How to Cite:
Chandran, P. P., Hema, R. N., & Jeyakarthic, M. (2022). Invasive weed optimization with stacked long short term memory for PDF malware detection and classification. *International Journal of Health Sciences*, *6*(S5), 4187–4204. https://doi.org/10.53730/ijhs.v6nS5.9540

Invasive weed optimization with stacked long short term memory for PDF malware detection and classification

P. Pandi Chandran
Department of Computer and Information Science, Faculty of Science, Annamalai University, Chidambaram, 608002, India
Corresponding author email: pandi.chandran@gmail.com

Hema Rajini. N
Department of CSE, Alagappa Chettiar Government College of Engineering and Technology, Karaikudi, 630003, India

M. Jeyakarthic
Department of Computer and Information Science, Annamalai University, Chidambaram, 608002, India

***Abstract*---**Due to high versatility and widespread adoption, PDF documents are widely exploited for launching attacks by cyber criminals. PDFs have been conventionally utilized as an effective method for spreading malware. Automated detection and classification of PDF malware are essential to accomplish security. Latest developments of artificial intelligence (AI) and deep learning (DL) models pave a way for automated detection of PDF malware. In this view, this article develops an Invasive Weed Optimization with Stacked Long Short Term Memory (IWO-S-LSTM) technique for PDF malware detection and classification. The presented IWO-S-LSTM model focuses on the recognition and classification of different kinds of malware that exist in PDF documents. The proposed IWO-S-LSTM model initially undergoes pre-processing in two stages namely categorical encoding and null value removal. Besides, autoencoder (AE) based outlier detection approach is presented to remove the existence of outliers. In addition, S-LSTM model is utilized to detect and classify PDF malware. Finally, IWO algorithm is applied to fine tune the hyperparameters involved in the S-LSTM model. To determine the enhanced outcomes of the IWO-S-LSTM model, a series of simulations were executed on two benchmark datasets. The experimental outcomes outperformed the promising performance of the IWO-S-LSTM technique on the other approaches.

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
Manuscript submitted: 27 Feb 2022, Manuscript revised: 9 April 2022, Accepted for publication: 18 June 2022

Keywords---PDFs, Malware detection, Outlier removal, Machine learning, Deep learning, Invasive Weed optimization.

1. Introduction

PDF is one of the popular and trusted extensions, whereas Adobe Reader is the more commonly used program to open these kinds of files [1]. The factor encourages attackers to research and search for vulnerabilities and new ways of making exploits that will execute arbitrary code while opened with this specific software. Attackers frequently embed malware in digital documents [2], including Microsoft Word and Adobe PDF. PDF document becomes a more commonly known file format exploits in targeted attacks and new vulnerability continues to be exploited by targeted attackers. Because system compromise is undesirable [3, 4], tools and methodologies to determine either malicious or benign, the disposition of a document, are required. The number of researches has proved that malicious document is widely employed in socially engineered phishing attack executed by group of targeted, sophisticated, and persistent attackers whose aim is espionage [5].

The usage of trojan PDF in targeted attacks includes those executed by a group of attackers named the Advanced Persistent Threat (APT), adding urgency to counter these kinds of malware delivery [6]. A proper method of delivery and exploitation used differs significantly: in certain cases, the document is merely utilized for exploiting vulnerabilities in the reader applications which provide small values interms of social engineering. For example, certain classes of web-based attacks, namely those leverage cross-site scripting, proceed accordingly [7]. There are several techniques for malicious document detection. Signature matching is extensively used and is efficient for identifying formerly recognized malware on a largescale. Signature matching is a dynamic analysis of document wherein the performance of whole system or the set of programmes is observed when the document is opened. In fact, matching system includes commodity antivirus scanner [8] provides specific functionality to PDF documents such that they could be decrypted to exploit document specific vulnerabilities and reveal malicious content.

Various technique for categorizing document has been pursued. One method is to check for anomalies in static features extracting from the document [9]. The abovementioned methods for detecting malevolent documents need that training or seeding with characterization of formerly encountered benign or malicious documents. But still, alternate method work by monitoring the runtime performance of a document observer for unpredicted action as it renders the document [10]. For example, popular antivirus system relies on curated database of byte-signature for identifying malicious documents and ML approach relies on model that is trained by the features (dynamic execution artifact, weighted byte n-grams, and so on.) extracted from a corpus that contains benign or malicious document.

This article develops an Invasive Weed Optimization with Stacked Long Short Term Memory (IWO-S-LSTM) model for PDF malware detection and classification.

The proposed IWO-S-LSTM model initially undergoes pre-processing in two stages namely categorical encoding and null value removal. Moreover, autoencoder (AE) based outlier detection approach is presented to remove the existence of outliers. Furthermore, IWO with S-LSTM model is utilized to detect and classify PDF malwares. In order to demonstrate the enhanced outcomes of the IWO-S-LSTM model, a series of simulations were performed using two benchmark datasets and distinct measures.

2. Related works

Yoo et al. [11] presented an ML based hybrid decision method which attains a maximum detection rate with minimum false positive rate. This hybrid method integrates an arbitrary forest and DL technique utilizing 12 hidden layers for determining malware and benign files correspondingly. Li et al. [12] projected an active-learning based malware detection method, utilizing mutual agreement analysis for choosing the uncertain instance as data augmentation. The detector was retrain based on the ground truth of uncertain instances before the entire test instances from the preceding epoch that is not only enhancing the detection performance, then also decreasing the trained time utilization of detectors. Jeong et al. [13] aimed to develop a CNN for tackling malware detection on PDF files. It intensively inspects the infrastructure of input data and showcases that it is proposal the presented network depends upon the features of data.

Zhang [14] examined a new method dependent upon multilayer perceptron (MLP) NN method, named MLPdf, for the recognition of PDF based malware. More particularly, the MLPdf technique utilizes a BP technique with stochastic gradient descent (SGD) search for technique updating. Ye et al. [15] presented a heterogeneous DL infrastructure collected of Auto-Encoder (AE) stacked up with multi-layer RBM and layer of associative memory for detecting recently unknown malware. The presented DL technique executes as greedy layer-wise trained function to unsupervised feature learning, then supervised parameter fine-tuned. Liu et al. [16] presented a new approach integrating Malware Visualization and Image Classification for detecting PDF files and recognizing that ones can be malicious. Moreover, dependent upon the signature of objects from the files, it is allocated various colors attained in SimHash for generating RGB images.

3. The Proposed Model

In this article, a new IWO-S-LSTM model has been developed for the recognition and classification of different kinds of malware that exist in the PDF documents. The proposed IWO-S-LSTM model is pre-processed in two stages namely categorical encoding and null value removal. Followed by, AE based outlier detection with S-LSTM based classification model is developed. At last, the IWO algorithm is used for fine tuning the hyperparameters involved in the S-LSTM model. Fig. 1 illustrates the overall process of IWO-S-LSTM technique.

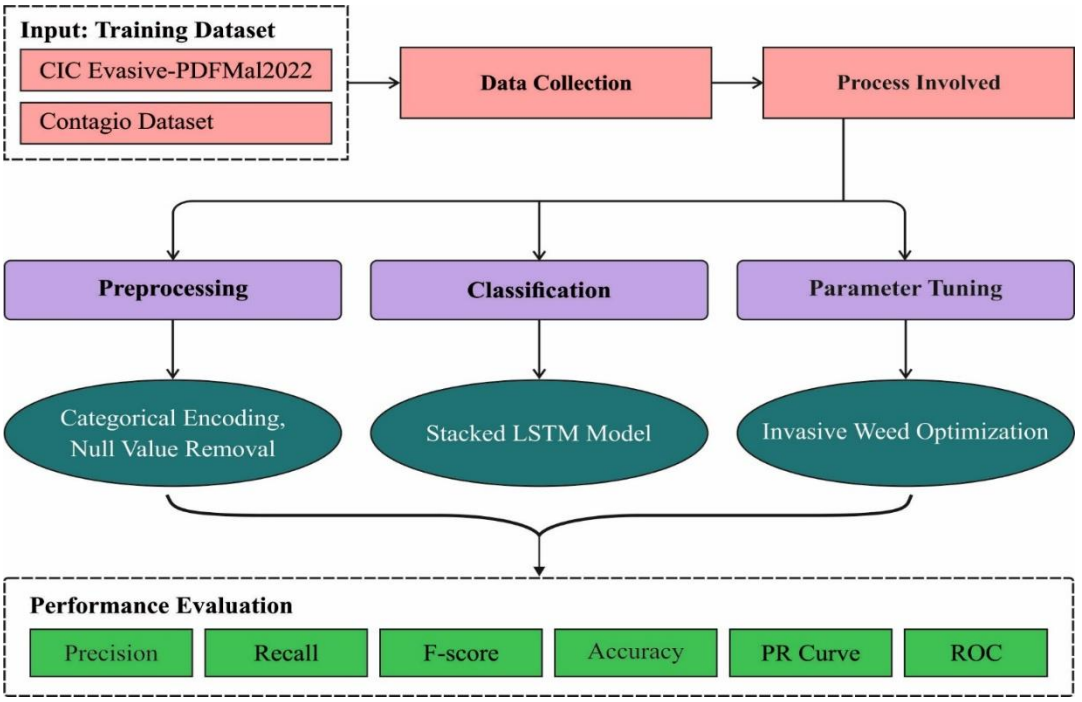

Fig. 1. Overall process of IWO-S-LSTM technique

3.1. Pre-processing

In a primary phase, the input data has been pre-processed in 2 phases of operations namely categorical encoded and null value elimination. Initially, the categorical values are encoded as numerical values. Secondly, the null value which exists from the dataset is removed.

3.2. AE based Outlier Detection Process

For the removal of outliers, the AE model is used in this study. An AE is an artificial neural network that is trained to learn an appropriate representation (coding) of the input information to assure desired property [17]. It enables us to compare input and output, says x and $\hat{x}$, and computes a loss function according to the distance, for training the net to regenerate the input vector. Fig. 2 showcase the structure of AE. Accurately, the encoding maps the input vector, $x \in \mathbb{R}^K$, to hidden representation, $h \in \mathbb{R}^H$, as follows

$$h = \phi_1(W_1 x + b_1) \tag{1}$$

whereas ϕ_1 $(\cdot)$ represent the activation function, W_1 denotes a $K \times H$ weighted matrix and b_1 indicates a bias vector. The decoder, sequentially, map the hidden representation to the output reconstruction as follows

$$\hat{X} = \phi_2(W_2 h + b_2) \tag{2}$$

In which $\phi_2, (\cdot)$ W_2 and b_2 have apparent meaning.

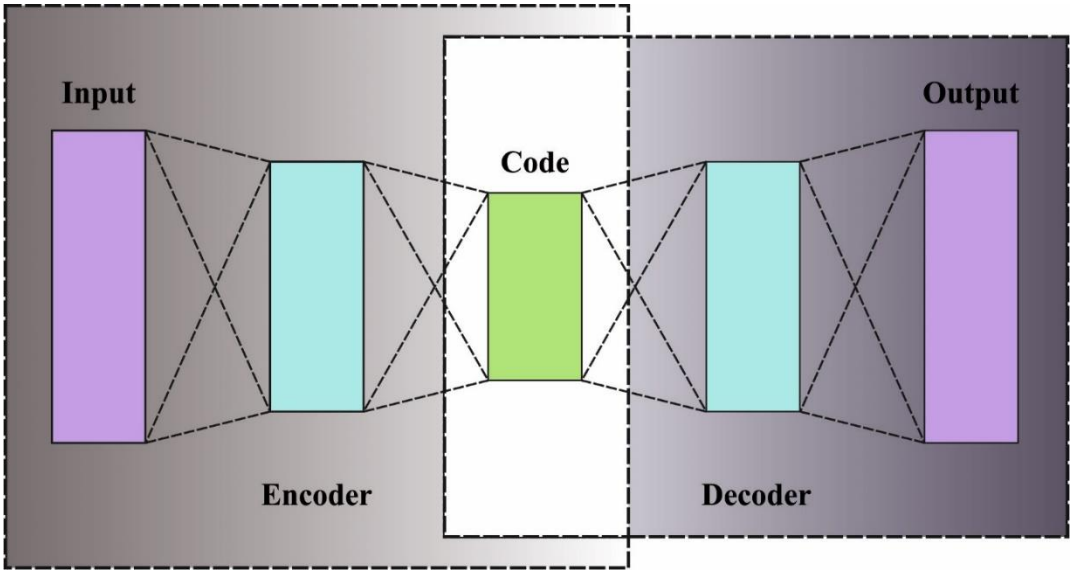

Fig. 2. Structure of AutoEncoder

Hence, All the training input $x^{(i)}$ is mapped to last representation $h^{(i)}$, and regenerated as $\hat{x}^{(i)}$. The parameter of the AE $\theta = \{W_1, b_1, W_2, b_2\}$ are learned by minimalizing the average recreation error amongst input and output typically evaluated by the average Euclidean distance

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} | \left|x^{(i)} - \hat{x}^{(i)}\right||^2 \tag{3}$$

3.3. S-LSTM based Classification

Once the outliers are removed, the next stage is to classify the PDF documents using S-LSTM model. The major concept behindhand LSTM lies in a gate that controls the data flow alongside time axis could capture long-term dependency at every time step. Especially, at every time step t, hidden layer h^t is upgraded by data fusion at the input gate i^t, similar step x^t, output gate 0^t, forget gate f^t, hidden layer at final time step h^{t-1} and memory cell c^t:

$$\begin{aligned} i^t &= \sigma(W^i x^t + V^i h^{t-1} + b^i)\ , \\ f^t &= \sigma(W^f x^t + V^f h^{t-1} + b^f)\ , \\ o^t &= \sigma(W^o x^t + V^o h^{t-1} + b^o), \\ c^t &= f^t \odot c^{t-1} + i^t \odot \tanh\,(W^c x^t + V^c h^{t-1} + b^c)\ , \\ h^t &= 0^t \odot \tanh\,(c^t)\ , \end{aligned} \tag{4}$$

In which, model parameter includes $W \in \mathbb{R}^{d\times k}$, $V \in \mathbb{R}^{d\times d}$, and $b \in \mathbb{R}^d$ are learned at the time of training and share at every time step, σ denotes sigmoid activation, $\odot$ indicates elementwise product, and k represent a hyperparameter that symbolizes the dimension of hidden states [18]. At first, LSTM is employed to handle time-series data. Also, the last output is at final time step, is employed for predicting the output through a linear regression layer, as follows:

$$\overline{y}_i = W^r h_i^T, \tag{5}$$

Whereas $W^r \in \mathbb{R}^{k \times z}$ and z denotes the dimensionality of output. Here, the cross-entropy is utilized as loss function amongst the target label distribution $p(x)$ and the forecasted label distribution $q(x)$. Hence it is expressed as follows

$$loss = H(p,q) = -\sum_x p(x) \log q(x). \tag{6}$$

The activation function allows the network to obtain a non-linear depiction of the input signal. The ReLU is expressed as follows:

$$a_i^{l+1}(j) = f\left(y_i^{l+1}(j)\right) = \max\left\{0, y_i^{l+1}(j)\right\}, \tag{7}$$

Here, $y_i^{l+1}(j)$ and $a_i^{l+1}(j)$ denotes the output of LSTM and activation value of $y_i^{l+1}(j)$, correspondingly. With the tremendous growth of computer hardware and a sequence of DL approaches that have been proposed, deep architecture has shown the effective ability to feature self-learning. Thus, stacking LSTM layer for a deep LSTM-based NN is useful. The major concept of DNN is that non-linear mapping layers among outputs and inputs are employed for learning features. The output of hidden state is propagated forward by time and utilized as input of following LSTM hidden state. This can be formulated by the following equation

$$\begin{aligned}
i_l^t &= \sigma(W_l^i h_{l-1}^t + V_l^i h_l^{t-1} + b_l^f), \\
f_l^t &= \sigma(W_l^f h_{l-1}^t + V_l^f h_l^{t-1} + b_l^f), \\
o_l^t &= \sigma(W_l^o h_{l-1}^t + V_l^o h_l^{t-1} + b_l^o), \\
c_l^t &= f_l^t \odot c_l^{t-1} + i_l^t \odot \tanh(W_l^c h_{l-1}^t + V_l^c h_l^{t-1} + b_l^c), \\
h_l^t &= 0_l^t \odot \tanh(c_l^t).
\end{aligned} \tag{8}$$

The input of initial layer is raw temporal signal that is $h_0^t = x^t$, whereas the output of inital layer is an abstraction of raw signal, i.e., considered as a hierarchical feature. The advantage of stacked LSTM is apparent: (1) Model parameter is distributed without improving storage capacity that allows to quicken convergence and refine non-linear operation of raw information. (2) stacking LSTM layer allows learning features of raw temporal signal from distinct aspects at every time step. Fig. 3 depicts the framework of Stacked LSTM Model.

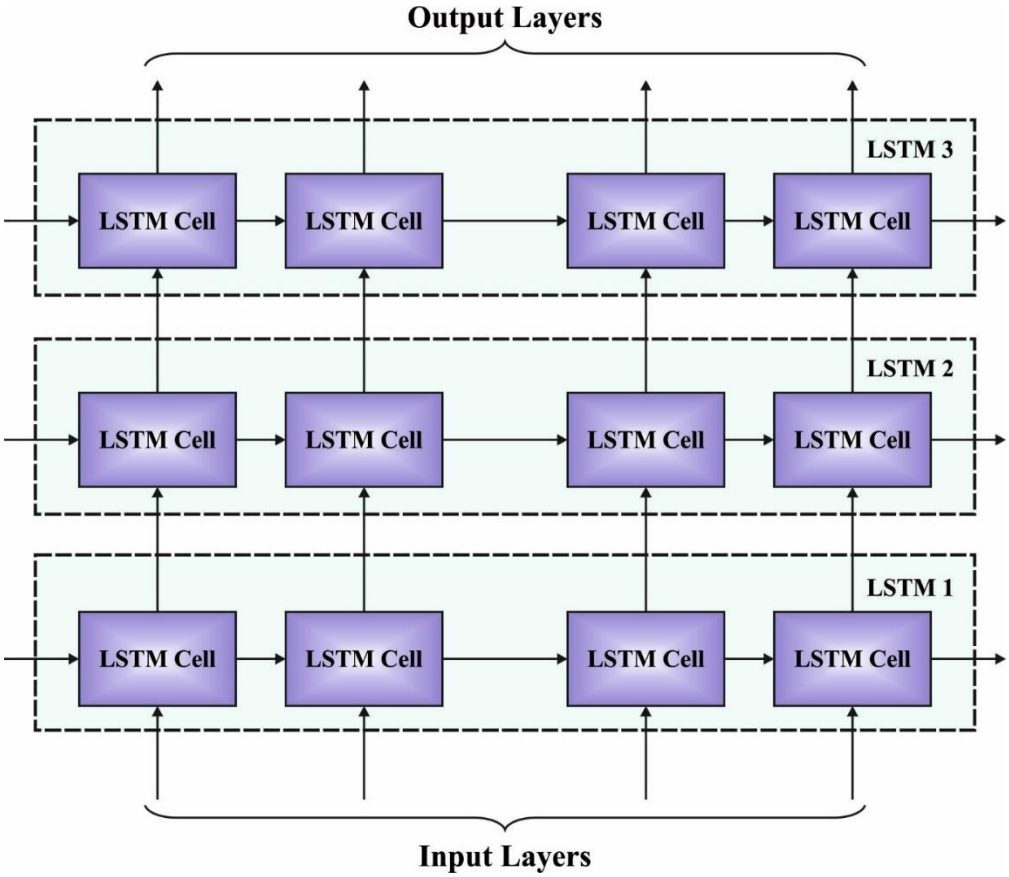

Fig. 3. Framework of Stacked LSTM Model

3.4. IWO based Hyperparameter Optimization

In order to optimally tuning the hyperparameters involved in the S-LSTM technique, the IWO algorithm is used. IWO is an arithmetical stochastic search technique that stimulates natural performance of weed colonizing to optimize the function. But it is proved that efficient in converging to optimum solution by using fundamental properties—for example, growth, seeding, and competition—in a weed colony [19]. The basic properties of the algorithm are given below

1. A fixed count of seeds are spread out through the searching region.
2. Each seed grows to a flowering plant and generates seed depending on its fitness.
3. The generated seed is dispersed randomly through the search region and grows to new plant.
4. This procedure repeats until the maximal amount of plants is obtained. Next, the plant with higher fitness could survive and generate seeds, and others are removed. The procedure repeats until the maximal amount of iterations is obtained and, the plant with the optimal fitness is close to the optimum solution. It can be discussed detailed in the following.

Population Initialization: A population initialization of solution is spread out through the n-dimension problem space with arbitrary position. Reproduction: A specific population of plants is permitted to generate seeds based on its own and the colony's highest and lowest fitness: The plant generates linearly increase from the minimal seed production level to its maximal level. Spatial Dispersal: Adaptation and Randomness is shown in the following. The produced seeds are distributed arbitrarily through the n-dimension search region by arbitrary number with a mean value equivalent to zero but differ with variance. This guarantees that the seeds are accumulated near the parent plant and distributed arbitrarily. But standard deviation (SD) σ of the random function reduce from a formerly determined initial value, $\sigma_{initia1}$, to the last value, σ_{final}, in all the steps. In simulation, non-linear variations show an outstanding performance, as follows

$$\sigma_{iter} = \frac{(iter_{\max} - iter)^n}{iter_{\max}^n}(\sigma_{initia1} - \sigma_{fina1}) + \sigma final \qquad (9)$$

Whereas $iter_{\max}$ represents the maximal amount of iterations, σ_{iter} indicates the SD at the existing step, and n signifies the non-linear modulation index. The experiment study suggests that the outcomes from IWO are superior to the outcomes from other models. The IWO performance is compared to other evolutionary mechanisms, and its outcomes are reasonable for each test function [19]. The IWO method develops a FF for gaining increased PDF malware classification performance of the S-LSTM technique. It resolves a positive integer to represent the optimal performance of the candidate solutions. Under this analysis, the minimization of classifier error rate has assumed as FF is provided in Eq. (10). The best result is a lesser error rate and least solution reaches a maximal error rate.

$$fitness(x_i) = ClassifierErrorRate(x_i) = \frac{number\ of\ misclassified\ PDF\ documents}{Total\ number\ of\ PDF\ documents} * 100 \qquad (10)$$

4. Experimental Validation

In this section, the performance validation of the IWO-S-LSTM model is carried out using two datasets namely CIC Evasive-PDFMal2022 [20] and Contagio [21] dataset. The first dataset holds 10023 instances with 31 attributes and two classes. After outlier removal process, the number of instances becomes 5007. The second Contagio dataset includes 9999 instances with 136 attributes and two classes. Once the outliers are removed, the number of instances become 4382

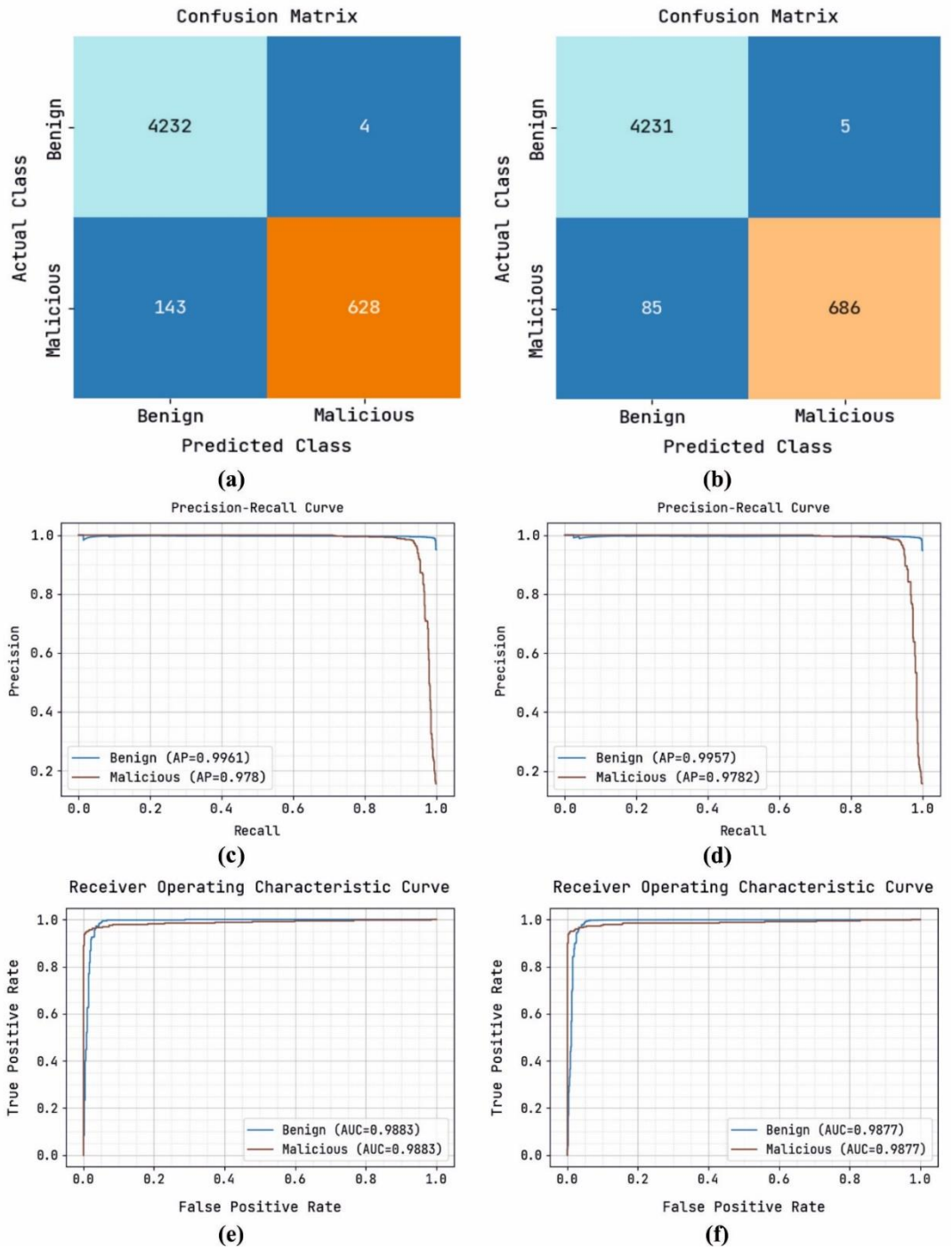

Fig. 4. CIC Evasive-PDFMal2022 Dataset a) CM of S-LSTM b) CM of IWO-S-LSTM c) PRC of S-LSTM d) PRC IWO-S-LSTM e) ROC of S-LSTM f) ROC of IWO-S-LSTM

Fig. 4 illustrates a set of results obtained by the IWO-S-LSTM model on the classification of PDF malware in CIC Evasive-PDFMal2022 dataset. Fig. 4a indicates that the S-LSTM model has identified 4232 instances into benign class and 628 instances into malignant. Next, Fig. 4b represents that the IWO-S-LSTM model has categorized 4231 instances into benign class and 686 instances into malignant class. Then, Figs. 4c-4d shows the precision-recall curves of the S-LSTM and IWO-S-LSTM models. The figures reported that the IWO-S-LSTM model has obtained enhanced performance over the LSTM model. Similarly, Figs. 4e-4f represents the ROC examination of the S-LSTM and IWO-S-LSTM models. The results indicated that the IWO-S-LSTM model has attained maximum ROC values.

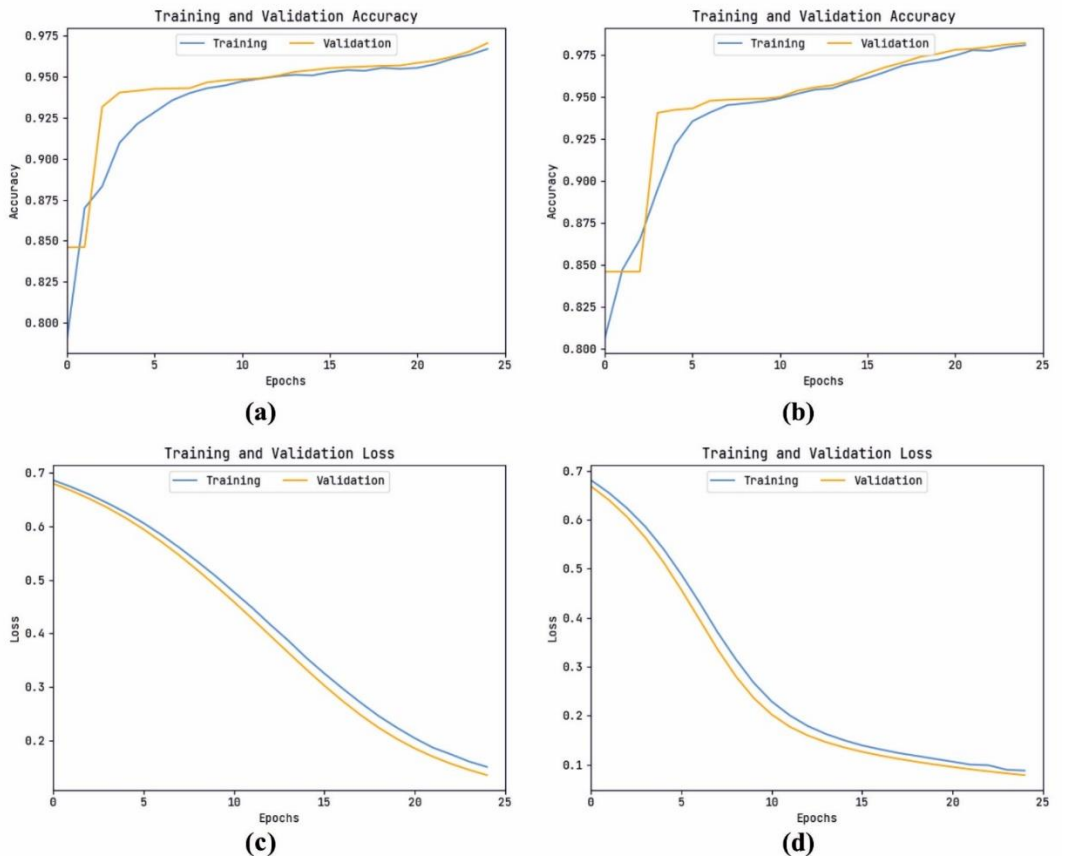

Fig. 5. CIC Evasive-PDFMal2022 Dataset a) Accuracy of S-LSTM b) Loss of S-LSTM c) Accuracy of IWO-S-LSTM d) Loss of IWO-S-LSTM

Fig. 5 illustrates a set of accuracy and loss analyses of the IWO-S-LSTM model on the classification of PDF malware in CIC Evasive-PDFMal2022 dataset. Figs. 5a-5b shows the accuracy and loss analysis of the S-LSTM models. The results show that the accuracy value tends to increase and loss value tends to decrease with an increase in epoch count. Similarly, Figs. 5c-5d shows the accuracy and loss analysis of the IWO-S-LSTM models. The figures show that the accuracy value tends to increase and loss value tends to decrease with an increase in epoch count.

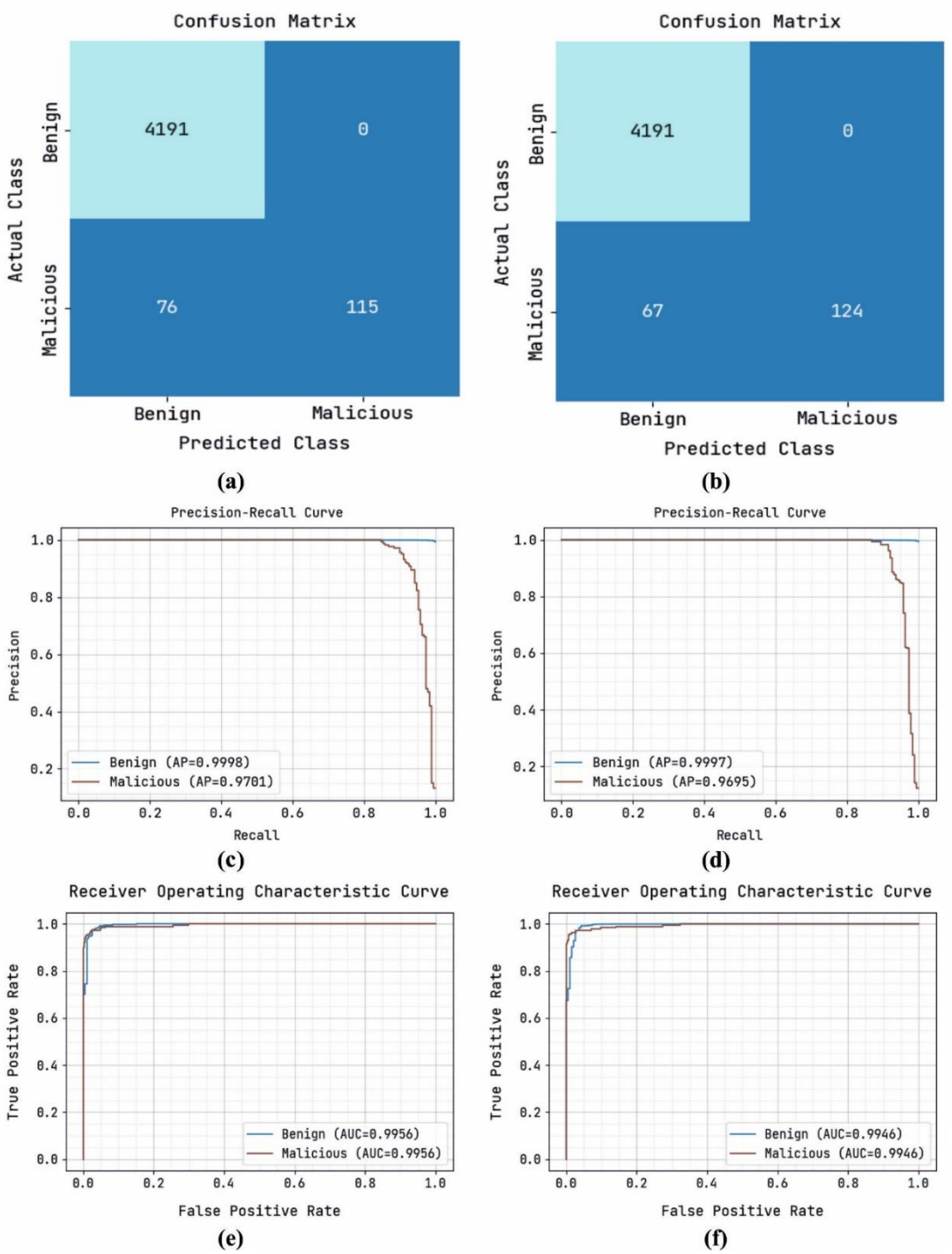

Fig. 6. Contagio Dataset a) CM of S-LSTM b) CM of IWO-S-LSTM c) PRC of S-LSTM d) PRC IWO-S-LSTM e) ROC of S-LSTM f) ROC of IWO-S-LSTM

Fig. 6 demonstrates a set of results attained by the IWO-S-LSTM technique on the classification of PDF malware in Contagio dataset. Fig. 6a shows that the S-LSTM approach has recognized 4191 instances into benign class and 115 samples into malignant. Then, Fig. 6b characterizes that the IWO-S-LSTM approach has categorized 4191 instances into benign class and 124 instances into malignant class. Next, Figs. 6c-6d displays the precision-recall curves of the S-LSTM and IWO-S-LSTM techniques. The figure reports that the IWO-S-LSTM technique has attained improved performance over the LSTM. Similarly, Figs. 6e-6f signifies the ROC examination of the S-LSTM and IWO-S-LSTM approaches. The result indicates that the IWO-S-LSTM technique has achieved maximal ROC values.

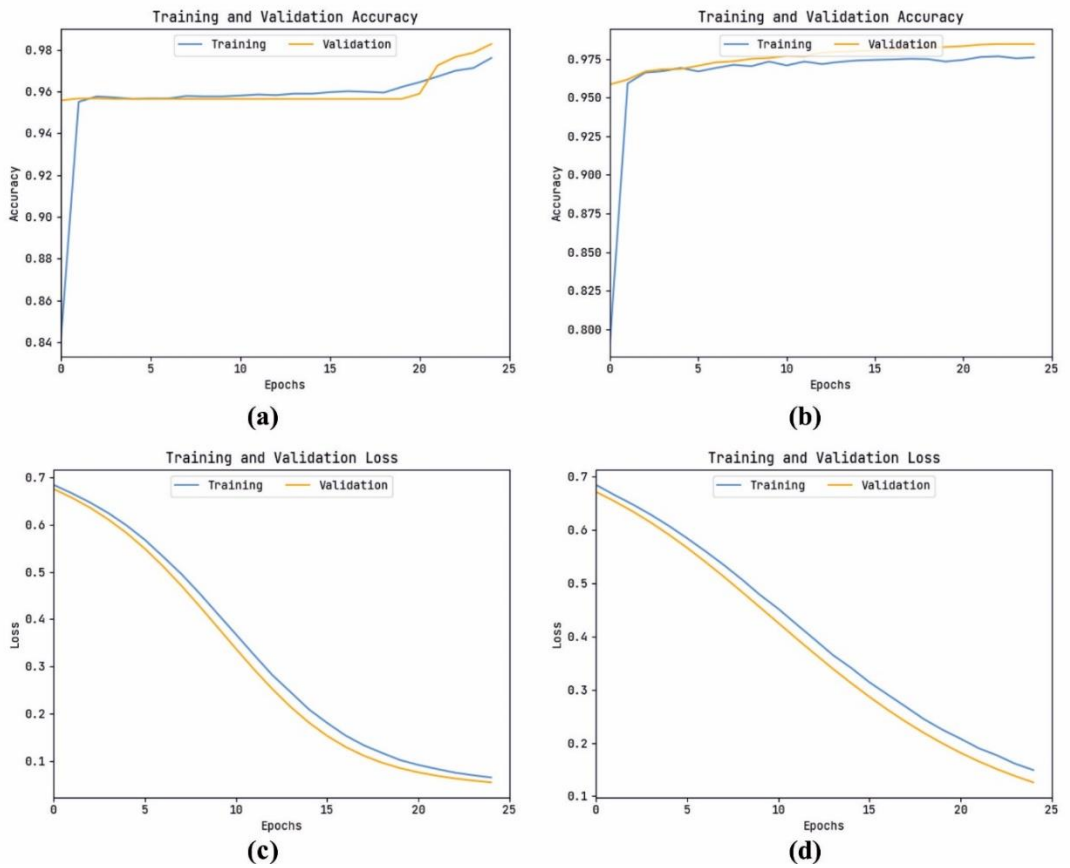

Fig. 7. Contagio Dataset a) Accuracy of S-LSTM b) Loss of S-LSTM c) Accuracy of IWO-S-LSTM d) Loss of IWO-S-LSTM

Fig. 7 shows a set of accuracy and loss analyses of the IWO-S-LSTM approach on the classification of PDF malwares in Contagio dataset. Figs. 7a-7b illustrates the accuracy and loss analysis of the S-LSTM models. The result shows that the accuracy value tends to increase and loss value tends to decrease with an increase in epoch count. Likewise, Figs. 7c-7d displays the accuracy and loss analysis of the IWO-S-LSTM models. The figure shows that the accuracy value tends to increase and loss value tends to decrease with an increase in epoch count.

Table 1 provides detailed PDF malware classification results of the S-LSTM and IWO-S-LSTM model on CIC Evasive-PDFMal2022 and Contagio dataset. The experimental values indicated that the IWO-S-LSTM model has obtained improved performance over the S-LSTM model.

Table 1 Result analysis of S-LSTM and IWO-S-LSTM technique on two datasets

Measures	CIC Evasive-PDFMal2022		Contagio Dataset	
	S-LSTM	IWO-S-LSTM	S-LSTM	IWO-S-LSTM
Accuracy	97.06	98.20	98.27	98.47
Precision	98.05	98.65	99.11	99.21
Recall	90.68	94.43	80.10	82.46
F1-Score	93.91	96.40	87.13	88.97
AUC-Score	98.83	98.77	99.56	99.46

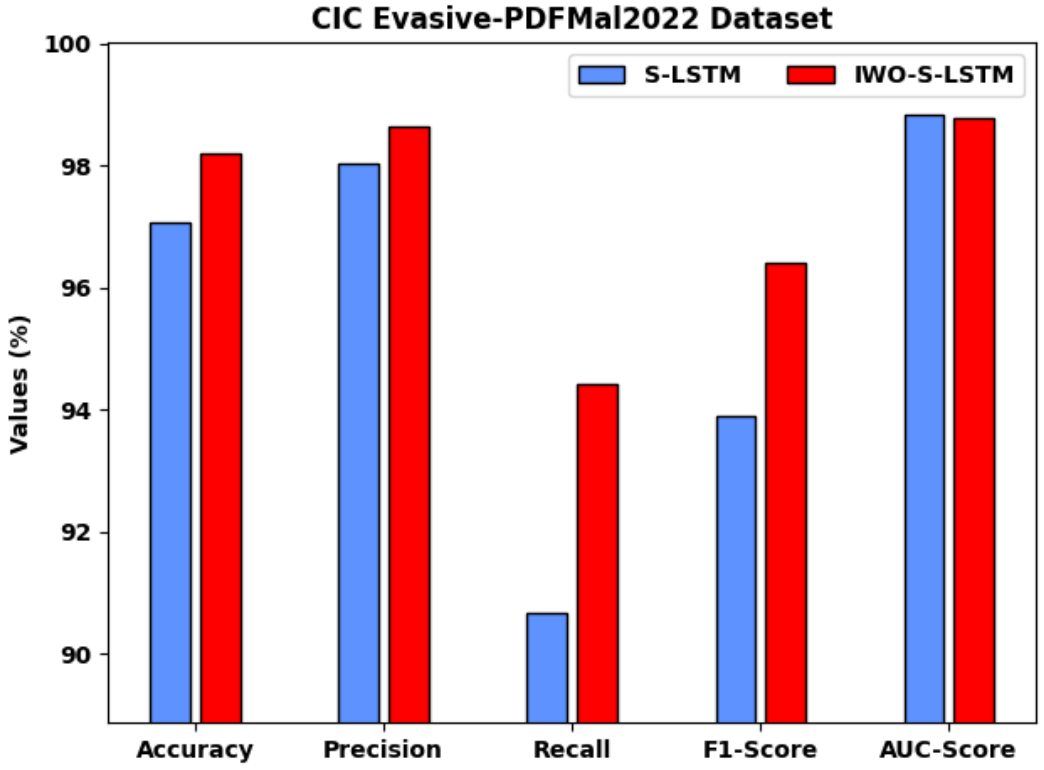

Fig. 8. Result analysis of S-LSTM and IWO-S-LSTM approach on CIC Evasive-PDFMal2022 dataset

Fig. 8 investigates a brief classification outcomes of the S-LSTM and IWO-S-LSTM model on the CIC Evasive-PDFMal2022 dataset. The results indicated that the S-LSTM model has obtained $accu_y$, $prec_n$,$reca_l$, $F1_{score}$, and AUC_{score} of 97.06%, 98.05%, 90.68%, 93.91%, and 98.83% respectively. However, the IWO-S-LSTM model has attained even improved $accu_y$, $prec_n$,$reca_l$, $F1_{score}$, and AUC_{score} of 98.20%, 98.65%, 94.43%, 96.40%, and 98.77% respectively.

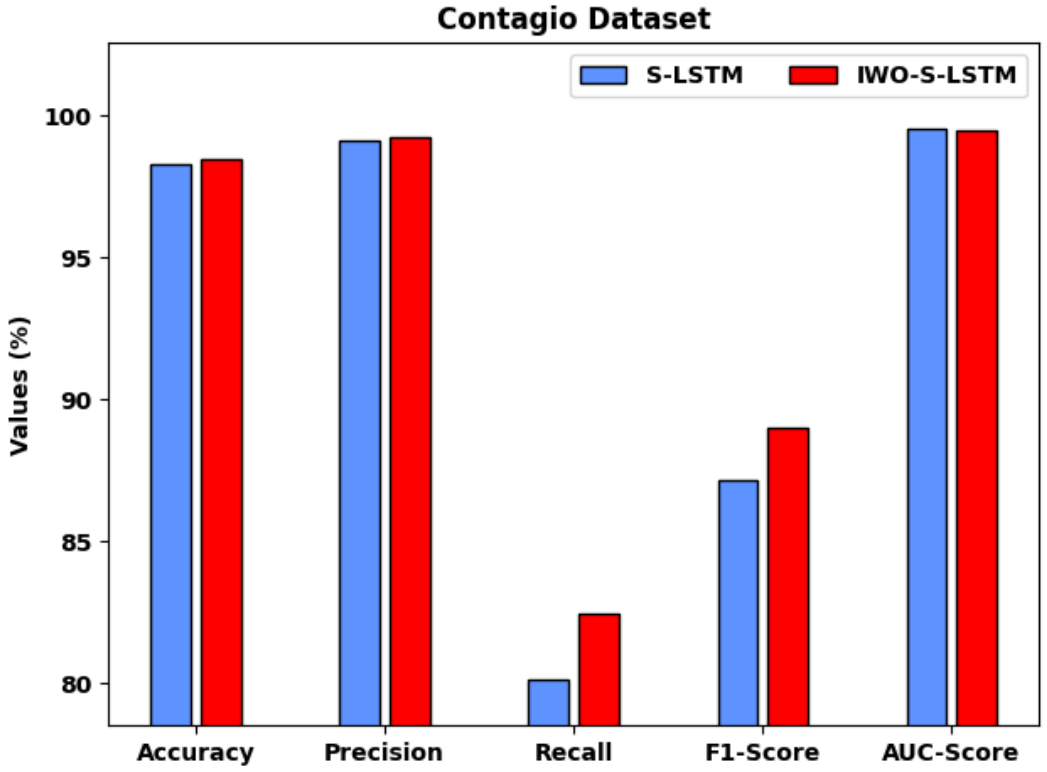

Fig. 9. Result analysis of S-LSTM and IWO-S-LSTM approach on Contagio dataset

Fig. 9 examines a classification outcomes of the S- LSTM and IWO-S-LSTM model on the Contagio dataset. The result indicates that the S-LSTM model has gained $accu_y$, $prec_n$,$reca_l$, $F1_{score}$, and AUC_{score} of 99.11%, 80.10%, 87.13%, 99.56%, and 98.83% correspondingly. But, the IWO-S-LSTM technique has accomplished even better $accu_y$, $prec_n$,$reca_l$, $F1_{score}$, and AUC_{score} of 98.20%, 99.21%, 82.46%, 88.97%, and 99.46% correspondingly.

In order to report the enhanced performance of the IWO-S-LSTM model, a brief comparison study with recent methods is made in Table 2 [22]. Fig. 10 investigates the $accu_y$ analysis of the IWO-S-LSTM model with recent models. The figure indicated that the ridge regression, DT model, and RF model have resulted to lower performance with $accu_y$ of 93.50%, 93.47%, and 93.19% respectively. Followed by, the AdaBoost and SGDC models have provided slightly enhanced outcomes with $accu_y$ of 95.82% and 95.68% respectively. Though the LR model has accomplished reasonably $accu_y$ of 96.33%, the IWO-S-LSTM model has outperformed the compared ones with higher $accu_y$ of 98.47%.

Table 2 Comparative analysis of IWO-S-LSTM technique with existing algorithms

Methods	Accuracy	Precision	AUC
DT Model	93.47	92.92	93.46
RF Model	93.19	93.54	92.82
AdaBoost	95.82	96.65	95.71
Logistic Regression	96.33	95.73	96.20
Ridge Regression	93.50	93.20	93.71
SGDC Model	95.68	95.57	95.76
IWO-S-LSTM	98.47	99.21	99.46

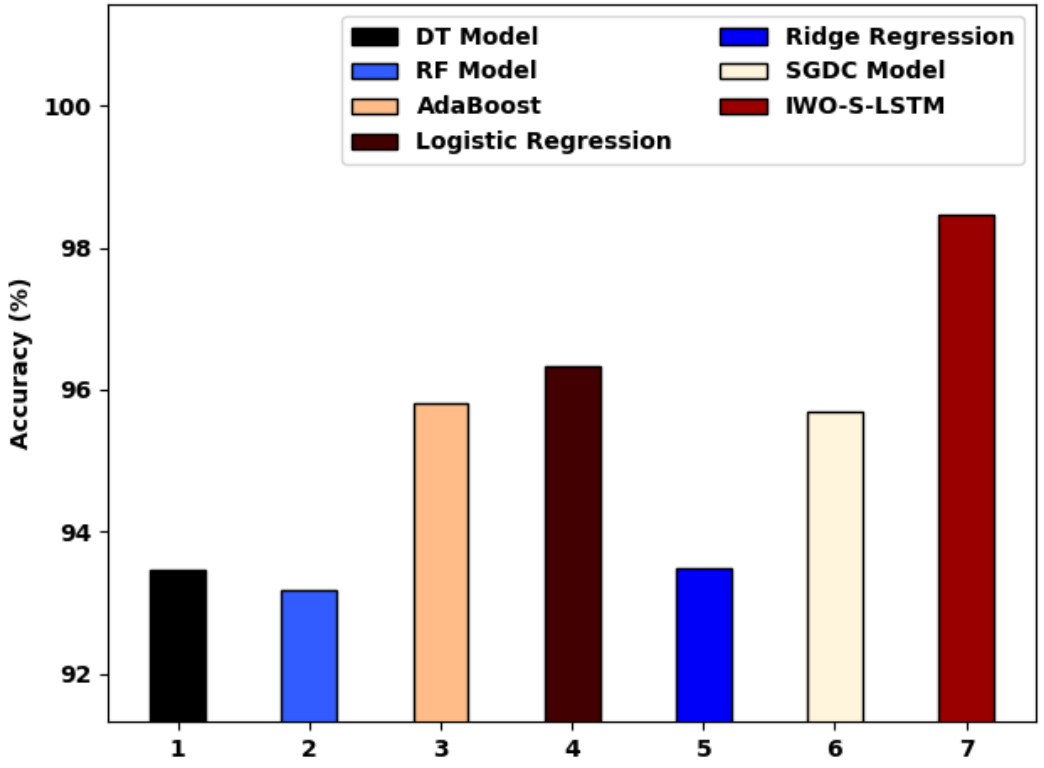

Fig. 10. Acc_y analysis of IWO-S-LSTM technique with existing algorithms

Fig. 11 examines the $prec_n$ and AUC analysis of the IWO-S-LSTM model with current models. The figure indicates that the ridge regression, DT model, and RF model have resulted in low performance with minimum values of $prec_n$ and AUC. Next, the AdaBoost and SGDC models have provided slightly improved values of $prec_n$ and AUC. Although the LR model has attained reasonable $prec_n$ and AUC of 95.73% and 96.20%, the IWO-S-LSTM model has outperformed the compared ones with high $accu_y$ of 99.21% and 99.46% correspondingly.

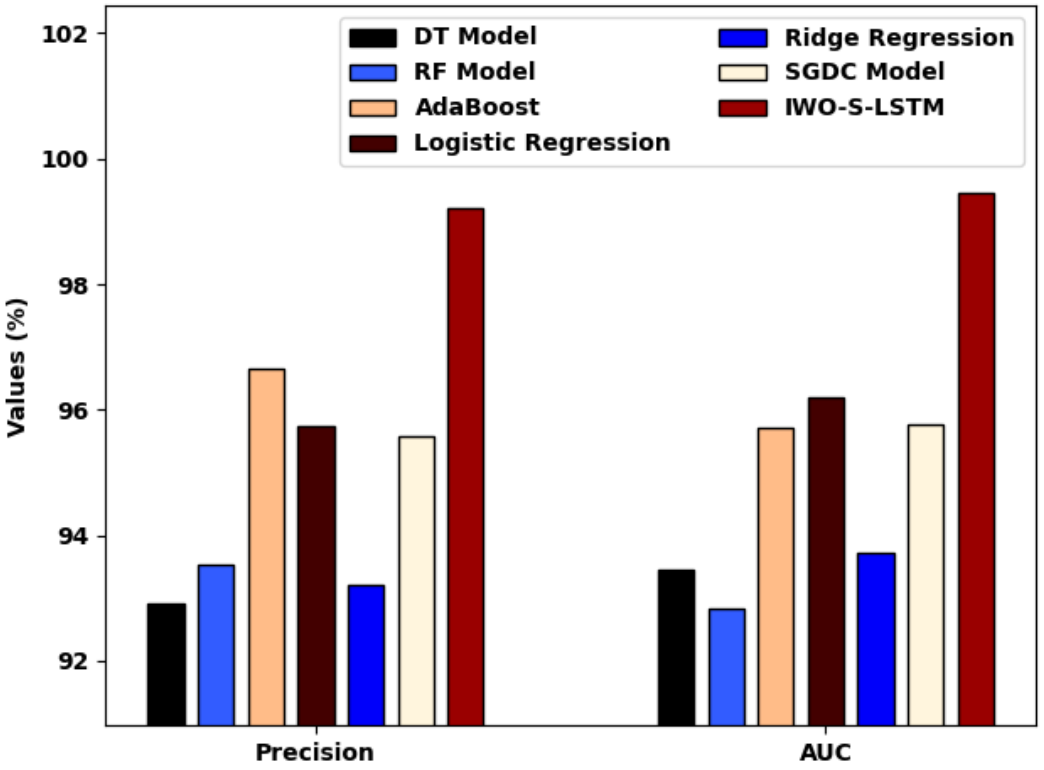

Fig. 11. Comparative analysis of IWO-S-LSTM technique with existing algorithms

The above mentioned results and discussion portrayed that the IWO-S-LSTM model has accomplished superior PDF malware detection and classification process.

5. Conclusion

In this article, a new IWO-S-LSTM model has been developed for the recognition and classification of different kinds of malware that exist in PDF documents. The proposed IWO-S-LSTM model is pre-processed in two stages namely categorical encoding and null value removal. Followed by, AE based outlier detection with S-LSTM based classification model is developed. At last, the IWO algorithm is used for fine tuning the hyperparameters involved in the S-LSTM model. To determine the enhanced outcomes of the IWO-S-LSTM technique, a series of simulations were executed on two benchmark datasets. The experimental outcomes highlighted the promising performance of the IWO-S-LSTM model over the other approaches. In future, hybrid metaheuristic algorithms are developed for improving the overall classification efficiency of the IWO-S-LSTM model.

References

[1] Rudra, B., 2021. Study of a hybrid approach towards malware detection in executable files. *SN Computer Science*, *2*(4), pp.1-7.

[2] Rathore, H., Agarwal, S., Sahay, S.K. and Sewak, M., 2018, December. Malware detection using machine learning and deep learning. In *International Conference on Big Data Analytics* (pp. 402-411). Springer, Cham.

[3] Mercaldo, F. and Santone, A., 2020. Deep learning for image-based mobile malware detection. *Journal of Computer Virology and Hacking Techniques*, *16*(2), pp.157-171.

[4] Yuxin, D. and Siyi, Z., 2019. Malware detection based on deep learning algorithm. *Neural Computing and Applications*, *31*(2), pp.461-472.

[5] Yen, Y.S. and Sun, H.M., 2019. An Android mutation malware detection based on deep learning using visualization of importance from codes. *Microelectronics Reliability*, *93*, pp.109-114.

[6] Singh, P., Tapaswi, S. and Gupta, S., 2020. Malware detection in pdf and office documents: A survey. *Information Security Journal: A Global Perspective*, *29*(3), pp.134-153.

[7] Cuan, B., Damien, A., Delaplace, C. and Valois, M., 2018, July. Malware detection in pdf files using machine learning. In *SECRYPT 2018-15th International Conference on Security and Cryptography* (p. 8p).

[8] Tian, D., Ying, Q., Jia, X., Ma, R., Hu, C. and Liu, W., 2021. MDCHD: A novel malware detection method in cloud using hardware trace and deep learning. *Computer Networks*, *198*, p.108394.

[9] Iadarola, G., Martinelli, F., Mercaldo, F. and Santone, A., 2021. Towards an interpretable deep learning model for mobile malware detection and family identification. *Computers & Security*, *105*, p.102198.

[10] Mohammed, T.M., Nataraj, L., Chikkagoudar, S., Chandrasekaran, S. and Manjunath, B.S., 2021, November. HAPSSA: Holistic Approach to PDF malware detection using Signal and Statistical Analysis. In *MILCOM 2021-2021 IEEE Military Communications Conference (MILCOM)* (pp. 709-714). IEEE.

[11] Yoo, S., Kim, S., Kim, S. and Kang, B.B., 2021. AI-HydRa: Advanced hybrid approach using random forest and deep learning for malware classification. Information Sciences, 546, pp.420-435.

[12] Li, Y., Wang, X., Shi, Z., Zhang, R., Xue, J. and Wang, Z., 2021. Boosting training for PDF malware classifier via active learning. International Journal of Intelligent Systems.

[13] Gandamayu, I. B. M., Antari, N. W. S., & Strisanti, I. A. S. (2022). The level of community compliance in implementing health protocols to prevent the spread of COVID-19. International Journal of Health & Medical Sciences, 5(2), 177-182. https://doi.org/10.21744/ijhms.v5n2.1897

[14] Jeong, Y.S., Woo, J. and Kang, A.R., 2019. Malware detection on byte streams of pdf files using convolutional neural networks. Security and Communication Networks, 2019.

[15] Rinartha, K., & Suryasa, W. (2017). Comparative study for better result on query suggestion of article searching with MySQL pattern matching and Jaccard similarity. In 2017 5th International Conference on Cyber and IT Service Management (CITSM) (pp. 1-4). IEEE.

[16] Zhang, J., 2018. MLPdf: an effective machine learning based approach for PDF malware detection. arXiv preprint arXiv:1808.06991.

[17] Ye, Y., Chen, L., Hou, S., Hardy, W. and Li, X., 2018. DeepAM: a heterogeneous deep learning framework for intelligent malware detection. Knowledge and Information Systems, 54(2), pp.265-285.
[18] Liu, C.Y., Chiu, M.Y., Huang, Q.X. and Sun, H.M., 2021, July. PDF Malware Detection Using Visualization and Machine Learning. In IFIP Annual Conference on Data and Applications Security and Privacy (pp. 209-220). Springer, Cham.
[19] Cozzolino, D. and Verdoliva, L., 2016, December. Single-image splicing localization through autoencoder-based anomaly detection. In *2016 IEEE International workshop on information forensics and security (WIFS)* (pp. 1-6). IEEE.
[20] Yu, L., Qu, J., Gao, F. and Tian, Y., 2019. A novel hierarchical algorithm for bearing fault diagnosis based on stacked LSTM. *Shock and Vibration*, *2019*.
[21] Sedighy, S.H., Mallahzadeh, A.R., Soleimani, M. and Rashed-Mohassel, J., 2010. Optimization of printed Yagi antenna using invasive weed optimization (IWO). *IEEE Antennas and Wireless Propagation Letters*, *9*, pp.1275-1278.
[22] https://www.unb.ca/cic/datasets/pdfmal-2022.html
[23] https://github.com/srndic/mimicus/tree/master/data
[24] Damaševi˘cius, R.; Ven˘ckauskas, A.; Toldinas, J.; Grigaliunas, Š. Ensemble-Based ¯ Classification Using Neural Networks and Machine Learning Models for Windows PE Malware Detection. Electronics 2021, 10, 485. https:// doi.org/10.3390/electronics10040485.