Human-centric Computing and Information Sciences

December 2024 | Volume 14



KCIA
Korea Computer Industry Association

www.hcisjournai.com

HCIS Human-centric Computing and Information Sciences

RESEARCH (Research Manuscript) Open Access

Human-centric Computing and Information Sciences (2024) 14:73
DOI: https://doi.org/10.22967/HCIS.2024.14.073
Received: March 25, 2024; Accepted: May 29, 2024; Published: December 30, 2024

Insider Threat Surveillance System via Multimodal-Based Physical Leak Detection and Optimal Imbalanced Augmentation in Research and Development Environments

Yuna Han[1] and Hangbae Chang[2,*]

Abstract

Increasing security incidents in research and development (R&D) demand enhanced measures to prevent technical leaks, especially against insider threats. Despite existing guidelines to mitigate external threats, current measures often overlook physical leaks in open collaborative environments. To the best of our knowledge, this study is the first to propose a surveillance system for detecting insider threats arising from physical paths. Security requirements are derived from a physical security perspective, and an audiovisual dataset is constructed. Additionally, this study presents a new augmentation technique to address overfitting by establishing an optimal point for augmenting imbalanced leakage versus normal behavior based on the proposed formula. Consequently, the proposed method could achieve 91.26% accuracy in leakage detection and 99% in distinguishing eight types of single events. This result validates the effectiveness of the proposed dataset and method in R&D settings, establishes a new academic standard for physical leakage detection, and introduces a universal augmentation formula for multiclass classification. In conclusion, this method enhances industrial surveillance systems and mitigates insider threats in R&D environments. Future studies should focus on developing a leakage behavior dataset based on big data using multiple criteria and building a real-time surveillance system suitable for industrial environments.

Keywords

Intelligent Surveillance System, Mitigating Insider Threat, Physical Security, Multimodality, Data Augmentation

1. Introduction

In the new era of technological hegemony, the heightened competition between major nations for research and development (R&D) dominance emphasizes the increasing importance of security in R&D settings. In this context, the frequent incidence of information leaks, primarily caused by insider threats, underscores the critical need for enhanced security measures.

For example, in 2020, the Federal Bureau of Investigation sought a Chinese Communist Party official

※ This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
***Corresponding Author:** Hangbae Chang (hbchang@cau.ac.kr)
[1]Department of Security Convergence, Chung-Ang University, Seoul, Korea
[2]Department of Industrial Security, Chung-Ang University, Seoul, Korea

disguised as a student at Boston University, who sent research documents to China [1]. In 2022, a Russian researcher working as an assistant at a German university was accused of disclosing Europe's aerospace launch program to Russian intelligence agents [2]. Many governments have established security guidelines for laboratories and research facilities to counter such security breaches, such as guidelines for academic integrity, research security implementation policies for researchers, and checklists for protecting research data [3–5]. These efforts aim to safeguard diverse R&D activities against security risks.

The current research on security requirements varies among countries, institutions, and organizations. These guidelines are often ineffective because the preventive benefits are difficult to guarantee. Most security experts have investigated external risk [6, 7]; hence, they may have overlooked insider threat in R&D settings. However, the open and collaborative nature of the R&D ecosystem and joint research activities results in many internal threats, complicating the security process [8].

Researchers have explored technical and physical techniques to overcome the nature of the R&D settings. Significant technical measures include analyzing signs of internal leakage and focusing on network firewalls and internet protocol management strategies using log-based data obtained via digital forensics [9] and monitoring systems [10]. However, detecting unapproved violations by visiting researchers and inadvertent leaks by internal researchers is challenging when relying on such data.

Furthermore, the managing entry and authentication procedures is the primary focus of physical security measures [11]. Research organizations advocate for surveillance technology with privacy and ethical considerations [12]. Closed-circuit television (CCTV) monitoring is employed in R&D settings for safety, controlling experimental equipment, securing restricted areas, and preventing information leaks [13]. Tracking and identifying electronic leakage channels is challenging with such surveillance systems. Intelligent security control systems driven by artificial intelligence (AI) have enhanced these capabilities while reducing human errors [14].

To the best of our knowledge, research has not been conducted on using CCTV and video surveillance systems to detect internal leaks in an R&D environment despite the increasing need for such technology. Given the contemporary advancements in computer vision [15, 16] and signal processing, this investigation concentrates on developing a leakage behavior detection method by exploiting the synergistic multimodality of audio and video. The R&D Leakage Behavior Detection dataset, supported by audiovisual data, was also constructed as a pivotal component of this research. Leakage behaviors were derived from security guidelines using the National Research Council of Science and Technology document to present a universal method not limited to specific laboratories or environments [17]. Furthermore, this study focuses on optimizing and combining a spatiotemporal-based three-dimensional (3D) convolutional neural network (CNN) and a one-dimensional (1D) CNN with a late-fusion strategy for video training. Beyond proposing a dataset, the main goal of this study is to discover a mechanism for training and verifying the dataset effectively.

Another problem is preventing overfitting due to the imbalanced dataset, which has approximately four times more normal data than leaked data. This study overcomes the difficulty of using data augmentation techniques and optimizing performance by controlling the augmentation of the ratio of leakage and normal behavior using the provided equation. The proposed method also enables the precise identification of specific leakage behaviors.

Therefore, this study makes the following contributions:

First, this study provides a foundational method for detecting leakage behavior in R&D settings to address the frequently overlooked problem of physical leaks in open and collaborative environments. Corresponding to this motivation, a method is proposed for detecting leakage behavior in an R&D environment with a focus on physical security for the first time. This study introduces an audiovisual-based method, detailing the derivation of the leakage behavior and construct the R&D Leakage Behavior Detection dataset in the outlined situations, and an audiovisual-based fusion model. This research establishes a method for detecting leakage behavior in R&D settings.

Second, this study demonstrates novel optimized data augmentation by determining the best performance point using the proposed formula. The proposed method is a simple, effective strategy that outperforms

the original method when on a full dataset or only a small dataset.

Last, the proposed method achieves an overall leakage detection rate of 91.26% and demonstrates 99% accuracy in distinguishing eight types of normal and leakage events for single video predictions. These results reveal that the suggested approach holds significant promise for future R&D security systems and has the potential for extended applications to broader multiclass classifications.

The remainder of this study is organized as follows. Section 2 reviews related studies. Next, Section 3 details the R&D leakage behavior detection method. Then, Section 4 provides deep insights into the experimental results and ablation studies. Finally, Section 5 discusses the results and concludes the study.

2. Related Work

This study introduces a method for performing preemptive physical security by detecting leaks in the R&D environment using video and audio sources. This study investigates previous research regarding the basis of research security, providing guidelines for management and operation, from the research process to the output. Furthermore, this study examined prior research on application methods and technologies in video-based intelligent surveillance systems. Recently, deep learning models have emerged based on multimodal techniques to recognize visual and audio information in videos as a strategy for training from multiple aspects. These strategies are effective; therefore, they can be applied to increase research security. This study focuses on investigating strategies to discover details about intelligent surveillance systems for R&D. In addition, the novelty of the proposed dataset is demonstrated by comparing and analyzing the current audio-visual surveillance dataset and the insider threat dataset to generate a dataset for this research.

2.1 Research Security

Research institutions are essential for generating R&D outcomes and creating solutions for societal problems across scientific and nonscientific fields [18]. The influence of R&D output on a nation's economic strength and security becomes more pronounced as science and technology evolve [3]. Thus, the importance of maintaining research security and the integrity of research data in these institutions has increased.

Frequent external interactions often occur when such institutions collaborate with external partners. Therefore, protecting sensitive data from internal and external parties is essential. Guidelines and studies are constantly developed to address these security challenges [19, 20]. For example, the National Protective Security Authority offers trusted academic research guidelines to protect intellectual property, personal data, and research outcomes [5]. These guidelines suggest creating a reliable research environment using a comprehensive checklist, highlighting the importance of assessing the separation of the cyber and physical realms.

Research security demands a combined approach that incorporates digital defense and physical strategies. The industry has proposed guidelines to protect sensitive research assets. Innovative solutions, such as Internet of Things-based biometrics and radio frequency identification, have been introduced for laboratory access control [21]. These procedures address access management and identify potential leakage points during research activities.

2.2 Video-based Surveillance System

Intelligent video surveillance systems have been employed to protect against hazards. For instance, home surveillance systems have been applied to detect unauthorized intrusions and weapon possession to allow, preventive action [22]. Beyond personal security, similar systems have also been applied to identify anti-social activities in offices, roads, and buildings aiding in post-event investigations [23]. Video surveillance systems have also been employed to protect users from harmful content and physical assaults online and offline [24–29]. Previous studies have indicated that these systems minimize human

resource use through deep learning techniques, such as faster regions with CNN features [30], which is beneficial for institutional video monitoring and home surveillance.

Additionally, intelligent video systems also have been employed for objectives other than protecting human life, such as environmental and traffic control [31, 32]. For example, drones equipped with cameras have been employed for forest fire surveillance because of their broad operational range and mobility using diverse strategies, such as optical image classification, semantic segmentation, and object detection [33]. Drones have also been deployed to monitor river flooding to determine whether the water level is at a dangerous threshold [34]. Most research on monitoring or video surveillance applications has focused on deep learning techniques suited to each task; however, scarce research has been conducted on intelligent surveillance systems in R&D environments. Security managers in R&D environments rely on traditional surveillance systems that manually detect abnormal behavior by referring to security regulations that are specific to research centers or laboratory environments; this study presents a method to mitigate leaks in the R&D environment rather than confining the method to certain organizations.

2.3 Audiovisual-based Multimodal Technique

Multimodal strategies that apply audiovisual data have been gaining momentum in computational analysis [35]. These strategies often involve extracting, segregating, and integrating audio and visual elements from video sources. Several studies have confirmed the efficacy of multimodal approaches in diverse applications, such as audiovisual localization, source separation, content classification, recognition tasks, and interactive question-answering systems [36–39].

This research focuses on identifying leakage behavior through advanced audiovisual classification and fusion techniques. These techniques are effective because they synergistically combine the unique characteristics of the audio and visual streams. Early and late-fusion methods have been identified as effective [40]. Late-fusion techniques are noteworthy, particularly those combining predictive values or feature concatenation [41]. Therefore, this study aims to analyze, and train audio and video datasets effectively by manipulating the strengths of feature concatenation in late-fusion paradigms.

A critical challenge in audiovisual fusion is precisely synchronizing audio and video components, which is critical when the classification process requires pinpointing audio events within a video timeline. Additionally, recent advancements have facilitated classifying videos into scenarios characterized by silent intervals or the presence of multiple audio sources [42]. Audiovisual-based models have been rigorously examined in theoretical and practical applications, particularly in surveillance systems, emotion recognition, speaker identification, and verification tasks [43–45]. A fusion method based on a multimodal model outperformed a unimodal model, such as an audio or video model, through diverse tasks. However, combining the multiple elements of a multimodal model might be challenging, making training in deep learning for real-world applications is more intricate.

This study presents a novel and comprehensive approach to applying audiovisual multimodal techniques. When detecting leakage behaviors in multimodal models R&D settings, the optimal combination method must be considered from a high-level application perspective. This unique strategy necessitates refining a CNN-based model tailored to the target dataset.

2.4 Audiovisual-based Surveillance Datasets

Recent advancements have underscored the necessity to create specialized audiovisual datasets, divided into two categories, for surveillance applications. The SoundNet-Flickr [46] and Kinetics-Sounds [47] datasets prioritize synchronizing sounds and visual information. Their core function is not their application but emphasizing repetitive tasks and enhancing the alignment between audio and video elements. The second category can be customized to specific surveillance applications. The XD-Violence and VGG-sound [48] datasets derive content from diverse sources, such as YouTube and movies, detecting violent actions. An in-house dataset curated to identify antisocial behavior in elevators

identified urination, vandalism, and other crimes [23]. Additionally, YouTube provides a plethora of datasets that delineate crowd behavior, ranging from chaotic riots to celebratory events [41]. Some datasets combine grayscale imagery with sensor data to predict natural disasters [49], whereas others focus on the auditory spectrum, such as capturing bird calls to monitor avian population shifts.

Although audiovisual datasets have been introduced in diverse categories, few datasets capture the subtle behavioral and vocal characteristics that emerge in leakage scenarios, as each research institute or organization manages them privately due to privacy and information leakage concerns. Table 1 compares the proposed dataset with existing datasets. This study presents an example of how leakage behavior can be collected from research institutions and organizations.

Table 1. Comparison of existing representative datasets and the proposed dataset

Dataset objective	Ref.	Type	Name	Source
Audiovisual sound localization/classification	[46]	Video (audiovisual)	SoundNet-Flickr	Website (Flickr)
	[47]	Video (audiovisual)	Kinetics-Sounds	Website (YouTube)
	[48]	Video (audiovisual)	VGG-Sound	Website (YouTube)
Emotional recognition	[43]	Video (visual)	AFEW 2016	Movies
Natural disaster detection/surveillance	[33]	Video (visual)	FLAME, Fire detection 360°	Drone filming
Violence detection	[24]	Video (audiovisual)	XD-Violence	Movies, YouTube
	[27]	Video (visual)	UCF-Crime	CCTV
	[28]	Video (audiovisual)	VSD	Movies
	[29]	Video (audiovisual)	CCTV-Fights	Mobile, CCTV
Crowded scene surveillance	[41]	Video (audiovisual)	DCASE Task1A	YouTube
Antisocial behavior detection	[22]	Video (visual)	-	CCTV
Insider threat detection	[52]	System logs	CERT	Synthetic
	[50]	Event logs	RUU	Synthetic
	[51]	Text (email)	Enron	Email
	[53]	Logs	TWOS	Synthetic
	[54]	Log (command)	Schonlau	Synthetic
Ecological environment monitoring	[31]	Audiovisual	Waterbirds	Camera
Leakage behavior detection in R&D settings	Ours	Audiovisual	R&D Leakage Behavior Detection	Camera/CCTV

2.5 Insider Threat Datasets

In cybersecurity, datasets have enabled the enhanced detection and development of countermeasures against insider threats. Renowned datasets, such as the RUU [50], Enron [51], CERT [52], TWOS [53], and Schonlau [54] datasets, have contributed to this progression. With its extensive array of scenarios, the CERT dataset offers a wealth of data from diverse sources such as emails, logins, devices, files, and HTTP, making it instrumental in uncovering potential security breaches. Similarly, the RUU dataset is designed to identify insider threats, offering a clear demarcation between standard insiders and potential masqueraders who obtain unauthorized insider identities [55]. Obtained via the CALO project, the Enron dataset provides an invaluable collection of more than half a million emails from top-tier executives, allowing it to identify potential acts of disloyalty. With its behavior-centric approach, TWOS emphasizes user dynamics, including complex attributes (e.g., mouse movement). Moreover, Schonlau focuses on Linux systems and highlights shell command patterns to identify anomalies.

These datasets are bound by the foundational reliance on electronic logs. Although this foundation offers a wide range of knowledge, it occasionally lacks the necessary level of detail for physical security applications. In addition, the datasets are simulations of potential scenarios; hence, their artificial nature can fail to the intricacies of the real world [56]. In contrast to current datasets, this research presents a

reliable dataset derived from real-life R&D scenarios. This approach increases the relevance of the dataset in practical applications and addresses the limitations of previous datasets.

3. Proposed Method

This study introduced a three-step method to identify and detect leakage behavior in an R&D environment from a physical security perspective. Fig. 1 presents the overall framework of the method comprising the stages of deriving leakage behavior in the R&D environment, constructing a dataset on leakage behavior, and building and predicting a multimedia-based model to detect leakage behavior using audio and visual materials. Specific descriptions of each step are provided in Sections 3.1–3.3.

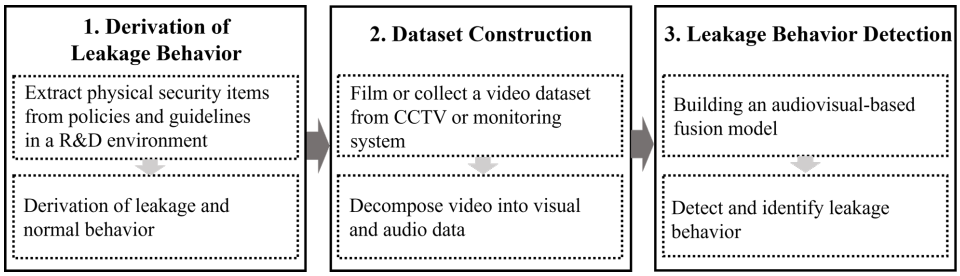


Fig. 1. Proposed physical security framework to detect leakage behavior in R&D environments.

3.1 Derivation of Leakage Behavior

This study aims to detect leakage behaviors proactively to mitigate potential insider threats in R&D environments. A list identifying these behaviors in terms of the primary objective was created to achieve this goal. The National Research Council of Science and Technology guidelines [17] are highly relevant to the research goals. These guidelines emphasize security protocols suited to researchers, aligning to identify researcher-specific leaks. The guidelines encompass administrative, technical, and physical security.

This study focuses on the leaks detected by video surveillance systems, underscoring the significance of physical security. This concentration facilitates detecting of visual and auditory leakage activities. The term "security guidelines" refers to the safeguards that researchers are required to observe during their studies in this context. Any action that violates these precautions is a "leak." This study describes the leakage behavior corresponding to each security guideline by employing a NOT operator to extract any leakage behavior contrary to each guideline. Table 2 presents a detailed breakdown of the physical security guidelines and the corresponding behaviors that could lead to security breach.

Comprehensive security measures in R&D are divided into human-centric protocols and research material-centric safeguards. In human-centric protocols, fundamental mechanisms were implemented, such as researcher access control and temporary visitor oversight, corresponding to Indexes 1 and 4 in Table 2. Protocols to protect research materials address the management of printouts relevant to R&D projects, governance of information and communication devices, and disposal management within research institutions (Indexes 2, 3, and 5 in Table 2). Each security guideline was generated based on the normal behavior of research institutes, and any deviation from these guidelines was identified as abnormal or potential leakage behavior.

The first type of negligence in research data access control refers to the potential threat of critical data disclosure through deliberate dissemination by unauthorized parties or unintended lapses. The second type involves access control management of researchers, regarding granting or handing over an identification (ID) card to another party without authorization, constituting a violation. The risks of negligence are multifaceted, including unauthorized access of security zones without ID cards, unauthorized entry with the intention of breaching sensitive data, and potential equipment failure by

unskilled researchers. The third type of negligence involves managing information and communication devices and refers to security risks in which vital data are transmitted to unauthorized electronic devices when media devices are introduced without strict security checks. The fourth type addresses managing external visiting researchers, referring to the security threat of divulging classified documents or information on ongoing research activities without approval. Finally, the fifth negligence type is regarding managing research data disposal protocols, referring to possible security threats caused by disposing of documents by tearing or crumpling critical research ideas and outputs without specialized equipment, such as shredders. For instance, individuals can intentionally take discarded data, and unintended exposure can occur via non-research staff, such as sanitation employees in research centers.

Table 2. Results of leakage behavior detection for physical security in an R&D environment

Index	Security type [17]	Security guidelines [17]	Proposed derivation of leakage behavior
1	Research data output management for R&D projects	Printed research material must be stored separately in a lockable physical storage device, and research material details should be recorded and managed once checked in and out.	Storing printed research data separately in a physical storage device and importing/exporting sensitive research data without permission.
2	Access control for researchers at a research institution	Researchers at research institutions must wear ID cards and not lend or transfer them to others.	Careless management of ID cards and the transfer or lending of ID cards to others.
3	Management of carrying information and communication devices in and out of a research institute	Cameras embedded in information and communication media should be sealed before being brought into a research center as needed.	Failure to attach security stickers to information and communication devices (laptops, cell phones, etc.)
4	Security management of temporary visitors to research institutions	A researcher or person in charge of the research institute should always be present to prevent unauthorized photography and recording.	Unauthorized photography or recording by visitors and outsiders.
5	Management of research data disposal at a research institution	Discarded research information should be physically destroyed and shredded to prevent research information leaks or breaches.	Discarding research information arbitrarily or without permission rather than fully destroying it.

3.2 Construction of the R&D Leakage Behavior Dataset

Generally, CCTV and cameras are employed to detect physical leaks in the R&D environment. Datasets were constructed using actions based on scenarios or specific behaviors indicated by designated labels in prior studies that created datasets using CCTV or cameras [20, 25, 27, 29]. The dataset was created to train leakage behavior in specific settings based on prior research analysis.

Due to security and privacy regulations, leaked activity data within R&D environments is unavailable to the public. Hence, this study collected original footage to create a dataset that closely aligns with the research objectives to facilitate model training while ensuring data privacy by deleting the footage after use. All participating researchers provided informed consent before recording, agreeing to the use of their audiovisual data for research. We applied a rigorous noise processing technique to recorded videos to protect individual identities, removing even the most minute details that could lead to personal

identification. This step guarantees that personal identification using the dataset is impossible, and the video and audio datasets were stored on a separate, secure hard drive, which was managed by a designated privacy officer who maintained the integrity of the privacy measures. The data we accessed only during the model training phase and were deleted immediately after use to prevent unauthorized access. All data collection and handling procedures complied with the ethical guidelines and standards for privacy protection in research.

The annotations for this method were based on the leakage behavior patterns in Table 2. The dataset comprises video collected based on scenario sequences matching each annotation label. For context, "Leak1" to "Leak5" represent specific leakage instances corresponding to the training and testing labels ranging from 1 to 5. In comparison, behavior associated with security requirements is classified as normal behavior.

For example, the action of storing printed research data separately in a physical storage device and importing/exporting sensitive research data without permission refers to leaving printed documents unsecured on a desk instead of storing them in a locked cabinet. Detection can involve monitoring document management logs and following proper storage protocols. The researcher's careless management of ID cards and the transfer or lending of ID cards to others can include such behaviors as handing over or receiving ID cards and whispering or making silent mouth movements in a manner inconsistent with normal speech patterns, which can be detected via audiovisual surveillance and behavior analysis. Moreover, the failure to attach security stickers to information and communication media (laptops, cell phones, etc.) can be detected by checking the presence of security stickers on external communication devices brought into the R&D environment. Further, unauthorized photography or recording by visitors and outsiders can be detected by the clicking sounds of a camera or recording from a person with a visitor tag who is taking photos or recording. This leakage behavior can be monitored via security cameras and visitor protocols. Finally, the act of discarding research information arbitrarily or without permission rather than fully destroying it can involve such actions as destroying or discarding documents without shredding them, especially when these documents are not logged in the management system or lack proper approval for disposal. Regular audits and checks are required to ensure adherence to document destruction policies.

Furthermore, the dataset distinguishes between leakage and normal behaviors using predefined actions and security protocols. For instance, Leak1 scenarios depict a researcher mishandling documents, which is a breach of protocol. The dataset captured leakage patterns, which were potentially augmented by additional scenes. Supplementary scenes can be integrated before or after a core leakage video sequence. Moreover, representing normal behavior requires adherence to a video sequence defined in the guidelines. This study evaluates five distinct scenarios, each corresponding to the security criteria specified in Table 2. Fig. 2 presents a representative sample of R&D information leakage videos in an R&D setting. When creating a video-based dataset, overfitting to a specific domain must be prevented by focusing on a single dataset sample, and the conditions and scenarios of the shooting environments must be examined. Thus, this study considers diverse data collection in R&D environments by altering the shooting site and focusing on training features connected to leaking rather than by being affected by people or places. The diversity of the video can be observed in terms of the characteristic apparel, hairstyle transitions, laboratory environment changes, role shifts, and distinct movement trajectories.

For instance, Fig. 2(a) represents a typical example of normal behavior, depicting a researcher storing confidential printed documents in a locked physical storage after use. Fig. 2(b) depicts a scenario in which the researcher violates the management guidelines by storing sensitive printed documents in a personal drawer instead of the designated physical storage. Fig. 2(c) illustrates two researchers lending an access card, a form of unauthorized access transfer, which is classified as an information leakage event. Fig. 2(d) presents a researcher using a laptop and mobile phone brought in from outside without security stickers, representing an unauthorized use classified as a leakage event. Fig. 2(e) displays two researchers in a discussion, where an outsider wearing a visitor badge is covertly taking photos of the research content with a mobile phone, an information leakage activity.

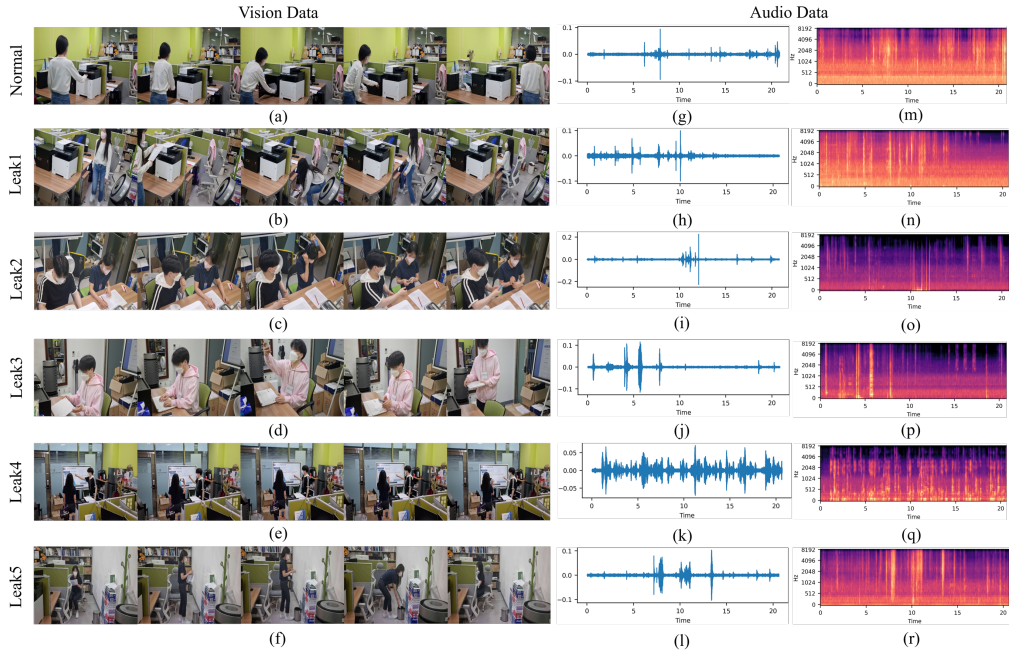


Fig. 2. Sample videos from the R&D Leakage Behavior Detection dataset consisting of visual and audio data with the following labels: Normal and Leak1 to Leak5. (a–f) Vision data contain multiple frames for each video. (g–l) Transformations of the original audio signals. (m–r) Audio data with spectrograms.

3.3 Audiovisual-based Leakage Behavior Detection Methodology

This section details the proposed leakage behavior detection process in an R&D environment. As illustrated in Fig. 3, the framework comprises four steps after loading the video dataset.

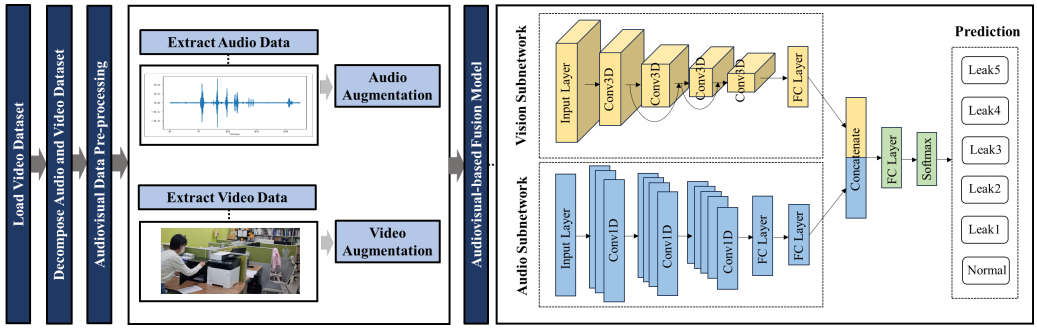


Fig. 3. Proposed audiovisual-based deep learning architecture to detect leakage behavior in R&D environments.

First, the visual and audio data from the video dataset D, encompassing leakage and normal activities, were decomposed into video D^V and audio D^A, respectively. This decomposition allows the independent storage and loading of the dataset. The labels associated with these datasets correspond to video sources 1, 2, 3, 4, 5, and 0. Formally, N, f, h, w, and m denote the number of data, data frames, height, width, and filterbank, respectively. In addition, dataset D is represented as $D = \{D_i^V, D_i^A, y\}_{i=0}^{N-1}$ for the ith snippet, where D_i^V denotes an element of $\mathbb{R}^{f \times h \times w \times 3}$, and $D_i^A \in \mathbb{R}^{m \times 1}$ with $y \in \{0, 1, 2, 3, 4, 5\}$.

The second phase involves a pre-processing technique that considers the distinctive features of each audio and visual dataset and converts them into a form that is appropriate for model training. Audio processing involves audio signal processing and log-Mel spectrogram transformations. In general, an audio signal represents a waveform that captures variations in the air pressure over time in digital or analog formats.

This study extracts data samples each second to analyze the amplitude–frequency relationship of the audio waveform. The audio signal is transformed into a domain that is aligned with its characteristics for efficient feature extraction. Although useful for decomposing signals into sine and cosine waveforms, the Fourier transform technique focuses on the frequency magnitude and ignores the phase information. This limitation makes it less effective for some audio types, such as music, speech, and audio events. Consequently, relying only on the Fourier approach for sounds indicative of leakage behavior (e.g., printing, walking, locking, and speaking) may not yield the best results. Instead, this study employs a Mel spectrogram, emulating human frequency perception by recognizing the enhanced sensitivity of the human auditory system to lower frequencies and the exponential nature of sound pitch perception. A logarithmic transformation was applied to the Mel spectrogram to better train on non-periodic information by taking advantage of human auditory characteristics.

Following video data preprocessing, the proposed data augmentation technique for audio and visual information was applied to address the overfitting problem in small datasets and input them into an audiovisual-based fusion model. The overall model structure was constructed using CNN-based audio and visual subnetworks, and each feature was concatenated at the fully connected layer stage to train for common features. Sections 3.3.1 and 3.3.2 explain the data augmentation technique and model construction in detail.

3.3.1 Audiovisual-based optimal imbalanced data augmentation

Due to the various scenarios and imbalanced properties of the provided R&D Leakage Behavior Dataset, a sophisticated reconstruction technique needs to be applied to the original data. Therefore, this study adopted an unsupervised autoencoder-based audiovisual augmentation technique divided into an encoder a decoder. The encoder phase extracts significant audio and visual information from the latent code, and the decoder employs an upsampling approach similar to that of the encoder.

Although other techniques, such as methods based on variational autoencoders and generative adversarial networks, have goals similar to the autoencoder, empirical evidence suggests that the autoencoder is more effective at reconstructing original data. For instance, the variational autoencoder approximates samples similar to the original data distribution and reconstructs them based on these samples, occasionally creating an optimal reconstruction. Similarly, models based on the generative adversarial network generate new samples while maintaining the underlying distribution of the original data. The autoencoder minimizes the reconstruction loss $L = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x_i})^2$ between the output data $\hat{x}$ and input data $x \in \mathbb{R}$, where n denotes the number of samples, resulting in remarkable reconstruction fidelity.

Fifty frames of concurrent audio and video signals were used as input, and individual video frames were processed using the 1D-CNN and 2D-CNN frameworks. Audio was reduced using a series of 1D-CNN layers (128, 64, and 32 channels), max-pooling, and rectified linear unit (ReLU) operations in the encoder and a symmetrical upsampling method in the decoder. The video segment uses a decoder identical to its encoder counterpart, downsampling 64- and 32-channel 2D-CNN layers, which were further enhanced using max-pooling and ReLU functionality. Finally, augmentation was conducted only for leakage behavior, and not normal behavior because this research aims to enhance an imbalanced and insufficient dataset of leakage behavior.

Furthermore, the proposed optimal imbalanced augmentation method in Equations (1)–(3) was employed in the autoencoder-based audiovisual augmentation model to determine the optimal point for using the data, which is simple but effective. In addition, D, $i, n,$ and A indicate the original dataset, i-th snippet, number of leakage data points, and data augmentation, respectively. Leakage data augmentation generates as much leakage as the original data, as expressed in Equation (1). Normal data augmentation can be

determined as the minimum number of original leakage datasets, expressed in Equation (2). Last, Equation (3) refers to the entire augmented data addition of normal data augmentation A_{normal} and the sum of the leakage augmentation data determined using Equations (1) and (2). These formulas were founded on the hypothesis that a large number of normal data, which are more abundant than leakage data, could be beneficial when combined with fewer leakage data. Furthermore, the optimal point of Equation (2) is mathematically described based on the experiment in Section 4.5.2, implying that stability is guaranteed when normal data are supplemented by the fewest number of leaks.

$$A_{leak_i} = D_{leak_i}, \tag{1}$$

$$A_{normal} < min\,(D_{leak_1}, D_{leak_2}, \ldots D_{leak_n}), \tag{2}$$

$$A_{Total} = \sum_{i=1}^{n} A_{leak_i} + A_{normal}. \tag{3}$$

3.3.2 CNN-based audiovisual fusion model

The proposed fusion approach detects leakage behavior by training on the properties of audio and video data. The audio model is introduced into the model as a 128 × 1 log-Mel spectrogram, which is transformed during pre-processing. The leakage behavior datasets contain time sequence information; thus, a 1D convolutional layer is applied for the audio model. Three convolutional layers were applied, with the number of filters increasing to 16, 32, and 64, respectively. Non-linearity is ensured by a kernel size of 3 and the ReLU activation function, boosting the model capacity. Given the sizes of the datasets, max-pooling and dropout layers were implemented to reduce overfitting. Dropout was set to a 0.25 ratio based on the finding that values under 0.5 reduce information loss, improving model performance. In addition, the convolutional block outputs were set to 63 × 16, 30 × 32, and 14 × 64, and were reshaped into a vector processed by one 64-unit fully connected layer. Second, video data with dimensions of 50 × 112 × 112 × 3 were processed using a 3D-convolutional layer. Given its effectiveness in human action recognition, the 3D-CNN extracted features from combined images and sequence data and fine-tuned the proposed dataset the following the basic structure from [57]. This study uses a video model based on the 3D-CNN, capitalizing on the effectiveness of 3D-convolutional layers in recognizing human actions by capturing spatiotemporal features to detect behavior in scenarios. An approach based on the existing research was adopted involving extracting spatial and temporal information through separate 3D-CNN layers to enhance the model learning capability.

This research improves performance by training on spatiotemporal data and using residual-layer prior information. The model comprises five convolutional blocks, with the initial spatial layer using eight filters and maintaining consistent padding for the subsequent temporal layers. In addition, the ReLU activation function was employed to guarantee nonlinearity, and layer normalization was applied to maintain performance across batch sizes. The convolutional process was designed to double the filter count for each block, enhancing the ability of the model to detect features in a limited dataset.

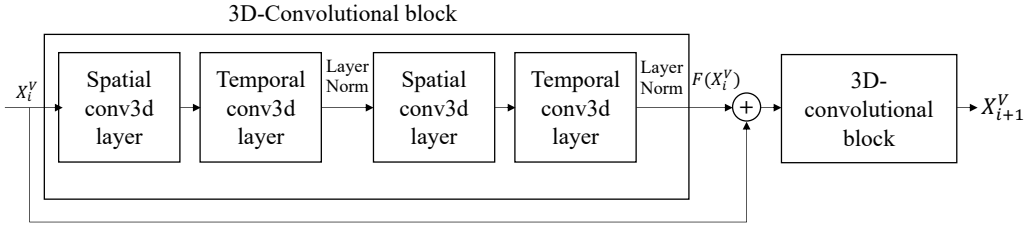


Fig. 4. Spatiotemporal three-dimensional convolutional block with residual mapping.

Fig. 4 illustrates the spatiotemporal convolutional layer within block F for the i-th snippet and the residual mapping by adding the output of the prior convolutional block X_i^V to the output of the current block $F(X_i^V)$, minimizing the information loss in the deep learning model. Thus, a feature vector was obtained with 64 channels after using 3D global max-pooling to derive the most representative feature vector among the frames.

Finally, the late-fusion technique integrates the independent audio and visual feature vectors. During this process, these vectors are fused via concatenation, performing better than the other techniques evaluated in this research, as demonstrated in Equation (4), where X^{AV} denotes the audiovisual concatenation output:

$$X_i^{AV} = Concatenate(X_i^V, X_i^A). \quad (4)$$

This study diverges from conventional late-fusion methods by focusing on the audiovisual model architecture and meticulously determining the integration point within the fully connected layers. The optimal layer size was experimentally calculated based on the label count, yielding two fully connected layers with 64 and 32 units. The softmax function was used to classify the data either leakage or normal behavior, using six additional layers. The refined model accurately detected leakages in the test dataset, resulting in proactive responses to such incidents within a research setting.

4. Experiments

4.1 Dataset Description

The R&D Leakage Behavior Dataset aims to differentiate between standard practices and leakage behavior in an R&D setting, detecting leakage incidents. As detailed in Section 3, the dataset was created by visualizing scenarios that contravene security protocols. Various factors, such as the filming environment, participants, movement trajectories, and camera angles, were changed to achieve a broad range of scenarios. For instance, filming was conducted at two venues, featuring two actors with interchangeable roles wearing different attire and varied hairstyles across several days. The camera angles were diverse to capture the participants' profiles and frontal views, reminiscent of typical surveillance footage. These settings were designed to vary the scenarios and serve as augmentation, focusing on the learning process of leakage-specific characteristics uninfluenced by individual or location biases. Nevertheless, when faces or research content are inadvertently captured, potential privacy and research confidentiality problems arise, necessitating de-identification procedures (e.g., introducing noise) (Fig. 2).

This dataset comprises 514 audiovisual records, categorized under six labels: one for standard behavior and five for assorted leakage types. Table 3 illustrates that each leakage type comprises approximately 50 records, whereas normal behaviors account for 254 records, presenting an imbalance, with the latter being about four times more frequent. For this study, the data were segmented into training and testing sets in an 8:2 ratio, as shown in Table 2. A notable aspect of this dataset is that each record lasts between 18 and 22 seconds, with most lasting approximately 21 seconds, reducing the likelihood of superfluous areas that might truncate vital information.

Table 3. Information on the R&D Leakage Behavior Detection dataset

Dataset type	Label	Number of data	Number of training datasets	Number of testing datasets
Leakage dataset	Leak1	50	39	11
	Leak2	50	35	15
	Leak3	53	45	8
	Leak4	55	44	11
	Leak5	52	38	14
Normal dataset	Normal	254	210	44

Entire dataset	Total	514	411	103

4.2 Experiment Environment

This section describes the process for detecting leakage behavior in an R&D environment. Fig. 3 shows the overall four-step framework after loading the video dataset. The vision and audio data from video dataset *D*, encompassing leakage and normal activities, were used in the audiovisual leakage detection experiments.

The experiments were executed on a system equipped with a GeForce RTX 3090, using CUDA 11.6 and NVIDIA 510.85.02 on an Ubuntu 20.04.6 LTS. This system has a 64-core CPU featuring an AMD Ryzen Threadripper PRO 3995WX. All experimental routines were scripted in Python, using the TensorFlow and Keras deep learning frameworks. The librosa and cv libraries facilitated audio and video data loading, and the graphics in Section 4 were rendered using matplotlib.

Hyperparameters were employed during model training and evaluation, encompassing the epoch, batch size, and validation split ratio, which were set to 150, 2, and 0.2, respectively. Typically, batch sizes in deep learning vary between 16 and 64, depending on the dataset size. Given the modest scale of this dataset, adjusting the batch size to 2 was appropriate. The training was optimized using the Adam optimizer with a learning rate of 0.001. When the data were augmented, the same hyperparameters persisted. Incorporating ReduceLROnPlateau reduced the learning rate when the performance was no longer progressing for adaptive learning rate adjustments. The validation loss was monitored, and if no decline was observed over 10 epochs, the learning rate was reduced to 80% of its prior value and capped at 0.0001.

The primary goal of this process was detection, and cross-entropy loss was used for multiclass categorization. For data augmentation, the mean squared error loss was employed to reduce any disparities between the original and modified values.

4.3 Evaluation Metrics

In this study, the primary objective of all experiments was detection. Consequently, the evaluation relies heavily on the metrics of accuracy and the F1-score, standard benchmarks in detection. The accuracy, as detailed in Equation (5), measures the proportion of correctly identified instances of normal and leakage behaviors relative to the overall count, where the number of true positives (TPs) correctly detected leakage behaviors, the number of the true negatives (TNs) is the correctly identified normal behaviors, the number of false positives (FPs) is the normal behavior incorrectly classified as leakage behavior, and the number of false negatives (FNs) is the leakage behavior missed by the model. Accuracy measures the overall effectiveness of the model in distinguishing between normal and leakage behaviors:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \quad (5)$$

The F1-score expressed in Equation (6), provides a more complicated evaluation by incorporating precision and recall, gauging over and under-detection. The precision measure is the proportion of TP findings among all positive results predicted by the model, which should be minimized to ensure the reliability of the detection system by reducing unnecessary alerts. Recall assesses the proportion of TP results among all positive cases by identifying actual leakage events, ensuring that few true leakage behaviors are missed. Furthermore, the F1-score is beneficial when the class distribution is imbalanced because it considers FPs and FNs. This metric is important in cases where accurately identifying of leakage behaviors and minimizing false alarms are equally important.

While achieving high overall accuracy is paramount for leakage detection, precisely categorizing leakage types and normal behavior is equally vital. This precision strengthens security and optimizes economic efficiency. Thus, the F1-score is a critical measure in specific scenarios, emphasizing the balance between precision and recall:

$$F1-score=\frac{2}{\frac{TP}{TP+FP}+\frac{TP}{TP+FN}}=\frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}=\frac{2\times Precision\times Recall}{Precision+Recall}. \quad (6)$$

4.4 Leakage Behavior Detection Results and Analysis

The proposed method, designed to detect leakage behavior in R&D environments, prioritizes discerning patterns that are characteristic of audiovisual data. Initially, when treated separately, the models with superior performance for video and audio data, were experimentally compared and analyzed. The rationale is that if an AI model excels in visual data interpretation but struggles with audio, the audio component might inhibit optimal learning. Therefore, the experiments were conducted separately on video and audio-centric deep learning models, as summarized in Table 4 [37, 41, 58–67].

The first through sixth rows of Table 4 present the outcomes of the experiments on the pretrained CNN-based models, sequence training via feature extraction, and models trained directly on the dataset. Both models use CNN in image feature processing and learning, making them adept at learning visual patterns. Direct CNN applications—i.e., Temporal3dConvNet, 2D-CNN, 2D-CNN + long short-term memory (LSTM), and the selected spatiotemporal CNN—exhibited accuracy improvements of 4.82% to 7.77% compared with the preliminary CNN methods. These findings underscore the effectiveness of applying the CNN directly to the data to capture visual patterns. The proposed spatiotemporal-based 3D-CNN model, a modified version discussed in Section 3, consistently outperformed other models, achieving an accuracy of 76.70% and an F1-score of 0.77.

Table 4. Audio- and video-based leakage behavior detection model performance

Type	Model	Accuracy (%)	F1-score
Vision model	InceptionV3+LSTM [58]	65.05	0.65
	MobileNetV2+GRU [59]	66.02	0.62
	Temporal3dConvNet [60]	70.87	0.71
	2D-CNN	72.82	0.73
	2D-CNN+LSTM [61]	71.84	0.70
	Ours (Small FTR-3D-CNN)	76.70	0.77
Audio model	DNN [62]	65.05	0.62
	VGGish+GRU [63]	42.72	0.42
	2D-CNN [64]	52.43	0.52
	CNN+LSTM [65]	42.72	0.42
	Att-CNN+Att-Bi-GRU [66]	62.14	0.60
	Ours (Small 1D-CNN)	77.22	0.78
Audiovisual model	L3-Net [37]	42.71	0.42
	VGG-15 [41]	57.28	0.55
	OpenL3 [67]	57.28	0.57
	Ours	86.41	0.86

FTR-3D-CNN= fine-tuned spatiotemporal three-dimensional convolutional neural network, Att = self-attention mechanism.

In Table 4, the seventh through twelfth rows present the results of evaluating the audio-based models designed for sound classification. Audio data can be modelled as either 1D or 2D signals, depending on the pre-processing. Several architectures, including deep neural networks (DNNs) and time-series-based models (CNN, gated recurrent unit, and LSTM), have also been explored. The 1D-CNN audio model introduced in Section 3 offers superior performance, with 77.22% accuracy and an F1-score of 0.78, followed by the DNN model at 65.05% accuracy and an F1-score of 0.62. Lightweight deep learning models are more effective for smaller datasets.

The first to twelfth rows of Table 4 collectively confirm the effectiveness of the CNN-based deep learning models for visual and audio data. Consequently, this study advocates a hybrid approach that

merges a fine-tuned spatiotemporal 3D-CNN visual model with a 1D-CNN audio model, both of which provide superior performance.

The last four rows in Table 4 compare the performance of the proposed model and existing CNN-based audiovisual models. The proposed model achieved an accuracy of 86.41%, with marked improvements from 29.13% to 43.7% over the established methods. Furthermore, the F1-score was 0.86, indicating advancements between 0.29 and 0.44. Notably, the hybrid approach adopted by the proposed model, which capitalizes on optimal feature extraction for audio and video, is designed for compact datasets with diverse scenarios. The F1-score further indicates the potential to improve security and reduce economic costs by curtailing FPs.

These results appear to stem from the ability of the model to learn the visual and auditory characteristics specific to each leakage type (Leak1–Leak5). The model detects the absence of visual information related to specific physical storage devices and the associated audio signals of locking and unlocking for Leak1, identifying when printed confidential documents are not stored in a designated device. This detection involves observing variations in audio signals, such as whispering or silence between researchers, and capturing the brief visual moment when an access card is handed over for Leak2, indicating an unauthorized transfer of access. For Leak3, the model detects the absence of security stickers on communication devices, such as laptops or phones, learning from the auditory cues of sticker application or the visual absence of the sticker, classifying it as a leakage event. The model identifies irregular sounds, such as a beep at the start of a recording or camera clicks, or visually detects covert actions, such as pressing a recording button or lifting a camera to take photos for Leak4, classifying this as unauthorized recording or photography. For Leak5, the model learns to recognize the absence of shredding sounds or visually detects crumpling or tearing and discarding documents, differentiating these from normal document disposal behavior.

A single prediction experiment was conducted for the R&D environmental monitoring system using a pre-trained model combined with the proposed audio and visual techniques (Fig. 5). This experiment achieved an approximate accuracy of 99% in distinguishing eight types of single events, including specific types of normal (1, 2, 3, 4) and leakage behaviors (2, 3, 4, 5). Fig. 5 presents examples of the best-scoring videos, serving as a starting point for developing advanced surveillance systems in industrial settings. These results highlight the efficacy of the model in accurately identifying leakage behavior types, demonstrating its potential for a more granular behavioral analysis in normal and leakage scenarios. This high accuracy underscores the applicability of the model in developing advanced insider threat surveillance systems.

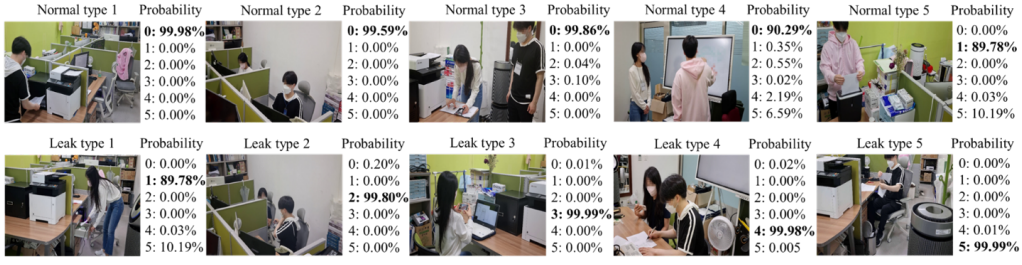


Fig. 5. Single-event video prediction results for leakage and normal behaviors using the proposed method.

4.5 Leakage Behavior Detection Results and Analysis

4.5.1 Ablation study on multimodality

Although specifications have outlined the need for surveillance systems and protection targets in R&D environments, research has yet to focus on effective detection methods. In Table 4, the results indicate that the multimodal approach outperformed the individual visual and auditory models. Although the independent

spatiotemporal 3D-CNN and 1D-CNN models achieved 76.7% and 77.23% accuracy, the third bar graph in Fig. 6 illustrates that their combined application increased the accuracy to 86.41%.

The general efficacy of the multimodal approaches was further assessed by merging high-performance models from each video and audio model (Table 4). Fig. 6 illustrates that the proposed method yields superior results in unimodal and multimodal contexts. For instance, the combination of a DNN-based audio model and 3D-CNN-based video model improved the combined accuracy to 71.84% from the individual accuracy values of 65.05% and 62.7%. Similarly, the combination of the 1D-CNN-based audio and 3D-CNN-based visual models improved the combined accuracy to 73.78% from the individual rates of 70.87% and 71%, respectively.

These findings underscore the fact that multimodal models predominantly outperform their unimodal counterparts. Additionally, the CNN-based audiovisual combinations exhibited consistently high performance, as evident in the second and third bar graphs.

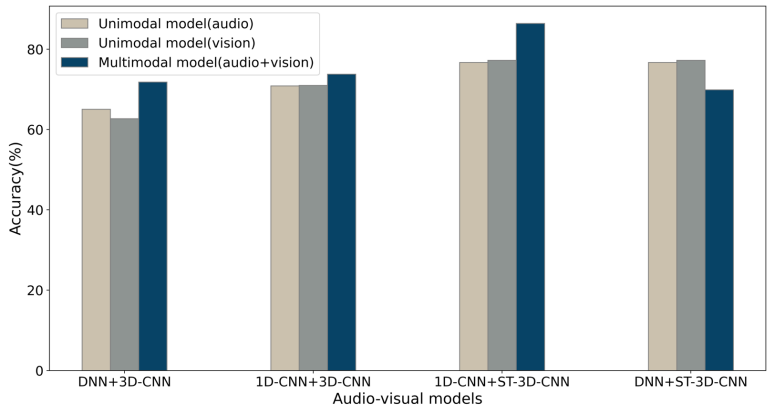


Fig. 6. Performance comparison of the multimodal and unimodal models.

4.5.2 Ablation study on data augmentation

This study employs an autoencoder-driven data augmentation technique to counteract overfitting caused by imbalanced and limited data. Table 5 reveals that this approach boosted the accuracy from 86.41% to 87.37% and improved the F1-score from 0.86 to 0.87.

Table 5. Performance comparison before and after applying the data augmentation technique

Data augmentation technique	Accuracy (%)	F1-score
Before application	86.41%	0.86
After application	87.37%	0.87

Although many studies have sought performance enhancements using diverse data augmentation strategies, the volume of potential technique combinations renders this approach time-consuming. Moreover, the combinations of optimal techniques can vary with the dataset. This study applies a singular technique to enhance performance based on the application ratios between leakage and normal behaviors, as presented in Fig. 7.

In Table 6, from an initial count of 411 training data points, the leakage activity dataset was augmented by adding 35 to 44 data entries for each leakage type, reaching 612. Traditionally, augmentation targets either deficient datasets or entire datasets. The performance improved when the excluded "normal behavior" dataset was moderately augmented. For instance, augmenting it by 20, 30, 40, or 50 units yielded optimal results at 30 additional units, achieving an accuracy of 91.26% and an F1-score of 0.91 (Fig. 7). The precise performance results of this optimal imbalanced augmentation were represented in Fig. 8 by a confusion matrix, highlighting the classification performance across leakage types and normal

behavior. This confusion matrix illustrates the distribution of the predicted labels versus the actual labels, indicating that the model accurately identified most leakage types.

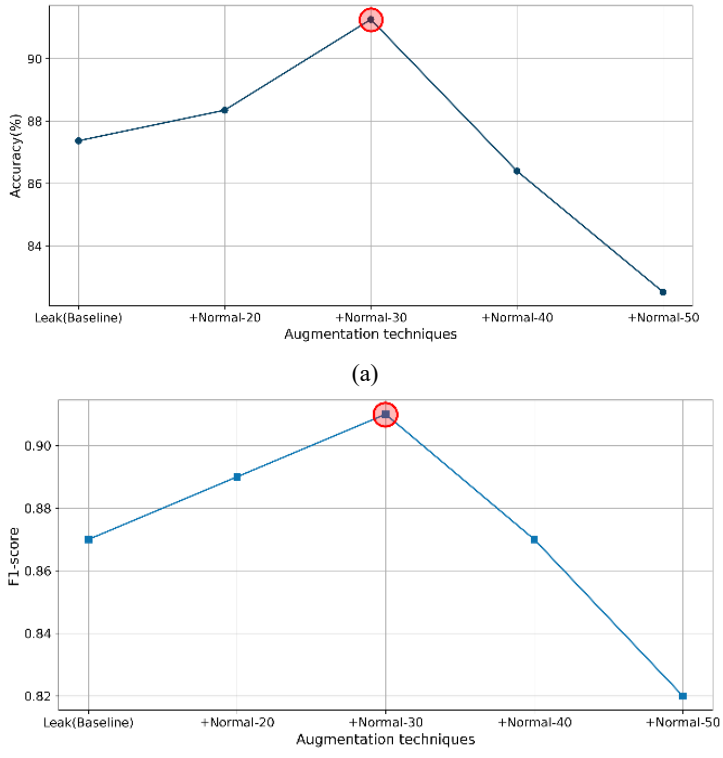


Fig. 7. Performance comparison of the proposed augmentation technique in terms of (a) accuracy and (b) F1-score. The red circle represents the optimal point for each metric. The baseline, Normal-20, Normal-30, Normal-40, and Normal-50 leaks refer to the data augmented with only the leak and those with the leakage augmented data and 20, 30, 40, and 50 normal data points, respectively.

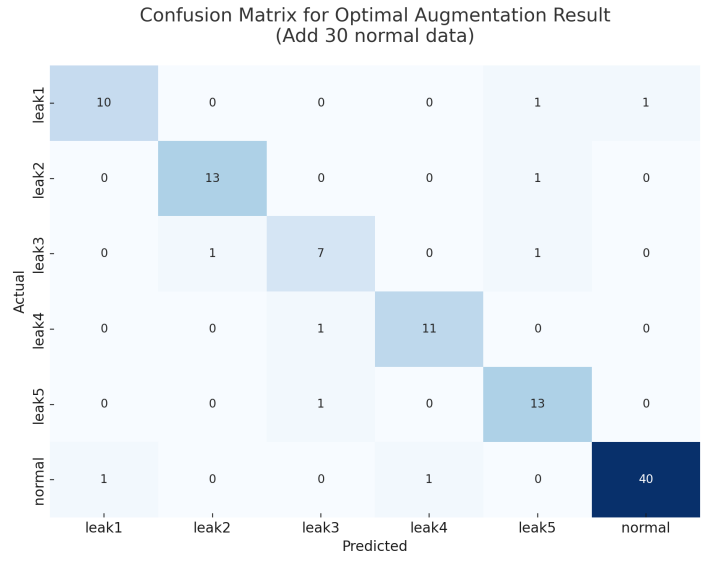


Fig. 8. Confusion matrix results for optimal augmentation, adding 30 normal data to an imbalanced dataset.

Table 6. Information on the proposed data augmentation

Augmentation technique	Number of training samples
Original data	411
Leak augmentation	612
Add 20 normal data	632
Add 30 normal data	642
Add 40 normal data	652

However, performance declined when normal behavior augmentation exceeded the leakage behavior augmentation (Fig. 7). This outcome suggests that skewed imbalanced datasets may lead to domain bias at some point. Consequently, the red circle in Fig. 7 represents the proposed optimal balance point according to Equation (3). The results in Figs. 7 and 8 demonstrate that augmenting a relatively small subset of the dataset is more effective than augmenting the entire dataset. This result is evident from the baseline and +Normal 50 results in Fig. 7, revealing a decline in performance when excessive augmentation is applied. Conversely, an improvement in accuracy and the F1-score occurs with +Normal 20 and +Normal 30, confirming that a balanced augmentation approach yields better results. Specifically, the confusion matrix in Fig. 8 provides a detailed view of the optimal model performance, balancing correctly identifying leakage types and normal behavior, validating the approach of moderate augmentation.

4.5.3 Ablation study on fusion techniques

Traditionally, multimodal learning uses two primary methods: early and late fusion. Although early fusion has predominantly been applied in initial image-based detection techniques, more recent studies have noted the superiority of late fusion. Early fusion entails integrating various types of data (e.g., audio and visual data) at the inception of the model. However, to capture the distinct characteristics of each data type, late fusion, combining features at a later stage, is more appropriate.

As detailed in Table 7, this study focuses on employing late-fusion techniques centered on concatenation and prediction. Prediction involves merging the probability values (outputs of the softmax function) from the audio and visual models. Potential strategies involve calculating the maximum, product, or mean of two probability values. The concatenation approach uses a composite vector of audio and visual features as input into the fully connected layer.

Table 7 reveals that the highest performance was recorded using the proposed concatenation technique (with 128 units) before and after data augmentation. Although the prediction technique applying the product displayed subpar performance before augmentation, it reached an accuracy of 85.43% post-augmentation. However, its F1-score of 0.5 suggests a susceptibility to FPs. Given the criticality of detecting leaks in this study, this vulnerability renders this technique less suitable.

Table 7. Performance results of late-fusion techniques on R&D Leakage Behavior Detection

Late fusion technique	Original data		Augmented data	
	Accuracy (%)	F1-score	Accuracy (%)	F1-score
Prediction (max)	7.86	0.43	42.71	0.64
Prediction (prod)	42.71	0.42	85.43	0.51
Prediction (mean)	12.52	0.10	14.56	0.0
Concatenate (unit 128)	86.41	0.86	87.37	0.87

5. Discussion and Conclusion

Although guidelines have been proposed to mitigate security threats posed by insiders and external visitors in R&D environments, established managerial and technical measures have limitations, particularly in accurately tracking the physical routes of data breaches by collaborating researchers. This study introduces an audiovisual-based detection method for leakage activities by applying a fusion strategy that trains audio and visual models from CNN in an end-to-end manner to address this challenge.

To support this approach, we outlined the extraction of the leakage activity lists, data construction and augmentation, and techniques throughout the learning process of the fusion model. This comprehensive study establishes a baseline method for leakage activity detection in such environments. The proposed strategy demonstrated excellent performance using a combination of CNN-based audiovisual models and achieved an accuracy of 91.26% by integrating the proposed optimal imbalanced data augmentation.

This research provides significant insights into presenting a foundational detection framework for leakage activities in the R&D context. Moreover, it presents a universally applicable and efficient data augmentation technique. These exemplary prediction results for individual images underscore the potential of integrating this method into future video surveillance systems. As this study represents an initial exploration of leakage detection via physical security measures, future research should focus on building a leakage behavior dataset based on big data by referencing various guidelines and implementing a real-time surveillance system that applies to industrial environments.

Author's Contributions

Conceptualization, YH, HC; Investigation, YH; Methodology, YH, HC; Software and validation, YH; Data curation, YH; Writing of the original draft, YH; Writing of the review and editing, YH; Visualization, YH; Supervision, HC; Project administration, HC; Funding acquisition, HC.

Acknowledgements

We are grateful to Masters Woojoo Kwon and Jiwon Moon of Chung-Ang University for their support in firming this research.

Funding

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the Convergence security core talent training business support program (No. IITP-2024-RS-2023-00266605) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

Competing Interests

The authors declare that they have no competing interests.

References

[1] A. Cromar, "Former Boston university student wanted by the FBI for alleged counterintelligence," 2020 [Online]. Available: https://www.boston.com/news/crime/2020/01/31/former-boston-university-student-wanted-by-the-fbi-for-alleged-counterintelligence/.

[2] M. Fitzpatrick, "University spy in Germany 'passed Ariane rocket details to Russia'," 2022 [Online]. Available: https://www.themoscowtimes.com/2022/01/27/university-spy-in-germany-passed-ariane-rocket-details-to-russia-a76170.

[3] National Science and Technology Council, "Recommended practices for strengthening the security and integrity of America's science and technology research enterprise," 2021 [Online]. Available: https://trumpwhitehouse.archives.gov/wp-content/uploads/2021/01/NSTC-Research-Security-Best-Practices-Jan2021.pdf.

[4] National Science and Technology Council, "Guidance for implementing National Security Presidential Memorandum 33 (NSPM-33) on national security strategy for United States government-supported

research and development," 2022 [Online]. Available: https://www.whitehouse.gov/wp-content/uploads/2022/01/010422-NSPM-33-Implementation-Guidance.pdf.
[5] National Protective Security Authority, "Trusted Research Guidance for Academia," c2023 [Online]. Available: https://www.npsa.gov.uk/trusted-research-academia.
[6] O. E. Llerena Castro, X. Deng, and J. H. Park, "Comprehensive survey on AI-based technologies for enhancing IoT privacy and security: trends, challenges, and solutions," *Human-centric Computing and Information Sciences*, vol. 13, no. 39, 2023. https://doi.org/10.22967/HCIS.2023.13.039
[7] S. Sengan, A. Mehbodniya, J. L. Webber, A. Bostani, A. Almusharraf, M. Alharbi, and S. B. Khan, "Improved LSTM-based anomaly detection model with cybertwin deep learning to detect cutting-edge cybersecurity attacks," *Human-centric Computing and Information Sciences*, vol. 13, no. 55, 2023. https://doi.org/10.22967/HCIS.2023.13.055
[8] National Institute of Standards and Technology (NIST), "Safeguarding International Science: Research Security Framework," 2023 [Online]. Available: https://www.nist.gov/publications/safeguarding-international-science-research-security-framework.
[9] M. A. Baballe, A. Hussaini, M. I. Bello, and U. S. Musa, "Online attacks types of data breach and cyberattack prevention methods," *Current Trends in Information Technology*, vol. 12, no. 2, pp. 1-5, 2022. https://doi.org/10.5281/zenodo.7144657
[10] M. Sheeraz, M. A. Paracha, M. U. Haque, M. H. Durad, S. M. Mohsin, S. S. Band, and A. Mosavi, "Effective security monitoring using efficient SIEM architecture," *Human-centric Computing and Information Sciences*, vol. 13, article no. 17, 2023. https://doi.org/10.22967/HCIS.2023.13.017
[11] T. A. Ricks, B. E. Ricks, and J. Dingle, *Physical Security and Safety: A Field Guide for the Practitioner.* Boca Raton, FL: CRC Press, 2014.
[12] Eclipse CCTV, "Lab research security cameras," 2024 [Online]. Available: https://www.eclipsecctv.com/lab-research-security-cameras.
[13] REVADER, "Laboratory CCTV & Hazardous Environments," 2024 [Online]. Available: https://www.revader.com/laboratory-cctv.
[14] S. L. Ullo and G. R. Sinha, "Advances in smart environment monitoring systems using IoT and sensors," *Sensors*, vol. 20, no. 11, article no. 3113, 2020. https://doi.org/10.3390/s20113113
[15] Y. Chen, R. Xia, K. Yang, and K. Zou, "DNNAM: image inpainting algorithm via deep neural networks and attention mechanism," *Applied Soft Computing*, vol. 154, article no. 111392, 2024. https://doi.org/10.1016/j.asoc.2024.111392
[16] Y, Chen, R. Xia, K. Yang, and K. Zou, "MICU: image super-resolution via multi-level information compensation and U-Net," *Expert Systems with Applications*, vol. 245, article no. 123111, 2024. https://doi.org/10.1016/j.eswa.2023.123111
[17] National Research Council of Science & Technology, "Research security management guide for researchers," 2023 [Online]. Available: https://www.nst.re.kr/www/selectBbsNttView.do?key=46&bbsNo=7&nttNo=39413&searchCtgry=&searchCnd=all&searchKrwd=&pageIndex=1&integrDeptCode=.
[18] M. Gulbrandsen, "Research institutes as hybrid organizations: central challenges to their legitimacy," *Policy Sciences*, vol. 44, pp. 215-230, 2011. https://doi.org/10.1007/s11077-011-9128-4
[19] I. Fushshilat and Y. Yogasmana, "IoT scheme for surveillance system and laboratory security access," *IOP Conference Series: Materials Science and Engineering*, vol. 850, no. 1, article no. 012014, 2020. https://doi.org/10.1088/1757-899X/850/1/012014
[20] F. Lind, A. Styhre, and L. Aaboen, "Exploring university-industry collaboration in research centres," *European Journal of Innovation Management*, vol. 16, no. 1, pp. 70-91, 2013. https://doi.org/10.1108/14601061311292869
[21] National Protective Security Authority, "Trusted Research Guidance for Industry," 2023 [Online]. Available: https://www.npsa.gov.uk/trusted-research-industry.
[22] D. Berardini, L. Migliorelli, A. Galdelli, E. Frontoni, A. Mancini, and S. Moccia, "A deep-learning framework running on edge devices for handgun and knife detection from indoor video-surveillance cameras," *Multimedia Tools and Applications*, vol. 83, pp. 19109-19127, 2024. https://doi.org/10.1007/s11042-023-16231-x
[23] E. M. Onyema, S. Balasubaramanian, C. Iwendi, B. V. V. S. Prasad, and C. D. Edeh, "Remote monitoring system using slow-fast deep convolution neural network model for identifying anti-social activities in

surveillance applications," *Measurement: Sensors*, vol. 27, article no. 100718, 2023. https://doi.org/10.1016/j.measen.2023.100718

[24] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: learning multimodal violence detection under weak supervision," in *Computer Vision–ECCV 2020*. Cham, Switzerland: Springer, 2020, pp. 322-339. https://doi.org/10.1007/978-3-030-58577-8_20

[25] V. D. Huszar, V. K. Adhikarla, I. Negyesi, and C. Krasznay, "Toward fast and accurate violence detection for automated video surveillance applications," *IEEE Access*, vol. 11, pp. 18772-18793, 2023. https://doi.org/10.1109/ACCESS.2023.3245521

[26] Z. Z. Jiang, S. Cui, N. He, K. Li, and L. Tian, L. (2023). Contract selection and piracy surveillance for video platforms in the age of social media. *Production and Operations Management*, 2023. https://doi.org/10.1111/poms.13990

[27] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6479-6488. https://doi.org/10.1109/CVPR.2018.00678

[28] C. H. Demarty, C. Penet, M. Soleymani, and G. Gravier, "VSD, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation," *Multimedia Tools and Applications*, vol. 74, pp. 7379-7404, 2015. https://doi.org/10.1007/s11042-014-1984-4

[29] M. Perez, A. C. Kot, and A. Rocha, "Detection of real-world fights in surveillance videos," in *Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 2662-2666. https://doi.org/10.1109/ICASSP.2019.8683676

[30] R. F. Mansour, J. Escorcia-Gutierrez, M. Gamarra, J. A. Villanueva, and N. Leal, "Intelligent video anomaly detection and classification using faster RCNN with deep reinforcement learning model," *Image and Vision Computing*, vol. 112, article no. 104229, 2021. https://doi.org/10.1016/j.imavis.2021.104229

[31] A. Sakhri, O. Hadji, C. Bouarrouguen, M. Maimour, N. Kouadria, A. Benyahia, E. Rondeau, N. Doghmane, and S. Harize, "Audio-visual low power system for endangered waterbirds monitoring," *IFAC-PapersOnLine*, vol. 55, no. 5, pp. 25-30, 2022. https://doi.org/10.1016/j.ifacol.2022.07.634

[32] M. A. Al-qaness, A. A. Abbasi, H. Fan, R. A. Ibrahim, S. H. Alsamhi, and A. Hawbani, "An improved YOLO-based road traffic monitoring system," *Computing*, vol. 103, pp. 211-230, 2021. https://doi.org/10.1007/s00607-020-00869-8

[33] A. Bouguettaya, H. Zarzour, A. M. Taberkit, and A. Kechida, "A review on early wildfire detection from unmanned aerial vehicles using deep learning-based computer vision algorithms," *Signal Processing*, vol. 190, article no. 108309, 2022. https://doi.org/10.1016/j.sigpro.2021.108309

[34] F. E. Fernandes, L. G. Nonato, and J. Ueyama, "A river flooding detection system based on deep learning and computer vision," *Multimedia Tools and Applications*, vol. 81, pp. 40231-40251, 2022. https://doi.org/10.1007/s11042-022-12813-3

[35] W. F. Pang, Q. H. He, Y. J. Hu, and Y. X. Li, "Violence detection in videos based on fusing visual and audio information," in *Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, 2021, pp. 2260-2264. https://doi.org/10.1109/ICASSP39728.2021.9413686

[36] R. Hu, S. Zhou, Z. R. Tang, S. Chang, Q. Huang, Y. Liu, W. Han, and E. Q. Wu, "DMMAN: a two-stage audio–visual fusion framework for sound separation and event localization," *Neural Networks*, vol. 133, pp. 229-239, 2021. https://doi.org/10.1016/j.neunet.2020.10.003

[37] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 609-617. https://doi.org/10.1109/ICCV.2017.73

[38] S. Mishra, A. K. Gupta, and P. Gupta, "DARE: deceiving audio–visual speech recognition model," *Knowledge-Based Systems*, vol. 232, article no. 107503, 2021. https://doi.org/10.1016/j.knosys.2021.107503

[39] A. Liu, H. Xie, X. Liu, Z. Yin, and S. Liu, "Revisiting audio visual scene-aware dialog," *Neurocomputing*, vol. 496, pp. 227-237, 2022. https://doi.org/10.1016/j.neucom.2021.08.151

[40] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 1725-1732. https://doi.org/10.1109/CVPR.2014.223

[41] L. Pham, D. Ngo, T. Nguyen, P. Nguyen, T. Hoang, and A. Schindler, "An audio-visual dataset and deep learning frameworks for crowded scene classification," in *Proceedings of the 19th International Conference on Content-Based Multimedia Indexing*, Graz, Austria, 2022, pp. 23-28. https://doi.org/10.1145/3549555.3549568

[42] G. Li, Y. Wei, Y. Tian, C. Xu, J. R. Wen, and D. Hu, "Learning to answer questions in dynamic audio-visual scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 19086-19096. https://doi.org/10.1109/CVPR52688.2022.01852

[43] L. N. Do, H. J. Yang, H. D. Nguyen, S. H. Kim, G. S. Lee, and I. S. Na, "Deep neural network-based fusion model for emotion recognition using visual data," *The Journal of Supercomputing*, vol. 77, pp. 10773-10790, 2021. https://doi.org/10.1007/s11227-021-03690-y

[44] N. Vryzas, L. Vrysis, and C. Dimoulas, "Audiovisual speaker indexing for Web-TV automations," *Expert Systems with Applications*, vol. 186, article no. 115833, 2021. https://doi.org/10.1016/j.eswa.2021.115833

[45] Y. Qian, Z. Chen, and S. Wang, "Audio-visual deep neural network for robust person verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1079-1092, 2021. https://doi.org/10.1109/TASLP.2021.3057230

[46] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: learning sound representations from unlabeled video," *Advances in Neural Information Processing Systems*, vol. 29. pp. 892-900, 2016.

[47] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, et al., "The kinetics human action video dataset," 2017 [Online]. Available: https://arxiv.org/abs/1705.06950.

[48] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "VGGSound: a large-scale audio-visual dataset," in *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 721-725. https://doi.org/10.1109/ICASSP40776.2020.9053174

[49] T. Wang and Z. Zhu "Vision-aided automated vibrometry for remote audio, visual, and range sensing," in *Optical Imaging Devices*. Boca Raton, FL: CRC Press, 2016, pp. 65-84.

[50] M. B. Salem and S. J. Stolfo, "Modeling user search behavior for masquerade detection," in *Recent Advances in Intrusion Detection*. Heidelberg, Germany: Springer, 2011, pp. 181-200. https://doi.org/10.1007/978-3-642-23644-0_10

[51] J. Shetty and J. Adibi, "The Enron email dataset database schema and brief statistical report," Information Sciences Institute, University of Southern California, Los Angeles, CA, USA, 2004.

[52] J. Glasser and B. Lindauer, "Bridging the gap: a pragmatic approach to generating insider threat data," in *Proceedings of 2013 IEEE Security and Privacy Workshops*, San Francisco, CA, USA, 2013, pp. 98-104. https://doi.org/10.1109/SPW.2013.37

[53] A. Harilal, F. Toffalini, J. Castellanos, J. Guarnizo, I. Homoliak, and M. Ochoa, "TWOS: a dataset of malicious insider threat behavior based on a gamified competition," in *Proceedings of the 2017 International Workshop on Managing Insider Security Threats*, Dallas, TX, USA, 2017, pp. 45-56. https://doi.org/10.1145/3139923.3139929

[54] M. Schonlau, W. DuMouchel, W. H. Ju, A. F. Karr, M. Theus, and Y. Vardi, "Computer intrusion: detecting masquerades," *Statistical Science*, vol. 16, no. 1, pp. 58-74, 2001.

[55] M. N. Al-Mhiqani, R. Ahmad, Z. Zainal Abidin, W. Yassin, A. Hassan, K. H. Abdulkareem, N. S. Ali, and Z. Yunos, "A review of insider threat detection: classification, machine learning techniques, datasets, open challenges, and recommendations," *Applied Sciences*, vol. 10, no. 15, article no. 5208, 2020. https://doi.org/10.3390/app10155208

[56] S. Yuan and X. Wu, "Deep learning for insider threat detection: review, challenges and opportunities," *Computers & Security*, vol. 104, article no. 102221, 2021. https://doi.org/10.1016/j.cose.2021.102221

[57] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6450-6459. https://doi.org/10.1109/CVPR.2018.00675

[58] A. M. Buttar, M. Bano, M. A. Akbar, A. Alabrah, and A. H. Gumaei, "Toward trustworthy human suspicious activity detection from surveillance videos using deep learning," *Soft Computing*, 2023. https://doi.org/10.1007/s00500-023-07971-x

[59] W. Ahmed, U. Naeem, M. H. Yousaf, and S. A. Velastin, "Lightweight CNN and GRU network for real-time action recognition," in *Proceedings of 2022 12th International Conference on Pattern Recognition Systems (ICPRS)*, Saint-Etienne, France, 2022, pp. 1-7. https://doi.org/10.1109/ICPRS54038.2022.9853854

[60] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. Van Gool, "Temporal 3d convnets: New architecture and transfer learning for video classification," 2017 [Online]. Available: https://arxiv.org/abs/1711.08200.

[61] A. Bousmina, M. Selmi, M. A. Ben Rhaiem, and I. R. Farah, "A hybrid approach based on GAN and CNN-LSTM for aerial activity recognition," *Remote Sensing*, vol. 15, no. 14, article no. 3626, 2023. https://doi.org/10.3390/rs15143626

[62] M. Bubashait and N. Hewahi, "Urban sound classification using DNN, CNN & LSTM a comparative approach," in Proce*edings of 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, Zallaq, Bahrain, 2021, pp. 46-50. https://doi.org/10.1109/3ICT53449.2021.9581339

[63] B. Bahmei, E. Birmingham, and S. Arzanpour, "CNN-RNN and data augmentation using deep convolutional generative adversarial network for environmental sound classification," *IEEE Signal Processing Letters*, vol. 29, pp. 682-686, 2022. https://doi.org/10.1109/LSP.2022.3150258

[64] F. Demir, D. A. Abdullah, and A. Sengur, "A new deep CNN model for environmental sound classification," *IEEE Access*, vol. 8, pp. 66529-66537, 2020. https://doi.org/10.1109/ACCESS.2020.2984903

[65] G. Petmezas, G. A. Cheimariotis, L. Stefanopoulos, B. Rocha, R. P. Paiva, A. K. Katsaggelos, and N. Maglaveras, "Automated lung sound classification using a hybrid CNN-LSTM network and focal loss function," *Sensors*, vol. 22, no. 3, article no. 1232, 2022. https://doi.org/10.3390/s22031232

[66] Z. Zhang, S. Xu, S. Zhang, T. Qiao, and S. Cao, "Attention based convolutional recurrent neural network for environmental sound classification," *Neurocomputing*, vol. 453, pp. 896-903, 2021. https://doi.org/10.1016/j.neucom.2020.08.069

[67] A. L. Cramer, H. H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: design choices for deep audio embeddings," in *Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 3852-3856. https://doi.org/10.1109/ICASSP.2019.8682475