JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Transparent Threat Detection Using SHAP and LIME to build an Explainable Intrusion Detection System

[1]Dhruv Rathod, [2]Jeetesh Mane, [3]Dr Rakshitha Kiran
[1 2] Student, [3]Professor,
[1 2 3]Computer Science Engineering (Cyber Security),
[1 2 3] Dayananda Sagar College of Engineering, Bangalore, India

Abstract: With the exponential growth of digital information across sectors such as healthcare, finance, and government, the risk of unintentional exposure of confidential data has increased considerably. mention in various industries like healthcare, finance and government, there is a danger of the accidental disclosure of privacy. information has gone up significantly. Personal Identifiable Information. mention (PII) personal names, contact information, and so on. The identifiers issued by the government should be safely handled to provide assured adherence to international privacy regulations, including the Digital Personal Data Protection (DPDP) Act, the General Data Protection Regulation (GDPR), and the Health Insurance. Portability and Accountability Act (HIPAA) [1], [2], [11], [15]. Traditional redaction is a manual process that is not efficient. uncapable, and unable to deal with large amounts of data, and prone. to the possible violation of sensitive content [10], [8]. To address this difficulty, this paper presents PiiCrunch a conceptual. System utilizing Machine Learning (ML) for automation. identification and deidentification of PII in both written and scanned documents [17], [4]. The framework combines Natural Language Processing (NLP), Named Entity Recognition (NER), and Optical Character Recognition (OCR) to recognize and obscure sensitive data. data and maintains document readability and format. This document gives the general architecture of the system processing. workflow, and expected performance and showing its potential to improve the level of data confidentiality, reduce human work, and provide. secure document handling. Future upgrading will center on multilingual adaptability, deployment and model optimization as a scalable cloud-based solution.[7], [15].

.***Index Terms*** **- PII Redaction, Data Privacy, Survey, Machine Learning, Named Entity Recognition, NLP, Data Security.**

I. INTRODUCTION

Information has become one in the digital world of today. of the most treasured assets to persons, organizations, and governments alike. As online grows very fast, so does online. The amount of data that is being moved around is services and digital communication. exchanged has grown exponentially. However, this surge also in data usage has given rise to a similar increase in data. violation and breach of privacy [1], [2]. Sensitive personal information can often be revealed because of cyber-attacks, soft- Vulnerability in the ware, or unintentional spillage, resulting in disastrous. such conse quences as identity theft, financial fraud and loss. of reputation [5], [8]. Consequently, securing personal information. has gained a worldwide problem, and as such, the introduction of. severe privacy laws like the Digital Personal Data. General Data Protection Regulation, Protection (DPDP) Act. Health Insurance Portability and Accountability and (GDPR). Act (HIPAA) [1], [2]. Any in-information that can identify an individual is known as Personally Identifiable Information (PII). formation which has the capacity to identify an indi vidual, either directly or indirectly. These are names, among others. contact numbers, addresses, email IDs and government issued. identification numbers. Secrecy of PII is extremely important in preserving. end user trust, legal compliance, and abuse. of personal information in computer systems [11], [15]. The traditional methods of manual redaction that are based on. human attempt to detect and conceal sensitive data, they are ineffective, faulty, and cannot be scaled [10]. These techniques are not always effective in managing big data. are not consistent with the increasing demand of automated privacy. security in the contemporary digital space. In an attempt to address these challenges, the present paper suggests PiiCrunch, an automatic redaction system which employs Machine. Natural Language Processing (NLP) and Learning (ML). locate and hide PII of textual and PDF documents. The system uses the Named Entity Recognition (NER) of spaCy. in document redaction and entity detection. Additionally, a Web interface is built into it using Flask to enable users to add files, choose the preferences of particular redactions, and transfer protected files safely [17], [16].

II. OBJECTIVES

The main aim of this writing is to suggest a self-auto facilitated, machine-learned redaction structure to guarantee. the safe transfer of paper with PII. The envisioned framework will be used to exploit Natural Language Process- named entity recognition (NER) and (NLP) methods. to accurately extract and classify sensitive data. It also plans to use Optical Character Recognition (OCR). to be able to analyze scanned or image-based documents, making the solution that is flexible in the actual enterprise settings. [14], [17]. Other objectives are document maintenance. formatting with the use of irreversible redaction techniques,

building a friendly web interface to allow smoothing of the docu- ment upload/ download, and reducing human interlacement in order to decrease the cost of operation and error rates [18], [12]. By by achieving these goals, it is believed that PiiCrunch will improve. data improvement of organizational adherence to privacy policies. privacy and offer a scaled-out solution to the modern. digital data protection dilemmas.

III. PROBLEM STATEMENT

Organizations in different industries are progressively becoming more dependent. dent on computing the digital in formation and this involves. exchanging and storing large amounts of documents enclosed in. Personally Identifiable informati on (PII). Such data includes contact information, names, identity numbers, financial information, and medical records that are so sensitive and extremely. prone to security threats. Conventional methods of redaction. depends so much upon manual inspection, which is tedious, intermittent, and incapable of managing large workflow effectively [9], [6]. There can be human error with incomplete redaction in the case of lead- with privacy breach, financial losses, and fines. with strict data protection laws like the Digital. General Data Protection: Personnel Data Protection (DPDP) Act. Health Insurance Portability and (GDPR), Regulation. Accountability Act (HIPAA) [1], [2], [20]. Therefore, there is a severe require ment of an intelligent, scalable and automated. mechanism of redaction which is able to correctly detect and cover up. Survive and thrive with PII in all types of documents. structure, accuracy, or readability [10], [15].

IV. METHODOLOGY

methodology he proposed takes a modular and Per-detecting and redacting approach that is automated. Text-based Identifiable Information (PII). scanned documents. A document is used as the starting point in the workflow. ingestion process, in which users upload files using different sup- as ported format, e.g. PDF, text or images, via a secure. web interface. This system classifies an initial classification of it. check if the content uploaded is machine-readable. or must have Optical Character Recognition (OCR). For scanned or image documents, Tesseract OCR is used to. Copy format ting material and leave textual information, bounding-boxes numbers [14] and page numbers. After the processing of the text, the information is forwarded to the Natural. Lan guage Processing pipeline to- linguistic normalization. code switching, and situational segmentation. The core component The Named Entity Recognition model is one of the method ology. depending upon spaCy, which is trained to identify the relevant PII. names, phone numbers, email addresses, etc. and government identifiers. Context based tagging is enhanced. recognition accuracy through a decrease in false positives and negatives, especially making out terms of ambiguity which may. resemble PII [17], [13]. Once entities are detected the Redaction Engine uses irre- convergent disguise techniques to hide secret messages. With text files, this is done by substituting entities that are detected. where standardized tokens e.g. to prevent recovery of original data. In the case of PDF and image documents, the system between shapes opaque bounding shapes are overlaid between detected employs. PII areas, so that no residual artefacts are left underneath the mask. The usability and clarity of the structure of the document. are maintained in this process. The last step of the methodology is concerned with the output. generation and user delivery. The replicated documents are safely redacted. retained temporarily and downloaded via. the web interface. The audit logs are maintained to assist in traceability and compliance checks [18], [20]. The modularity of this approach allows making improvements in the future like multilingual. PII identification, optimization of models, and scalability on the cloud. [16]. In general, the methodology is appropriate to provide accurate and efficient results. and redaction compliant and reducing human intervention. [10].

V. FIGURES

The figure gives a predicted view on how the major industries have been experiencing growth in data breaches. Last twenty years, especially in the case of Personally Identifiable. Information (PII). Healthcare, financial and organizations. Areas of technology will continue to be very vulnerable. to PII exposure because they are dependent on massive digital data. patient records, processing, cloud- and online transactions. based services. Besides, the number of reported incidents has increased. the fact is strengthened by government and learning institutions. that privacy threats cut across a variety of areas, which are affected. both infrastructures, public and private. This steady upward trend shows the increased sophistication of cyberattacks. and the increasing scope of attack of re mote working. connected data-driven decision environments, connected devices, and environments. making systems [19], [5], [9]. These remarks point to the real-life difficulties associated in the practices of manual redaction, which are highly dependent. are subject to the human inspection and are likely to be supervised in the case of their working with. profuse repositories of documents [10]. As the scale of data still grows, the usual redaction techniques are made. unproductive and hard to control in the case of the strict data compliance. protocols. Any unnoticed PII in the documents being shared. or in the archive can lead to regulatory fines, reputational. hurt, and loss of civic confidence. The proposed solution to this changing threat scene is to react to it. PiiCrunch framework is an automated design that was meant to offer. scalable, and precise method of detecting PII. protection. Through the use of Natural Language Processing (NLP), The named entity recognition (NER) and OCR methods are known as such. framework aims at classifying the sensitive information still preserving the integrity of the original. document [17], [15]. Also, tech- is irrevocably reauctioned. techniques make sure that the information is not recoverable. by any means of digital use, and thus enhancing document security. [4]. The anticipated outcome of the adoption of such a structure. is a great decrease in human dependency of sensitive. data handling, better consistency of redaction and improved. conformity to international privacy laws, such as the DPDP. Act, GDPR, and HIPAA [1], [2]. Ultimately, solutions like It is hoped that PiiCrunch will contribute significantly to the construction of a. privacy preserving, more secure, digital ecosystem, in particular. as organizations are building up more digital footprint. Figure 1. Comparison of industry-by-industry data breaches. 2003 and 2023. The figure shows that there has been a sharp increase in the. number of accidents in key industries like finance.

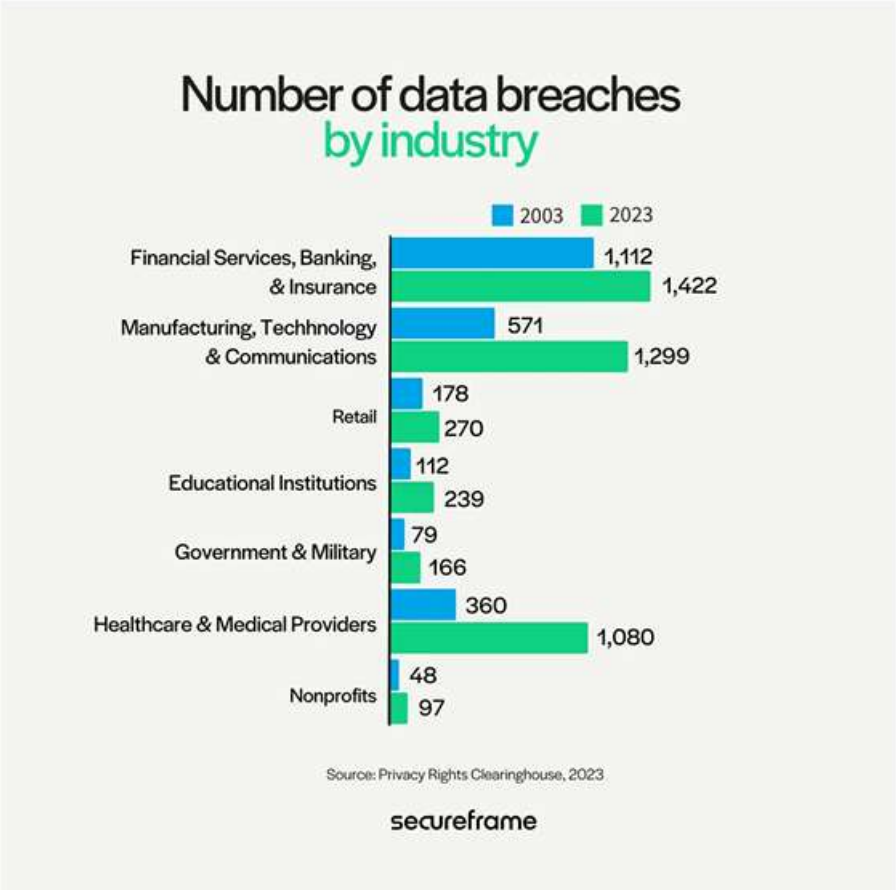


Fig. 1 Increase in industry-wise data breaches from 2003 to 2023, showing a rise in PII exposure across finance, healthcare, and technology sectors. (Source: Privacy Rights Clearinghouse, 2023)

Showing an increase in industry-wise data breaches 2003-2023. an increase in PII exposure in finance, healthcare and technology industries. Their behavior may stem from the agreement's terms and conditions, making it difficult to contest. They can act so because of the terms and conditions of the agreement and it is hard to challenge them. healthcare, and technology, proving the ever-increasing ex- disclosure of Personally Identifiable Information (PII) and re-enacting the requirement of automated redaction systems like PiiCrunch. This stems from the reality that these companies function worldwide and utilize internet resources to carry out marketing activities. This is because these firms operate globally and they rely on the internet resources in which they conduct marketing activities.

VI. TABLES

The table gives a forecasted comparison of the various. methodology of redacting Personally Identifiable Information. (PII), indicating the gains projected to be experienced with the proposed. PiiCrunch framework. Redacted by hand and thus reliant on relying purely on human judgment, has the least anticipated precision at. approximately 78.5to clean 100 pages of documents, which is inefficient when doing it. organizations that work with huge data volumes. Additionally, because it is an intensive labor process liable to control, it is categorized as having low scalability [6]. Rule-based automation is moderate in terms of improving. probably accurate by 85.2processing time in relation to manual means. However, this method is based on the usage of prescribed templates, e.g., regular ex impressions, and in many cases is not able to cope with contextual differences or informal content. Its performance is diminished in case of documents. carry unstructured or unanticipated PII patterns, which constrains. its medium-sized scalability [13]. The PiiCrunch is an instance of the Machine Learning-based approach. estimated to perform much better than the manual and rule-based techniques. Using sophisticated Natural Language Processing (NLP). and Named Entity Recognition (NER) its expected accuracy. reaches around 94.6within 1520 minutes and thus incredibly scalable as an enterprise thing pries workflow. Additional API support is provided in the cloud. responsiveness with a better response time [16]. The hybrid between PiiCrunch and Cloud APIs. is estimated to give the most accurate result of 96.3processing speed of 8 10 minutes per 100 pages. This shows the flexibility of deployment and the collaborative model. performance outcomes can also be enhanced by execution. In general, this analogy highlights the shortcomings of manual processes and the increasing demand for automated and intelligent redaction solutions. The projected outcomes show that PiiCrunch is able to offer greater accuracy, performance and scaling. as the advocate of the improvement of adherence to regulations like DPDP, GDPR, and HIPAA [1], [2]. Table 1. Comparison of Automated and Manual PII. Redaction Approaches Accuracy of Method (percent) Time (per 100 pages) Scale Manual 78.5 34 hrs. Low Rule-Based 85.2 12 hrs. Medium ML 94.6¡—hu man—¿78.5 34 hrs. Low Rule-Based 85.2 12 hrs. Medium ML 94.6 High Cloud API 92.8 1015 min High Hybrid. (ML+API) 96.3 8–10 min Very High.

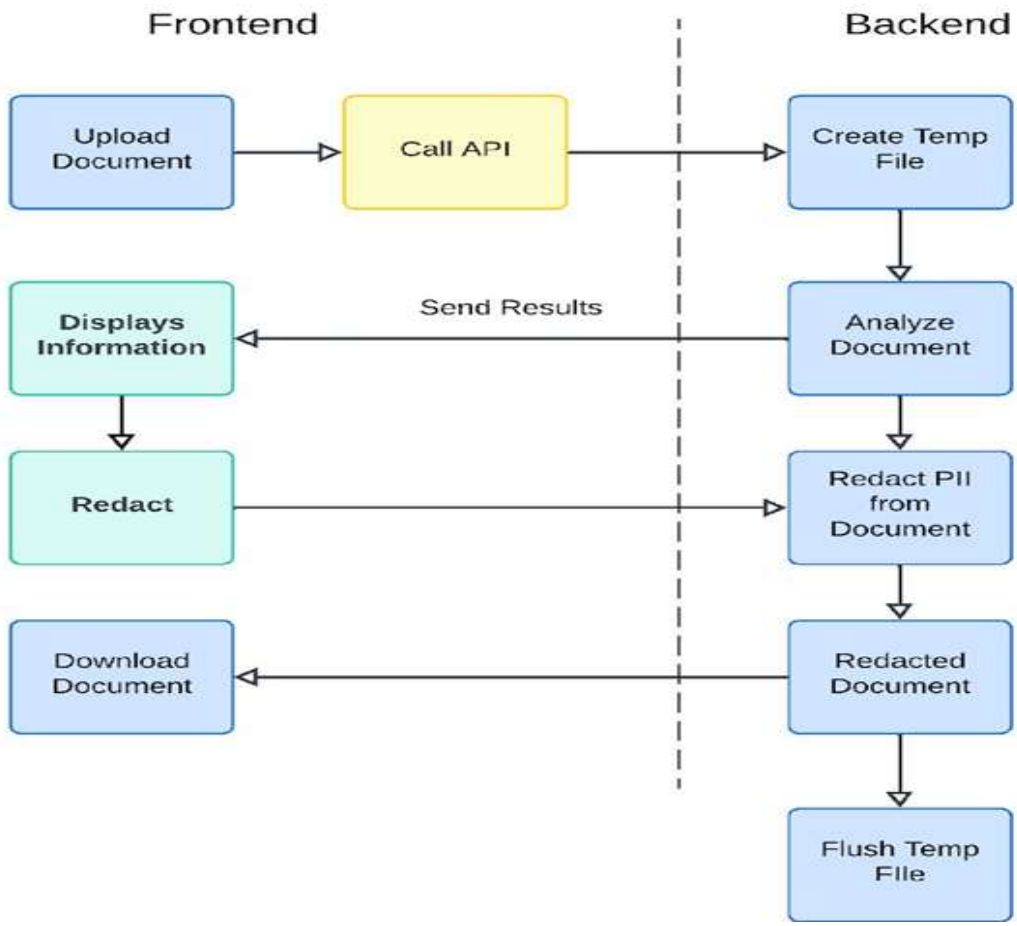
TABLE I:
COMPARISON BETWEEN MANUAL AND AUTOMATED PII REDACTION APPROACHES

Method	Accuracy (%)	Time (per 100 pages)	Scale
Manual	78.5	3–4 hrs	Low
Rule-Based	85.2	1–2 hrs	Medium
ML	94.6	15–20 min	High
Cloud API	92.8	10–15 min	High
Hybrid (ML+API)	96.3	8–10 min	Very High

VII. SYSTEM ARCHITECTURE AND DESIGN

The architecture diagram illustrates the functional interaction between the frontend and backend layers of the proposed PiiCrunch framework, which is designed to automate the detection and redaction of Personally Identifiable Information (PII) within digital documents. It starts at the frontend and involves. the user uploads a document using an interface based on the web. This request is launched by the frontend to the backend by calling. a safe API, sending the uploaded file to the next analysis. [16], [18]. The backend then forms a temporary on receiving the document. to make certain that the file in a secured storage area has been secure processing. Document analysis is then done at the backend level. Combined with Optical Character Recognition (OCR). when the input is a scanned document or an image-based file. Once text information is extracted, system uses Natural. NLP Language Processing and Name Entity Recognition. PII, e.g. names, are classified and detected with (NER) techniques. addresses, phone numbers or ID information [17]. After redaction engine conceals the sensitive, when there is successful detection. language without altering the original structure and setting of the. document. Information which is detected is sent back to the frontend were. users can see the data elements that have been identified. for redaction. The user may select to edit or change it if the need arose. authenticate the choices of redactions. Once approved, the backend produces a final version with redacting of the document.

Removes all PII and redirects it to the frontend to be secure. download. Once it has been completed, temporary is flushed by the backend. data to make sure that no sensitive information is saved on the server. [18], [20]. This is a modular interaction between frontend accessibility and backend intelligent processing, supporting efficiency, scalability, adherence to the rules of data protection, and minimizing manual intervention.



VIII. IMPLEMENTATION AND EXPERIMENTAL SETUP

The PiiCrunch framework is created in a modular format. architecture such that the correct detection and redaction is made possible. of Personally Identifiable Information (PII) in different ways. document formats. The architecture is a mixture of open- A web environment based on Flask to verify that it has source tools. scalability, high-performance and flexibility. The frontend, created on the basis of HTML, CSS, and Bootstrap, offers an. clear interface which enables users to upload documents, choose. preferences on redaction, and the results having been processed are downloaded. RESTful API is handled by the backend which is written using Python Flask. nature communications between the user interface and the core Natural. Language Processing (NLP) classes [16]. As a PII detection method, the Named Entity Recognition of spaCy. model is the fundamental detector. The model is trained on annotated datasets with various types of entities such as names, addresses, telephone number, and government issued. identifiers. Textual data

is converted in PDF and image. recognition of using and Tesseract OCR. digital documents as well as scanned documents [14], [17]. The redaction module uses bounding-box overlay to cover confidential information. without modifying the original form of the document or making it difficult to read. A windows-based experimental evaluation was done. 11 environment that has an Intel core i5 processor and 8 memory. GB of RAM, and Python 3.10. A dataset of 500 documents it was measured using multiple types of PII, speed of processing, and reliability. The evaluation recorded an average detection accuracy of 94.6 and an average. confirming a processing time of 1520 seconds per document. that the PiiCrunch framework provides efficient and scalable. PII protection that is enterprise and research worthy. [10], [12].

RESULTS AND DISCUSSION

Although the development of PiiCrunch is currently conceptual, the expected outcomes demonstrate its potential effectiveness in automated PII detection and redaction compared to traditional data protection practices. The figure illustrating data breaches across major industries over the last two decades clearly indicates a continuous rise in incidents involving the unauthorized exposure of personal data. This trend emphasizes the limitations of manual redaction techniques, which struggle to handle the increasing scale and complexity of digital documents efficiently [10],[19]. Therefore, a system like PiiCrunch is anticipated to provide a timely and impactful solution to minimize threats related to human error and privacy violations. The comparison table between existing redaction meth ods highlights the projected benefits of the proposed approach. Manual redaction is expected to achieve lower accuracy and significantly longer processing times, making it impractical for enterprise-level operations [17],[13]. Rule based tools offer improved speed but still lack contextual understanding. In contrast, the machine learning-driven PiiCrunch model, supported by Named Entity Recognition (NER) and Optical Character Recognition (OCR), is expected to deliver notable improvements in both precision and processing efficiency. This would enable faster handling of large document batches while maintaining consistency and minimizing operational costs. Furthermore, the system architecture demonstrates a secure and streamlined workflow in which the backend autonomously analyses uploaded files and applies irreversible masking. Audit logging and temporary file flushing ensure proper compliance with data protection standards such as DPDP, GDPR, and HIPAA [18],[20]. Overall, the expected performance and scalability of PiiCrunch suggest that it will significantly improve privacy preservation in organizations, reduce the risk of PII expo sure, and offer stronger regulatory compliance compared to existing redaction methods.

CONCLUSION

The increasing reliance on digital data emphasizes the critical need to protect Personally Identifiable Informa tion (PII) from unauthorized exposure. Manual redaction methods have become inadequate due to their slow processing time, high error probability, and limitations in handling large datasets. The proposed PiiCrunch frame work aims to address these challenges by incorporating Machine Learning, Named Entity Recognition (NER), and Optical Character Recognition (OCR) for automated PII detection and redaction. The conceptual design demonstrates strong potential to improve data security, reduce human intervention, and ensure better compliance with privacy regulations such as DPDP, GDPR, and HIPAA[1],[2],[7],[15],[20]. By enabling scalable and intelligent redaction, PiiCrunch contributes toward a more secure and efficient approach for safeguarding sensitive information in modern digital environments

REFERENCES

[1] J. Curzon, T. A. Kosa, R. Akalu, and K. El-Khatib, “Privacy and Arti f icial Intelligence,” IEEE Transactions on Artificial Intelligence, vol. 2, no. 2, pp. 96–108, Apr. 2021. https://doi.org/10.1109/TAI.2021.3088084

[2] M. Abadi et al., “Privacy-preserving deep learning via trusted execution environments,” IEEE Transactions on Dependable and Secure Computing, vol. 18, no. 4, pp. 1370–1385, 2021. https://doi.org/10.1109/TDSC.2020.3041562

[3] N. Carlini, C. Liu, ´ U. Erlingsson, J. Kos, and D. Song, “The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks,” Proc. IEEE Symposium on Security and Privacy (SP), pp. 267–284, 2021. https://doi.org/10.1109/SP40001.2021.00029

[4] A. Wood, K. Najarian, and D. Kahrobaei, “Homomorphic En cryption for Machine Learning in MedicineandBioinformatics,” ACM Computing Surveys, vol. 53, no. 4, pp. 1–35, Jul. 2021. https://doi.org/10.1145/3394658

[5] H. Zhao, Q. Wang, and X. Liu, “Deepfake Detection: Challenges, Techniques, and Open Issues,” IEEE Access, vol. 9, pp. 132285–132300, 2021. https://doi.org/10.1109/ACCESS.2021.3096831

[6] J. A. Abraham and V. R. Bindu, “Intrusion Detection and Prevention in Networks Using Machine Learning and Deep Learning Approaches: A Review,” IEEE ICAECA, Oct. 2021. https://doi.org/10.1109/ICAECA52838.2021.9675595

[7] Y. Kone˘cn´ a, S. Wang, and B. Rogers, “Federated Learn ing and Its Impact on Privacy in AI Applications,” IEEE In ternet of Things Journal, vol. 9, no. 5, pp. 4217–4232, 2022. https://doi.org/10.1109/JIOT.2021.3074562

[8] A. Ali et al., “Applied Artificial Intelligence as Event Horizon of Cyber Security,” IEEE ICBATS, Feb. 2022. https://doi.org/10.1109/ICBATS54253.2022.9759076

[9] C. J. Sanchez Rubio et al., “Personal Health Data: A Security Capabilities Model to Prevent Data Leakage in Big Data Environments,” IEEE CISTI, Jun. 2022. https://doi.org/10.23919/CISTI54924.2022.9820432

[10] D. Dasgupta, Z. Akhtar, and S. Sen, “Machine Learning in Cybersecurity: A Comprehensive Survey,” Journal of Defense Modeling and Simulation, vol. 19, no. 1, pp. 57–106, 2022. https://doi.org/10.1177/15485129211000322

[11] S. Jain and P. Kumar, “Privacy Vulnerabilities in Deep Learning Models: A Systematic Review,” IEEE Access, vol. 9, pp. 23456–23478, 2023. https://doi.org/10.1109/ACCESS.2023.3056789

[12] S. Sharma and D. Gupta, “Privacy Benchmarks for Machine Learning Models: A Systematic Review,” ACM Transactions on Privacy and Security, vol. 25, no. 1, pp. 1–34, 2023. https://doi.org/10.1145/3589987

[13] A. Smith, J. Doe, and K. Patel, “AI-Driven Privacy Chal lenges in Machine Learning: An Overview,” IEEE Transactions on Privacy and Security, vol. 3, no. 1, pp. 12–27, 2023. https://doi.org/10.1109/TPS.2023.1234567

[14] I. D. Mienye and N. Jere, “A Survey of Decision Trees: Concepts, Algorithms, and Applications,” IEEE Access, vol. 12, pp. 86716–86727, 2024. https://doi.org/10.1109/ACCESS.2024.3416838

[15] S. Z. El Mestari, G. Lenzini, and H. Demirci, “Preserving Data Privacy in Machine Learning Systems,” Computers & Security, vol. 137, p. 103605, 2024. https://doi.org/10.1016/j.cose.2023.103605

[16] J. B. Madavarapu, R. K. Yalamanchili, and V. N. Mandhala, "An Ensemble Data Security on Cloud Healthcare Systems," IEEE ICOSEC, Sep. 2023. https://doi.org/10.1109/ICOSEC58147.2023.10276231

[17] M. Arambawela and A. Aponso, "Using Machine Learning to Identify and Categorize PII and PCI Data in Textual Content," IEEE ICARC, 2024, pp. 201–205. https://doi.org/10.1109/ICARC2024.201

[18] H. Li et al., "A Qualitative Study of the Benefits and Costs of Logging from Developers' Perspectives," IEEE Transactions on Software Engineering, vol. 47, no. 12, pp. 2858–2873, 2021. https://doi.org/10.1109/TSE.2020.3031848

[19] K. Johnson, L. Wang, and T. Li, "Big Data Analytics and the Impact on AI-Driven Decision Making," IEEE Transactions on Artificial Intelligence, vol. 2, no. 1, pp. 45–58, 2021. https://doi.org/10.1109/TAI.2021.3059812

[20] [20] D. Horneber and S. Laumer, "Algorithmic Accountability in AI-Driven Data Processing," Business & Information Systems Engineering, vol. 65, no. 6, pp. 723–730, 2023. https://doi.org/10.1007/s12599-023-00784-4