, Western Sudan," *Emerging Infectious Diseases*, vol. 25, no. 11, p. 2126, 2019. https://doi.org/10.3201/eid2511.181766

[6] G. B. M. Wirajaya, A. A. P. P. D. Sutanegara, and D. N. D. Lestari, "Variations of dengue shock syndrome cases and their management: Report of three cases," *Intisari Sains Medis*, vol. 13, no. 3, pp. 625–631, 2022. https://doi.org/10.15562/ism.v13i3.1507

[7] K. M. Senthilkumar and R. Hema Harini, "Clinical profile of dengue fever in children presented at a tertiary care hospital," *International Journal of Contemporary Pediatrics*, vol. 6, no. 2, pp. 761–764, 2019. https://doi.org/10.18203/2349-3291.ijcp20190726

[8] Z. Ahmad and C. L. Poh, "The conserved molecular determinants of virulence in dengue virus," *International Journal of Medical Sciences*, vol. 16, no. 3, pp. 355–365, 2019. https://doi.org/10.7150/ijms.29938

[9] Q. T. Islam, H. T. Hossain, M. A. Khandaker, H. N. Ahasan, M. Majumder, and T. Jabeen, "Dengue expanded syndrome: An unusual presentation," *Bangladesh Journal of Medicine*, vol. 29, no. 1, pp. 45–47, 2018. https://doi.org/10.3329/bjmed.v29i1.35408

[10] J. A. Potts *et al.*, "Prediction of dengue disease severity among pediatric Thai patients using early clinical laboratory indicators," *PLOS Neglected Tropical Diseases*, vol. 4, no. 8, p. e769, 2010. https://doi.org/10.1371/journal.pntd.0000769

[11] H. Zhang *et al.*, "Predictive symptoms and signs of severe dengue disease for patients with dengue fever: A meta-analysis," *BioMed Research International*, vol. 2014, no. 1, p. 359308, 2014. https://doi.org/10.1155/2014/359308

[12] A. E. Laureano-Rosario *et al.*, "Application of artificial neural networks for dengue fever outbreak predictions in the Northwest Coast of Yucatan, Mexico and San Juan, Puerto Rico," *Tropical Medicine and Infectious Disease*, vol. 3, no. 1, p. 5, 2018. https://doi.org/10.3390/tropicalmed3010005

[13] D. T. Zade, D. K. Srinivas, and D. A. Berad, "The study of detection of dengue cases by NS1 antigen and IGM antibody in RIMS, Adilabad, India," *International Journal of Medical and Biomedical Studies*, vol. 3, no. 11, pp. 103–106, 2019. https://doi.org/10.32553/ijmbs.v3i11.720

[14] D. N. Anggraini Ningrum, Y.-C. J. Li, C.-Y. Hsu, M. Solihuddin Muhtar, and H. Pandu Suhito, "Artificial intelligence approach for severe dengue early warning system," *Stud Health Technol Inform*, vol. 310, pp. 881–885, 2024. https://doi.org/10.3233/SHTI231091

[15] A. Schols, J. Donkers, M. Voorend, D. Verstegen, H. Hoogland, and P. Kubben, "The use of smartphones and mobile clinical decision support systems in clinical clerkships: A pilot study," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 7, no. 2, pp. 80–84, 2013. https://doi.org/10.3991/ijim.v7i2.2446

[16] H. Yu, G. Zhang, J. Liu, and K. Li, "Intelligent knowledge service system based on depression monitoring of college students," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 14, no. 12, pp. 71–84, 2019. https://doi.org/10.3991/ijet.v14i12.10702

[17] G. N. Georgieva-Tsaneva and I. Serbezova, "Research on the impact of innovative interactive technologies in the education of health care students," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 17, no. 20, pp. 283–291, 2022. https://doi.org/10.3991/ijet.v17i20.32903

[18] P. Guo *et al.*, "Developing a dengue forecast model using machine learning: A case study in China," *PLOS Neglected Tropical Diseases*, vol. 11, no. 10, p. e0005973, 2017. https://doi.org/10.1371/journal.pntd.0005973

[19] S. K. Dey, K. M. M. Uddin, H. Md. H. Babu, Md. M. Rahman, A. Howlader, and K. M. A. Uddin, "Chi2-MI: A hybrid feature selection based machine learning approach in diagnosis of chronic kidney disease," *Intelligent Systems with Applications*, vol. 16, p. 200144, 2022. https://doi.org/10.1016/j.iswa.2022.200144

[20] Y. Chen, A. Huang, Y. Sui, X. Tong, and F. Yu, "Progress and development of three types of live attenuated vaccines for dengue fever," *Highlights in Science, Engineering and Technology*, vol. 8, pp. 497–504, 2022. https://doi.org/10.54097/hset.v8i.1204

[21] Z. Mumtaz, Z. Rashid, R. Saif, and M. Z. Yousaf, "Deep learning guided prediction modeling of dengue virus evolving serotype," *Heliyon*, vol. 10, no. 11, p. e32061, 2024. https://doi.org/10.1016/j.heliyon.2024.e32061

[22] M. A. Majeed, H. Z. M. Shafri, A. Wayayok, and Z. Zulkafli, "Prediction of dengue cases using the attention-based long short-term memory (LSTM) approach," *Geospatial Health*, vol. 18, no. 1, 2023. https://doi.org/10.4081/gh.2023.1176

[23] A. Y. Saleh and L. Baiwei, "Dengue prediction using deep learning with long short-term memory," in *2021 1st International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, Sana'a, Yemen, 2021, pp. 1–5. https://doi.org/10.1109/eSmarTA52612.2021.9515734

[24] H. Gunasekaran, K. Ramalakshmi, A. R. M. Arokiaraj, S. D. Kanmani, C. Venkatesan, and C. S. G. Dhas, "Analysis of DNA sequence classification using CNN and Hybrid models," *Computational and Mathematical Methods in Medicine*, vol. 2021, p. e1835056, 2021. https://doi.org/10.1155/2021/1835056

[25] Q. Zhang, Z. Shen, and D. S. Huang, "Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network," *Sci. Rep.*, vol. 9, p. 8484, 2019. https://doi.org/10.1038/s41598-019-44966-x

[26] M. N. Melaugh, S. Coleman, and D. Kerr, "A computational approach to uncertainty in DNA sequences," in *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2023, pp. 1043–1048. https://doi.org/10.1109/SSCI52147.2023.10371838

[27] S. Juneja, A. Dhankhar, A. Juneja, and S. Bali, "An approach to DNA sequence classification through machine learning: DNA sequencing, K Mer counting, thresholding, sequence analysis," *International Journal of Reliable and Quality E-Healthcare (IJRQEH)*, vol. 11, no. 2, pp. 1–15, 2022. https://doi.org/10.4018/IJRQEH.299963

[28] Z. B. Ozger, "A robust protein language model for SARS-CoV-2 protein–protein interaction network prediction," *Artificial Intelligence in Medicine*, vol. 142, p. 102574, 2023. https://doi.org/10.1016/j.artmed.2023.102574

[29] R. Ren, C. Yin, and S. S.-T. Yau, "kmer2vec: A novel method for comparing DNA sequences by word2vec embedding," *Journal of Computational Biology*, vol. 29, no. 9, pp. 1001–1021, 2022. https://doi.org/10.1089/cmb.2021.0536

[30] Q. Geng, R. Yang, and L. Zhang, "A deep learning framework for enhancer prediction using word embedding and sequence generation," *Biophysical Chemistry*, vol. 286, p. 106822, 2022. https://doi.org/10.1016/j.bpc.2022.106822

[31] C. Moeckel *et al.*, "A survey of k-mer methods and applications in bioinformatics," *Computational and Structural Biotechnology Journal*, vol. 23, pp. 2289–2303, 2024. https://doi.org/10.1016/j.csbj.2024.05.025

[32] B. K. Sarkar, A. R. Sharma, M. Bhattacharya, G. Sharma, S.-S. Lee, and C. Chakraborty, "Determination of k-mer density in a DNA sequence and subsequent cluster formation algorithm based on the application of electronic filter," *Sci. Rep.*, vol. 11, no. 1, p. 13701, 2021. https://doi.org/10.1038/s41598-021-93154-3

[33] G. Van Houdt, C. Mosquera, and G. Nápoles, "A review on the long short-term memory model," *Artif. Intell. Rev.*, vol. 53, pp. 5929–5955, 2020. https://doi.org/10.1007/s10462-020-09838-1

[34] X.-H. Le, H. V. Ho, G. Lee, and S. Jung, "Application of long short-term memory (LSTM) neural network for flood forecasting," *Water*, vol. 11, no. 7, p. 1387, 2019. https://doi.org/10.3390/w11071387

8 AUTHORS

Lailil Muflikhah received B.Sc. degree in Computer Science from the Institut Teknologi Sepuluh Nopember (ITS) M.Sc. degree in Computer Science from the Universiti Teknologi Petronas (UTP), Malaysia, and a Ph.D. degree in Biology Engineering from Brawijaya University. She is currently an Associate Professor with the Department of Informatics Engineering in the Faculty of Computer Science, Brawijaya University. Her research interests include biomedical engineering, bioinformatics, soft computing, machine learning, and intelligent systems (E-mail: lailil@ub.ac.id).

Agustin Iskandar obtained a Bachelor's degree in Medical Science from the Faculty of Medicine at Airlangga University in Surabaya, Indonesia. She pursued her Master's degree in Biomedical Interest at Brawijaya University, specializing in Clinical Pathology. She also pursued PhD in Doctoral Program of Medical Science, Faculty of Medicine Brawijaya University. Currently, she serves as a Lecturer in the Department of Parasitology and Clinical Pathology at Universitas Brawijaya in Malang, Indonesia. Her research interests encompass various areas, including Clinical Pathology and Laboratory Medicine, Infectious Diseases, and disease prognosis modeling.

Novanto Yudistira is a Lecturer and Researcher at the Faculty of Computer Science, Universitas Brawijaya, Indonesia. He earned his Bachelor's degree in Informatics Engineering from Institut Teknologi Sepuluh Nopember (2007) and a master's in computer science from Universiti Teknologi Malaysia (2011). He is a Doctorate in Information Engineering from Hiroshima University, Japan (2018). He has collaborated with renowned institutions like AIST, RIKEN, and Osaka University, focusing on informatics, data analytics, and deep learning. His research interests include deep learning, multi-modal computer vision, medical informatics, integration of large language and visual models, and big data analytics. He is the founder of Deep Learning Indonesia, an active community for advancing deep learning algorithms. He has published and reviewed works in esteemed journals and conferences and collaborates with various institutions on research and community projects.

Bambang Nurdewanto is an academician affiliated with Universitas Merdeka Malang, focusing on Information Systems. His research interests cover various innovative topics, including the application of Fuzzy methods in decision-making for Small and Medium Enterprises exports and the development of raw material stock monitoring systems for printing. He completed his postgraduate studies at Institut Teknologi Sepuluh Nopember (ITS), with a thesis centered on dynamic and multi-project software development using a heuristic approach. Through his work, Bambang actively contributes to advancing science and technology to support SMEs and other industries.