ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika

ISSN(p): 2338-8323 | ISSN(e): 2459-9638 | Vol. 14 | No. 3 | Halaman 358 – 373
DOI : http://dx.doi.org/10.26760/elkomika.v14i3.358 July 2026

An HHO-Optimized LSTM Framework for Predicting Adverse Effects Associated with Reproductive and Breast Disorders

Article History:
Received
23 April 2026
Revised
4 June 2026
Accepted
2 July 2026

ANGEL METANOSA AFINDA[1], IGA NARENDRA PRAMAWIJAYA[2], FAUZAN FIRDAUS[3]

[1]Computer Engineering, School of Electrical Engineering, Telkom University, Indonesia
[2]Biomedical Engineering, School of Electrical Engineering, Telkom University, Indonesia
[3]Software Engineering, School of Informatics, Telkom University, Indonesia
Email: angelmetanosa@telkomuniversity.ac.id

ABSTRAK

Prediksi toksisitas reproduksi merupakan tantangan dalam pengembangan obat karena efek samping sering sulit terdeteksi dini. Representasi SMILES digunakan untuk pemodelan berbasis urutan. Penelitian ini mengusulkan model LSTM yang dioptimasi dengan Harris Hawks Optimization (HHO) untuk meningkatkan prediksi efek samping pada sistem reproduksi dan payudara. Empat skema arsitektur dievaluasi, termasuk L (LSTM saja), CL (Convolution + LSTM), LD (LSTM + Dense), dan CLD (Convolution + LSTM + Dense). Hasil menunjukkan skema L hasil tuning memberikan performa terbaik, dengan akurasi meningkat dari 0.6304 menjadi 0.6739 dan F1-score dari 0.6792 menjadi 0.7097. Temuan ini menegaskan efektivitas optimasi metaheuristik dalam pemodelan toksikologi komputasional.

***Kata kunci**: sistem reproduksi, SMILES, LSTM, Harris Hawks Optimization, prediksi efek samping*

ABSTRACT

Reproductive toxicity prediction is a major challenge in drug development as side effects are often difficult to detect early. SMILES representations provide a compact sequential format suitable for deep learning. This study proposes an HHO-optimized LSTM model to predict reproductive and breast-related side effects. Four architectural schemes were evaluated including L (LSTM only), CL (Convolution + LSTM), LD (LSTM + Dense), and CLD (Convolution + LSTM + Dense). Results show that the tuned L scheme achieved the best performance with accuracy increasing from 0.6304 to 0.6739 and F1-score from 0.6792 to 0.7097. These findings highlight the effectiveness of metaheuristic optimization in computational toxicology modeling.

***Keywords**: reproductive system, SMILES, LSTM, Harris Hawks Optimization, side effects prediction*

1. INTRODUCTION

Reproductive toxicity has become an increasingly complex global health issue with broad implications for societal quality of life. The World Health Organization reports that one in six individuals of reproductive age, or approximately 17.5% of the global adult population, experiences infertility at some point in their lives **(*Infertility*, n.d.)**. Exposure to pesticides, industrial chemicals, and pharmaceutical residues contributes to endocrine disruption, elevated oxidative stress, and impaired reproductive organ function, thereby increasing the risk of reproductive system disorders **(Nik Hazlina et al., 2022)**. A global assessment covering the period from 1990 to 2021 indicates a rising prevalence of infertility and disability-adjusted life years associated with reproductive conditions, in line with lifestyle changes and growing exposure to chemical agents **(Feng et al., 2025)**. Furthermore, the toxicological effects of drugs on the reproductive system present additional challenges, as the biological responses to therapeutic compounds are not always detectable through conventional testing methods. This underscores the need for efficient computational approaches to support broad compound evaluation and enable earlier, more systematic risk prediction **(Fuadah et al., 2024)(Miller et al., 2023)**.

Conventional computational models that rely on handcrafted descriptors often fail to comprehensively capture stereochemistry, molecular flexibility, and long-range interactions that influence toxicological responses **(Karimah et al., 2023)**. Such approaches require predefined feature sets and dataset-specific preprocessing, thereby limiting generalization capability and reducing reproducibility when applied to heterogeneous chemical structures **(Yap, 2011)**. As an alternative, the Simplified Molecular Input Line Entry System (SMILES) has been widely adopted as a compact, informative, and scalable sequence-based molecular representation since its introduction by David Weininger **(M. Veselinovic et al., 2015)(Weininger, 2002)**. Advances in deep learning have further strengthened the utility of SMILES, as recurrent, convolutional, and transformer architectures are capable of extracting meaningful chemical patterns directly from raw sequences without reliance on manual feature engineering **(Goh et al., 2017).** A study by Afinda and colleagues in 2023 demonstrated that SMILES-based deep learning models consistently outperform traditional machine-learning methods in predicting molecular properties and evaluating toxicity, thus reinforcing sequence-based molecular representation as a methodological foundation in modern computational toxicology **(Afinda et al., 2023)**.

Recent advances in computational toxicology indicate a methodological shift toward analytical frameworks that integrate deep learning to model the complex relationships between chemical structures and biological responses more comprehensively. For example, the SMILES2Vec model developed by Goh and colleagues reported an accuracy of up to 88% in predicting chemical properties and demonstrated greater performance stability compared to conventional multilayer-perceptron–based artificial neural networks **(Goh et al., 2018)**. Two-dimensional image–based approaches, such as the Chemception framework, have likewise shown competitive performance relative to traditional QSAR models without requiring explicit chemical descriptors, thereby reducing dependence on manual feature engineering **(Goh et al., 2017)**. In a study conducted by Fan and colleagues on the detection and extraction of drug adverse events from open data sources, the deep learning–based BERT-ADE model achieved an F1-score of 0.89 in identifying adverse event occurrences **(Fan et al., 2020)**. Meanwhile, research by Lee and Chen employing a hybrid GCNN–BiLSTM architecture for drug side-effect prediction reported an accuracy of 0.84 and an AUC of 0.88 **(Lee & Chen, 2021)**. Collectively, these findings demonstrate that deep learning approaches can enhance the

accuracy of drug adverse-effect identification and prediction by capturing complex patterns within both clinical and molecular data.

Various deep learning architectures, including convolutional, transformer-based, and hybrid graph–sequence models, have demonstrated competitive performance in toxicity prediction. However, in studies where molecular structures are represented as Simplified Molecular Input Line Entry System (SMILES) sequences, sequential learning approaches are especially suitable due to the ordered nature of the representation. This representation encodes atoms and chemical bonds linearly, requiring predictive models to capture inter-token dependencies and long-range contextual information to accurately reflect the structural characteristics of a molecule **(Goh et al., 2018)**. Given the sequential nature of SMILES representations, LSTM is particularly suitable for capturing long-range dependencies while mitigating the vanishing-gradient problem. Nevertheless, the performance of LSTM models is highly sensitive to hyperparameter configurations such as the number of hidden units, dropout rate, learning rate, and sequence length. Suboptimal settings may lead to overfitting, slow convergence, and training instability, particularly when dealing with heterogeneous chemical datasets **(Zhou et al., 2024)**. Moreover, the high-dimensional parameter space renders manual tuning or grid search inefficient, thereby necessitating adaptive optimization strategies to enhance model robustness in SMILES-based toxicity prediction **(Igani & Kurniawan, 2025)(Sofian & Kurniawan, 2025)**.

The limitations of manual tuning and grid search in navigating high-dimensional parameter spaces have driven the development of metaheuristic optimization algorithms as automated solutions for hyperparameter tuning in deep learning models, particularly within search landscapes that are broad, complex, and nonlinear. One widely studied method is Harris Hawks Optimization (HHO), introduced by Heidari and colleagues, which employs exploration and exploitation mechanisms inspired by cooperative hunting behavior **(Kaur et al., 2021)**. Several studies have reported the effectiveness of integrating HHO with LSTM architectures for various sequence-based predictive tasks. For example, Wu and collaborators demonstrated that an HHO–LSTM model achieved lower MSE and RMSE values than conventional time-series models in water-quality prediction **(Wu et al., 2024).** Similarly, Mnassri and colleagues showed that HHO-optimized Attention-based LSTM achieved an RMSE of 269 Wh and a MAPE of 9.43%, outperforming both standard LSTM and PSO-based variants in residential energy forecasting **(Mnassri et al., 2025)**. These findings indicate that integrating LSTM with metaheuristic optimization not only improves accuracy but also enhances training stability and convergence efficiency in sequence-modeling tasks.

Based on the literature review, most studies combining deep learning with SMILES-based molecular representations remain limited in exploring adaptive hyperparameter optimization for predicting reproductive toxicity and breast-related adverse effects. Although LSTM and metaheuristic approaches have been applied in various domains, the integration of LSTM and Harris Hawks Optimization (HHO) for reproductive toxicity prediction remains underexplored. As a result, the effectiveness of HHO in optimizing LSTM hyperparameters to improve predictive performance, robustness, and generalization capability for heterogeneous SMILES-based reproductive toxicity datasets remains unclear. This study aims to address this problem by investigating the effectiveness of HHO-based hyperparameter optimization in enhancing the performance of LSTM models for SMILES-based reproductive toxicity prediction and evaluating whether the proposed HHO-LSTM framework can outperform conventional non-optimized LSTM models.

2. METHODS

2.1 Research Workflow

The research methodology is designed to provide a systematic framework for developing a SMILES-based reproductive-toxicity prediction model integrating Harris Hawks Optimization (HHO) and Long Short-Term Memory (LSTM). The workflow consists of five main stages, which are illustrated schematically in Figure 1. This study focuses on the prediction of reproductive toxicity and breast-related adverse effects using SMILES-based molecular representations obtained from the SIDER database. The scope is limited to the development of LSTM models and the application of HHO for hyperparameter tuning. The effectiveness of the proposed framework is evaluated in terms of predictive performance, robustness, and generalization capability.

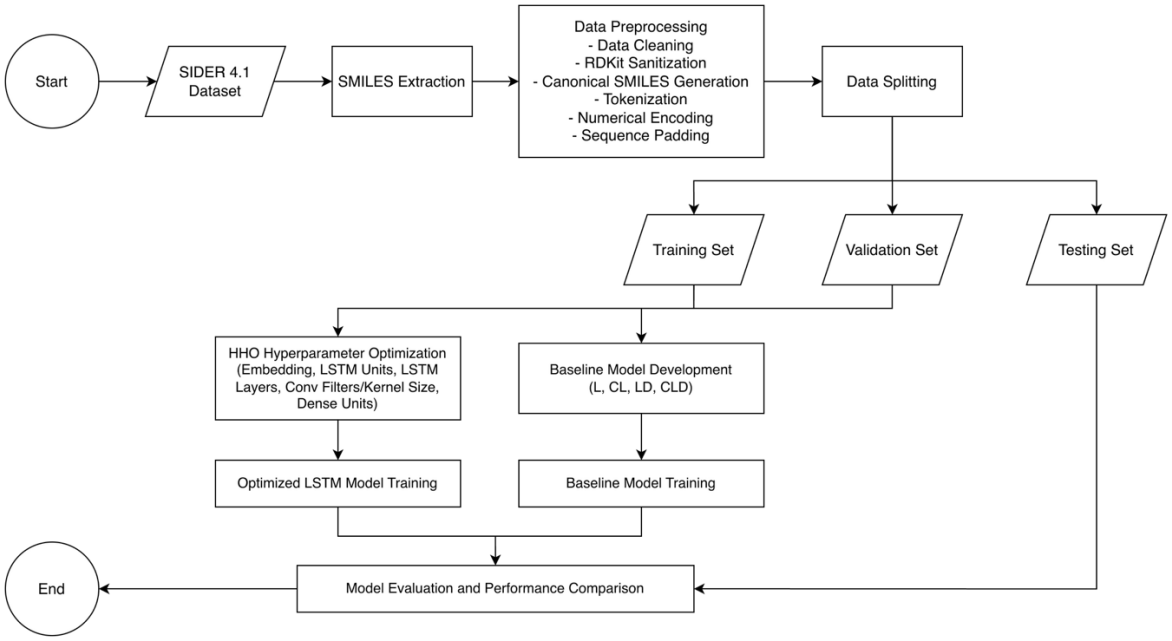

Figure 1. Methodological Workflow of the HHO–LSTM Framework.

The first stage begins with extracting molecular representations in the form of SMILES from drugs associated with reproductive-system adverse effects and breast-related disorders, obtained from the SIDER Side Effect Resource 4.1 (*SIDER Side Effect Resource*, n.d.). The SMILES representation is selected due to its ability to express chemical structures in a concise linear format, thereby facilitating sequence-based modeling in deep learning.

The second stage involves a series of preprocessing steps aimed at ensuring chemical validity, data consistency, and input suitability for deep learning architectures. These steps include data cleaning to remove missing or duplicate entries, SMILES character tokenization, numerical encoding, and sequence padding to a fixed length. A systematic preprocessing pipeline is essential for reducing bias, preserving data integrity, and enabling the model to capture relevant structural patterns. Subsequently, the dataset is divided into training, validation, and testing subsets using a stratified sampling approach. This strategy maintains class distribution across subsets, allowing the model to learn representatively and ensuring unbiased performance evaluation.

The modeling stage consists of two main pipelines. The first pipeline constructs a baseline LSTM model using manually defined hyperparameters, serving as a reference point for assessing the impact of optimization. The second pipeline applies the Harris Hawks Optimization (HHO) algorithm to adaptively explore the hyperparameter space, producing an automatically optimized LSTM configuration. This optimization process is designed to maximize model performance, accelerate convergence during training, and enhance robustness against data variability.

In the final stage, the performance of both the baseline model and the HHO–LSTM model is evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and AUC. This evaluation aims to assess predictive improvements, model stability, and generalization capability on previously unseen data. Overall, the research workflow provides a comprehensive and systematic framework for the development, optimization, and validation of reproductive-toxicity prediction models.

2.2 Dataset Overview

This study employs the SIDER Side Effect Resource 4.1, which provides structured information on approved drugs and systematically reported clinical adverse effects. The dataset was accessed through DeepChem and includes 1,427 drugs, each represented by its molecular structure in SMILES format and annotated with adverse-effect categories classified according to the MedDRA system. For the purposes of this research, the focus is placed on the Reproductive System and Breast Disorders endpoint, as it is directly relevant to reproductive-toxicity prediction.

Out of the 1,427 SMILES sequences, 727 samples are labeled as positive and 700 as negative, resulting in a relatively balanced class distribution. This balance enables the model to be trained without applying resampling techniques or additional adjustments for class imbalance, thereby minimizing potential bias arising from uneven class representation. With a balanced dataset, model performance can be evaluated more fairly and accurately, particularly on metrics such as recall and precision, which are critical for detecting toxic samples.

2.3 Pre-Processing Data

Before being used in sequence-based modeling, all SMILES sequences were processed through a structured preprocessing pipeline to ensure chemical validity, data consistency, and compatibility with the LSTM architecture. The process begins with data cleaning, during which missing or duplicate entries are removed to preserve dataset quality. Each molecule is then parsed and sanitized using RDKit, where the SanitizeMol function eliminates chemically invalid or inconsistent molecular structures. Subsequently, Canonical SMILES representations are generated for each molecule to standardize structural encoding and ensure consistent comparability across the dataset.

To reduce the influence of structural outliers, molecules with more than 200 characters were filtered out, resulting in the removal of 56 sequences and leaving 1,371 valid samples. The cleaned dataset was then divided into training, validation, and testing subsets using a stratified ratio of 70:20:10, yielding 959, 274, and 138 samples, respectively. In addition, a character-level vocabulary was constructed from the 56 unique SMILES tokens present in the training set, enabling each character in the sequence to be numerically encoded. All sequences were subsequently padded to a maximum length of 100 characters to ensure uniform input dimensions, thereby allowing the LSTM to process the data consistently and effectively.

This preprocessing pipeline not only ensures molecular validity and data consistency but also supports the model in capturing key sequential patterns essential for reproductive-toxicity prediction. As a result, potential bias due to outliers or data inconsistencies can be minimized, leading to more reliable and stable model performance during evaluation.

2.4 Long Short-Term Memory

Long Short-Term Memory (LSTM) is an advanced variant of recurrent neural networks (RNNs) designed to address the vanishing and exploding gradient problems, which often limit the ability of standard RNNs to model long-range dependencies **(Sofian & Kurniawan, 2025).** Instead of relying solely on a single hidden state, LSTM maintains a persistent cell state, enabling information to be selectively retained or discarded at each time step. This behavior is governed by three gating mechanisms, illustrated in Figure 2. The input gate regulates how much new information is added to the memory cell, the forget gate removes information that is no longer relevant, and the output gate determines which parts of the updated cell state contribute to the hidden output. Through this gated memory structure, LSTM is able to maintain stable gradient propagation and effectively capture long-sequence patterns, making it highly suitable for modeling SMILES-based molecular representations **(Igani & Kurniawan, 2025)**.

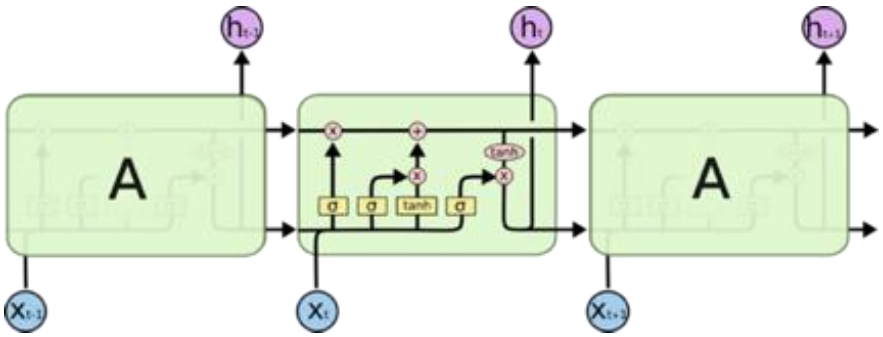

Figure 2. Illustration of the LSTM Architecture.

Mathematically, the behaviors of the LSTM gates and states are described by Equations (1)–(6).

$$f_t = \sigma\left(w_f \left[h_{t-1}, x_t\right] + b_f\right) \quad (1)$$

$$i_t = \sigma\left(w_i \left[h_{t-1}, x_t\right] + b_i\right) \quad (2)$$

$$\hat{C}_t = tanh(W_c \left[h_{t-1}, x_t\right] + b_c) \quad (3)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \hat{C}_t \quad (4)$$

$$o_t = \sigma\left(W_o \left[h_{t-1}, x_t\right] + b_0\right) \quad (5)$$

$$h_t = o_t \; tanh(C_t) \quad (6)$$

where $\boldsymbol{x_t}$ is the input vector, $\boldsymbol{h_{t-1}}$ is the previous hidden state, $\boldsymbol{C_{t-1}}$ is the previous cell state, $\boldsymbol{C_t}$ is the current cell state, and $\boldsymbol{h_t}$ is the current hidden state or output. $\boldsymbol{W}$and $\boldsymbol{b}$ denote the

learnable parameters, σ represents the sigmoid activation function, and $\odot$ indicates element-wise multiplication.

The ability of LSTM networks to retain information across long sequences makes them well-suited for modeling SMILES strings, in which important structural patterns may occur far apart within the sequence. In this study, the LSTM serves as the primary sequence-learning component prior to hyperparameter optimization.

2.5 Model Architecture Scheme

This study was designed to evaluate how variations in architectural components influence the prediction of reproductive toxicity from SMILES-based representations. Four model schemes were selected to highlight the specific role of each architectural block, as summarized in Table 1.

Table 1. Model Architecture Schemes

Scheme	Term
LSTM	L
Convolution + LSTM	CL
LSTM + Dense	LD
Convolution + LSTM + Dense	CLD

The single LSTM scheme (L) focuses on the fundamental capability of LSTM networks to capture sequential dependencies, serving as the baseline for understanding molecular sequences. The Convolution + LSTM scheme (CL) introduces a convolutional layer to extract local motifs prior to sequential modeling. The LSTM + Dense scheme (LD) explores the impact of higher-level feature transformations through a Dense layer, enhancing representations and shaping clearer decision boundaries. The Convolution + LSTM + Dense scheme (CLD) combines all components, resulting in a composite architecture with the richest representation capacity, though more challenging in terms of training stability and optimization.

Table 2. Hyperparameters Optimized by HHO

Parameter	Objective	Range
Number of Layers	Controls architectural depth and prevents overfitting	[1–3]
Embedding Dimension	Balances vector representation capacity and computational cost	[32–128]
LSTM Units	Determines the model's capacity to learn sequential dependencies	[32–128]
Kernel Size (Conv)	Extracts local motifs in SMILES with an appropriate receptive field	[2–10]
Number of Conv Filters	Controls the feature extraction capacity of thesconvolutional layer	[32–128]
Dense Units	Specifies the number of neurons for final representation learning	[32–128]

To maximize the potential of each scheme, every architecture was optimized using the Harris Hawks Optimization (HHO) algorithm, as summarized in Table 2. HHO searches for the best configuration for each architectural block, including the number of layers, embedding dimension, LSTM units, convolutional filters and kernel size, as well as dense-layer neurons.

These parameters influence the model's ability to extract informative features from SMILES sequences, capture long-range dependencies, and discriminate toxic from non-toxic samples.

With carefully designed parameter ranges, HHO systematically explores the search space and adjusts the model's capacity according to the specific needs of each scheme. This approach underscores that the alignment between LSTM, Convolution, and Dense components, combined with proper hyperparameter optimization, serves as a key determinant of predictive performance while also providing critical insights into the contribution of each architectural block.

2.6 Harris Hawks Optimization

Harris Hawks Optimization (HHO) is a population-based metaheuristic algorithm inspired by the cooperative hunting strategies of Harris hawks **(Wu et al., 2024)**. The algorithm alternates between exploration and exploitation phases to identify optimal solutions within a high-dimensional search space, making it well suited for hyperparameter optimization in deep learning models. In this study, HHO is employed to optimize the hyperparameters of the LSTM-based model architecture introduced in Section 2.5. Each hawk in the population represents a candidate configuration that includes the number of layers, embedding dimension, LSTM units, convolutional kernel size, number of convolutional filters, and dense units.

During the optimization process, the position of each hawk is updated relative to the prey, which represents the best solution found so far. The transition between exploration and exploitation is controlled by the escaping energy E, computed using Equation (7).

$$E = 2E_0(1 - \frac{t}{T}) \tag{7}$$

where $E_0$is an initial value randomly sampled from the interval $[-1,1]$, $t$is the current iteration, and $T$is the maximum number of iterations. When $E \geq 1$, the hawk performs global exploration to avoid premature convergence. When $E < 1$, the hawk shifts to exploitation and refines the search through soft or hard besiege mechanisms depending on adaptive probability conditions.

Table 3. HHO Parameter Settings

Parameter	Description	Value
Iterations	Maximum number of optimization iterations	30
Population Size	Number of hawks (candidate solutions)	15
Fitness Epochs	Training epochs used for fitness evaluation	8
Fitness Batch Size	Batch size used for fitness evaluation	32

The HHO parameters used in this study are summarized in Table 3, including the number of iterations, population size, epochs for fitness evaluation, and batch size. Each candidate configuration is evaluated through partial training on the training subset, allowing the fitness value to reflect the model's early performance and guide HHO in identifying optimal hyperparameters. The parameter settings were selected based on preliminary experiments and computational resource considerations.

2.7 Model Evaluation

Model evaluation was conducted to assess the classification performance of both the baseline model and the HHO-optimized model. A confusion matrix was used to categorize predictions into four groups: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These values were subsequently used to compute several evaluation metrics, namely accuracy, precision, recall, and F1-score, which together offer a clear and interpretable assessment of the model's predictive capabilities. The formulas for the evaluation metrics are provided in Equations (8)–(11).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{10}$$

$$\text{F1-Score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{11}$$

This evaluation approach enables a transparent assessment of model performance, highlighting the model's ability to accurately identify toxic samples while minimizing misclassification errors.

3. RESULTS AND DISCUSSION

3.1 Results and Discussion of Non-Tuned Models

Before applying hyperparameter optimization using the HHO algorithm, the baseline architectures were evaluated to understand the intrinsic capabilities of each model without additional tuning. This analysis provides an essential perspective on how each architecture processes the SMILES dataset and how effectively the default settings capture structural patterns associated with toxicity.

3.1.1 Comparative Effectiveness Across Schemes

The baseline performance of the L, CL, LD, and CLD architectures is presented in Table 4. The results demonstrate clear and meaningful differences in predictive capability across the models. These variations highlight the distinct functional contributions of the LSTM, Convolution, and Dense components when processing SMILES-based molecular sequences.

Table 4. Baseline Performance of Each Architecture

Scheme	Accuracy	Recall	Precision	F1-score
L	0.6304	0.7606	0.6136	0.6792
CL	0.6087	0.6197	0.6197	0.6197
LD	0.6377	0.6620	0.6438	0.6528
CLD	0.6304	0.6620	0.6351	0.6483

The LD architecture, which integrates LSTM with a Dense layer, achieves the highest accuracy of 0.6377 and the highest F1-score of 0.6528. This indicates that the Dense layer effectively

consolidates LSTM outputs into more stable and discriminative feature spaces. The improved precision suggests that this additional transformation reduces false positives and sharpens the decision boundary, producing a more balanced classifier even before hyperparameter tuning is applied.

The CLD scheme shows the next best performance with accuracy 0.6304 and F1-score 0.6483. This demonstrates that convolutional filters are capable of extracting meaningful local motifs from SMILES sequences. However, the lack of optimized filter sizes and kernel configurations results in residual noise, which reduces the consistency of feature maps passed to the LSTM. Although the architecture is deeper, its potential is not fully realized without tuning.

The L architecture yields accuracy 0.6304 and the highest recall of 0.7606. This indicates that a single LSTM layer already performs well in identifying toxic samples, likely due to its ability to capture long-range dependencies in SMILES strings. Despite this strength, the low precision (0.6136) reveals that the model tends to overpredict toxicity. The absence of additional feature refinement limits its ability to form clearer decision boundaries, causing an increase in false positives.

The CL scheme performs the worst among all architectures, with accuracy 0.6087 and F1-score 0.6197. The convolutional layer produces unstable local features when not properly tuned, and the absence of a Dense layer to reorganize the representations results in inconsistent inputs to the LSTM. These factors reduce classification reliability and show that convolutional preprocessing requires careful calibration to avoid amplifying irrelevant or noisy fragments of the molecular sequence.

Overall, the baseline results lead to several important conclusions. Increasing architectural complexity does not automatically improve performance. A well-balanced architecture such as LD can outperform deeper models when the components are better aligned with the structure of the data. The L architecture demonstrates strong sensitivity but requires additional layers to address false positives. Meanwhile, convolutional modules show potential but need careful adjustment to avoid degrading representational quality.

These findings establish a strong motivation for applying systematic hyperparameter optimization. Each architecture contains strengths that can be enhanced and weaknesses that can be mitigated through targeted adjustments of embedding size, layer depth, convolutional kernel patterns, and representational capacity. Careful tuning is therefore expected to improve not only accuracy but also prediction consistency and robustness in toxicity classification.

3.1.2 Training Curves of Baseline Models

The training curves provide essential insights into the learning dynamics and convergence stability of each architecture prior to hyperparameter tuning. Figures 3(a)–3(d) illustrate the progression of training and validation loss over 20 epochs for the L, CL, LD, and CLD schemes.

The curves in Figures 3(a)–3(d) highlight clear differences in the learning behavior of each baseline architecture. The L model exhibits consistent loss trajectories with steadily increasing accuracy, indicating that the single LSTM architecture is capable of capturing SMILES sequence patterns reasonably well without showing signs of overfitting. The LD model demonstrates a similar pattern but with slightly improved stability. The presence of a Dense layer refines the representations produced by the LSTM, resulting in more consistent validation accuracy.

In contrast, the CL model displays a substantial gap between training and validation loss. This suggests that the non-optimized convolutional filters extract noisy or less relevant local patterns, causing the LSTM to struggle in forming meaningful representations and reducing generalization performance. The CLD model shows the most unstable dynamics, with notable fluctuations in both loss and validation accuracy. The increased complexity arising from combining Convolution, LSTM, and Dense layers makes the training process more sensitive to noise and suboptimal baseline hyperparameter settings, resulting in less stable optimization and greater performance fluctuations during training.

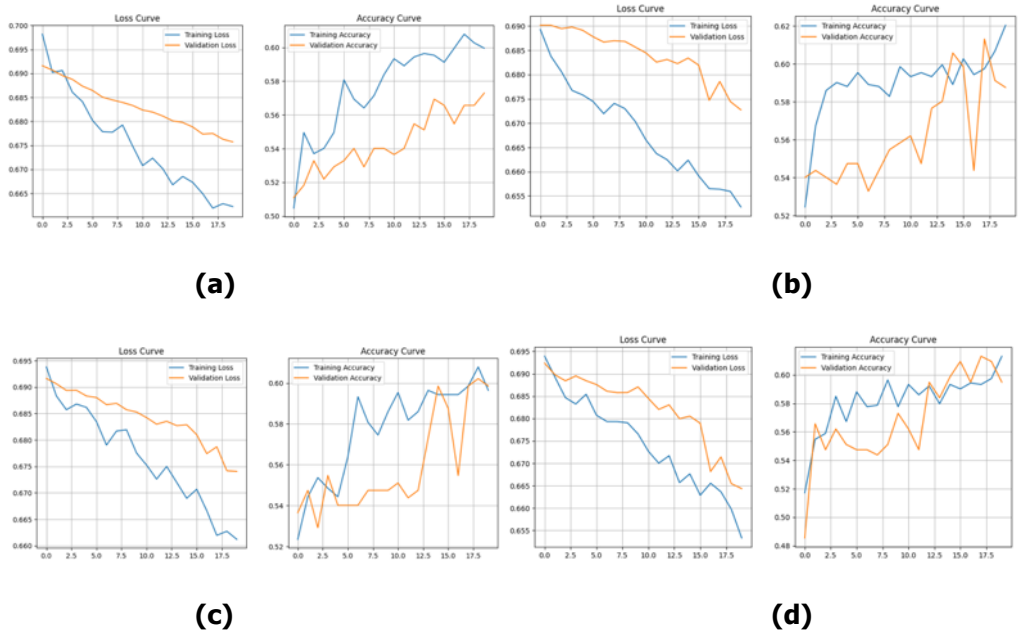

Figure 3. Training and Validation Loss and Accuracy Curves for the Four Baseline Architectures: (a) L, (b) CL, (c) LD, and (d) CLD.

Overall, these trends explain why the L and LD schemes outperform the CL and CLD schemes in Table 4. These findings emphasize that architectures involving convolutional layers require careful hyperparameter tuning to achieve stable training and reliable toxicity prediction.

3.2 Results and Analysis of Tuned Models

After establishing the baseline performance of all architectural schemes, the next step was to evaluate the effect of hyperparameter tuning using the HHO algorithm. This process aimed to identify the optimal configuration for each architecture and assess how systematic optimization influences the predictive performance of the models. The analysis not only highlights the best-performing hyperparameter sets discovered by HHO but also provides insight into how each architectural block contributes to the model's ability to process SMILES data.

3.2.1 Optimal Hyperparameter Configuration

In developing SMILES-based toxicity prediction models, appropriate hyperparameter selection plays a critical role in determining the model's ability to capture sequential patterns and molecular structures. Each LSTM-based architecture, whether standalone or combined with Convolution or Dense layers, requires different capacity levels, and careful optimization can substantially improve predictive performance. HHO was employed to explore the broad

hyperparameter search space and identify the optimal configuration for each architectural scheme.

Table 5 presents the optimal hyperparameter configurations obtained by HHO. Several notable patterns can be observed. First, three of the four architectures (L, LD, and CLD) converged toward large embedding dimensions (126–128), suggesting that rich token representations are consistently important for capturing molecular information encoded in SMILES sequences. Second, architectures containing additional Dense layers (LD and CLD) tended to require deeper recurrent structures, with HHO selecting three LSTM layers for both schemes. This finding indicates that increased architectural complexity benefits from stronger sequential modeling capacity. Third, the CLD architecture was assigned substantially larger convolutional filters and kernel sizes than the CL architecture, implying that more complex hybrid models benefit from capturing longer local molecular patterns before recurrent processing.

The optimal hyperparameters obtained from HHO reveal several consistent patterns across the four architectural schemes. For the L scheme, the optimizer selected a large embedding dimension of 128 and a moderate LSTM size of 41 units. This outcome indicates that rich token representations provide substantial benefit in capturing sequential characteristics of SMILES strings, even without additional layers. The model relies heavily on embedding expressiveness, while extensive recurrent capacity is not as crucial in this configuration.

Table 5. Optimal Hyperparameters Obtained from HHO

Scheme	Embed Dim	LSTM Units	LSTM Layers	Conv Filters	Kernel Size	Dense Units
L	128	41	1	–	–	–
CL	37	37	1	46	3	–
LD	127	128	3	–	–	127
CLD	126	32	3	126	9	85

Overall, the hyperparameter configurations identified by HHO reveal that effective SMILES-based toxicity prediction depends not only on model complexity but also on selecting an appropriate balance between representation learning and sequential modelling capacity. The consistent preference for larger embedding dimensions and deeper recurrent structures in more complex architectures suggests that rich molecular representations and sufficient sequence-learning capability play key roles in improving predictive performance. These findings demonstrate that HHO is able to adapt the hyperparameter configuration to the specific requirements of each architecture, resulting in more effective and robust toxicity prediction models.

3.2.2 Model Performance After Tuning

After determining the optimal hyperparameter configuration through HHO, the next step was to evaluate the impact of tuning on toxicity prediction performance for each architectural scheme. The tuning results summarized in Table 6 show substantial and meaningful improvements across most evaluation metrics when compared to the baseline models.

The L scheme shows the most consistent improvements across nearly all metrics. Accuracy increased from 0.6304 to 0.6739 (+0.0435) and F1-score from 0.6792 to 0.7097 (+0.0305). Recall rose from 0.7606 to 0.7746 (+0.0140), while precision increased from 0.6136 to 0.6548 (+0.0412). The relatively larger gain in precision compared to recall indicates that the

optimization of the embedding layer and LSTM units effectively reduced false positives without compromising the model's ability to detect toxic samples. This confirms that the single-layer LSTM, which is already capable of capturing sequential dependencies in SMILES, became more balanced in terms of sensitivity and precision, resulting in more stable and reliable predictions.

The LD scheme, which integrates an LSTM with a Dense layer, exhibits balanced improvements across all metrics. Accuracy increased from 0.6377 to 0.6667 (+0.0290), F1-score from 0.6528 to 0.7013 (+0.0485), recall from 0.6620 to 0.7606 (+0.0986), and precision from 0.6438 to 0.6506 (+0.0068). The notable increase in recall highlights that tuning the Dense layer strengthened the high-level feature transformations derived from the LSTM, resulting in clearer decision boundaries and improved discrimination between toxic and non-toxic samples. This shows that the Dense layer is not merely an "additional layer," but plays a critical role in boosting the discriminatory power of the model after tuning, particularly for complex and variable SMILES data.

Table 6. Performance of HHO-Optimized Architectures

Scheme	Accuracy	Recall	Precision	F1-score
L	0.6739 (+0.0435)	0.7746 (+0.0140)	0.6548 (+0.0412)	0.7097 (+0.0305)
CL	0.6522 (+0.0435)	0.7465 (+0.1268)	0.6386 (+0.0189)	0.6883 (+0.0686)
LD	0.6667 (+0.0290)	0.7606 (+0.0986)	0.6506 (+0.0068)	0.7013 (+0.0485)
CLD	0.6304 (+0.0000)	0.5070 (–0.1550)	0.6923 (+0.0572)	0.5854 (–0.0629)

The CL scheme demonstrates the most pronounced improvement in recall, increasing from 0.6197 to 0.7465 (+0.1268), alongside a +0.0686 gain in F1-score. This indicates that the optimized filters and convolutional kernels were able to extract more relevant local SMILES features. The baseline model tended to capture noisy or irrelevant patterns, reducing generalization. Tuning allowed the model to highlight molecular substructures truly associated with reproductive toxicity. However, although recall increased substantially, precision improved only modestly, suggesting that the model continues to produce some false positives. This highlights the inherent challenge in optimizing local feature extraction: increasing sensitivity may introduce trade-offs in specificity.

The CLD scheme presents a more complex behaviour. Precision improved to 0.6923 (+0.0572), but recall dropped sharply from 0.6620 to 0.5070 (–0.1550), resulting in a decrease in F1-score to 0.5854 (–0.0629). The decline in recall suggests that simultaneous optimization of the Convolution, LSTM, and Dense blocks requires a larger search space or a multi-stage strategy. The higher architectural complexity made the model more conservative: false positives decreased, but many toxic samples were missed. This indicates that a more complex architecture is not inherently superior; without appropriate optimization, complexity may hinder detection capability.

Overall, Table 6 demonstrates that HHO consistently improves the performance of the L, LD, and CL schemes over their respective baselines, producing models that are more stable, sensitive, and precise. The L and LD schemes achieve the most consistent gains due to their architectural balance relative to optimization capacity. These findings emphasize that successful SMILES-based toxicity prediction depends not only on architectural complexity but also on an appropriate balance between model capacity, layer structure, and hyperparameter optimization. HHO effectively enhances the model's ability to extract critical features from SMILES sequences, yielding more reliable predictions and reinforcing the importance of metaheuristic approaches in improving computational toxicology models.

4. CONCLUSION

This study demonstrates that Harris Hawks Optimization (HHO) effectively improves the performance of LSTM-based models for SMILES-based reproductive toxicity prediction. The best results were achieved by the L and LD architectures, indicating that strong sequential modeling is more important than increasing architectural complexity through additional convolutional components. A notable finding is that HHO consistently selected large embedding dimensions (126–128) across most architectures, highlighting the importance of rich molecular representations. Furthermore, the LD architecture benefited from deeper recurrent modeling with three LSTM layers and 128 hidden units, whereas the more complex CLD architecture did not achieve superior performance despite employing larger convolutional filters and kernel sizes. These results suggest that predictive performance is primarily influenced by the quality of molecular representations and recurrent learning capacity rather than by architectural complexity alone. Future work may explore alternative optimization strategies and advanced architectures to further improve model generalization and predictive capability.

REFERENCES

Afinda, A. M., Karimah, A. F., & Kurniawan, I. (2023). Gated Recurrent Unit with SMILES2Vec-based Descriptor for Predicting Drug Side Effects: Case Study of Hepatobiliary Disorders. *2023 International Conference on Data Science and Its Applications (ICoDSA)*, (pp. 426–431). https://doi.org/10.1109/ICoDSA58501.2023.10276594

Fan, B., Fan, W., Smith, C., & Garner, H. "Skip." (2020). Adverse drug event detection and extraction from open data: A deep learning approach. *Information Processing & Management*, *57*(1), 102131. https://doi.org/10.1016/j.ipm.2019.102131

Feng, J., Wu, Q., Liang, Y., Liang, Y., & Bin, Q. (2025). Epidemiological characteristics of infertility, 1990–2021, and 15-year forecasts: An analysis based on the global burden of disease study 2021. *Reproductive Health*, *22*(1), 26. https://doi.org/10.1186/s12978-025-01966-7

Fuadah, Y. N., Pramudito, M. A., Firdaus, L., Vanheusden, F. J., & Lim, K. M. (2024). QSAR Classification Modeling Using Machine Learning with a Consensus-Based Approach for Multivariate Chemical Hazard End Points. *ACS Omega*, *9*(51), 50796–50808. https://doi.org/10.1021/acsomega.4c09356

Goh, G. B., Hodas, N. O., Siegel, C., & Vishnu, A. (2018). *SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties* (arXiv:1712.02034). arXiv. https://doi.org/10.48550/arXiv.1712.02034

Goh, G. B., Siegel, C., Vishnu, A., Hodas, N. O., & Baker, N. (2017). Chemception: A Deep Neural Network with Minimal Chemistry Knowledge Matches the Performance of Expert-developed QSAR/QSPR Models. *ArXiv*.

Igani, K. D. A., & Kurniawan, I. (2025). Implementation of LSTM Optimized by Simulated Annealing for Toxicity Prediction: Case Study AR-LBD Toxicity. *2025 International Conference on Data Science and Its Applications (ICoDSA)*, (pp. 231–236). https://doi.org/10.1109/ICoDSA67155.2025.11157576

Infertility. (n.d.). Retrieved December 6, 2025, from https://www.who.int/news-room/fact-sheets/detail/infertility

Karimah, A. F., Afinda, A. M., & Kurniawan, I. (2023). Implementation of SMILES2Vec-based LSTM for Predicting Drug Side Effects: Case Study of Hepatobiliary Disorder. *2023 International Conference on Data Science and Its Applications (ICoDSA)*, (pp. 333–338). https://doi.org/10.1109/ICoDSA58501.2023.10277245

Kaur, N., Kaur, L., & Cheema, S. S. (2021). An enhanced version of Harris Hawks Optimization by dimension learning-based hunting for Breast Cancer Detection. *Scientific Reports*, *11*(1), 21933. https://doi.org/10.1038/s41598-021-01018-7

Lee, C. Y., & Chen, Y.-P. P. (2021). Descriptive prediction of drug side-effects using a hybrid deep learning model. *International Journal of Intelligent Systems*, *36*(6), 2491–2510. https://doi.org/10.1002/int.22389

M. Veselinovic, A., B. Veselinovic, J., V. Zivkovic, J., & M. Nikolic, G. (2015). Application of SMILES Notation Based Optimal Descriptors in Drug Discovery and Design. *Current Topics in Medicinal Chemistry*, *15*(18), 1768–1779.

Miller, L. G., Demny, M., Tamamis, P., & Contreras, L. M. (2023). Characterization of epitranscriptome reader proteins experimentally and *in silico*: Current knowledge and future perspectives beyond the YTH domain. *Computational and Structural Biotechnology Journal*, *21*, 3541–3556. https://doi.org/10.1016/j.csbj.2023.06.018

Mnassri, A., Mansouri, N., Nasri, S., Lashab, A., Vasquez, J. C., & Cherif, A. (2025). An Intelligent Framework for Energy Forecasting and Management in Photovoltaic-Integrated Smart Homes in Tunisia with V2H Support Using LSTM Optimized by the Harris Hawks Algorithm. *Energies*, *18*(21), 5635. https://doi.org/10.3390/en18215635

Nik Hazlina, N. H., Norhayati, M. N., Shaiful Bahari, I., & Nik Muhammad Arif, N. A. (2022). Worldwide prevalence, risk factors and psychological impact of infertility among women: A systematic review and meta-analysis. *BMJ Open*, *12*(3), e057132. https://doi.org/10.1136/bmjopen-2021-057132

SIDER Side Effect Resource. (n.d.). Retrieved June 2, 2022, from http://sideeffects.embl.de/

Sofian, H. A., & Kurniawan, I. (2025). Implementation of LSTM Tuned by Camel Algorithm for Predicting Toxicity of Ionic Liquids toward Acetylcholinesterase Enzymes. *2025*

International Conference on Data Science and Its Applications (ICoDSA), (pp. 26–31). https://doi.org/10.1109/ICoDSA67155.2025.11157538

Weininger, D. (2002, May 1). *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules* [Research-article]. American Chemical Society. (world). ACS Publications. https://doi.org/10.1021/ci00057a005

Wu, M., Blancaflor, E., Ren, F., Wang, Y., & Dong, T. (2024). Enhanced Water Quality Prediction in the Yellow River Basin: The Application of the HHO-LSTM Model. *International Journal of Online and Biomedical Engineering (iJOE)*, *20*, 4–14. https://doi.org/10.3991/ijoe.v20i05.48225

Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, *32*(7), 1466–1474. https://doi.org/10.1002/jcc.21707

Zhou, Y., Ning, C., Tan, Y., Li, Y., Wang, J., Shu, Y., Liang, S., Liu, Z., & Wang, Y. (2024). ToxMPNN: A deep learning model for small molecule toxicity prediction. *Journal of Applied Toxicology*, *44*(7), 953–964. https://doi.org/10.1002/jat.4591