Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 28s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Deep Learning Combined with A Technique for Detecting Viruses in Pdfs and Urls

Mrs. J. Jayapradha[1], R. Dhinesh[2], R. Balasubramanian[3], D. Madhurakavi[4], R. Swathi[5], S. Tejaswini[6]

[1,2,3,4,5]Department of Computer Science and Engineering, Pondicherry University, Puducherry
[1,4]Department of Computer Science and Engineering, Manakula Vinayagar Institute of Technology
Email ID: jpanandhi@gmail.com 1, dhineshofficial1203@gmail.com 2, balar152004@gmail.com 3, madhurakavi9787@gmail.com4, swathir1025@gmail.com 5, tejasundar1234@gmail.com 6

Cite this paper as: Mrs. J. Jayapradha, R. Dhinesh, R. Balasubramanian, D. Madhurakavi, R. Swathi, S. Tejaswini, (2025) Deep Learning Combined with A Technique for Detecting Viruses in Pdfs and Urls, *Journal of Neonatal Surgery*, 14 (28s), 22-31

ABSTRACT

PDF malware is becoming a more serious cybersecurity risk as hackers use malicious payloads and embedded URLs to avoid detection. These complex dangers frequently cause traditional machine learning classifiers to fail. For improved PDF virus detection, we suggest a hybrid RNN-BiLSTM model in order to solve this. BiLSTM improves contextual awareness by processing data in both directions, while the RNN component records temporal dependencies. Furthermore, to detect malicious URLs, we incorporate a BiLSTM-BiGRU architecture, in which BiLSTM improves contextual analysis and BiGRU records sequential dependencies. This hybrid technique increases the efficiency and accuracy of detecting hidden linkages and malware.Our system efficiently identifies new threats while cutting down on training time by utilizing sequential modeling capabilities. According to experimental results, the suggested model performs more accurately and efficiently than conventional techniques, making it a reliable and expandable solution for PDF virus detection.

Keywords: *PDF malware detection, RNN, BiLSTM, BiGRU, hybrid model, malicious URLs, cybersecurity, sequential data analysis, phishing detection, deep learning*

1. INTRODUCTION

Advanced detection techniques, like hybrid algorithms and image-based analysis, have been developed to battle PDF malware. These techniques examine both the structural and content characteristics of PDF files, uncovering hidden threats more efficiently than standard detection systems. Malicious code that is embedded in Portable Document Format (PDF) files and takes advantage of flaws in PDF readers or deceives users through social engineering techniques is known as PDF malware. These infected PDFs may contain harmful scripts, links, or executable code that activate upon opening, leading to severe consequences such as data theft, security breaches, or unauthorized system access. While hybrid approaches use deep learning models to identify patterns linked to malicious PDFs, image-based analysis looks for irregularities in the document's visual components. Cybersecurity experts can increase malware detection and system security by employing these advanced tactics. Malware, in a broader sense, refers to a variety of harmful software, such as viruses, worms, trojans, ransomware, and spyware. In order to corrupt systems, cybercriminals use phishing emails, malicious websites, and compromised software downloads. Once inside, malware can carry out destructive tasks like stealing confidential information, interfering with regular business activities, or giving attackers unapproved access

2. LITERATURE SURVEY

A literature survey is a thorough analysis of previous investigations, studies, and advancements pertaining to a certain subject. It aids researchers in comprehending earlier work, spotting gaps, and investigating approaches from earlier investigations. Reviews of the literature look at a range of cybersecurity strategies, including those used to detect PDF malware, including deep learning, machine learning classifiers, and traditional signature-based approaches.Because malware is always changing, studies have shown that traditional detection systems frequently fail to keep up with new threats. As a result, researchers are looking into hybrid models like RNN, BiLSTM, and BiGRU to increase accuracy. The efficiency of deep learning in identifying sequential data patterns in malicious PDFs and embedded URLs has been demonstrated by recent studies. Compared to standard models, hybrid techniques that include multiple neural networks have shown higher detection rates. Researchers can improve current methods, provide fresh approaches, and support the continuous developments in cybersecurity by examining these papers. A solid literature review serves as a basis for creating malware detection algorithms that are more reliable and effective. *Masao Kubo, Hiroshi Sato,* and *Tuan Van Dao[1]* Since malware is becoming more

complex, effective detection techniques are crucial.Conventional methods, which depend on behavior analysis and signatures, demand a lot of processing power. Although machine learning has shown great promise, choosing the best set of methods is still difficult. Deep neural networks require a lot of computing resources even though they increase classification accuracy. This research suggests a lightweight architecture that combines attention mechanisms, a Variational Autoencoder (VAE), and tiny Convolutional Neural Networks (CNNs). Superior performance on both balanced and unbalanced Malimg datasets is demonstrated by experimental findings, providing a practical and efficient malware detection approach while lowering computing complexity. *Di Niu, Husam Kinawi, Hongwen Zhang, and Qikai Lu [2]* Cybersecurity depends on real-time malware categorization, which demands minimal inference latency and excellent accuracy. These demands are difficult for conventional Convolutional Neural Network (CNN) models to reconcile. *SeqConvAttn* and *ImgConvAttn*, two self-attention transformer-based classifiers, are presented in this study as better substitutes for CNNs. It also suggests a two-stage methodology that integrates these models in a file-size-aware manner to maximize accuracy-latency tradeoffs. Transformer-based models reduce inference latency and exceed CNNs in accuracy, according to extensive experiments conducted on the BIG 2015 and BODMAS datasets. With better classification performance and less computing cost, our method offers a practical, real-time malware detection solution that tackles contemporary network issues. *[3] Liang Ge* Machine learning-based solutions are required since traditional signature-based malware detection is unable to keep up with developing evasion strategies. In order to categorize malware with anomaly-free training data, this study suggests a semi-supervised malware detection method that uses energy-based models. The technique successfully separates harmful code by using density estimations as normality scores. Furthermore, malware localization within the code is facilitated by a revolutionary gradient-based technique. Comprehensive testing on a PE dataset with 20,000 samples shows performance on par with the most advanced techniques. By improving interpretability and giving security analysts a better understanding of malware behavior, the suggested paradigm advances the development of reliable and intelligible cybersecurity solutions. *W. Hardy, L. Chen, S. Hou, Y. Ye, and X. Li [4]* Data security has become essential due to the rise in virus threats. This study introduces a machine learning-based behavioral malware detection solution for Portable Executable (PE) files.The system employs a two-phase methodology, utilizing the Blue Hexagon Open Dataset for Malware Analysis (BODMAS). First, a random forest-based binary classifier has a 99.48% accuracy rate in determining whether a PE file is dangerous or benign. The malware family is then identified with 92.49% accuracy by an ensemble-based multi-class classifier that combines KNN, SVM, random forest, decision tree, and gradient boosting.

This work demonstrates a reliable, highly accurate method for enhancing cybersecurity malware detection and categorization. *Hazem Qattous, Ammar Odeh, and Qasem Abu Al-Haija [5]* Due to their widespread use in digital document sharing, PDF files are frequently the target of virus assaults. Advanced detection techniques are required because malicious actors insert malicious code into PDFs that appear to be safe. This study presents a new detection system that makes use of a hyperparameter-optimized AdaBoost decision tree.With a minimal prediction interval of 1.894 µSec and an excellent 98.88% precision, the system was trained using the Evasive-PDFMal2022 dataset. Its high-performance, lightweight design outperforms current models and offers a useful way to identify PDF viruses. By providing a dependable and effective solution to protect against PDF-based threats, our work strengthens cybersecurity efforts. *S. Alshamrani, Sultan [6]* Malicious PDF files are difficult for traditional antivirus software to identify, so sophisticated security measures are required. This study presents a machine learning-based approach for classifying PDF malware that makes use of both dynamic and statistical analysis. This technique successfully detects unknown and zero-day threats, in contrast to signature-based techniques. After evaluating five classifier methods, random forest (RF) outperformed the other models with the greatest F1-measure of 0.986. Simulated obfuscation assaults are used to further validate the system's resilience.By improving accuracy and resilience against sophisticated malware, this research significantly advances cybersecurity and offers a powerful tool for detecting and mitigating PDF-based threats in modern computer systems.

Yang Fan, Meijin Li, Zhengyang Mao, and Zhiyang Fang [7] The detection of PDF malware is becoming more difficult as adversaries create evasion strategies. This paper presents EvadeRL, a new framework for creating adversarial PDF examples that uses a double deep Q-Network. By selecting the optimal modifications based on categorization feedback, the EvadeRL agent leverages reinforcement learning to enhance its methodology. As opposed to earlier studies that expose detection shortcomings, our approach enhances evasion sustainability by online fine-tuning. Experimental results show that EvadeRL outperforms existing methods in terms of evasion rate, execution cost, and adaptability to evolving malware and detectors.This research advances cybersecurity defences against advanced malware threats concealed within PDFs by offering crucial insights into hostile strategies.

3. PROPOSED SYSTEM

For sophisticated PDF virus detection, the suggested system presents a hybrid strategy that combines Recurrent Neural Networks (RNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks. Because of its ability to examine sequential data, this model can spot minute trends that could point to harmful information in PDF files.

3.1 RNN-BiLSTM Model for PDF Malware Detection :

BiLSTM examines data both forward and backward, improving its comprehension of the document's structure and context,

the RNN component records temporal dependencies. The system can identify hidden malware, including embedded URLs and malicious elements frequently utilized by attackers to launch assaults, thanks to this dual-directional processing.To further enhance the detection of dangerous URLs, the model is expanded by including a hybrid BiLSTM and BiGRU (Bidirectional Gated Recurrent Unit) architecture. While BiLSTM provides context by examining both past and future sequences, the BiGRU component records sequential dependencies inside the URL language. The accuracy and effectiveness of identifying harmful links that might be utilized for malware downloads or phishing attacks are improved by this hybrid technique. In the quickly changing cybersecurity landscape, the hybrid RNN-BiLSTM and BiGRU model performs better than conventional detection techniques, providing increased accuracy and shorter training times while being flexible enough to respond to new threats and attack methods.

3.2 BiLSTM-BiGRU Model for URL Analysis :

A URL (Uniform Resource Locator) is a web address that is used to locate and access resources on the internet. Among its components are the protocol (like HTTP or HTTPS), domain name (like www.example.com), and path (like /page).URLs facilitate file downloads, website navigation, and online service access. HTTPS, which encrypts data to safeguard user privacy and stop online risks like man-in-the-middle attacks, is used by secure URLs. In order to fool consumers into installing malware or disclosing private information, malicious actors frequently use URLs to their advantage by inserting dangerous links into phishing emails, phony websites, or compromised PDF files. Cybersecurity systems use threat intelligence databases, link structure analysis, and suspicious pattern detection to identify and stop dangerous URLs.Advanced detection methods like machine learning and deep learning models like BiLSTM and BiGRU improve the detection of harmful URLs by uncovering hidden patterns in link topologies.Users can defend themselves from URL-based threats by utilizing security tools, avoiding unverified links, and updating their browsers on a regular basis.

3.3Malicious URL Link and PDF:

A malicious link is a URL created with the intent to trick people and carry out destructive tasks like distributing malware, collecting private information, or sending users to phony websites. To fool people into clicking on these links, cybercriminals insert them into phishing emails, phony websites, social media posts, and even PDF files. A malicious link can result in identity theft, financial fraud, or unauthorized access, among other cyberthreats, once it is clicked. In order to make it harder for consumers to recognize these links as threats, attackers frequently employ abbreviated URLs or significantly altered domain names that mimic authentic websites. Cybersecurity solutions examine link architectures, look for banned URLs, and keep an eye out for questionable trends in order to identify and stop fraudulent links. By identifying hidden patterns in URL text, advanced methods like machine learning and deep learning models like BiLSTM and BiGRU enhance detection. To stop dangerous connections before they reach users, organizations also utilize firewalls, email filters, and safe browsing software. Reducing cyber dangers requires educating people on how to spot phishing attempts and steer clear of unconfirmed links. In order to protect against malicious links, regular security upgrades, browser safeguards, and URL reputation services are essential. To increase security, users should avoid downloading files from unidentified sources, double-check URLs before clicking, and use multi-factor authentication (MFA).In order to reduce the dangers connected with harmful links, it is crucial for individuals and businesses to stay educated and implement proactive cybersecurity measures, as cybercriminals are always changing their strategies.

4. DATA COLLECTION

The Stratosphere Laboratory at Czech Technical University, which keeps open-source malware and network-traffic datasets for cutting-edge cybersecurity research, provided the dataset URL used in this project.The dataset URL utilized in this study came from the Stratosphere Laboratory at the Czech Technical University, which offers network traffic data and open-source malware samples for cutting-edge cybersecurity research. In particular, the "pdfmal-2022" dataset serves as the basis for the training, verification, and testing of the suggested malware classification method. The CIC-curated dataset is probably made up of a variety of PDF files that have malware implanted in them. For the algorithm to successfully learn and generalize patterns suggestive of malicious software, real-world, representative samples must be included. It is anticipated that the dataset would include differences in malware kinds, reflecting the dynamic character of cyberthreats.Making use of this information is essential for assessing the system's functionality in real-world scenarios and making sure it is resilient and flexible enough to adjust to the ever-changing landscape of malware based on PDFs. The project's commitment to developing a trustworthy and efficient malware categorization solution is demonstrated by its use of datasets from credible sources, such as the CIC.

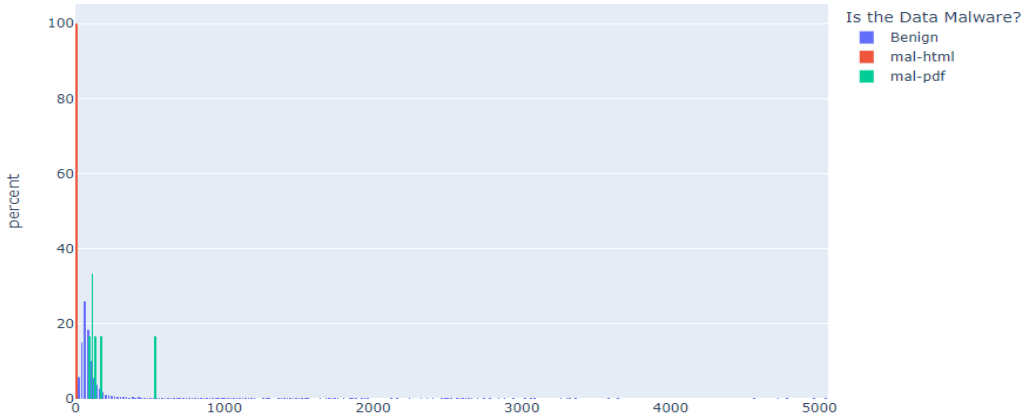

Fig. 4.1 Data Source: University of New Brunswick - CIC pdfmal-2022 Dataset

The prevalence of malware in PDF files in comparison to both benign PDFs and malware in HTML files is graphically represented in the graph. According to the data, malware in PDFs makes up a sizable percentage of the threats found, which may point to criminal actors specifically taking advantage of the PDF format. On the other hand, although malware in HTML files is likewise a significant risk, it seems to occur less frequently than virus in PDF documents. This realization emphasizes how crucial it is to have strong detection systems designed especially for PDF files because of their extensive usage and vulnerability to abuse.In order to lessen the threats caused by malware, it also emphasizes the necessity of thorough cybersecurity procedures across a variety of file formats. By recognizing these patterns, effective defenses against malicious activity that targets both PDF and HTML files can be developed, strengthening digital security overall.

5. PRE-PROCESSING

Preparing the raw data from the "pdfmal-2022" dataset—which was acquired from Stratosphere Laboratory – Czech Technical Universityor efficient usage in the malware classification system requires pre-processing. This stage entails a number of crucial steps to improve the data's appropriateness and quality. Data cleaning, which takes care of any anomalies, missing numbers, or discrepancies in the dataset, is usually one of the first processes. To guarantee consistency in the feature scale, the data may then undergo normalization or standardization.

	filename	obj	end obj	stream	endstream	xref	trailer	startxref	Page	Encrypt	...	AA	OpenAction	AcroForm	JBIG2Decode	RichMedia	Launch	EmbeddedFile	XFA	Colors_2_24	label
0	999875	73	73	25	25	2	2	2	7	0	...	0	0	0	0	0	0	0	0	0	Benign
1	999870	122	122	46	46	2	2	2	22	0	...	0	0	0	0	0	0	0	0	0	Benign
2	999858	66	66	19	19	2	2	2	12	0	...	0	0	0	0	0	0	0	0	0	Benign
3	999854	72	72	22	22	2	2	2	12	0	...	0	0	0	0	0	0	0	0	0	Benign
4	999331	70	70	19	19	2	2	2	11	0	...	0	0	0	0	0	0	0	0	0	Benign

Fig. 5.1 Dataset of PDF File Structural Features for Malware Classification

A)Data Cleaning and Standardization : Key characteristics or patterns that indicate malware can also be recorded using feature extraction techniques. The dataset is most likely separated into training, validation, and testing sets in order to evaluate the classification system's performance.In order to minimize potential biases, minimize noise, and optimize the dataset for machine learning model training, this pre-processing stage is essential. The overall efficacy and generalizability of the malware classification system are greatly enhanced by the careful data handling that takes place at this phase.

B)Feature Extraction: Duplicate values are found and eliminated during the feature extraction process to improve the analysis's effectiveness and precision. When duplicate values are present in the dataset, they can introduce redundancy into the information or inflate the significance of particular features, which can distort the results. The feature extraction stage guarantees that every distinct trait or attribute is accurately documented and taken into account during analysis by removing duplicates[17].

*C) Model Creation:*The model for the proposed system combines Recurrent Neural Networks (RNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks to enhance PDF virus detection.A thorough examination of the structure of the PDF file is made possible by the RNN component's ability to identify sequential dependencies in the data and the BiLSTM's ability to process information both forward and backward. This two-pronged method aids in the detection of hidden malware, such as malicious URLs and links that are embedded. The model combines a hybrid BiLSTM and Bidirectional Gated

Recurrent Unit (BiGRU) architecture, which focuses on finding patterns in URL structures, to further increase detection accuracy. BiLSTM processes both past and future data to provide contextual understanding, whereas BiGRU records sequential dependencies inside URLs.This hybrid technique is designed to effectively identify harmful links and PDF malware, guaranteeing high accuracy and flexibility in response to new threats.

*D) Architecture Diagram:*The integration of the Bidirectional Long Short-Term Memory (BiLSTM) method, which improves the system's capacity to evaluate sequential input in both forward and backward directions, is highlighted in the suggested architecture diagram for PDF virus prediction. BiLSTM is very good at spotting intricate temporal relationships and patterns in PDF file structures.

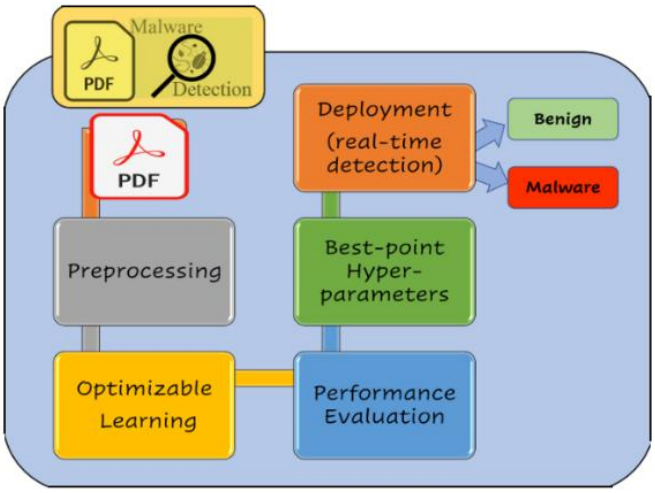

Fig. 5.2 Architecture diagram of PDF and URL malware detection

This design guarantees accurate detection of both known and unknown threats by allowing the system to learn from past data and adjust to changing malware strategies. The design provides a strong and advanced method of malware categorization by integrating BiLSTM into the framework, greatly enhancing the overall effectiveness and dependability of PDF virus detection.

6. TEST DATA ANALYSIS FOR PDF AND URL

A PDF file that may contain malware makes up the test data. These files frequently have hidden payloads that, when opened, carry out harmful operations, obfuscated URLs, or embedded JavaScript. These files are disguised as authentic documents by attackers, who then insert scripts that take advantage of software flaws or trick users into performing destructive tasks. In order to get beyond conventional security measures, the PDF can have encoded JavaScript that launches automatically when it is opened. Phishing URLs concealed within the file, which drive users to phony websites intended to steal passwords, are another frequent danger. Additionally, some PDFs could have embedded files with malware payloads, such as pictures or attachments.Attackers alter the file's structure using obfuscation techniques, which makes it challenging for conventional antivirus software to identify. To determine the test PDF's level of risk, the system takes out and examines its textual, behavioral, and structural components. Using both known and novel threats, the goal is to assess the model's ability to differentiate between safe and malicious PDFs.The system makes sure that even sophisticated malware that uses evasion techniques to get around security measures is detected by analysing patterns, abnormalities, and possible attack paths.

6.1 Prediction:Examining the document's structure and searching for anomalies like malformed headers, excessively strong encryption, or dubious metadata patterns is the first stage in the PDF malware detection process. Bidirectional Long Short-Term Memory (BiLSTM), which is used by the system to retain consecutive dependencies, aids in the detection of potentially dangerous patterns such as hidden scripts or illegal system modifications.By examining URLs inside the PDF and identifying obfuscated or dubious links that could result in phishing attempts, the Bidirectional Gated Recurrent Unit (BiGRU) improves this. Higher accuracy is ensured by the system's ability to detect deeply implanted malware components thanks to the dual-directional processing of BiLSTM and BiGRU.The PDF is marked as dangerous if harmful components are discovered, which initiates additional security measures. The model's explainability—highlighting the precise passages of the document that went into its classification—is one of its main characteristics. For instance, the system will identify the location of an embedded JavaScript function and provide risk evaluations if it is determined to be dangerous. Security analysts can comprehend the nature of the danger and act quickly thanks to this understanding. The model is a useful cybersecurity tool since it can also identify new and changing malware strains, which improves defence against zero-day attacks.

7. PERFORMANCE METRICS

A) Accuracy : One important performance parameter for assessing how well the suggested hybrid RNN-BiLSTM-BiGRU model detects harmful URLs and PDF malware is accuracy. It calculates the percentage of cases—both malicious and benign files—that are correctly classified out of all the instances. The accuracy formula is:

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

Where:

- **TP (True Positives)** – The number of correctly identified malicious PDF files.
- **TN (True Negatives)** – The number of correctly identified benign PDF files.
- **FP (False Positives)** – The number of benign files incorrectly classified as malware.
- **FN (False Negatives)** – The number of malware files incorrectly classified as benign.

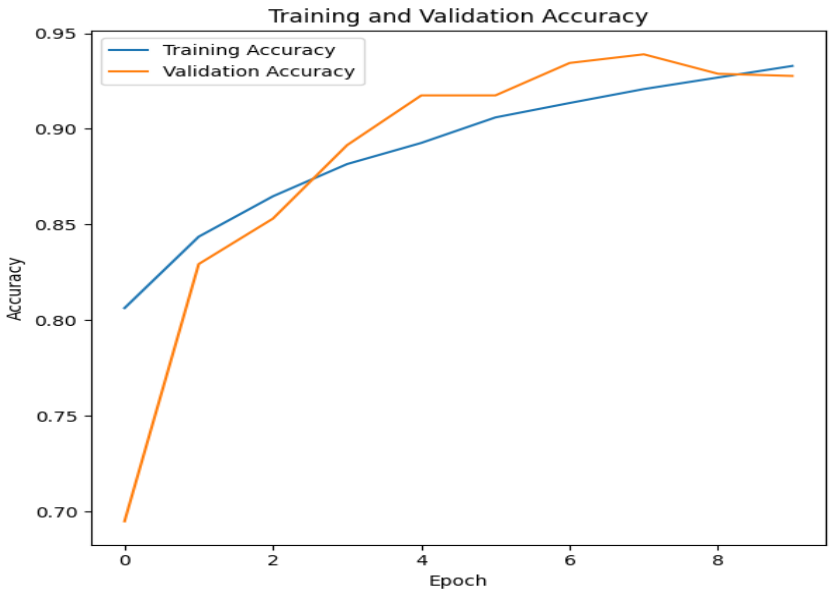

Fig. 7.1 Training vs. Validation Accuracy Over Epochs

The RNN-BiLSTM model uses bidirectional context processing and sequential pattern recognition to increase accuracy. RNN-BILSTM effectively identifies harmful URLs encoded in PDFs, whereas BiLSTM improves feature extraction from malware patterns. Compared to conventional machine learning classifiers, this hybrid technique ensures improved accuracy by minimizing FP and FN rates. The model provides a dependable cybersecurity solution for detecting PDF-based threats, as demonstrated by experimental findings that confirm its consistent high detection accuracy.

B) Precision: An important criterion for assessing the RNN-BiLSTM model's efficacy in identifying dangerous URLs and PDF viruses is precision. It lowers false positives, which could result in needless alarms, by calculating the proportion of projected harmful files that are indeed malicious. The following formula is used to calculate precision:

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

Where:

- **TP (True Positives)** – The number of correctly identified malicious PDF files.
- **FP (False Positives)** – The number of benign files incorrectly classified as malware.

By efficiently examining sequential patterns and bidirectional context, the RNN-BiLSTM model increases precision and guarantees that the malware files it detects are actually harmful. By capturing intricate contextual linkages, the BiLSTM component improves detection accuracy and lowers false positives. The methodology guarantees that fewer innocuous files are incorrectly identified by reducing FP, which improves the accuracy of malware detection. The suggested method is an effective cybersecurity solution for precisely identifying PDF-based risks because it has a high precision score, which reduces false alarms.

C) Recall : An important parameter for assessing how well the RNN-BiLSTM model detects dangerous URLs and PDF viruses is recall. It assesses the model's accuracy in identifying every real hazardous file, guaranteeing that no dangerous files are overlooked. The recall formula is:

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

Where:

- TP = True Positives
- FN = False Negatives (the number of actual positive cases that were incorrectly predicted as negative)

By successfully identifying sequential patterns and bidirectional dependencies in PDF file structures, the RNN-BiLSTM model improves recall. The approach minimizes the chance of undetected attacks by lowering FN, which guarantees that the majority of malware occurrences are appropriately identified. The model's great capacity to identify even complex and concealed malware is demonstrated by its high recall score, which makes it a reliable defense against changing online threats for PDF files.

*D)F1 Score:*The F1 Score, which provides a single indicator of how well the model detects PDF malware, is a crucial parameter that balances precision and recall. It is particularly useful when the dataset is unbalanced between benign and harmful PDF files. The F1 Score is calculated using the formula below:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

When you need to ensure that the number of false positives and false negatives is kept to a minimum, such as in fraud detection or medical diagnosis, this measure is highly helpful. The primary focus of precision is the caliber of positive forecasts. F1 Score combines the two metrics to provide a comprehensive performance assessment, whereas Recall emphasizes the model's ability to find all relevant samples. These metrics are commonly employed to offer a more thorough understanding of a model's performance, especially when dealing with tasks that involve uneven class distributions.The RNN-BiLSTM model maximizes recall and precision to guarantee a high F1 Score. It minimizes False Positives (FP) and False Negatives (FN) while successfully identifying dangerous material encoded in PDF files.A high F1 Score attests to the model's ability to strike a solid balance between accurately identifying threats and preventing false alarms, making it a dependable cybersecurity strategy.

E)LOSS: The suggested RNN-BiLSTM model for PDF malware detection evaluates the model's capacity to distinguish between harmful and benign PDF files using the loss function, a crucial statistic. Because the classification is binary (benevolent or malevolent), Binary Cross-Entropy (BCE) Loss is employed to assess the difference between predicted probability and actual labels.While a higher loss denotes more misclassification and forces the model to modify its weights during training, a lower loss suggests that the model's predictions are more accurate.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right]$$

Where:

- N = Total number of samples
- y_i = Actual label (1 for malicious, 0 for benign)
- $\hat{y}_i$ = Predicted probability of the sample being malicious

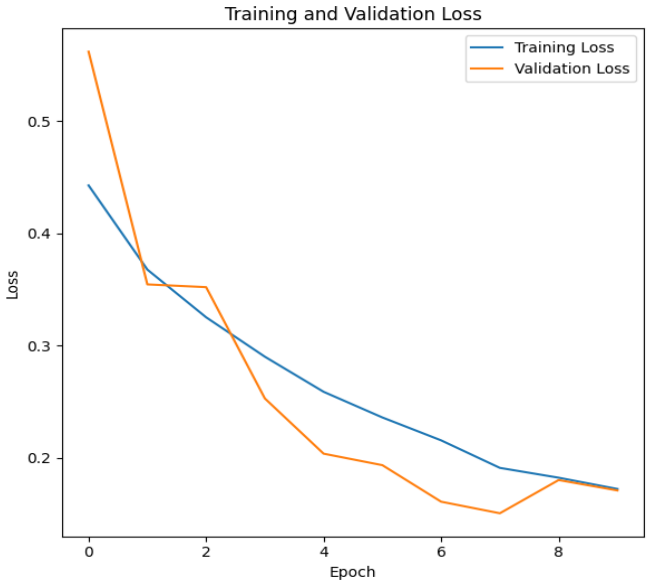

Fig. 7.2 Evaluation of Training and Validation Loss to Monitor Model Convergence

Inaccurate predictions are effectively penalized by the BCE loss function, which motivates the model to gradually improve its classification. The logarithmic function reinforces learning adjustments by assigning a greater penalty when a harmful PDF is mistakenly categorized as benign, or vice versa. By using backpropagation and an optimization technique such as Adam, the RNN-BiLSTM model reduces this loss and guarantees improved detection accuracy. The model adjusts its settings as training goes on to minimize loss and enhance malware detection capabilities.

F) Confusion matrix:

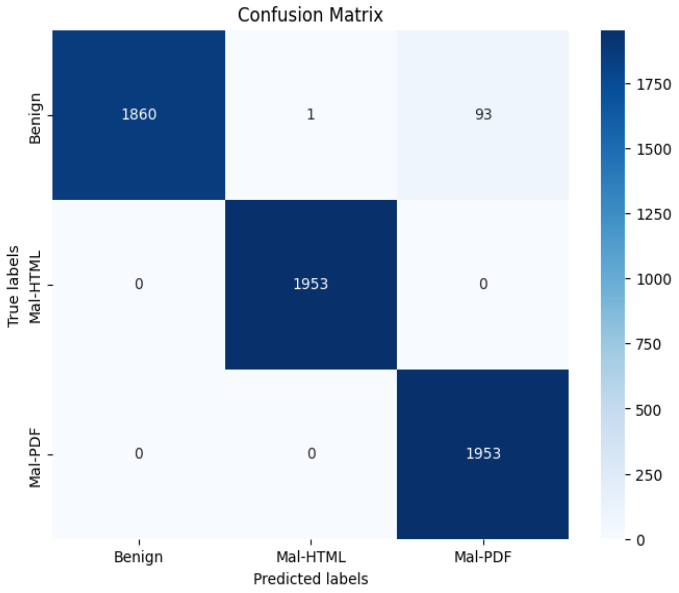

Fig. 7.3 Confusion Matrix for Multi-Class Classification of PDF and HTML Files

Benign, Mal-HTML, and Mal-PDF are the three categories that the image's confusion matrix illustrates how well a model classifies data. The number of accurate and inaccurate forecasts for each category is displayed in the matrix. Instances when the true labels and the anticipated labels match are indicated by the diagonal values (1860, 1953, and 1953). Misclassifications are represented by the off-diagonal values. Notably, the model accurately identified all Mal-HTML and Mal-PDF samples without error, however it incorrectly identified 93 "Benign" samples as Mal-PDF and one as Mal-HTML. This implies that while the model does a good job of identifying dangerous files, it struggles to tell some benign samples apart from malicious ones. The heatmap's color intensity, where deeper hues indicate more counts and lighter hues indicate occurrences that were incorrectly identified, supports the classification accuracy.

8. RESULT AND DISCUSSION

Performance metrics are crucial instruments for evaluating machine learning models' efficacy. The most intuitive statistic, accuracy, measures the percentage of correctly classified occurrences to provide an overall assessment of correctness. However, when there are unequal class distributions, accuracy may not accurately represent the model's performance, leading to inaccurate findings. Precision and recall offer more in-depth information by focusing on certain classes. By determining the proportion of true positive predictions among all positive predictions, precision demonstrates the model's ability to avoid false positives. Conversely, recall, also known as sensitivity, highlights the model's ability to identify relevant examples of a class by calculating the proportion of genuine positive predictions among all actual positive cases.By measuring the difference between expected and actual values, loss, on the other hand, directs the optimization process during training and helps to improve the model. A thorough grasp of a model's performance and fit for the task at hand can be obtained by evaluating it using a mix of these metrics, which guarantees solid and trustworthy results.

9. CONCLUSION

In summary, a major development in cybersecurity is represented by the incorporation of the Bidirectional Long Short-Term Memory (BiLSTM) algorithm into the suggested method for identifying PDF malware. The system's ability to evaluate sequential data and better capture temporal dependencies is increased by utilizing BiLSTM's special ability to process data both forward and backward. This dual processing power enhances detection accuracy and robustness by enabling the model to identify minute variations and patterns suggestive of harmful materials within PDF files. In addition to improving the system's predictive performance over time, the strategic integration of BiLSTM fortifies defences against the constantly changing array of cyberthreats. BiLSTM enables a proactive and flexible approach to thwarting PDF-based malware by providing a more thorough grasp of the context around possible infection. In the end, the application of BiLSTM marks a significant advancement in creating advanced answers to the enduring problems of protecting digital data and networks, guaranteeing that cybersecurity defences continue to be successful against new threats in the digital sphere. Future research must investigate ensemble approaches, improve training strategies using a variety of datasets, integrate explainable AI strategies, and put in place systems for ongoing observation and adjustment. Diverse datasets would improve generalization, while ensemble approaches might combine several detection models for better performance. Explainable AI has the potential to increase trust by offering insights into the system's decision-making process. Last but not least, features for continuous upgrades and feedback loops would guarantee that the system continues to be successful against changing threats from PDF malware

REFERENCES

[1] Y. Liu, W. Lin, J. Wang, and Z. Chen, "A novel approach for malicious PDF detection using deep neural networks," Computers & Security, vol. 92, p. 101760, 2020.

[2] S. Tobiyama, Y. Yamaguchi, H. Shimada, T. Ikuse, and T. Yagi, "Malware detection with deep neural network using process behavior," in Proc. IEEE Trustcom/BigDataSE/ISPA, 2016, pp. 261–267.

[3] W. Hardy, L. Chen, S. Hou, Y. Ye, and X. Li, "DL4MD: A deep learning framework for intelligent malware detection," in Proc. Int. Conf. Data Mining Workshops, 2016, pp. 61–68.

[4] Y. David, N. Partush, and E. Yahav, "Statistical similarity of binaries," in Proc. ACM SIGPLAN Notices, vol. 50, no. 6, pp. 266–280, 2015.

[5] Y. Zhang, L. Wang, Y. Wang, and J. Liu, "Malicious PDF detection using convolutional neural network," IEEE Access, vol. 8, pp. 158131–158140, 2020.

[6] E. Raff, J. Barker, J. Sylvester, R. Brandon, B. Catanzaro, and C. K. Nicholas, "Malware detection by eating a whole EXE," in Proc. AAAI Workshops, 2018.

[7] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," Expert Systems with Applications, vol. 41, no. 4, pp. 1690–1700, 2014.

[8] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: Learning to detect malicious web sites from suspicious URLs," in Proc. WWW '09, pp. 1245–1254.

[9] L. Bilge, D. Balzarotti, W. Robertson, E. Kirda, and C. Kruegel, "Disclosure: Detecting botnet command and control servers through large-scale DNS graph analysis," in Proc. ACSAC, 2014.

[10] Y. Wang, L. Wang, and Y. Zhang, "PDF malware detection via hierarchical learning model," Computers & Security, vol. 89, p. 101682, 2020.

[11] A. Shabtai, R. Moskovitch, Y. Elovici, and C. Glezer, "Detecting unknown malicious applications using machine learning techniques," Computers & Security, vol. 30, no. 4, pp. 325–337, 2012.

[12] T. Nguyen, T. Hung, and N. Pham, "uitPDF-MalDe: Malicious PDF document detection based on machine learning," Journal of Information Security and Applications, vol. 65, p. 103140, 2022.

[13] Z. Li, X. Zhang, Y. Zhu, and J. Liu, "A PDF malware detection model using CNN and multi-layer features," IEEE Access, vol. 9, pp. 10401–10410, 2021.

[14] W. Zhou, Z. Qin, and J. Zhang, "Ensemble learning for PDF malware detection," Security and Communication Networks, vol. 2021, Article ID 6627631.

[15] A. Kirichenko, A. Skuratovskii, and A. Sychev, "Static analysis-based feature engineering for malicious document classification," in Proc. MMM-ACNS, 2020.

[16] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Sok: Security and privacy in machine learning," in Proc. EuroSP, 2018.

[17] Seetharaman, K., and N. Palanivel. 2013. “Texture Characterization, Representation, Description, and Classification Based on Full Range Gaussian Markov Random Field Model with Bayesian Approach.” International Journal of Image and Data Fusion 4 (4): 342–62. doi:10.1080/19479832.2013.804007.

[18] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in Proc. IEEE SP, 2017