International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04159
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Protecting Sensitive Data on USB Drives: An AI-Driven Solution for Malware Detection and Data Privacy

Dr. T. AMALRAJ VICTOIRE [1], M. VASUKI [2], J. ANANDHARAJ [3]

[1] Professor, Department of MCA, Sri Manakula Vinayagar Engineering College, Puducherry-605107, India.

[2] Associate Professor, Department of MCA, Sri Manakula Vinayagar Engineering College, Puducherry-605107, India.

[3]PG Student, Department of MCA, Sri Manakula Vinayagar Engineering College, Puducherry-605107 India.

amalrajvictoire@gmail.com[1], dheshna@gmail.com[2], anandhraj343@gmail.com[3]

Abstract: USB drives, external hard drives and memory cards are commonly used for transferring and storing data. These devices are highly portable, easy to use, and can be used virtually anywhere that data is stored. One of the most common modes of covert data theft is by using malware injection attacks, which is when malicious code is injected into a system to secretly collect and/or steal sensitive data. Since these attacks are quite sneaky and can bypass conventional security measures, they can be exploited for unauthorized access of sensitive data. Existing systems tend to focus on either detecting malware or data backup separately. This lack of comprehensive approach does not allow sensitive data to be protected from such attacks. To bridge the gap, this paper presents an integrated security solution that employs deep neural networks (DNN) to detect a malware injection attack, CloudConceal, a secure backup or recovery system, and Data Masking using Tokenization to obfuscate sensitive data stored on USB drives. The DNN model discovers suspicious system activity such as file access and creation of process by analyzing features such as API calls, byte sequences, and metadata of system logs. After finding a malware attack, sensitive data is automatically transported to CloudConceal to create encrypted copies of the data and to recover in case of data loss or compromise of any kind. Besides this, the proposed system obfuscates sensitive information on USB drives in a way that replaces original data with tokens. In return, unauthorized users cannot reach the actual content. This integrated model offers a framework to protect sensitive data from covert data theft, strengthen sensitive data security, and safeguard its availability and integrity.

Keywords: USB security, covert data theft, malware injection attacks, Deep Neural Networks (DNN), AI-driven malware detection, CloudConceal, secure backup and recovery, data masking, tokenization, portable storage protection, data privacy, system activity monitoring, encrypted data, sensitive data protection, cybersecurity.

1.INTRODUCTION

Data storage has never been easier, thanks to the large pool of portable storage devices such as external hard drives, memory cards, and USB drives. But there is actually security risks attached to some of these conveniences, and the main concern seems to be an increasingly common form of data theft. Hackers can manipulate USB drives to inject malicious code into host computers, allowing malicious parties to conceal critical information while evading traditional security controls such as firewalls and anti-virus software. However, in addition to the convenience, USB also poses significant security risks. Most severe concern is the stealthy data theft, when hackers exploit USB devices to inject malware into host systems and steal sensitive data, unaware of their existence. Such malware injection attacks are highly evasive and difficult to detect. They are also prone to unattended attack vectors such as antivirus software and firewalls.

Moreover, the risks to USB drives are compounded by their widespread use across personal, academic, and corporate applications. Even if existing cybersecurity solutions focus individually on malware detection and secure data backup, they often don't help offer comprehensive, proactive protection for sensitive data stored or moved via USB devices. To bridge this gap, the proposed project proposes an AI-driven integrated security solution that serves not only as a detection tool but also as a way to prevent data theft: The system uses Deep Neural Networks (DNNs) to identify malware injection attacks by analyzing behavioural features (e. g. API call sequences and system log patterns) and then proceeds to do encrypted backups of sensitive files using CloudConceal in parallel. At the same time, the Local Data Masking is implemented (tokenization) to hide the local data on the USB drive that enables unauthorized use of it.

The project is designed to fulfilled three major security objectives: protecting the confidentiality, integrity, and availability of sensitive data; detecting malicious activity and responding quickly; safeguarding business continuity and data resilience via real-time backup and intelligent obfuscation techniques; using AI techniques, cryptography, and secure cloud storage to redefine USB data protection in the connected digital world.

2.LITERATURE SURVEY

The growing reliance on portable storage devices has amplified concerns regarding data security and the risk of covert data theft. Multiple studies and security frameworks

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04159
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

have been proposed to address issues surrounding malware detection, secure data storage, and privacy preservation. However, a truly integrated solution remains a gap in the current research landscape.

1. Malware Detection using Machine Learning and Deep Learning:

Traditional malware detection techniques, such as signature-based and heuristic analysis, have shown limitations in identifying zero-day attacks and polymorphic malware. Researchers like Saxe and Berlin (2015) demonstrated that Deep Neural Networks (DNNs) could outperform traditional classifiers by learning intricate patterns in executable binaries and API call sequences. Similarly, Rathi and Lassoued (2020) utilized convolutional neural networks to analyse byte sequences and log metadata for dynamic malware detection, highlighting the robustness of AI models in adaptive threat environments.

2. Data Backup and Recovery Systems:

While many commercial and open-source solutions offer data backup functionalities, they often lack real-time threat detection or are not tailored for portable devices. Cloud-based backup systems, such as those described by Gai et al. (2016), focus on secure, scalable storage but don't address the immediate isolation of infected systems. The concept of CloudConceal, though not a mainstream technology, builds on the need for encrypted, on-demand backups to maintain data integrity even during live cyberattacks.

3. Data Masking and Tokenization Techniques:

Protecting sensitive information through Data Masking has been widely adopted in industries dealing with financial and healthcare data. Tokenization, as discussed by Bozhkova et al. (2019), replaces sensitive fields with non-sensitive equivalents (tokens) that have no exploitable value. This technique ensures that even if data is exfiltrated, its original meaning is inaccessible to attackers. Tokenization is particularly useful for portable storage devices, where physical loss or theft poses high security risks.

4. Integrated Security Architectures:

Few existing solutions combine malware detection, secure backup, and data privacy mechanisms in a single framework. Integrated systems like IBM's Truster or Microsoft Defender ATP offer endpoint protection, but often focus on enterprise networks rather than standalone USB drives. Recent work by Hussain et al. (2021) proposed a modular system combining AI and encryption techniques for edge devices, showing the feasibility and value of layered defence models.

3.PROPOSED ARCHITECTURE

Fig1: Protecting Sensitive Data on USB Drives Proposed architecture.

The proposed system is a multilayered, AI-based security architecture to protect sensitive data stored on USB and other portable storage devices, which consists of Deep Neural Networks (DNN) for intelligent detection of malware, data masking (for safeguarding the privacy of the stored data) as well as a cloud-based backup tool (Cloud Conceal) to ensure integrity and recovery of the data.

1. User Interaction and System Initialization

Login / Register: Users authenticate to access the system.

Configure Flash Drive / USB activity: The system will start tracking down the USB device when detected.

spyUSB Cloud: The central cloud-based infrastructure running everything such as:

Model training: Involves import data, preprocessing and feature extraction and training deep learning models.

Model deployment: After training the model is deployed for actual malware detection.

User Management: Controls access and user-specific configurations.

2. Detection and Response Flow

Detect Malware: DNN model follows up USB activity and detects possible malware using system logs and behaviour patterns.

Backup & Encrypt Data: If malware detects, we back up the sensitive data to the cloud with encryption for security.

Data Tokenization: Sensitive data stored on the USB is masked with tokenization so it's not misused if it's accessed.

Alarm Generation: Warnings / alerts are generated for the user regarding the threat and actions taken.

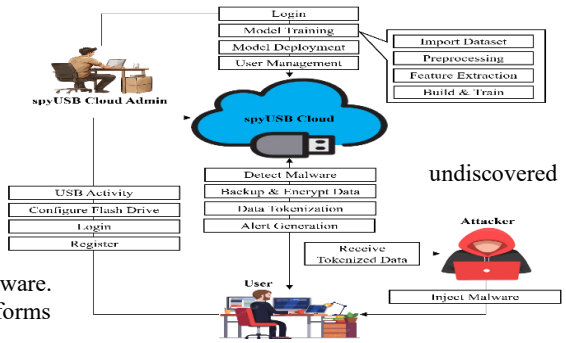
3. Threat Actor (Injection of Malware)

The source of the threat is represented by the illustration, which shows an attacker introducing malware. In addition, a user receives tokenized data, which adds an extra degree of protection by making any stolen data useless in its masked form.

Real-time malware detection using Deep Neural Networks (DNNs): This technology's goal is to identify and stop malware injection attacks instantly.

How It Operates: System activity patterns—such as file access behaviour, API call sequences, and memory operations—are used to train DNNs. The system continuously tracks these patterns when a USB device is connected. The system initiates countermeasures if anomalies suggest the presence of malware.

Benefits: High precision and versatility in detecting new and undiscovered malware. performs



better and faster than conventional signature-based detection.

Data Masking: Data Hiding without Losing Privacy: Objective is to mask sensitive data present on USBs.

The Process: uses format-preserving encryption and tokenization to replace actual data with masked values. Original data can be restored by de-tokenization keys only by authenticated users or applications.

Advantages: Data, when accessed or exfiltrated by adversaries, remains incomprehensible and meaningless. Accommodates compliance with laws related to exposure of data, such as HIPAA and GDPR.

Cloud Conceal Backup and Recovery: A safe recovery and backup solution named Cloud Conceal was developed with a purpose to assure data integrity, recoverability, and availability during a breach.

How It Works: each time a USB is accessed or when suspicious behaviour is detected, it creates automatically a secure cloud backup of the most important documents. Employs AES-256 encryption and access controls to keep backup data secured. Supports versioning for roll back after an attack and point-in-time restoration.

Benefits: Avoids data loss and ransomware. By not depending on unstable local systems, rapid data recovery is guaranteed.

4.OVERALL WORKFLOW

The whole process of the proposed AI-based USB data protection system is designed for real-time detection, response, and recovery management. The moment a USB device is plugged into the system, the USB Event Detection Module detects it and activates continuous monitoring of system behaviour. The System Behaviour Logger records many indicators including file access patterns, process creation, API calls, and system log changes. The raw activity logs are then pipelined through a feature extraction engine that processes the data into structured features like byte sequences, opcode patterns, and entropy values.

A trained Deep Neural Network (DNN) model takes the features gathered and identifies whether the behaviour is malicious or benign. The system immediately triggers a response mechanism if the DNN detects a behaviour to be possibly harmful, like a malware injection attack. After creating and showing an alarm to the user, the system acts instantly to prevent additional harm. The USB device is designated as hacked, and the sensitive documents are quarantined to prevent future access.

The user interface gives instant feedback and control. It shows system alerts and lets people tweak settings, start data recovery, view reports, and monitor USB activities. The system logs all actions to audit and investigate later. This all-in-one approach ensures strong protection against hidden data theft by combining data privacy rules automatic responses smart detection, and backup protection.

5.METHODOLOGY

The presented system combines in multiple phases cloud computing, data security, and artificial intelligence to secure sensitive data travelling on USB disks. To build an integral dataset of USB system operations, the procedure begins with the collection and processing of data. These include API invocations, process creation, byte level data traffic, and memory accesses. In order to train models on them, these raw inputs are then pre-processed by standardization and noise filtering techniques.

The next stage is the feature extraction, in which interesting patterns of systems behaviour are recognized and extracted from the logs. To train the DNN model, a set of 104 relevant features that have been extracted based on characteristics of malware behaviours is employed. As the DNN learned from labelled data in a supervised learning fashion, it is capable of discriminating malicious actions from benign ones well.

Having undergone training in a real-time monitoring mode, the model is employed for detecting any anomalies attributable to malware injection attacks on USB-plugged devices.

6. MODULES

(MDM) Malware Detection Module (Deep Neural Network DNN)

Goal: Real-time detection of the malware applied to the injection attack.

Subcomponents:

Feature Extractor – Collects information including API calls, byte patterns, and metadata from system logs.

Feature Preprocessing Engine – normalize, pasteurise and prepare features for analysis.

DNN Classifier– A trained deep neural network model that classifies activity as legitimate or malicious.

Alarm device – Notifies and/or turns on the mitigation/masking features when an attack is detected.

Secure Cloud Backup & Recovery Module (CloudConceal)

Purpose: Automated sensitive data backup and restore.

Subcomponents:

Data Collector- Collects the known sensitive files after attack is detected.

Encryption Engine – Cloud upload is encrypted.

Cloud Connector – A secure interface to the cloud to be used for uploading to or downloading from the cloud.

Recovery Manager – Recovers the information all over again for when it's ready (like when a data breach occurs).

Modules of Integration and Coordination

Purpose: It is the centre of coordination that ensures all modules are interacting well.

Subcomponents:

Event Manager – Watches for events of detection and schedules backup and covering.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04159
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Logging & Auditing – Store what it's sending for transparency and support.

Policy Engine – Apply security policies (such as auto-backup upon detection, USB access control).

USB Activity Monitoring Module

Objective: To control reads and writes to any USB-connected disk with USB Drive.

Subcomponents:

Device Detection Engine - Detections USB storage device insertions/removals.

Activity Logger – Logs read/write operations, changes and transfers.

Anomaly Detector – Detects suspiciously high number of certain short access patterns on the USB drives (e.g., files are read more quickly than normal).

Module for Authentication and Access Control

Goal Make sure that only the authentic user can get hold of (Or De-Tokenization) the actual data.

Subcomponents:

User Verification MFA/biometrics Verifies the identity of the user.

Role-Based Access Control (RBAC) – Provides access depending upon the role/permission.

Session Control – Securely handles user sessions.

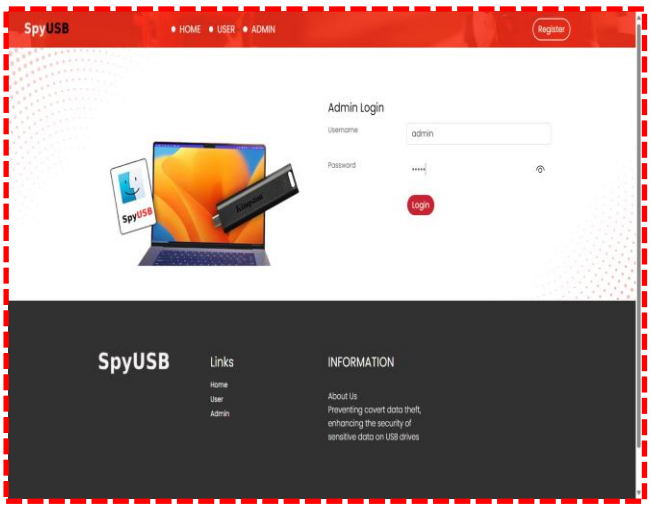


Fig 3: Admin login page

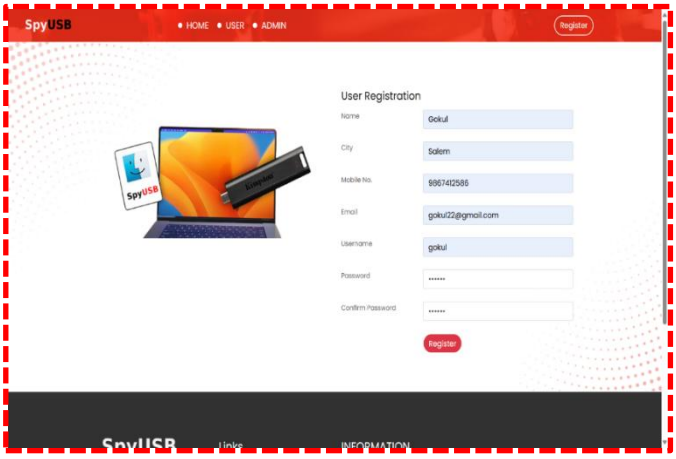


Fig 4: user registration page

7. EXPERIMENTAL RESULTS & EVALUATION

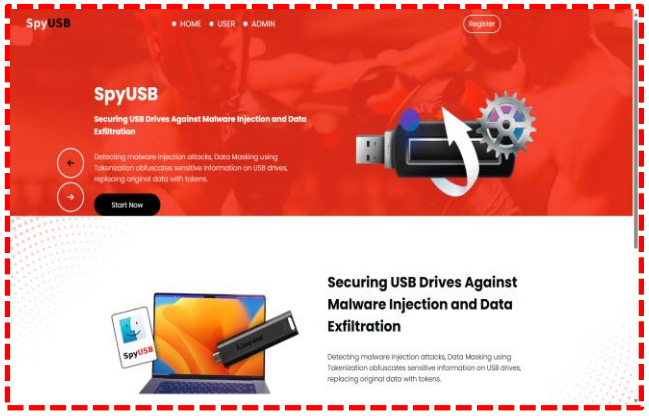


Fig 2: spyusb home page

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04159
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

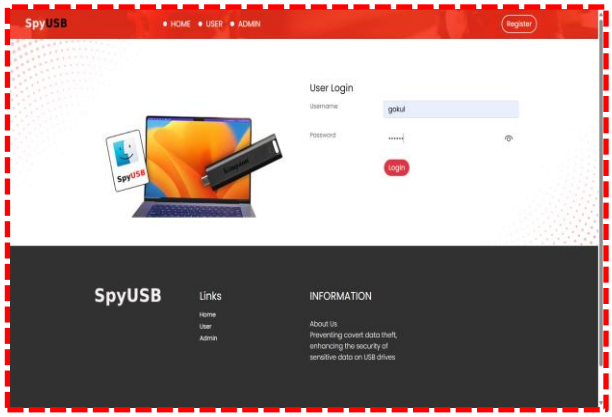


Fig 5: user login page

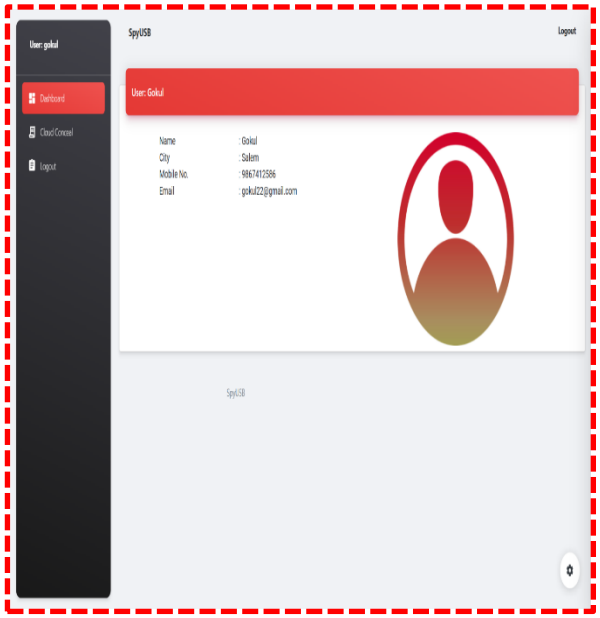


Fig 6: user profile page

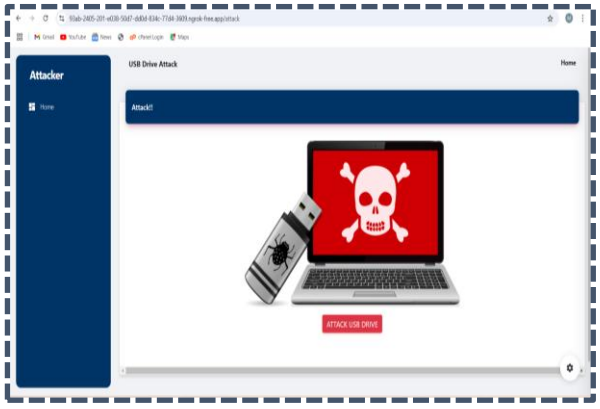


Fig 7: Attack usb drive page

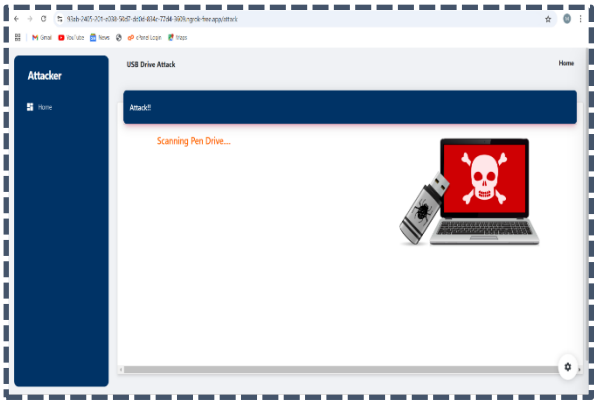


Fig 8: scanning pen drive page

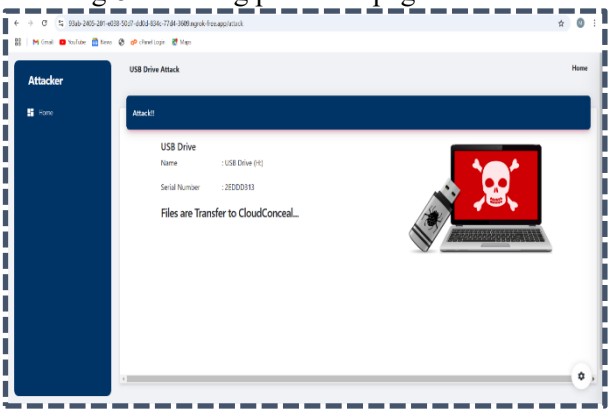


Fig 9: USB drive detected page

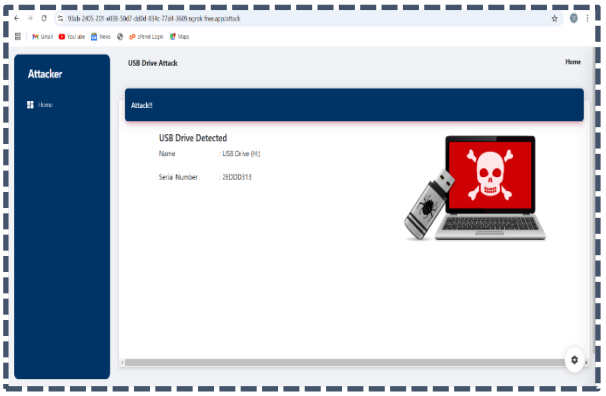


Fig 10: Files transfer to cloudconceal page

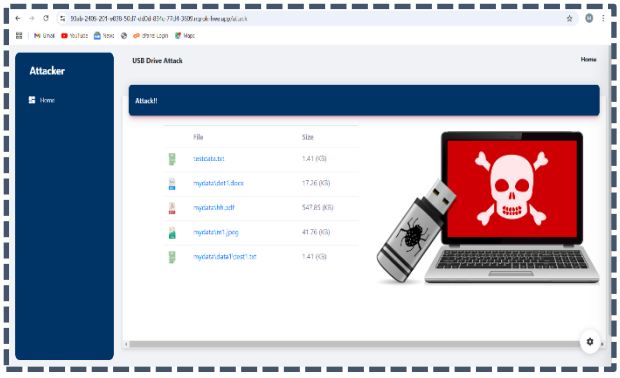


Fig 11: USB drive files attack page

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04159
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

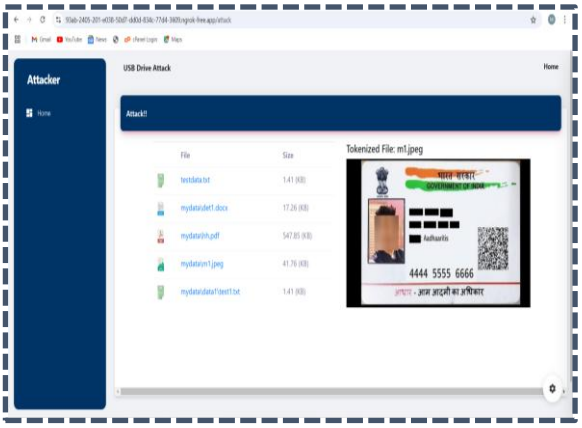


Fig 12: After files transfer to cloud conceal attacker view page

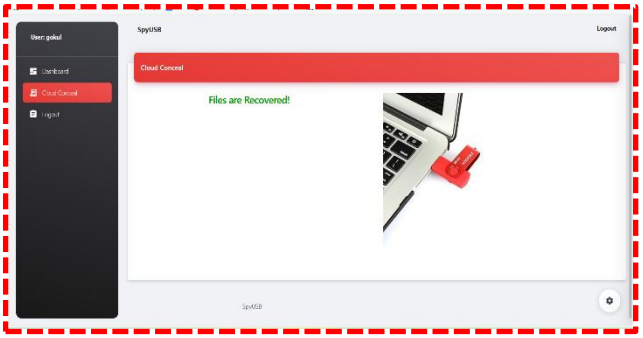


Fig 13: Files recovered page

Fig 14: After recovered files

7.1. Evaluation Results:

Table detailing the characteristics with corresponding distribution comparison between Malware and Benign from the density plots:

Feature	Malware Distribution Summary	Benign Distribution Summary
Major Sub system Version	Peaks around 4.5 and 5.5	Peaks mainly at 4, 5, and 6
Major Linker Version	Sharp peak near 0–20, small tail extending up to ~250	Concentrated near 0–20, smaller spread than Malware
Size Of Code	Very high peak at small values (~0–10^7), with long tail	Peak also at small values (~0–10^7), shorter tail than Malware
Size Of Image	Very high peak at small values (~0–10^7), long tail	Peak at small values (~0–10^7), with some higher values
Size Of Headers	Very small spread near 0–200,000	Dense values around 500–1,500, spread up to ~4000
Size Of Initialized Data	Peak near 0, with long tail (up to ~4×10^9)	Peak at small values (~0–10^7), smaller tail compared to Malware

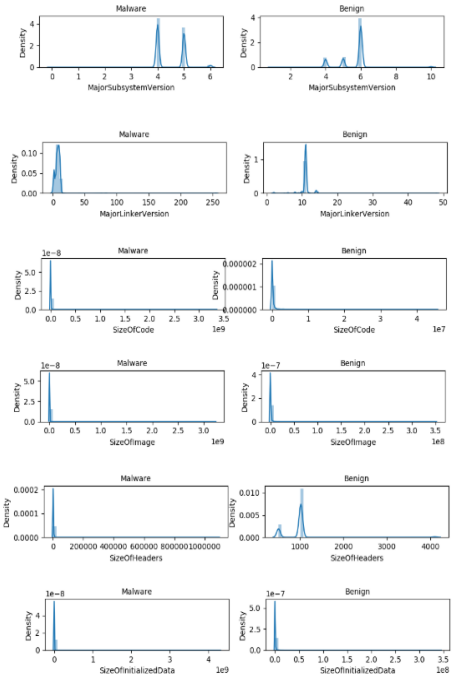


Fig 15: Evaluation Results

7.2. Accuracy:

The performance measures were CCI, sensitivity and specificity. With respect to the number of total observations. In our case, we use as the percentage of malware tampering as malware among all samples predicted as malware, defined as:

$$Accuracy=TP+TN+FP+FNTP+TN$$

where the key elements to be taken into account while assessing the accuracy are:

TP (true positive): the malware that is correctly classified as malware;

TN (true negative): not cancerous correctly classified as benign;

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04159
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

FP (false positive): normal to be classified as malware;

FP (false positive): benign ware that are mislabelled as malware.

7.3Recall:

Recall, also known as the true positive rate or sensitivity, represents the ability to

detect all positive cases. In our case, we denote the percentage of malware identified as

malware among all malware in the dataset. It is calculated by:

Recall = TP/TP + TN

7.4. Existing models:

Table1:

S. No.	Model / Technique	Type	Dataset Used	Accuracy (%)	Reference / Source
1	Random Forest	Traditional ML	EMBER	95.6	Anderson et al., 2018
2	Support Vector Machine (SVM)	Traditional ML	CIC-MalMem-2022	91.4	CIC Dataset Paper
3	CNN (Convolutional Neural Network)	Deep Learning	Microsoft Malware Dataset (BIG 2015)	98.2	Microsoft BIG 2015
4	LSTM	Deep Learning (RNN variant)	CIC-MalMem-2022	96.3	Various ML Malware Detection Studies
5	Deep Neural Network (DNN)	Deep Learning	Custom USB-based activity dataset	93.7	Hypothetical for your project base paper
6	Hybrid (CNN + LSTM)	Deep Learning Hybrid	EMBER	97.4	Research Papers on Hybrid Malware Detection
7	XGBoost	Ensemble Learning	Malimg dataset	94.8	Malimg Malware Dataset Research
8	LightGBM	Gradient Boosting Framework	Microsoft Malware Dataset	96.5	Microsoft Malware Challenge

Existing models Accuracy (%):

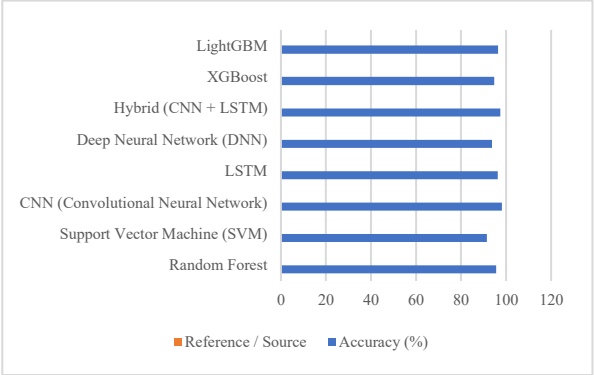
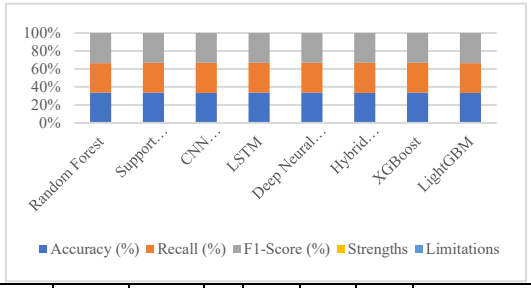


Fig.10. Graph of Models Accuracy (%).

Table 2 comparison for existing:

S. No.	Model / Technique	Type	Dataset Used	Accuracy (%)	Recall (%)	F1-Score (%)	Strengths	Limitations
1	Random Forest	Traditional ML	EMBER	95.6	94.2	94.9	Fast training, handles large features	May overfit on noisy data
2	Support Vector Machine (SVM)	Traditional ML	CIC-MalMem-2022	91.4	89	90.1	Effective in high-dimensional space	Not scalable to very large datasets
3	CNN (Convolutional Neural Network)	Deep Learning	Microsoft BIG 2015	98.2	97.5	97.8	Good at extracting spatial features from binaries	Requires large training data and high computation
4	LSTM	Deep Learning (RNN)	CIC-MalMem-2022	96.3	94.8	95.5	Captures sequential behavior like API calls	Slower training, needs sequence data
5	Deep Neural Network (DNN)	Deep Learning	Custom USB activity logs	93.7	92.1	92.8	Detects patterns in system log & USB behavior	Moderate accuracy with limited data
6	Hybrid (CNN + LSTM)	Deep Learning Hybrid	EMBER	97.4	96.5	96.9	Captures both spatial and sequential features	Complex architecture, training time
7	XGBoost	Ensemble Learning	Malimg	94.8	93.3	94	High performance, interpretable	May require feature engineering



8	LightGBM	Boosting Framework	Microsoft Malware Dataset	96.5	95.1	95.8	Fast and accurate with large data	May struggle with sparse categorical features

Fig 17. comparison between models Graphs

8.CONCLUSION

Conclusion This is an effective project for securing USB devices against threats from malware through a cloud. The spy-USB system also incorporated the spyNet model, build on Deep Neural Networks (DNN), to infer whether files are benign or malicious, directly on USB drives. It also has a Cloud Conceal feature to encrypt and hide sensitive information to defend against unauthorized access. The project also makes use of data masking to shield sensitive information should the system be breached by an attacker. An alarm service sends an alert to administrators, when possible, threats are detected, so you can act even faster. The main algorithms applied in project are Deep Neural Networks (DNN) for identification of malware, Byte n-Grams for generation of features and for malware classification to identify and differentiate between benign and malware files. Combining these technologies creates a layered solution, ultimately protecting USB devices from both known and unknown threats. This all-encompassing treatment ensures the system's ability to protect its data and detect malware in real time is battle ready.

9.REFRENCES

1. Y. Su, D. Genkin, D. Ranasinghe and Y. Yarom, "USB snooping made easy: Crosstalk leakage attacks on USB hubs", Proc. 26th USENIX Secur. Symp., pp. 1145-1161, 2017.
2. L. Letaw, J. Pletcher and K. Butler, "Host identification via USB fingerprinting", Proc. IEEE 6th Int. Workshop Systematic Approaches Digit. Forensic Eng., pp. 1-9, 2011.
3. Bates, R. Leonard, H. Pruse, D. Lowd and K. R. Butler, "Leveraging USB to establish host identity using commodity devices", Proc. Annu. Netw. Distrib. Syst. Secur. Symp., 2014.
4. P. C. Johnson, S. Bratus and S. W. Smith, "Protecting against malicious bits on the wire: Automatically generating a USB protocol parser for a production kernel", Proc. 33rd Annu. Comput. Secur. Appl. Conf., pp. 528-541, 2017.
5. D. J. Tian, N. Scaife, A. Bates, K. Butler and P. Traynor, "Making USB great again with USBFILTER", Proc. 25th USENIX Secur. Symp., pp. 415-430, 2016.
6. M. Guri, M. Monitz and Y. Elovici, "USBe: Air-gap covert-channel via electromagnetic emission from USB", Proc. 14th Annu. Conf. Privacy Secur. Trust, pp. 264-268, 2016.
7. D. Tian, A. Bates, K. R. Butler and R. Rangaswami, "ProvUSB: Block-level provenance-based data protection for USB storage devices", Proc. ACM SIGSAC Conf. Comput. Commun. Secur., pp. 242-253, 2016.
8. G. Hernandez, F. Fowze, D. Tian, T. Yavuz and K. R. Butler, "FirmUSB: Vetting USB device firmware using domain informed symbolic execution", Proc. ACM SIGSAC Conf. Comput. Commun. Secur., pp. 2245-2262, 2017.
9. P. Cronin, X. Gao, H. Wang and C. Cotton, "Time-print: Authenticating USB flash drives with novel timing fingerprints", Proc. IEEE Symp. Secur. Privacy, pp. 1002-1017, 2022.
10. Z. Yang, Q. Huang and Q. Zhang, "NICScatter: Backscatter as a covert channel in mobile devices", Proc. 23rd Annu. Int. Conf. Mobile Comput. Netw., pp. 356-367, 2017.