Al-Noor Journal for Information Technology and Cyber Security
December (2024); 1 (0): 1- 22
DOI: https://doi.org/10.69513/jncs.v1.i0.a1

LSTM-Based Analysis of De-Identification Techniques for Protecting Sensitive Data

Cik Feresa Mohd Foozy[1 a)], K. Ravindran[1 b)], Naqliyah Zainuddin[2 c)], Ahmad S. Mohd Rozi[2 d)], Muhammad H. A. Fakhrudin[2 e)]

[1]Fakulti Sains Komputer & Teknologi Maklumat, Information Security Interest Group (ISIG), Industry Centre of Excellence Railway (ICoE-REL),
Universiti Tun Hussein Onn Malaysia,
Batu Pahat, Malaysia
[2]Proactive Technology & Services Division, CyberSecurity Malaysia,
Selangor, Malaysia
[a)] Corresponding author: feresa@uthm.edu.my
[b)] kugastinaa@gmail.com, [c)] naqliyah@cybersecurity.my,
[d)] ahmadsyuhaidi@cybersecurity.my, [e)] haziqaiman@cybersecurity.my
ORCID: 0000-0002-9085-6819

Received: 15 / 10/ 2024 Accepted: 20 / 11/ 2024
Published: 17 / 12 / 2024

Abstract

This research examines the efficiency of de-identification techniques in enhancing privacy protections for sensitive data using Long Short-Term Memory (LSTM) models. Following a structured five-step methodology such as Dataset Collection, Data Preparation, Feature Extraction, Classification, and Performance Evaluation. The study evaluates LSTM's performance of dataset based on Resume, Construction, and medical domains. The primary goal is to examine the ability of de-identification methods to hide certain information based on classification accuracy.

Results indicate that LSTM achieves accuracy levels 97.14% on unmodified data, explaining its success detecting sensitive information. However, after applying de-identification using Java Programming at pre-processing phase to eliminate sensitive keyword, the accuracy drops to 78.30%. These findings highlight the effectiveness of de-identification techniques to enhance data privacy, especially in fields that require strict confidentiality.

Introduction

Data security mainly covers three aspects such as confidentiality and integrity are closely linked to data privacy out of the three elements mentioned. Confidentiality is important for ensuring that authorized individuals can access data securely and maintain trust by access attempts. Access integrity involves safeguarding the accuracy and consistency of data over time to prevent any alterations during its lifespan. Ensuring availability means making sure that authorized users can easily access information when necessary to enable the smooth flow of data.

In today's world of progress and transformations, in technology and society it is crucial to protect information as access could harm one's reputation and invade their privacy [1-2]. Besides methods like fingerprints or iris scans that uniquely and biologically identify individuals, sensitive data encompasses company records containing employee information or customer details and financial dealings, with business associates. Educational organizations manage an amount of information including student records, from enrolment to achievements as well as research findings and financial records which makes them susceptible to cyber-attacks. This has led to an increased emphasis on protecting the privacy of data being a topic of conversation.

Al-Noor Journal for Information Technology and Cyber Security
December (2024); 1 (0): 1- 22
DOI: https://doi.org/10.69513/jncs.v1.i0.a1

Unprotected privacy data can result in outcomes, such as damage and repercussions for both individuals and businesses alike. In scenarios where companies may be held accountable for breaching data protection regulations and educational establishments could expose student records and research outputs from faculty members or confidential origins. The internal risks posed by employees with access present a danger to all parties as they have the potential to misuse the data to them. An effective approach to reduce these risks result in educating employees on how to identify phishing emails and other harmful links.

The statistics regarding data breaches highlight the importance of implementing measures to control and protect it from access or misuse. Yearly reports consistently reveal several data breaches occurring annually. The data breach incidents in 2020 exemplify this concern, with factors playing a role. Organized criminal groups were involved in 70% while internal actors accounted for 4%. 55% of the breaches were linked to groups. The importance of implementing strategies to reduce the likelihood of security breaches is underscored by these figures. To enhance data privacy effectively one can, employ deidentification tactics to protect individuals' identities by concealing information. There are three deidentification approaches; anonymization, pseudonymization and data masking [3]. These techniques work by replacing elements with symbols or alternative representations to uphold the secrecy and confidentiality of data.

The research adopts a five steps approach to assess how de-identification privacy methods impact the accuracy of identifying data using Long Short-Term Memory (LSTM). It commences by gathering data through Dataset Acquisition that includes information from Resume Writing Services Construction Companies and Medical Institutions. Data cleaning and

restructuring for classification purposes are carried out as part of Data Preprocessing. During the Feature Extraction phase, in the Natural Language Processing (NLP) process techniques are utilized to identify features that affect the classification procedure. In the Classification Algorithm phase implementation process involves using the LSTM model to analyze data while preserving confidentiality through data masking methods, like anonymization and pseudonymization to test its sorting accuracy capabilities without compromising privacy. In the Performance Evaluation stage of the study results analysis focused on comparing LSTMs performance after data deidentification for any differences, in accuracy and efficiency. The results indicate that the LSTM model achieved an accuracy of 97% which dropped to 78 % after applying the deidentification process highlighting the need to balance privacy protection with maintaining classification accuracy. When studying how de-identification techniques impact enhancing privacy, in identifying information across sectors like healthcare and finance the study offers insights into data privacy concerns.

Literature Review

The rapid advancement of digital technology has significantly heightened the vulnerability of sensitive data, necessitating effective de-identification strategies to protect personal information. This literature review evaluates research and frameworks that explore de-identification techniques for safeguarding data privacy.

In Malaysia, the Personal Data Protection Act 2010 (PDPA) authorities handle personal data responsibly and uphold individuals' privacy. The PDPA outlines requirements for organisations to adhere to data protection principles, including data accuracy, secure storage, and controlled access to personal data.

These regulations align with global standards, such as the European Union's General Data Protection Regulation (GDPR), which emphasises robust anonymisation and pseudonymisation practices to protect sensitive information adequately [4]. Both the PDPA and GDPR prioritise de-identification as a key method to address unauthorised access to critical data, reflecting an international standard in data protection.

Data security is founded on the triad of confidentiality, integrity, and availability, which are essential for managing personal information. Confidentiality restricts data access to authorised users only, integrity preserves the data's accuracy and consistency, and availability ensures data is accessible when required [5]. In Malaysia, these principles form the basis of compliance with PDPA, reinforcing organisational obligations to prevent breaches of personal information and uphold stringent data security standards.

A. De-Identification Techniques

De-identification techniques, such as anonymisation, pseudonymisation, and data masking, are instrumental in securing sensitive data. Anonymisation entails removing identifiable information to prevent data from being traced back to individuals, thereby enhancing privacy by preventing re-identification [6]. Pseudonymisation replaces identifiers with pseudonyms, thus protecting individuals' identities while retaining the data's utility for analysis [7]. Data masking, on the other hand, involves obfuscating sensitive elements within datasets to safeguard privacy while maintaining sufficient usability [8]. These de-identification methods are designed to balance privacy with data utility, allowing for the sharing of de-identified data while minimising privacy risks. Table 1 presents a comparative analysis of three primary de-identification techniques of Anonymisation, Pseudonymisation

and Data Masking for evaluating their respective strengths, limitations, and de-identification risks. Each technique serves to protect privacy by modifying identifiable information in a way that aligns with regulatory standards, such as the Personal Data Protection Act (PDPA) and General Data Protection Regulation (GDPR).

Table 1 shows a comparison of Anonymization, Pseudonymization and Data Masking. Anonymisation is described as removing or simplifying identifiers to prevent data linkage, thereby significantly reducing the risk of re-identification. This method effectively protects privacy but may limit data utility and can be complex to implement in certain cases. With a very low de-identification risk, anonymisation is optimal for situations where strong privacy protection is required, even at the expense of some data usability. Pseudonymisation involves replacing identifiers with pseudonyms, which balances privacy protection with data utility, allowing for some continued analysis of data while protecting individual identities. However, pseudonyms can be reverse engineered if additional information is available, making the de-identification risk moderate. This method is suitable when retaining data utility is essential, though it requires careful handling to mitigate re-identification risks. Data Masking disguises specific data elements, effectively reducing the risk of identification while preserving data for use in applications where exact identifiers are unnecessary. While masking provides low de-identification risk, it may impact usability for specific analytical applications, depending on the degree and nature of the masking applied. Masking is useful when a lower level of risk reduction suffices, and some usability is still required. Overall, this table illustrates the strengths of each method and provides a framework for selecting a

Al-Noor Journal for Information Technology and Cyber Security
December (2024); 1 (0): 1- 22
DOI: https://doi.org/10.69513/jncs.v1.i0.a1

suitable de-identification technique based on privacy requirements and data utility needs.

TABLE 1: DE-IDENTIFICATION COMPARISON

	Anonymization	**Pseudonymization**	**Data Masking**
Description	Removes or generalizes identifiers to prevent data linkage.	Replaces identifiers with pseudonyms to protect privacy	Obfuscates sensitive data elements to protect privacy
Strength	Reduces risk of de-identification significantly.	Protects privacy while maintaining data utility	Effective in obscuring data and reduces de-identification risk
Limitation	May reduce data utility and complex to implement.	Pseudonyms can potentially be reverse engineered	Masking may affect data usability for certain applications
Evaluation on Metrics	De-Identification Risk, Data Usability.	De-Identification Risk, Data Utility	Masking Effectiveness, Data Utility
De-Identification Risk	Very low and identifiers removed.	Moderate and pseudonyms	Low and masking

B. Deep Learning in De-Identification

The combination of deep learning models with de-identification techniques represents an evolving approach to improving data privacy. Deep learning, as a subset of artificial intelligence, includes various neural network architectures, such as Convolutional Neural Networks (CNN), Recurrent Long Short-Term Memory (LSTM) networks, which offer robust abilities for pattern recognition in large datasets [10, 11]. Advances in deep learning increased computation powers [12], also new methods has been introduced for refining de-identification processes, where LSTM, a type of RNN, demonstrates superior accuracy in handling sequential data, making it well-suited for sensitive information detection and privacy protection [13]. By capturing complex data

patterns, LSTM-based de-identification frameworks provide scalable solutions for data privacy, aligning with evolving privacy regulations [14].

Despite these advancements, several challenges persist in de-identification, including establishing optimal levels of de-identification, mitigating re-identification risks, and achieving a balance between privacy and data utility [15]. In Malaysia, the increasing volume of digital data, coupled with evolving cyber threats, underscores the need for frameworks that incorporate advanced de-identification techniques to meet modern data protection demands. Continued research in this area is essential for supporting organisations in navigating the complexities of data security and enhancing their cybersecurity resilience.

C. Long Short-Term Memory (LSTM) Models in Privacy Protection

Recent research has discovered the integration of LSTM models within de-identification frameworks, determining their utility in privacy protection. L. Wu and M. Pan [16] investigated the combined application of LSTM networks and Conditional Random Fields (CRF) models, highlighting their effectiveness in feature extraction and data processing for de-identification tasks. The LSTM-CRF model leverages LSTM's capacity to handle sequential data and CRF's feature constraint capabilities, enhancing the accuracy and efficiency of data processing, thus improving de-identification performance. By refining feature selection and reducing re-identification risks, the LSTM-CRF model contributes to more robust privacy protection.

D. Comparative Analysis of Deep Learning Models

Table 2 summarizes a comparative analysis of four deep learning models such as LSTM-CRF[16], FS-WOA-DNN[17], RNN[18], and LSTM[19] each applied to sensitive information classification

with distinct datasets, preprocessing methods, and evaluation metrics. Existing models often use datasets such as KDD CUP 99 and corpus datasets; however, study introduces a dataset collected based on sensitive and non-sensitive [19], encompassing sensitive and non-sensitive data across real-world, industry-based contexts. Preprocessing techniques vary, with LSTM-CRF employing word segmentation, FS-WOA-DNN using sentiment analysis, and this study's LSTM model enhancing sensitivity recognition through labelling and classification.

This study's LSTM model uniquely incorporates de-identification, a critical advancement over existing models that enhances privacy protection within sensitive data classification. While previous research applies LSTM models for sensitive data classification, the lack of integrated de-identification limits their applicability in privacy-centric contexts. By embedding de-identification processes, this study bridges the gap between data classification and privacy protection, underscoring the potential of LSTM as a scalable, privacy-preserving solution.

TABLE 2: DEEP LEARNING ALGORITHM COMPARISON

Deep Learning	LSTM-CRF [16]	FS-WOA-DNN [17]	RNN [18]	LSTM [19]
Dataset	Not mentioned	KDD CUP 99 [20]	Corpus Dataset [18]	Sensitive and Non-Sensitive dataset [19]

Al-Noor Journal for Information Technology and Cyber Security
December (2024); 1 (0): 1- 22
DOI: https://doi.org/10.69513/jncs.v1.i0.a1

Preprocessing Stage	Word Segmentation, Character Digitization, Vector Construction	Paraphrasing, Sentiment Analysis, Image Sentence Ranking	Normalization	Tag and label sentences, identify sensitive and insensitive information
Feature Sensitive Information	Not mentioned	Example: Prepay Transactions, Letters of Credit	Not mentioned	Example: Tender, Procurement, Audit
Classification in Deep Learning	RNN	RNN	DNN	LSTM and RNN
Detection Evaluation	Accuracy, Return Efficiency, F1 Index	Not mentioned	Accuracy, Specificity, Sensitivity, Error, False Positive Rate	Accuracy, Recall, Precision, F1 Score
Applied De-Identification	None	None	None	None

Methodology

This study implements LTSM approach to evaluating de-identification techniques aimed at protecting sensitive information. The methodology consists of five phases such as Dataset Acquisition, Data Preprocessing, Feature Extraction, Classification Algorithm Application, and Performance Evaluation. Each phase will contribute to the accuracy result that will be done at classification phase.

Figure 1 shows the methodology of this study investigates the framework for sensitive information detection using Long Short-Term Memory (LSTM) focusing on the combination of de-identification privacy concerns. There are 4 phases in LSTM such as Forget Gate, Remember Gate, Input Gate and Output Gate. The Forget Gate is responsible for eliminating data that is no longer

Al-Noor Journal for Information Technology and Cyber Security
December (2024); 1 (0): 1- 22
DOI: https://doi.org/10.69513/jncs.v1.i0.a1

relevant, helping to ensure that only significant information is maintained for processing.

Remember Gate: The Remember Gate stores and fill in data to create new features based on the information stored within the gate. This allows the model to keep necessary data while removing unrelated information.

Input Gate: The Input Gate manages the entry of new data into the system, ensuring that fresh information is properly integrated and utilized in the processing pipeline.

The classification of sensitive data has been conducted based on Figure 1. These results indicate that the classification LSTM achieves high accuracy using the sensitive words in Table 3. Consequently, this dataset will go through the de-identification process using a de-identification Java program. After de-identification, the dataset will be trained and tested with LSTM model to determine the accurateness of post-de-identification

The dataset used in this study includes both sensitive and non-sensitive information across diverse categories such as Resumes, Construction, and Medical fields, sourced from project documentation, medical records, and job resumes. This selection

Al-Noor Journal for Information Technology and Cyber Security
December (2024); 1 (0): 1- 22
DOI: https://doi.org/10.69513/jncs.v1.i0.a1

requires a comprehensive foundation for testing de-identification techniques on different types of sensitive information.

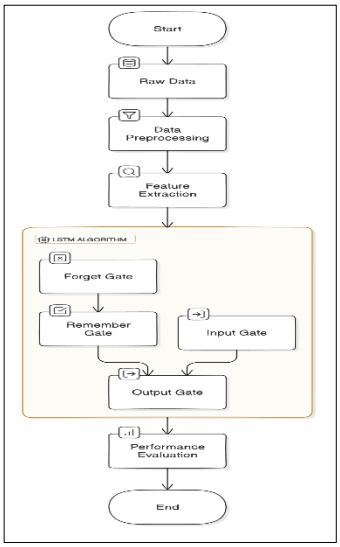

Figure 1: Framework For Sensitive Information Detection Using Long Short-Term Memory (Lstm).

A. Dataset Acquisition

Data acquisition complies with strict privacy principles, including the Personal Data Protection Act (PDPA) and the General Data Protection Regulation (GDPR), to ensure personal data is protected against data breach. After collection, the dataset undergoes a thorough preprocessing phase, where it is cleaned and organised into a structured format. Each entry is categorised into two columns such as one containing the sentence and the other indicating whether the information is classified as sensitive or non-sensitive.

To facilitate a robust evaluation, the dataset is ready in two versions. The first remains unaltered to provide a baseline measurement, while the second applies de-identification techniques to mask identifiable information, ensuring privacy. This dual-dataset approach allows for a direct comparison of model

Al-Noor Journal for Information Technology and Cyber Security
December (2024); 1 (0): 1- 22
DOI: https://doi.org/10.69513/jncs.v1.i0.a1

accuracy with and without de-identification, providing insights into the impact of privacy measures on classification performance.

B. Data Pre-Processing

Data preprocessing (Figure 2) is where the raw dataset is cleaned and transformed into structured data usable in an algorithm. In this research, data preprocessing has several stages which are Data Integration, Data Validation and Data Transformation. The dataset is pre-processed into two types of datasets. One will be ongoing to de-identification, which the sensitive data will be hidden. By applying the data cleaning methods, it will help to remove irrelevant or erroneous information. Use de-identification techniques to protect sensitive information during preprocessing. Implement text processing steps such as tokenization and normalization to facilitate feature extraction while ensuring that identifiable information is not exposed.

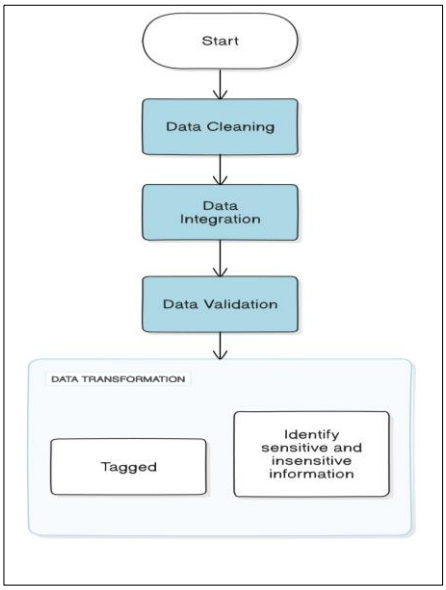

Figure 2: Data Preprocessing Flowchart

Al-Noor Journal for Information Technology and Cyber Security
December (2024); 1 (0): 1- 22
DOI: https://doi.org/10.69513/jncs.v1.i0.a1

In Data Integration is where all the datasets are combined and unified while, Data Validation is where the dataset is checked to ensure data is complete and accurate then, Data Transformation is to refine the dataset. The dataset is refined by labelling whether the sentence is sensitive or insensitive based on features of classification. So, when the features are found in the sentence, the data will be labelled as sensitive information. Hence cleaning the dataset requires a few runs through to ensure sentences are understood leading to a better dataset, as well as better outputs when classification of sensitive and insensitive information are made clearly.

C. Feature Extraction

Extract relevant features from the pre-processed data that are essential for identifying sensitive information while maintaining privacy. Employ feature extraction techniques, including Natural Language Processing (NLP) to identify key attributes related to sensitive information [20-21]. Ensure that feature extraction methods do not compromise the privacy of the individuals represented in the data. The sensitive word according to the dataset will be identified. After that, it will apply de-identification techniques. Table 3 below shows an example of the sensitive word in a dataset.

Table 3: EXAMPLE SENSITIVE WORD IN A DATASET [19]

Sensitive words	Description
Tenders	It is a formal documented offer involving money to a client. Most of the information contained in tender responses should be kept confidential
Procurement	It is a highly competitive list involving great care and attention before proceeding with the project that needs to specifically and efficiently ensure the confidentiality of these documents according to

Al-Noor Journal for Information Technology and Cyber Security
December (2024); 1 (0): 1- 22
DOI: https://doi.org/10.69513/jncs.v1.i0.a1

Responsible	The party responsible for creating a contract can detail any information they wish to make confidential
Cost	Cost is closely related with cash flow when involved with project documentation and contracts
Quotation	A formal statement of an estimated cost for a project that is agreed upon holding such information in strict confidence
Audit	Auditing must be conducted within a framework of complete trust and strictly confidential
Report	There are some reports that cannot be made public due to privacy of who may be involved
Contractual	Contractual confidentiality obligations are fundamental and necessary to help protect the parties that disclose information in these situations
Allocation	Allocation is usually done in consultation with the borrower, who are interested in relationship banks receiving the largest allocations

D. Classification

Implement and evaluate classification algorithms to detect and categorize sensitive and non-sensitive information effectively. Apply classification algorithms, including Long Short-Term Memory (LSTM) networks and other deep learning models, to classify data based on the extracted features. Ensure that the models are trained to recognize and protect sensitive information while minimizing the risk of de-identification. Employ techniques such as data masking or encryption where necessary to enhance privacy protection during model training.

E. Performance Evaluation

Performance evaluation will assess the efficiency of the de-identification techniques and classification algorithms in terms of accuracy or classification. The model LTSM will evaluate the performance using metrics such as Accuracy to determine the effectiveness of de-identification techniques. Additionally, this model will examine the accuracy, whether by implementing de-identification techniques to the datasets it can maintain privacy and

prevent data breach. By performing a comparative analysis of different datasets, it can identify the most effective strategies for privacy protection.

Result And Discussion

This study investigates the impact of de-identification on the accuracy of sensitive information classification using Long Short-Term Memory (LSTM) networks. Through a structured methodology, comprising Dataset Acquisition, Data Preprocessing, Feature Extraction, Classification, and Performance Evaluation, this research has evaluated the delicate balance between classification efficacy and data privacy. The findings, with specific emphasis on privacy enhancement through de-identification, are summarized below.

A. Dataset and De-identification Approach

The dataset contains various categories including Resume, Construction, and medical fields, each rich in sensitive and non-sensitive classes. This diversity offers a robust basis for assessing sensitive data classification across multiple domains, where keywords like “Tenders,” “Procurement,” and “Audit” signify potentially identifiable information. The de-identification of these elements serves as a key privacy-preserving measure, aimed at mitigating risks associated with the exposure of sensitive data. The first dataset will remain as it, meanwhile the other dataset will be de-identification. The sensitive keyword that has been identified based on literature review will be hidden by using de-identification algorithms Java Program.

B. Impact of De-identification on Classification Performance

Two datasets were employed: one containing visible sensitive keyword, and another where sensitive identifiers were masked. The de-identification led to a notable shift in classification accuracy,

showcasing both the strengths and limitations of privacy-enhanced models. As summarised in Tables 4 and 5, the LSTM network attained an accuracy of 97.14% on the unaltered dataset, a rate attributed to the availability of distinct sensitive identifiers. After examining the second dataset, the accuracy decreased to 78·30%. This result shows that de-identification techniques effectively reduce accuracy for second dataset that does not contain privacy keyword.

Table 4: Results Before De-Identification

Performance Metrics	Accuracy
LSTM	97.14%

Table 5: Results After De-Identification

Performance Metrics	Accuracy
LSTM	78.30%

C. Benefits of De-identification for Privacy Protection

The de-identification approach in this study confers several benefits such as Reduction in Re-identification Risk. By hiding certain elements, the model adheres to privacy standards, such as GDPR and PDPA, reducing the risk that individuals could be re-identified from the dataset.

De-identified data allows for secure sharing and collaboration without compromising individual privacy. This facilitates use in collaborative research, regulatory reporting, and external audits while maintaining confidentiality.

The inclusion of de-identification within the model ensures that privacy is preserved during classification, supporting ethical standards in machine learning by reducing exposure to sensitive details.

D. Practical Implications and Research Contribution

This study's integration of de-identification techniques into an LSTM-based classification framework is particularly valuable for industries requiring stringent data privacy, such as healthcare, finance, and government. The research methodology provides a replicable model for privacy-preserving machine learning, one that upholds classification accuracy while aligning with privacy protection principles. By addressing privacy risks at the model level, this study offers a practical framework that mitigates privacy concerns and broadens the applicability of AI-driven data processing within regulated sectors.

Conclusion

This study has presented a comprehensive evaluation of de-identification techniques within a deep learning framework for sensitive data classification, emphasising the balance between data utility and privacy. By employing a five-phase methodology such as Dataset Acquisition, Data Preprocessing, Feature Extraction, Classification, and Performance Evaluation, this research systematically assessed the impact of de-identification on classification accuracy using Long Short-Term Memory (LSTM) networks. The approach was tested on a dataset, covering sectors such as Resume, Construction, and Medical, and highlighted the model's performance in identifying sensitive versus non-sensitive information across these domains.

The findings indicate that LSTM achieves high accuracy (97.14%) in classifying sensitive data when de-identification is not applied, underscoring its efficacy in recognising distinctive, identifiable information. However, following de-identification, accuracy reduced to 78.30%, illustrating the trade-off inherent in privacy protection. This reduction confirms that while de-identification

protects individual privacy by hiding sensitive identifiers, it also slightly reduces classification accuracy.

The study's methodology and results underscore the benefits of de-identification for privacy compliance, especially in regulated industries such as healthcare and finance. De-identification minimises risks and enhances the security of data sharing, aligning the framework with privacy laws like GDPR and PDPA. These contributions make this research model an adaptable solution for organisations requiring privacy-preserving data processing.

In summary, this study validates the practical value of combining de-identification within deep learning models. By achieving a significant result between data accuracy and privacy protection, this research contributes to the future studies of privacy-preserving techniques in sensitive data classification.

Acknowledgment

This work paper was sponsored by CyberSecurity Malaysia an agency under Ministry of Digital through project on Data Privacy Management and Universiti Tun Hussein Onn Malaysia (UTHM) through Tier1 (Vot Q508).

References

[1] Majeed and S. O. Hwang, "When AI Meets Information Privacy: The Adversarial Role of AI in *Data Sharing Scenario*," *IEEE Access*, vol. 11, pp. 76177–76195, 2023, doi: 10.1109/ACCESS.2023.3297646.

[2] J. Pool, S. Akhlaghpour, F. Fatehi, and A. Burton-Jones, "A Systematic Analysis of Failures In *Protecting Personal Health Data: A Scoping Review*," *Int J Inf Manage*, vol. 74, p. 102719, 2024, doi: https://doi.org/10.1016/j.ijinfomgt.2023.102719.

[3] M. M. Silveira *et al.*, "Data Protection based on Searchable Encryption and Anonymization Techniques," in *NOMS 2023-2023 IEEE/IFIP Network Operations and Management*

Symposium, 2023, pp. 1–5. doi: 10.1109/NOMS56928.2023.10154280.

[4] E. M. Weitzenboeck, P. Lison, M. Cyndecka, and M. Langford, "The GDPR and Unstructured Data: Is Anonymization Possible?," *International Data Privacy Law*, vol. 12, no. 3, pp. 184–206, Aug. 2022, doi: 10.1093/idpl/ipac008.

[5] V. Software, "Cybersecurity: CIA Triad Explained," 2024. Accessed: Oct. 23, 2024. [Online]. Available: https://www.veeam.com/blog/cybersecurity-cia-triad-explained.html

[6] P. H. R. Emerick, S. C. Sampaio, B. L. Dalmazo, A. Riker, A. V. Neto, and R. Immich, "Enhancing Privacy in Healthcare: A Multilevel Approach to (Pseudo)Anonymization," in *2024 International Wireless Communications and Mobile Computing (IWCMC)*, 2024, pp. 1814–1819. doi: 10.1109/IWCMC61514.2024.10592397.

[7] E. Raso, P. Loreti, M. Ravaziol, and L. Bracciale, "Anonymization and Pseudonymization of FHIR Resources for Secondary Use of Healthcare Data," *IEEE Access*, vol. 12, pp. 44929–44939, 2024, doi: 10.1109/ACCESS.2024.3381034.

[8] M. Fotache, A. Munteanu, C. Strîmbei, and I. Hrubaru, "Framework for the Assessment of Data Masking Performance Penalties in SQL Database Servers. Case Study: Oracle," *IEEE Access*, vol. 11, pp. 18520–18541, 2023, doi: 10.1109/ACCESS.2023.3247486.

[9] N. I. of Standards and T. (NIST), "Security and Privacy Controls for Information Systems and Organizations (NIST SP 800-53 Revision 5)," 2020. doi: 10.6028/NIST.CSWP.04162018.

[10] A. Holzinger, I. Fister, I. Fister, H.-P. Kaul, and S. Asseng, "Human-Centered AI in Smart Farming: Toward Agriculture 5.0," *IEEE Access*, vol. 12, pp. 62199–62214, 2024, doi: 10.1109/ACCESS.2024.3395532.

[11] W. Li, Y. Chen, H. Hu, and C. Tang, "Using Granule to Search Privacy Preserving Voice in Home IoT Systems," *IEEE Access*, vol. 8, pp. 31957–31969, 2020, doi: 10.1109/ACCESS.2020.2972975.

[12] H. F. Nweke, Y. W. Teh, M. A. Al-garadi, and U. R. Alo, "Deep Learning Algorithms For Human Activity Recognition Using Mobile And Wearable Sensor Networks: State Of The Art And Research Challenges," *Expert Syst Appl*, vol. 105, pp. 233–261, 2018, doi: https://doi.org/10.1016/j.eswa.2018.03.056.

[13] G. Naveen Kumar and J. Anuradha, "Data Masking Techniques for Patient Privacy in Healthcare Systems: A Review," *J Med Syst*, vol. 45, no. 11, pp. 1–12, 2021, doi: 10.1007/s10916-021-01744-9.

[14] D. Kavitha, "Preserving Privacy of IoT Healthcare Data using Differential Privacy and LSTM," *Journal of Electrical Systems*, vol. 20, pp. 2483–2492, Sep. 2024, doi: 10.52783/jes.4071.

[15] V. Yogarajan, B. Pfahringer, and M. Mayo, "A review of Automatic end-to-end De-Identification: Is High Accuracy the Only Metric?," *Applied Artificial Intelligence*, vol. 34, no. 3, pp. 251–269, Feb. 2020, doi: 10.1080/08839514.2020.1718343.

[16] L. Wu and M. Pan, "English Grammar Detection Based on LSTM-CRF Machine Learning Model," *Comput Intell Neurosci*, vol. 2021, pp. 1–10, Aug. 2021, doi: 10.1155/2021/8545686.

[17] A. Agarwal, M. Khari, and R. Singh, "Detection of DDOS Attack using Deep Learning Model in Cloud Storage Application," *Wirel Pers Commun*, vol. 127, Mar. 2021, doi: 10.1007/s11277-021-08271-z.

[18] A. Kumar and N. Gupta, "Enhancing Data Security in Cloud Storage Using Advanced Encryption Techniques," *Journal of Engineering Science and Technology*, vol. 15, no. 5, pp. 27–40, 2020.

[19] N. I. M. Roslan and C. F. M. Foozy, "Penerbit UTHM," *Journal of Science and Computing Data Management*, 2024, Accessed: Oct. 24, 2024. [Online]. Available: https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/12805

[20] S. Choudhary and N. Kesswani, "Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 Datasets using Deep Learning in IoT," *Procedia Comput Sci*, vol. 167, pp. 1561–1573, 2020, doi: https://doi.org/10.1016/j.procs.2020.03.367.