DOI: 10.19884/j.1672-5220.202505004

Review of Research on Black-Box Model Inversion Attack

HE Shifei, WANG Chen, LUO Xin*, XU Guangwei*
School of Information and Intelligent Science, Donghua University, Shanghai 201620, China

Abstract: The development of artificial intelligence has given rise to the paradigm of machine learning as a service, enabling users to either train the requisite models or employ pre-existing models to make predictions based on the training data that users provide. Despite various measures taken to prevent privacy leakage in machine learning, such as data deletion and anonymization, model inversion attack (MIA) remains capable of inferring sensitive information from user data to a certain extent. This paper reviews the current state of research on black-box MIA, with a particular focus on confidence-based MIA and label-based MIA. It further analyzes the MIA methodologies applied to the emerging modes such as text and audio in recent years, filling the gap in the current review of such modes. Finally, the paper discusses the future challenges and research directions for black-box MIA.

Keywords: machine learning; model inversion attack; black-box model; privacy protection
CLC number: TP181 **Document code:** A
Article ID: 1672-5220(2026)04-0049-10

0 Introduction

Machine learning (ML) is a significant method in artificial intelligence (AI). This technology enables computers to automatically acquire and extract valuable information from data by employing appropriate methods and algorithms. ML has been extensively applied in various fields, including facial image analysis, medical records management, and financial information analysis[1-3]. In the context of the continuous generation of vast amounts of data, prominent Internet companies have not only employed ML methodologies for internal data analysis and research but also made ML as a service (MLaaS) available to the general public.

It has been demonstrated that small and medium-sized enterprises frequently encounter limitations in their capacity to train ML models due to the substantial computational demands of model training and the need for large-scale datasets. Consequently, these enterprises often resort to leveraging MLaaS platforms to train models using their proprietary data. Furthermore, users lacking training data can query pre-trained models on MLaaS platforms by paying a fee. However, while ML provides users with numerous valuable services, the training data used in ML often contains large amounts of private or sensitive information, leading to security issues associated with ML. Attackers can infer model or training data by accessing the platform query interface, thereby causing potential risks of data privacy leakage, such as the exposure of user medical information and facial images[4-5]. A paradigmatic example is the work of Fredrikson et al.[6], who have obtained prediction labels from the target model to infer the genotypes of patients in the training data. At present, a multitude of privacy attacks on ML exist, including member inference attacks[7], attribute inference attacks[8], gradient inversion attacks[9], and model capacity abuse attacks[10]. Consequently, defensive measures have emerged, such as the application of differential privacy[11-12] to introduce a small amount of noise to the data, thereby obfuscating the input and impeding the mapping from input to observed values. Additionally, ML methods such as dropout[13], regularization[14], and adversarial training[15] have been used to mitigate attacks. Despite the implementation of privacy protection measures for data, ML models have the capacity to retain training data, which may result in privacy leakage[16]. ML privacy and security issues have been identified as primary concerns, with model inversion attack (MIA) being a notable example.

MIA uses model outputs to infer sensitive attribute information present in the training data, resulting in a substantial compromise to data privacy[17-18]. At present, two primary categories of MIA have been identified in Ref. [19], i. e., confidence-based MIA and label-based MIA. This paper analyzes the fundamental concept of MIA and focuses on black-box MIA. Different from the existing literature, which mainly focuses on the research of image modes, this paper further analyzes the MIA methodologies applied to the emerging modes such as texts and audio in recent years, filling the gap in the current review of such modes and highlighting the challenges and research potential of MIA in multimodal intelligent systems. Finally, the future challenges and development trends of black-box MIA are

Received date: 2025-05-11
Foundation item: Natural Science Foundation of Shanghai, China (No. 19ZR1402000)
* Correspondence should be addressed to XU Guangwei, email: gwxu@dhu.edu.cn; LUO Xin, email: xluo@dhu.edu.cn
Citation: He S F, Wang C, Luo X, et al. Review of research on black-box model inversion attack[J]. *Journal of Donghua University (English Edition)*, 2026, 43(4): 49-58.

put forward.

The structure of the paper is as follows. Section 1 provides an overview of MLaaS and discusses the threats it faces. Section 2 introduces various black-box MIA schemes in detail. Section 3 describes the practical applications of MIA, as well as commonly employed datasets and evaluation metrics. Section 4 discusses challenges and future trends. Section 5 summarizes the paper.

1 Potential Threats of MLaaS

In the context of MLaaS, data owners upload training datasets to the operational environment of the cloud service provider and leverage its computing capabilities to train the target model. It is important to note that data proprietors have the prerogative to employ the trained target model to provide inference services for the uploaded datasets.

1.1 Potential threats

The release of trained ML models as public services enables malicious actors to access target models by leveraging known datasets. In the threat model, attackers abuse their access to the target model T to obtain sensitive information by inputting information from a subset of the dataset. The threat model process is outlined as follows.

1) Obtain the attack dataset. The attacker selects different sensitive attributes based on the original dataset provided by the user. Then, the attacker queries T and compares the prediction labels returned by T for the input records with the labels in the original dataset. The input records with matching labels are filtered out to construct the attack dataset.

2) Train the attack model A. The attacker treats the non-sensitive attributes and labels in the attack dataset as features and the sensitive attributes as labels to train A.

3) Infer the sensitive attributes in the target records. The attacker utilizes records containing non-sensitive attributes as input to query A. Then, the output of A is used to obtain the sensitive attributes. Through these steps, the attacker can successfully infer the sensitive attributes in the original dataset.

1.2 MIA

MIA leverages the target model output information of MLaaS to infer sensitive attributes in the training data. The objective of the attacker is to maximize the posterior probability or similarity to authentic samples. A typical MIA model is proposed as follows.

Given a model f_θ: $X \rightarrow Y$, where θ is the model parameter, the input space $X \subseteq \mathbf{R}^d$ is defined over d-dimensional real Euclidean space (d represents feature dimensionality, and Y represents the output space). The model is trained using a set of training samples $D_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in X$ represents the ith sample, $y_i \in Y$ is its label, and n is the number of training samples. Let the input features be divided into the public part $x_{i,\text{pub}} \in X_{\text{pub}}$ and the private part $x_{i,\text{priv}} \in X_{\text{priv}}$, i.e., $x_i = (x_{i,\text{pub}}, x_{i,\text{priv}})$. The objective of the attacker is to infer the approximate or true value of x_{priv}, given that x_{pub} is known and the attacker can query model f_θ. The MIA is represented as

$$\hat{x}_{\text{priv}} = \arg\max_{x \in X_{\text{priv}}} P_r[x \mid f_\theta(x_{\text{pub}}, x), x_{\text{pub}}, K_{\text{adv}}], \quad (1)$$

where $\hat{x}_{\text{priv}}$ represents the private feature inferred by the attacker via membership inference attack; P_r represents conditional posterior probability; K_{adv} represents the set of prior knowledge held by the attacker, including the distribution prior of the training data, the categories of the attack samples, the output format, and other information.

Although MIA shares certain similarities with other typical attacks, e.g., membership inference attacks and attribute inference attacks, there are fundamental differences in their core objectives, attack methods, and output results. MIA aims to recover some or all of the input features in training samples. Attackers access the model's prediction outputs and, given partial input information or category labels, approximate the missing sensitive features. Conversely, MIA prioritizes ascertaining whether a particular sample is utilized during training. Attribute inference attacks seek to extract specific global attributes of the entire training set, emphasizing statistical inference about the distribution of the training data[20].

The process of accessing the target model can be classified as either white-box or black-box[21]. Preliminary studies on MIA predominantly focused on white-box attacks. However, these studies made the unrealistic assumption that attackers could access all components of the target model, which was impractical in real-world scenarios. Consequently, many researchers have redirected their attention to black-box MIA[22], which are capable of obtaining access to the target classification black-box model. These attacks rely exclusively on the model's prediction labels to execute their malicious actions and to infer unknown sensitive attributes in the training data. For instance, the prediction of label results $\hat{y} = f_\theta(x)$ or the predicted probability distribution $f_\theta(x) = [p_1 \quad p_2 \quad \cdots \quad p_C] \in \Delta^{C-1}$, where p_C represents class-C predicted probability, and Δ^{C-1} represents $(C-1)$-dimensional probability simplex. Thus, the attack model can be expressed as

$$\hat{x}_{\text{priv}} = \arg\max_{x \in X_{\text{priv}}} S(f_\theta(x_{\text{pub}}, x), \tilde{y}), \quad (2)$$

where $S(\cdot, \cdot)$ represents a function to measure the similarity between the model output and the attacker's expected output (such as a specific confidence distribution $\tilde{y}$). During the attack process, the attacker first inputs carefully designed data into the target model to obtain the prediction labels returned by the model. Subsequently, employing these labels in conjunction with

prior knowledge or specialized algorithms, the adversary methodically constructs an understanding of the model's internal structure and sensitive attributes, ultimately achieving the inference of sensitive attributes. A typical MIA process is shown in Fig. 1, which generally encompasses three steps[16].

1) Obtain attack data D_S. Extract data records from the training dataset, perform an extending process on each record to generate candidate-sensitive-attribute variants. Then, based on the relationship between sensitive attributes and non-sensitive attributes in the training dataset, use different methods to identify strong correlations between sensitive and non-sensitive attributes in each record, and consequently generate an attack dataset.

2) Train attack model A using D_S. The adversary inputs a set of target records from D_S, which are d-dimensional vectors $[x_2 \quad x_3 \quad \cdots \quad x_d \quad y]$ composed of non-sensitive attributes and labels, and outputs predictions of the sensitive attribute x_1 to train the attack model. The key to this step lies in establishing an association between sensitive attributes and other non-sensitive attributes (including the predicted labels of the target model) based on the strong correlation between the x_1 and non-sensitive attribute vectors $[x_2 \quad x_3 \quad \cdots \quad x_d \quad y]$.

3) Infer sensitive attributes in target records using A. Take the non-sensitive attributes and labels of a target record as input, and infer the sensitive attributes in the target record after querying the attack model.

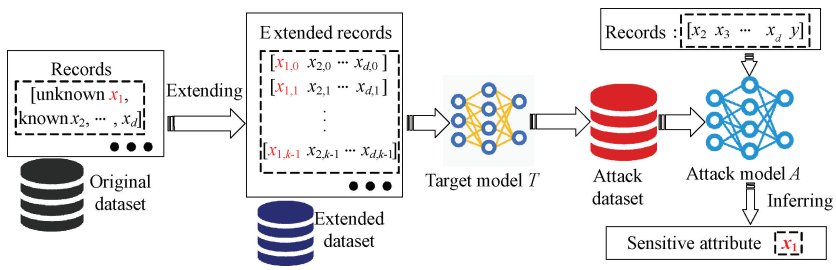


Fig. 1 Typical MIA process

2 Black-Box MIA Schemes

To systematically organize existing black-box MIA schemes and facilitate rapid identification of existing research, this section introduces a corresponding taxonomy of the black-box attack paradigm, as illustrated in Fig. 2. This taxonomy classifies existing solutions into two primary categories based on the output type exploited by the attack, i. e., confidence-based MIA and label-based MIA. Within the spectrum of confidence-based MIA, two distinct forms are identified: full confidence output and top-k confidence output. All these attacks are then classified into two categories based on their respective attack strategies: generative model-based attacks and optimization-based attacks.

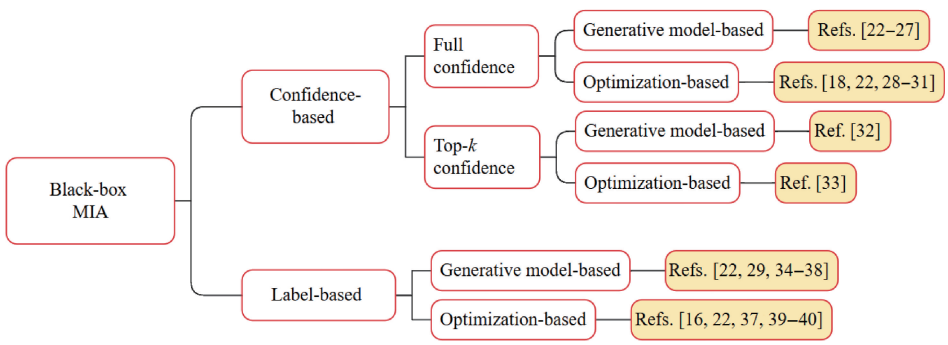


Fig. 2 Taxonomy of black-box MIA

2.1 Confidence-based MIA

Attackers gain access to the target model in order to obtain the corresponding output. Subsequently, the labels and confidence scores yielded by the target model are used to ascertain the maximum attainable values of sensitive attributes. This process enables the acquisition of sensitive attribute information. In 2015, Fredrikson et al.[18] proposed the use of confidence scores for MIA. The input images of a face recognition model were successfully reconstructed by attacking linear models and neural networks and leveraging the confidence outputs of the models. Yin et al.[28] investigated a confidence-based MIA within a black-box environment and demonstrated that, even with only confidence outputs (as opposed to full model access), attackers could effectively reconstruct the training data of the target model, particularly in the context of deep learning models. Yang et al.[23] proposed a generative MIA algorithm that integrated generative adversarial networks with confidence information to effectively reconstruct high-quality images by collecting

confidence outputs. Yin et al.[24] conducted a study to ascertain the significance of confidence in model outputs, with a particular focus on knowledge distillation scenarios. The loss function of the model was modified to enable the use of confidence outputs, thereby facilitating more effective MIA. In a related study, Zhang et al.[25] progressively relaxed assumptions regarding the capabilities of the attackers, proposed a variety of attack algorithms and demonstrated their effectiveness across multiple datasets. In addition, the authors proposed two defense mechanisms to mitigate the risk of MIA by reducing model overfitting. In their seminal work, Liu et al.[26] proposed a novel confidence-based MIA algorithm that effectively circumvented the limitations of black-box MIA on data distribution assumptions. Table 1 presents a comparative analysis of typical confidence-based MIA.

Table 1 Summary of typical confidence-based MIA

Reference	Feature	Limitation
[18]	Use of confidence scores	Low accuracy under defensive technologies such as dropout
[23]	Inversion of deep neural networks at a high success rate	Reliance on extracting prior knowledge from public datasets
[24]	Image synthesis without using additional training datasets	Reduction in image diversity
[25]	Realization of model and data-independent attack	Suitability for neural network models
[26]	Attack performance improvement under different data distribution scenarios	Requirement of frequent queries to target models
[28]	Restoration of original images from a large batch of average gradients	High consumption of computing resources

2.2 Label-based MIA

Label-based MIA represents a category of attacks that utilize the classification labels of a model to infer and reconstruct training data. The model's label values are then used to infer sensitive attributes. Xu et al.[34] proposed a label-based MIA algorithm by combining auxiliary information (e.g., publicly available similar datasets) with label information, and then leveraged the similarity between the data to infer and reconstruct the training data. Zhu et al.[35] proposed a strategy for launching attacks with minimal information. The strategy utilizes the discrepancy between sample data and actual data to calculate the most probable inferred sample results. Liu et al.[36] employed a conditional diffusion model to facilitate an efficient label-based MIA. This algorithm generates realistic training data samples through a diffusion process and relies exclusively on the classification labels of the model to complete the reconstruction of the training data. In comparison with conventional methodologies, conditional diffusion models are capable of capturing more intricate data distribution characteristics, thereby manifesting enhanced reconstruction capabilities on high-dimensional data (e.g., images). Ye et al.[37] proposed a novel two-stage black-box MIA algorithm. This improvement has the potential to enhance the efficacy of black-box MIA to a considerable degree. They also provided an efficient solution for black-box MIA, while also conducting robustness analysis and exploring possible defensive measures. Zhou et al.[39] conducted a thorough analysis of the factors contributing to the suboptimal accuracy of learning-based MIA. The proposed algorithm aims to enhance the efficacy of these attacks by incorporating adversarial examples and regularization losses. Mehnaz et al.[16] proposed two novel black-box model inversion attribute attack algorithms. Through a series of evaluation attacks on multiple datasets, they demonstrated the superiority of the proposed algorithms in terms of sensitive attribute inference. It is noteworthy that the prediction label-based MIA has achieved a comparable effect to the confidence-based MIA. Table 2 presents a comparative analysis of typical label-based MIA.

Table 2 Summary of typical label-based MIA

Reference	Feature	Limitation
[16]	Proposal of two new black-box model inversion attribute inference attacks	Susceptibility to quality of attack dataset and distribution of specific groups
[34]	Sample extraction from a more general data distribution to form an auxiliary set	Dependence on distribution difference between auxiliary set and training set
[35]	Association of MIA with model error rate	Dependence on distribution difference between auxiliary set and training set
[36]	Capture of more intricate data distribution characteristics	Requirement of long training time for diffusion models
[37]	Excellent performance in black-box scenarios and computational efficiency	Reliance on features of auxiliary data
[39]	Improvement of attack accuracy in black-boxes and classification accuracy	Dependence on adversarial sample generation and introduction of pseudo labels

3 Application and Evaluation

3.1 Application

MIA has made significant progress in various data domains[41], including images, texts, and audio, demonstrating cross-domain privacy threat capabilities. An examination of the existing attack methods and defense strategies for the three data scenarios is summarized as follows.

3.1.1 Black-box MIA on images

In the image domain, black-box MIA aims to reconstruct sensitive input images solely from model outputs, without accessing internal parameters or intermediate representations. Guo et al.[41] optimized targets using model confidence scores and reconstructed high-fidelity face images without accessing internal information. Liu et al.[29] proposed the prediction-to-image (P2I) framework, which enhanced the consistency and clarity of image reconstruction by aligning the prediction results with the latent space of the generator. Ye et al.[37] introduced a coarse-to-fine model inversion (C2FMI) attack process that used a coarse-to-fine architecture to gradually optimize image details in stages, significantly improving the quality of black-box reconstruction. In order to enhance the efficacy of attacks and minimize query costs, Han et al.[40] formulated black-box inversion as a reinforcement learning task and proposed the reinforcement learning-based black-box model inversion (RLBMI) attack framework, which facilitated a more efficient and strategic query process. Li et al.[33] incorporated long-tail distribution-aware initialization in structured model inversion via latent embedding (SMILE) attacks, effectively reducing the search space and achieving target image restoration with minimal queries.

In response to the above attacks, researchers have proposed various defense measures to improve the privacy protection capability of the models. Differential privacy mechanisms limit the model memory of the training samples by injecting noise during the training process[42]. Transfer learning methods could reduce the direct correlation between training samples and outputs by utilizing pre-trained models, thereby enhancing robustness[43]. In addition, solutions based on the principle of information bottleneck, such as sensitive-sample defense (SSD), suppress sensitive feature leakage from an information theory perspective[44].

3.1.2 Black-box MIA on texts

In the text domain, black-box MIA reconstructs sensitive textual inputs by exploiting correlations between model outputs and underlying data representations. Petrov et al.[45] proposed the data-aware gradient extraction and reconstruction (DAGER) algorithm, which leveraged the low-rank properties of self-attention gradients in large language models and discrete word embeddings to achieve accurate batch reconstruction of input texts, overcoming the limitation of earlier methods that could only reconstruct short sequences. Zhang et al.[27] introduced a system that integrated external data and generative models to effectively reconstruct training text under black-box feedback. The efficacy and versatility of this category of attack on a range of text data were exemplified in Ref. [27]. These methods demonstrate that black-box MIA in the text domain poses a significant privacy threat.

To counter such attacks, researchers have proposed various effective defense strategies. The masking mechanism effectively enhances privacy protection in cross-linguistic scenarios by randomly masking multilingual embedding vectors[46]. At the same time, differential privacy mechanisms have also been introduced into the text domain to achieve selective protection of sensitive information, weakening the effectiveness of MIA while maintaining the model performance[47].

3.1.3 Black-box MIA on audio

In the audio domain, black-box MIA can recover sensitive speech information by leveraging correlations between model outputs and acoustic features. Pizzi et al.[30] first extended MIA to automatic speech recognition (ASR) systems. They proposed the "sliding model inversion" method, which effectively reconstructed complete speech samples by gradually inverting overlapping audio segments. Chen et al.[31] investigated a combination of feature conversion and adversarial training defense techniques in a voiceprint recognition system, and found that feature-level conversion combined with adversarial training was effective in improving defense performance and could effectively block black-box query-based inversion attacks. These studies demonstrate that despite the highly dynamic and complex structure of audio data, black-box MIA can still recover highly sensitive information by exploiting the mapping relationship between the model behavior and acoustic characteristics, posing a significant threat to the privacy protection of intelligent speech systems.

To address threats, researchers have also proposed various defense mechanisms. One type of mechanism reduces the effectiveness of inversion through input transformation to interfere with reconstruction paths and assist in attack detection[12]. Another type of mechanism introduces adversarial training, which can effectively suppress identity-related privacy leakage[48]. In addition, differential privacy technology has also been applied to end-to-end ASR systems, achieving the goal of significantly improving the system's robustness against MIA[49].

3.2 Datasets

Some representative datasets are selected to evaluate attack effectiveness, defense mechanism robustness, and applicability to real-world scenarios, as shown in Table 3.

Table 3 Summary of datasets used for MIA

Dataset	Type	Description	Reference
CIFAR-10	Image	32 × 32 pixels RGB images from 10 classes	[22–23,28]
MNIST	Image	28 × 28 pixels grayscale digit images, 60 000 train/10 000 test	[18, 22, 25, 50]
CelebA	Image	200 000 celebrity faces with 40 binary attributes	[22, 25, 34, 51]
ImageNet	Image	1.4×10^7 labeled images across 20 000 categories	[52]
Social Survey	Text	Data on demographics, opinions, and social behavior	[18, 53]
FiveThirtyEight	Text	Journalism-curated datasets on politics, sports, and more	[18, 54]
SST-2	Text	Stanford Sentiment Treebank with binary labels	[55–56]
LibriSpeech	Audio	1 000 h of English audiobook data	[30, 57]
CommonVoice	Audio	Multilingual speech dataset with speaker metadata	[31, 58]

3.3 Evaluation metrics

The evaluation metrics range from traditional pixel-level error metrics, e.g., mean square error (MSE), to more advanced metrics in feature and perceptual spaces, e.g., learned perceptual image patch similarity (LPIPS) and Fréchet inception distance (FID), as shown in Table 4.

Table 4 Summary of evaluation metrics

Metric	Modality	Advantage	Reference
Accuracy	Image/audio/text	General purpose	[20, 26]
Precision	Image/audio/text	General purpose	[18]
Recall	Image/audio/text	General purpose	[18]
MSE	Image/audio	Simple and efficient, with direct quantification of reconstruction deviation	[23]
LPIPS	Image	Capture of high-level semantic differences	[26, 28]
Peak signal-to-noise ratio (PSNR)	Image/audio	Traditional measure of fidelity	[22, 28]
FID	Image	Generation quality at distribution level	[26, 34]

4 Research Challenges and Future Directions

The application of ML involves a large amount of personal privacy data, which poses a huge risk of sensitive attribute information privacy leakage in the training data when threatened by MIA. Therefore, further research is needed to effectively carry out MIA on sensitive attributes, and to establish corresponding effective defense measures against MIA for resisting sensitive attribute leakage.

4.1 Inference of multiple values for single sensitive attribute

Current MIA is mainly applied to classification machine learning models. Attackers use the labels output by the model, combined with the strong relationship between the labels and sensitive attribute, to infer the sensitive attribute. However, existing label-based MIA can only infer a sensitive attribute when the attribute has two values[38, 59-61]. When there are multiple values for the sensitive attribute, MIA is less effective because there is no one-to-one correspondence between the output labels and sensitive attribute when querying the black-box target model. In practical scenarios, a large number of data attributes have multiple values. It is of great practical significance to implement effective MIA when sensitive attributes have multiple values.

4.2 Inference of values for multiple sensitive attributes

The existing body of research on MIA has predominantly centered on scenarios where a single sensitive attribute is present in the training data. However, in certain datasets, there may be more than one sensitive attribute, meaning that multiple attributes may contain sensitive information. In the context of MIA, the objective is to establish a correlation between the output information and a specific sensitive attribute. This is typically achieved by leveraging the output information of the target model, such as confidence scores or labels. However, when there are multiple sensitive attributes, it may be difficult to infer the value of each sensitive attribute by using the connection between the output information of the target model and the multiple sensitive attributes. Consequently, the inference of multiple sensitive attributes through MIA poses significant privacy risks to the target model's output information, thereby making this a prominent research direction.

4.3 Deep learning-based MIA

Current MIA predominantly targets ML classification models. This is primarily due to the fact that ML models are characterized by simplicity in application, a clear structure, and the ability to more effectively determine the relationship between output information and sensitive attributes. Nevertheless, when certain training data is employed in the context of deep learning models, the

resulting outputs may, in such cases, reveal sensitive attribute information inherent in the training data. In such cases, the utilization of deep learning models as attack models may prove to be a more effective approach for the inference of sensitive attributes. In the case of a deep learning model, the target model utilizes the training data to acquire feature information through the multi-layer structure of the neural network, subsequently producing the corresponding classification results. Consequently, modeling the neural network structure between the output information and sensitive attributes to improve the effectiveness of MIA has become a significant research direction.

4.4 Defense measures

Although many defense measures have performed well against MIA in specific scenarios, there are still many problems to be solved. First, current defense methods are mostly targeted at specific modalities or tasks, and lack generalizability. In the future, a unified defense strategy suitable for multimodal and multitask scenarios should be explored. Second, how to protect privacy while minimizing the impact on model performance remains an important research problem. Finally, in the face of the increasing black-box attack capability, dynamic defense mechanisms, such as game theory or adaptive learning methods, can be introduced to automatically adjust defense strategies according to attack characteristics.

5 Conclusions

This paper presents a summary of the existing literature on MIA, with a discussion of the application of MIA in MLaaS scenarios, and a classification and analysis of different algorithms. In order to achieve a more comprehensive understanding of MIA, a summary of typical applications in fields such as images, texts, and audio is given. Summaries of the datasets that are extensively utilized in related research and evaluation metrics that are commonly employed are also provided. These summaries will serve as a foundational reference for subsequent algorithm evaluation and comparison. Furthermore, this paper has explored several prospective research, including the implementation of MIA in the inference of single sensitive attribute with multiple values, the inference of values for multiple sensitive attributes, and the use of deep learning-based MIA as both a challenge and a future direction for attack models. These endeavors not only promote the application and development of MIA, but also introduce novel challenges for security and privacy protection in the field of AI.

References

[1] Raju K, Rao B C, Saikumar K, et al. An optimal hybrid solution to local and global facial recognition through machine learning [M]//A Fusion of Artificial Intelligence and Internet of Things for Emerging Cyber Systems. Cham: Springer International Publishing, 2021: 203-226.

[2] Wood A, Najarian K, Kahrobaei D. Homomorphic encryption for machine learning in medicine and bioinformatics [J]. *ACM Computing Surveys*, 2020, 53(4): 1-35.

[3] Bello O A, Folorunso A, Ejiofor O E, et al. Machine learning approaches for enhancing fraud prevention in financial transactions [J]. *International Journal of Management Technology*, 2023, 10(1): 85-108.

[4] Weiss J C, Natarajan S, Peissig P L, et al. Machine learning for personalized medicine: predicting primary myocardial infarction from electronic health records [J]. *AI Magazine*, 2012, 33(4): 33-45.

[5] Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2015: 815-823.

[6] Fredrikson M, Lantz E, Jha S, et al. Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing [J]. *Proceedings of the USENIX Security Symposium*, 2014, 2014: 17-32.

[7] Shokri R, Stronati M, Song C Z, et al. Membership inference attacks against machine learning models [C]//2017 IEEE Symposium on Security and Privacy (SP). Piscataway, NJ: IEEE, 2017: 3-18.

[8] Melis L, Song C Z, De Cristofaro E, et al. Exploiting unintended feature leakage in collaborative learning [C]//2019 IEEE Symposium on Security and Privacy (SP). Piscataway, NJ: IEEE, 2019: 691-706.

[9] Zhu L G, Liu Z J, Han S. Deep Leakage from Gradients [C]//Proceedings of the 32nd International Conference on Advances in Neural Information Processing Systems. Vancouver: NeurIPS, 2019: 14747-14756.

[10] Song C Z, Ristenpart T, Shmatikov V. Machine learning models that remember too much [C]// Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017: 587-601.

[11] Dwork C. Differential privacy [C]//Automata, Languages and Programming. Berlin, Heidelberg: Springer, 2006: 1-12.

[12] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy [C]// Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2016: 308-318.

[13] Murphy K, Schölkopf B, Srivastava N, et al. Dropout: a simple way to prevent neural networks

from overfitting [J]. *Journal of Machine Learning Research*, 2014, 15(1): 1929-1958.
[14] Wang T H, Zhang Y H, Jia R X. Improving robustness to model inversion attacks via mutual information regularization [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35(13): 11666-11673.
[15] Wang X B, Hou R, Zhu Y F, et al. NPUFort: a secure architecture of DNN accelerator against model inversion attack [C]//Proceedings of the 16th ACM International Conference on Computing Frontiers. New York: ACM, 2019: 190-196.
[16] Mehnaz S, Dibbo S V, De Viti R, et al. Are your sensitive attributes private? Novel model inversion attribute inference attacks on classification models[C]//31st USENIX Security Symposium. Berkeley: USENIX Association, 2022: 4579-4596.
[17] Liu R X, Chen H, Guo R Y, et al. Survey on privacy attacks and defenses in machine learning [J]. *Journal of Software*, 2020, 31(3): 866-892. (in Chinese)
[18] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures [C]// Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2015: 1322-1333.
[19] Kahla M, Chen S, Just H A, et al. Label-only model inversion attacks via boundary repulsion [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2022: 15025-15033.
[20] Ma W L, Wang D R, Song Y L, et al. TrapNet: model inversion defense via trapdoor [J]. *IEEE Transactions on Information Forensics and Security*, 2025, 20: 4469-4483.
[21] Yang W C, Wang S, Wu D, et al. Deep learning model inversion attacks and defenses: a comprehensive survey [J]. *Artificial Intelligence Review*, 2025, 58(8): 242.
[22] Bao H, Wei K M, Wu Y D, et al. Distributional black-box model inversion attack with multi-agent reinforcement learning[J]. *IEEE Transactions on Information Forensics and Security*, 2025, 20: 5425-5437.
[23] Yang Z Q, Zhang J Y, Chang E C, et al. Neural network inversion in adversarial setting via background knowledge alignment [C]// Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2019: 225-240.
[24] Yin H X, Molchanov P, Alvarez J M, et al. Dreaming to distill: data-free knowledge transfer via DeepInversion [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2020: 8712-8721.
[25] Zhang Y H, Jia R X, Pei H Z, et al. The secret revealer: generative model-inversion attacks against deep neural networks [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2020: 250-258.
[26] Liu R K, Wang D, Ren Y Z, et al. Unstoppable attack: label-only model inversion via conditional diffusion model [J]. *IEEE Transactions on Information Forensics and Security*, 2024, 19: 3958-3973.
[27] Zhang R S, Hidano S, Koushanfar F. Text revealer: private text reconstruction via model inversion attacks against transformers [PP/OL]. arXiv (2022-09-21) [2025-04-21]. https://arxiv.org/abs/2209.10505.
[28] Yin H X, Mallya A, Vahdat A, et al. See through gradients: image batch recovery via GradInversion [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2021: 16332-16341.
[29] Liu Y F, Zhang W Q, Wu D Y, et al. Prediction exposes your face: black-box model inversion via prediction alignment [C]// Computer Vision – ECCV 2024. Cham: Springer, 2025: 288-306.
[30] Pizzi K, Boenisch F, Sahin U, et al. Introducing model inversion attacks on automatic speaker recognition [PP/OL]. arXiv (2023-01-09) [2025-04-21]. https://arxiv.org/abs/2301.03206.
[31] Chen G K, Zhao Z, Song F, et al. Towards understanding and mitigating audio adversarial examples for speaker recognition [J]. *IEEE Transactions on Dependable and Secure Computing*, 2023, 20(5): 3970-3987.
[32] Li O X, Hao Y B, Wang Z C, et al. Model inversion attacks through target-specific conditional diffusion models [PP/OL]. arXiv (2024-07-16)[2025-04-21]. https://arxiv.org/abs/2407.11424.
[33] Li Z A, Zhang H G, Wang J, et al. From head to tail: efficient black-box model inversion attack via long-tailed learning [C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2025: 29288-29298.
[34] Xu Q, Arafin M T, Qu G. An approximate memory based defense against model inversion attacks to neural networks [J]. *IEEE Transactions on Emerging Topics in Computing*, 2022, 10(4): 1733-1745.
[35] Zhu T Q, Ye D Y, Zhou S, et al. Label-only model inversion attacks: attack with the least

information [J]. *IEEE Transactions on Information Forensics and Security*, 2023, 18: 991-1005.
[36] Liu Z T, Chen S T. Trap-mid: trapdoor-based defense against model inversion attacks [J]. *Advances in Neural Information Processing Systems*, 2024, 37: 88486-88526.
[37] Ye Z P, Luo W J, Naseem M L, et al. C2FMI: corse-to-fine black-box model inversion attack [J]. *IEEE Transactions on Dependable and Secure Computing*, 2024, 21(3): 1437-1450.
[38] Nguyen N B, Chandrasegaran K, Abdollahzadeh M, et al. Label-only model inversion attacks via knowledge transfer [J]. *Advances in Neural Information Processing Systems*, 2023, 36: 68895-68907.
[39] Zhou S, Zhu T Q, Ye D Y, et al. Boosting model inversion attacks with adversarial examples [J]. *IEEE Transactions on Dependable and Secure Computing*, 2024, 21(3): 1451-1468.
[40] Han G, Choi J, Lee H, et al. Reinforcement learning-based black-box model inversion attacks [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2023: 20504-20513.
[41] Guo P X, Zeng S, Chen W H, et al. A new federated learning framework against gradient inversion attacks [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, 39(16): 16969-16977.
[42] Zhang Q C, Ma J, Xiao Y H, et al. Broadening differential privacy for deep learning against model inversion attacks [C]//2020 IEEE International Conference on Big Data. Piscataway, NJ: IEEE, 2021: 1061-1070.
[43] Ho S T, Hao K J, Chandrasegaran K, et al. Model inversion robustness: can transfer learning help [C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2024: 12183-12193.
[44] Zhuang T Q, Yu H Y, Qiu Y X, et al. Stealthy shield defense: a conditional mutual information-based approach against black-box model inversion attacks [C]//The Thirteenth International Conference on Learning Representations (ICLR). Cham: Springer Nature Switzerland, 2025: p0DjhjPXl3.
[45] Petrov I, Dimitrov D I, Baader M, et al. Dager: exact gradient inversion for large language models [J]. *Advances in Neural Information Processing Systems*, 2024, 37: 87801-87830.
[46] Chen Y Y, Lent H, Bjerva J. Text embedding inversion security for multilingual language models [PP/OL]. arXiv(2024-01-22)[2025-04-21]. https://arxiv.org/abs/2401.12192.
[47] Shi W Y, Shea R, Chen S, et al. Just fine-tune twice: selective differential privacy for large language models [PP/OL]. arXiv(2022-04-15)[2025-04-21]. https://arxiv.org/abs/2204.07667.
[48] Gharib S, Tran M, Luong D, et al. Adversarial representation learning for robust privacy preservation in audio[J]. *IEEE Open Journal of Signal Processing*, 2024, 5: 294-302.
[49] Yang C H, Chen I F, Stolcke A, et al. An experimental study on private aggregation of teacher ensemble learning for end-to-end speech recognition [C]//2022 IEEE Spoken Language Technology Workshop (SLT). Piscataway, NJ: IEEE, 2023: 1074-1080.
[50] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. *Proceedings of the IEEE*, 1998, 86(11): 2278-2324.
[51] Norc at the University of Chicago. The general social survey[EB/OL]. (2025-05-22)[2026-06-25]. https://gss.norc.org/.
[52] Walt H. Fivethirtyeight. com datalab: how Americans like their steak. https://fivethirtyeight.com/features/how-americanslike-their-steak/.
[53] Liu Z W, Luo P, Wang X G, et al. Deep learning face attributes in the wild [C]// Proceedings of IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2016: 3730-3738.
[54] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database [C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 248-255.
[55] Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank [C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: Association for Computational Linguistics, 2013: 1631-1642.
[56] Carlini N, Tramer F, Wallace E, et al. Extracting training data from large language models[C]//30th USENIX Security Symposium (USENIX Security 21). Berkeley, CA: USENIX Association, 2021: 2633-2650.
[57] Panayotov V, Chen G G, Povey D, et al. Librispeech: an ASR corpus based on public domain audio books [C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2015: 5206-5210.
[58] Ardila R, Branson M, Davis K, et al. Common voice: a massively-multilingual speech corpus [PP/OL]. arXiv(2019-12-13)[2025-04-21]. https://arxiv.org/abs/1912.06670.
[59] Jaber A, Fritsch L. Towards AI-powered cybersecurity attack modeling with simulation

tools: review of attack simulators [C]// Advances on P2P, Parallel, Grid, Cloud and Internet Computing. Cham: Springer, 2023: 249-257.
[60] He Z C, Zhang T W, Lee R B. Model inversion attacks against collaborative inference [C]// Proceedings of the 35th Annual Computer Security Applications Conference. New York: ACM, 2019: 148-162.
[61] Zhang S J, Yuan W, Yin H Z. Comprehensive privacy analysis on federated recommender system against attribute inference attacks [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2024, 36(3): 987-999.

黑盒模型反演攻击的研究综述

何世飞，王　晨，罗　辛*，徐光伟*
东华大学 信息与智能科学学院，上海 201620

摘　要：人工智能技术的发展催生了“机器学习即服务”这一新模式，使用户能够训练所需模型，或利用已有模型为其提供的训练数据进行预测。尽管已采取数据删除、匿名化等多种措施来防止机器学习中的隐私泄露，但模型反演攻击（model inversion attack，MIA）仍能在一定程度上从用户数据中推断出敏感信息。本文综述了现有的黑盒 MIA 的研究进展，重点分析了基于置信度的 MIA 和基于标签的 MIA。进一步梳理了近年来针对文本、音频等新兴数据模态所开展的 MIA 攻击方法，填补了现有综述对该类模态研究梳理的空白。最后，对黑盒 MIA 的未来挑战及研究方向进行了探讨。

关键词：机器学习；模型反演攻击；黑盒模型；隐私保护