A Survey on Privacy-Preserving Machine Learning: Toward Feature-Level Selective Protection

Bilal ABDULHADI
Computer Engineering Department
Biruni University
Istanbul, Turkey
251132903@st.biruni.edu.tr

Güray YILMAZ
Computer Engineering Department
Biruni University
Istanbul, Turkey
gurayy@biruni.edu.tr

***Abstract*—Privacy-preserving machine learning has become an important research direction as machine learning models are increasingly trained and deployed using sensitive and private data. Existing studies propose different protection mechanisms, including cryptographic techniques, differential privacy, and federated learning. However, these approaches often introduce important trade-offs between privacy guarantees, computational cost, communication overhead, and model utility. This survey reviews major privacy risks in machine learning, including membership inference, reconstruction, and gradient leakage attacks, and categorizes existing privacy-preserving approaches into cryptographic, perturbation-based, and distributed or hybrid methods. In addition, the paper focuses on selective and feature-level privacy protection as an emerging direction for reducing unnecessary overhead while maintaining useful model performance. By analyzing the limitations of full-data protection and noise-based mechanisms, this survey positions feature-importance-based selective protection as a promising research direction for future privacy-preserving machine learning systems.**

***Index Terms*—privacy-preserving machine learning, homomorphic encryption, differential privacy, federated learning, feature selection, selective encryption**

I. Introduction

Machine learning systems are now used in many domains where data may include personal, financial, medical, or behavioral information. Although these systems provide strong predictive capabilities, they may also expose sensitive information through the trained model, its outputs, or the learning process itself. Previous studies have shown that machine learning models are vulnerable to privacy attacks such as membership inference, model inversion, reconstruction, and white-box inference attacks [1]–[3]. These risks make privacy-preserving machine learning (PPML) an important research direction for both centralized and distributed learning settings.

Existing PPML methods are commonly grouped into cryptographic, perturbation-based, and distributed or hybrid approaches. Cryptographic techniques, including secure multi-party computation and homomorphic encryption, allow learning or inference to be performed without directly exposing sensitive inputs [4]–[6]. Differential privacy protects individual records by injecting controlled noise into the training process or output mechanism [7]–[9]. Federated learning reduces the need to centralize raw data, but shared gradients, model updates, and parameters may still leak private information if no additional protection is used [2], [10], [11].

Despite their benefits, these approaches introduce different trade-offs. Cryptographic methods can provide strong protection, but they usually increase computation, memory usage, and communication overhead [4], [5], [12]. Differential privacy is often more practical, but the added noise may reduce accuracy, especially under stronger privacy requirements [7], [8]. Federated learning improves data locality, yet it does not automatically guarantee privacy because model updates can still be analyzed by attackers [10], [11]. Therefore, PPML requires not only privacy mechanisms, but also careful consideration of utility, efficiency, and threat models.

One limitation of many existing approaches is that data or model components are often protected uniformly. For example, full encryption-based methods may encrypt all input features or computations, even when not all features contribute equally to the model decision. This can lead to unnecessary overhead. Recent studies have begun to explore more selective forms of protection, such as feature selection before encrypted learning, chunk-based encryption for image classification, and privacy mechanisms focused on sensitive components such as labels [9], [13], [14]. Other work also shows that privacy-preserving mechanisms can affect feature importance, feature relationships, and model performance [15].

The guiding research question of this survey is: how can feature importance and feature sensitivity be used to guide selective privacy protection in machine learning while balancing privacy, model utility, and computational efficiency? This question narrows the broad PPML area toward a feature-level perspective and provides a basis for comparing existing methods according to their privacy guarantees, practical limitations, and performance trade-offs.

Based on this direction, this survey reviews key privacy risks, categorizes existing PPML techniques, compares their limitations, and positions feature-importance-based selective protection as a potential direction for reducing unnecessary overhead while preserving useful model performance. Rather than proposing a full implementation, the paper provides a structured literature review and conceptual positioning that can support future thesis-level research.

II. Background on Privacy-Preserving Machine Learning

Privacy-preserving machine learning (PPML) refers to techniques that allow machine learning to be trained, evaluated, or deployed while reducing the exposure of sensitive data, model parameters, intermediate computations, or prediction outputs. This need has grown as machine learning is applied to data that may contain personal, financial, medical, behavioral, or organizational information. The main challenge is that models require useful patterns from data, while privacy mechanisms try to limit what can be revealed from the same data [3], [16].

The relationship between machine learning and privacy is not one-dimensional. Machine learning can be the target of privacy attacks when adversaries attempt to extract information from trained models or their outputs. It can also support privacy-related tasks, such as detecting abnormal access patterns or supporting privacy-aware decisions. At the same time, machine learning can be used by attackers to infer sensitive attributes or reconstruct private information [16]. For this reason, PPML is concerned not only with raw datasets, but also with the learning process, trained models, predictions, gradients, and shared updates.

Existing PPML methods can be grouped into several categories. Cryptographic methods protect data during computation. Techniques such as secure multi-party computation and homomorphic encryption allow computations to be performed without directly revealing the underlying data to other parties [4], [17]. These methods provide strong privacy guarantees, but often introduce high computational and communication overhead, especially for complex models such as deep neural networks [5], [6], [12].

Perturbation-based methods protect privacy by modifying data, gradients, outputs, or model behavior through controlled noise. Differential privacy is the most common example in this category. Differentially private stochastic gradient descent limits the influence of individual training samples by clipping gradients and adding noise during training [7]. Private aggregation of teacher ensembles applies a different strategy by transferring knowledge from private teacher models to a public student model through noisy aggregation [8]. These methods are often more practical than fully encrypted computation, but stronger privacy can reduce model accuracy.

Distributed and hybrid approaches reduce direct data exposure by changing how learning is organized. Federated learning is a common example, where data remain distributed across clients and only model updates are shared with a server or aggregator. However, federated learning is not automatically privacy-preserving, because gradients and updates may still leak information about local data [2], [10], [11]. Therefore, it is often combined with secure aggregation, differential privacy, or encryption.

Across these approaches, the central design issue is the balance between privacy, utility, and efficiency. Cryptographic protection can preserve confidentiality but may be computationally expensive. Differential privacy gives formal guarantees but can reduce accuracy because of noise. Federated learning avoids direct data centralization but remains vulnerable to update-based leakage [6], [7], [10], [11]. These limitations motivate selective forms of protection, where stronger privacy mechanisms are applied only to specific data components, features, labels, or image regions [9], [13]–[15].

In this survey, PPML is treated as a design problem: what should be protected, against which threat, and at what cost? This view is important for feature-level selective protection, where feature importance and sensitivity may guide where stronger privacy protection is most useful.

III. Privacy Risks and Attacks in Machine Learning

Privacy risks in machine learning arise because trained models may reveal information about the data used during training or inference. Even when raw datasets are not directly shared, attackers may exploit model outputs, confidence scores, gradients, parameters, or intermediate updates to infer sensitive information. Prior studies identify several major attack types, including membership inference, model inversion, reconstruction, property inference, and white-box inference attacks [1]–[3]. Understanding these attacks is necessary before comparing privacy-preserving techniques, because each protection mechanism assumes a different threat model.

A. Membership Inference Attacks

Membership inference attacks aim to determine whether a specific data record was included in the training set of a machine learning model. This can reveal sensitive information in domains such as healthcare, finance, or behavioral analysis, where participation in a dataset may itself be private. Shokri et al. introduced a black-box membership inference attack in which an adversary queries a target model and uses the prediction vector to infer whether a sample was part of the training data [1]. Their work showed that overfitted models and highly confident predictions are more vulnerable to this type of attack.

Membership inference is not limited to black-box access. Nasr et al. extended the analysis to white-box settings, where the attacker may access model parameters, gradients, or intermediate computations [2]. Their study showed that deeper access to the model can increase privacy leakage, especially in centralized and federated learning environments. This finding is important for PPML because it shows that hiding raw data alone is not enough; the model behavior and training process may also expose private information.

B. Model Inversion and Reconstruction Attacks

Model inversion and reconstruction attacks attempt to recover sensitive information about training data or input features from a trained model. In these attacks, the adversary may use model outputs, confidence scores, or gradients to reconstruct representative features or approximate private records. Rigaki and Garcia identify reconstruction and model inversion

as important attack categories because they can expose sensitive attributes even when the attacker does not directly access the original dataset [3].

These attacks are especially relevant when models are deployed through prediction APIs or machine-learning-as-a-service platforms. If a model returns detailed confidence scores, an attacker may repeatedly query the system and use the responses to infer private information. This means that privacy protection should not only focus on the training dataset, but also on model outputs and inference-time interactions. Depending on the scenario, possible defenses may include output restriction, differential privacy, secure inference, or encrypted computation.

C. Gradient and Parameter Leakage

Gradient and parameter leakage is particularly important in distributed and federated learning. Federated learning keeps raw data on local devices, but clients still share model updates or gradients with a server. Studies show that these shared updates may leak information about local training samples if they are not protected [2], [10], [11]. Therefore, federated learning should not be considered automatically privacy-preserving.

White-box inference attacks further show that access to gradients, parameters, and intermediate model states can increase the attacker's ability to infer private information [2]. In practical systems, this creates a need for additional protection methods such as secure aggregation, differential privacy, homomorphic encryption, or hybrid approaches. However, these defenses also introduce trade-offs in computation, communication, accuracy, and privacy strength.

Overall, privacy attacks show that PPML must be evaluated according to the type of information that may leak and the attacker's level of access. Membership inference focuses on whether a record was part of the training set, reconstruction attacks attempt to recover sensitive data or features, and gradient leakage exposes risks during training or distributed learning. These risks motivate privacy-preserving techniques that protect sensitive information while maintaining practical model performance.

IV. Cryptographic Approaches for Privacy-Preserving Machine Learning

Cryptographic approaches are a major category of privacy-preserving machine learning because they protect data during computation, rather than only before or after learning. These methods allow parties to train or evaluate models without directly exposing sensitive inputs, intermediate values, or model parameters. The main techniques include secure multi-party computation (SMPC), homomorphic encryption (HE), fully homomorphic encryption (FHE), and multi-key homomorphic encryption (MKHE). Although these methods can provide strong privacy protection, they often introduce high computational cost, memory overhead, and communication complexity [4], [6], [17].

A. Secure Multi-Party Computation

Secure multi-party computation allows multiple parties to jointly compute a function over private inputs without revealing those inputs to each other. In PPML, SMPC can support training or inference when data are distributed among entities that cannot share raw data. SecureML is a key example in this direction. Mohassel and Zhang proposed a scalable privacy-preserving machine learning system that supports linear regression, logistic regression, and neural networks using secure computation techniques [4].

SecureML combines secret sharing, secure two-party computation, multiplication triples, garbled circuits, and oblivious transfer to protect training data during learning [4]. This makes it an important baseline for cryptographic PPML. However, the approach also illustrates a common limitation of secure computation: stronger protection usually increases protocol complexity and computational cost. In addition, such systems often depend on assumptions such as non-colluding parties or semi-honest adversaries, which may not always match real deployment conditions.

B. Homomorphic Encryption

Homomorphic encryption allows computations to be performed directly on encrypted data. After decryption, the result should correspond to the result of the same computation on plaintext data. This property is attractive for PPML because it enables inference or learning without exposing sensitive inputs to a model provider or service operator [5], [17].

Approximate homomorphic encryption schemes, such as CKKS, are especially relevant for machine learning because they support approximate arithmetic over real numbers [17]. However, HE-based machine learning requires careful adaptation of model operations. Many standard operations, including nonlinear activations, comparisons, divisions, and normalization, are not directly efficient under HE. As a result, models often require polynomial approximations, modified architectures, or encryption-friendly operations [6], [12].

The main strength of HE is strong confidentiality during computation, but its main limitation is efficiency. Encrypted operations are much slower than plaintext operations, and ciphertexts are larger than raw data. These limitations make scalability difficult, especially for deep learning models or large datasets [5], [6]. Therefore, the practical use of HE depends on reducing overhead and adapting model design.

C. Fully Homomorphic Encryption for Deep Learning

Fully homomorphic encryption extends homomorphic computation by supporting both addition and multiplication over encrypted data, which theoretically enables arbitrary computation over ciphertexts. In deep learning, FHE is mainly used for encrypted inference, where the input remains encrypted and the model produces an encrypted prediction. Lee et al. explored FHE-based deep neural network inference and showed that encrypted deep learning can maintain useful prediction performance while protecting input data [5].

Despite this potential, FHE-based deep learning remains difficult in practice. Deep neural networks contain nonlinear operations such as ReLU, sigmoid, softmax, pooling, and normalization, which are not naturally efficient under FHE. These operations usually require approximation or replacement with polynomial functions [6], [12]. Deeper models also increase multiplicative depth, which can increase ciphertext noise and require expensive bootstrapping.

Several studies show that FHE can preserve privacy while keeping accuracy close to plaintext models, but encrypted inference is still much slower and more resource-intensive than standard inference [5], [12]. This trade-off motivates selective protection: if all features or computations are encrypted uniformly, the cost may become unnecessarily high. Protecting only the most important or sensitive parts of the data may therefore reduce overhead while keeping meaningful privacy protection.

D. Multi-Key Homomorphic Encryption for Classical Models

Multi-key homomorphic encryption allows computation over ciphertexts encrypted under different keys. This is useful when multiple data owners participate in computation without sharing a common secret key. In PPML, MKHE can protect both client data and model-owner information, especially in outsourced inference. Petrean and Potolea studied random forest evaluation using multi-key homomorphic encryption and lookup tables, showing how encrypted decision tree and random forest inference can protect both input features and model structure [18].

This direction is relevant because many PPML studies focus on neural networks, while classical models such as decision trees and random forests are still widely used, especially for tabular data. However, protecting decision paths, feature values, and model structure during encrypted inference remains challenging. MKHE and lookup-table-based techniques provide one way to support private evaluation of such models [18].

The cost of this approach can increase with model size, feature representation, number of trees, and bit-level encoding requirements. This connects cryptographic PPML with feature-level selective protection: if encrypted evaluation cost depends on the number and representation of features, then feature selection and feature importance may help reduce unnecessary encrypted computation.

Overall, cryptographic approaches provide some of the strongest privacy guarantees in PPML, but efficiency remains their main challenge. SMPC supports privacy-preserving training, HE and FHE support encrypted computation, and MKHE supports multi-party or multi-key inference. However, these methods often require model adaptation, system assumptions, and significant resources. These limitations motivate more selective approaches, where encryption is applied according to data sensitivity, feature importance, model structure, or application requirements.

V. Differential Privacy Approaches

Differential privacy (DP) is one of the most widely studied perturbation-based approaches in privacy-preserving machine learning. Unlike cryptographic methods, which protect data during computation, DP limits how much information about an individual record can be inferred from a computation output or trained model. Its central idea is that the presence or absence of a single data sample should not significantly change the released result, making it useful for individual-level privacy protection [7], [16].

In machine learning, DP is commonly applied by adding calibrated noise during training, prediction, or aggregation. The level of privacy is usually controlled by a privacy budget, commonly denoted as ϵ. A smaller ϵ generally provides stronger privacy, but it may require more noise and reduce model accuracy. Therefore, DP introduces a direct privacy-utility trade-off [7]–[9].

A. Differentially Private Stochastic Gradient Descent

Differentially private stochastic gradient descent (DP-SGD) is a major method for training deep learning models with differential privacy. Abadi et al. proposed a practical DP-SGD approach that clips per-example gradients and adds Gaussian noise during training [7]. Gradient clipping limits the influence of each training sample, while noise addition makes it harder to infer whether a specific record was included in the training set. Privacy accounting is then used to track the accumulated privacy loss over training iterations [7].

DP-SGD directly addresses the risk that neural networks may memorize information from training data. By limiting individual sample influence, it reduces the possibility that the model reveals information about specific records, making it relevant to membership inference attacks [1], [7]. However, DP-SGD also introduces practical challenges. Per-example gradient computation increases training complexity, and the added noise can reduce accuracy, especially for complex datasets or strong privacy requirements.

For this survey, DP-SGD serves as a strong perturbation-based baseline. It provides formal privacy guarantees, but protection is achieved by modifying the training process and adding noise. This differs from selective encryption, where the goal is to decide which features or components should receive stronger protection. Thus, DP-SGD highlights a different type of privacy trade-off: accuracy loss caused by noise rather than computational cost caused by encryption.

B. Private Aggregation of Teacher Ensembles

Private Aggregation of Teacher Ensembles (PATE) is another influential DP approach. Papernot et al. proposed a framework where multiple teacher models are trained on disjoint private subsets, and a student model learns from the noisy aggregated predictions of these teachers [8]. Since the student does not directly access the private training data, privacy is provided through the aggregation mechanism and noise added to teacher votes.

PATE is useful because it separates private training data from the final deployed student model. Instead of training one model directly on all sensitive data, private knowledge is distributed across teacher models, while the noisy voting process reduces the chance of revealing individual training examples [8]. However, PATE requires enough private data to train multiple teachers and often depends on public or non-sensitive unlabeled data for student training. Its privacy budget is also affected by the number of queries made to the teacher ensemble.

In relation to feature-level selective protection, PATE shows that privacy does not always need to be applied uniformly across the whole learning pipeline. Privacy can be controlled through selected interactions, such as teacher voting and public student learning. This supports the broader idea that PPML can be designed around the points where sensitive information is most exposed.

C. Label Differential Privacy

Label differential privacy focuses on scenarios where labels are sensitive while input features may be considered non-sensitive or publicly available. Ghazi et al. studied deep learning under label differential privacy and proposed mechanisms that protect label information while allowing the model to use input features more directly [9]. This differs from standard DP settings, where both features and labels may require protection.

This direction is important because not all parts of a dataset have the same privacy sensitivity. In applications such as recommendation systems, surveys, medical diagnosis, or user behavior prediction, the label may reveal private information even when input attributes are less sensitive. Label DP therefore provides an example of selective privacy protection, where a specific data component receives stronger protection [9].

Ghazi et al. showed that label DP can achieve better utility than applying full differential privacy when only labels need protection [9]. This finding is relevant to feature-level selective protection: if protecting only labels can preserve more utility than protecting the whole dataset, then protecting only important or sensitive features may also help balance privacy and performance. However, label DP is not sufficient when input features themselves contain sensitive information. It is best understood as evidence that privacy mechanisms should match the actual sensitive components of the data.

Overall, differential privacy approaches provide formal privacy guarantees and are generally more computationally practical than fully encrypted computation. DP-SGD protects training by perturbing gradients, PATE protects knowledge transfer through noisy teacher aggregation, and label DP protects sensitive labels specifically. These methods show the value of perturbation-based privacy, while also emphasizing the importance of deciding what to protect and how much privacy to apply.

VI. Federated Learning and Privacy Preservation

Federated learning (FL) is a distributed machine learning approach in which multiple clients collaboratively train a model without directly sharing their raw data. Instead of centralizing datasets on one server, each client trains locally and sends model updates, gradients, or parameters to an aggregation server. This design makes FL attractive for privacy-sensitive domains such as healthcare, mobile applications, finance, and Internet of Things systems, where raw data may be difficult or legally restricted to share [10], [11].

However, FL should not be considered fully privacy-preserving by default. Although raw data remain local, shared updates can still contain information about local training samples. Studies have shown that gradients, parameters, and intermediate updates may be exploited to infer sensitive information, reconstruct data, or perform membership inference attacks [2], [10], [11]. Therefore, FL reduces direct data exposure, but it does not eliminate privacy risks.

A. Privacy Leakage in Federated Learning

Privacy leakage in FL mainly occurs through the information exchanged during training. In a typical FL process, clients compute local updates from private datasets and send them to a central server. If an attacker has access to these updates, the attacker may infer properties of local data or reconstruct sensitive information. This risk increases when updates are highly specific to small local datasets or when the attacker has white-box access to model parameters and gradients [2].

Truong et al. analyzed privacy preservation in FL from the GDPR perspective and emphasized that keeping data local is not sufficient for either technical privacy protection or privacy compliance [10]. Similarly, Ge et al. reviewed secure federated learning and identified leakage threats from gradients, model parameters, and aggregation processes [11]. These studies show that FL privacy risks may involve not only external attackers, but also curious servers, compromised clients, or malicious participants.

B. Privacy-Preserving Techniques in Federated Learning

Several techniques have been proposed to improve privacy in FL. Secure aggregation prevents the server from seeing individual client updates by allowing it to learn only the aggregated result [10], [11]. This reduces exposure from single clients, but it may also make malicious or low-quality updates harder to detect, creating a trade-off between privacy and robustness.

Differential privacy can also be applied by adding noise to local updates, gradients, or aggregated results. This limits the influence of individual clients or training records on the final model, but too much noise may reduce model utility or slow convergence [7], [11]. Cryptographic methods, including homomorphic encryption and secure multi-party computation, can protect updates during aggregation or computation, but they may increase communication and computational cost, especially for resource-limited clients [4], [11], [17].

From the perspective of this survey, FL represents a distributed or hybrid PPML approach. It reduces the need for centralized raw data collection, but still requires additional mechanisms to protect shared updates. It is therefore closely

related to the privacy-utility-efficiency trade-off. Selective protection may also be useful in FL by applying stronger privacy mechanisms only to the most sensitive features, gradients, updates, or clients, instead of protecting all exchanged information uniformly.

Overall, federated learning provides an important foundation for privacy-aware distributed machine learning, but it is not a complete privacy solution by itself. Its privacy depends on how updates are shared, how aggregation is performed, what adversarial model is considered, and whether additional techniques such as secure aggregation, differential privacy, or encryption are applied. These characteristics make FL an important comparison point for feature-level selective protection.

VII. Selective and Feature-Level Privacy Protection

Selective and feature-level privacy protection aims to avoid applying the same level of protection to every data component. In many PPML approaches, the entire dataset, all input features, or all model computations are protected uniformly. While this may strengthen privacy, it can also increase computation, communication, or accuracy costs. Selective protection addresses this issue by identifying which parts of the data or learning process require stronger protection and which parts can be treated with lighter mechanisms [9], [13]–[15].

The motivation is that features differ in both predictive value and privacy sensitivity. Some features may be highly useful for prediction but not sensitive, while others may be sensitive but less important for the model. Features that are both important and sensitive are the strongest candidates for selective protection. This creates a more fine-grained design question: what should be protected, why should it be protected, and how much protection is appropriate?

A. Selective Encryption in Image-Based Models

A clear example of selective encryption appears in image-based privacy-preserving classification. Jia et al. proposed an approach that combines homomorphic encryption with chunk-based convolutional neural networks [14]. Instead of encrypting an entire image, the method divides the image into smaller chunks and applies encryption selectively to feature-rich chunks. This reduces encrypted computation while still protecting important visual information.

This work is relevant because it demonstrates that full encryption is not always necessary. If only the most informative regions are encrypted, the system may reduce computational cost compared with encrypting the whole input, while still protecting regions that strongly contribute to classification [14]. This provides a useful analogy for tabular or structured data, where important features could be identified and protected more strongly than less important ones.

However, chunk-based selective encryption is mainly designed for image data. The selection process depends on image-specific characteristics, such as feature-rich regions or gradient-based chunk importance. This differs from feature-importance-based protection in tabular machine learning, where features may represent age, income, location, diagnosis, transaction behavior, or other structured attributes. Therefore, while this work supports the idea of selective encryption, more research is needed to generalize selective protection across data types and model families.

B. Feature Selection and Privacy-Preserving Mechanisms

Feature selection is closely related to selective privacy protection because it identifies which features are most useful for a learning task. In PPML, feature selection can reduce the amount of data that needs to be protected or processed under expensive privacy mechanisms. Lo et al. studied practical considerations of fully homomorphic encryption in PPML and used feature selection to reduce the number of features processed in encrypted machine learning experiments [13]. This shows that feature selection can have a practical role in reducing encrypted learning overhead.

Alishahi et al. further analyzed the mutual impact of feature selection and privacy-preserving mechanisms [15]. Their work is important because it shows that privacy mechanisms can affect feature importance, feature ranking, feature correlation, and classification performance. This means privacy protection should not be treated as a separate step applied after feature analysis. Instead, feature importance and privacy mechanisms can influence each other.

This finding is directly relevant to feature-level selective protection. If privacy mechanisms change the usefulness or ranking of features, selecting features before applying protection may lead to different results than selecting them after privacy transformation. A privacy-aware feature selection process should therefore consider predictive importance, privacy sensitivity, and the effect of protection mechanisms on model utility.

C. Feature Importance as a Privacy-Aware Criterion

Feature importance measures are commonly used to understand which input variables contribute most to model predictions. In selective protection, feature importance can help decide which features should receive stronger privacy protection. For example, if a small subset of features provides most of the predictive value, protecting only those features may reduce computational overhead compared with full-data encryption. This idea is supported by studies showing that feature selection can reduce encrypted processing and that selective encryption can avoid unnecessary computation [13], [14].

However, feature importance alone is not enough. A feature may be important but not sensitive, or sensitive but not highly important. Label differential privacy provides a related example of selective protection, where privacy focuses on labels because labels are the sensitive component in certain scenarios [9]. This supports the broader idea that privacy mechanisms should target the parts of the dataset that carry the highest privacy risk.

A feature-importance-based selective protection approach could rank features by predictive contribution, identify sensitive attributes or feature groups, and apply stronger privacy

mechanisms to features that are both important and sensitive. Depending on the application, this protection could involve encryption, perturbation, masking, or a hybrid method. The goal is not to weaken privacy, but to apply protection more strategically so that unnecessary overhead is reduced while useful model performance is preserved.

This direction also introduces risks. Protecting only selected features may leave unprotected features vulnerable if they are correlated with sensitive information. Feature importance may also change across models, datasets, and training stages. A feature that appears unimportant in one model may become important in another, and combinations of weak features may still reveal private information. For this reason, selective protection requires careful threat modeling, feature analysis, and evaluation of both privacy leakage and model performance.

Overall, selective and feature-level privacy protection provides a promising middle ground between full protection and no protection. It builds on ideas from cryptographic methods, differential privacy, feature selection, and selective privacy mechanisms, while focusing on the practical challenge of reducing unnecessary overhead. For this reason, it represents the main conceptual direction of this survey and supports the positioning of feature-importance-based selective protection as a future research topic.

VIII. Discussion

The reviewed literature shows that PPML is not based on a single solution, but on several families of techniques with different assumptions, strengths, and limitations. Cryptographic approaches protect data during computation, differential privacy limits individual-level leakage through noise, federated learning reduces direct data centralization, and selective approaches focus protection on specific components such as labels, image chunks, or selected features [4], [7], [9], [10], [14], [15].

A. Comparison of Existing Approaches

Table I summarizes the main categories of PPML methods discussed in this survey. The comparison shows that each category protects different parts of the learning process and introduces different trade-offs. This supports the view that feature-level selective protection should be considered as a complementary direction, not as a replacement for existing PPML methods.

Cryptographic methods such as SMPC and HE are suitable when confidentiality during computation is the main requirement. SecureML shows that secure computation can support privacy-preserving training for several model types, while HE-based work demonstrates the possibility of encrypted inference for neural networks [4], [5]. However, these methods often require significant resources and may need model adaptations, especially when nonlinear operations are involved [6], [12].

Differential privacy provides a different form of protection by limiting the effect of individual records on the final model or output. DP-SGD protects training by clipping gradients and adding noise, while PATE transfers knowledge through noisy aggregation of teacher models [7], [8]. Label differential privacy further shows that privacy can be targeted toward a specific sensitive component, such as labels [9]. This selective aspect makes label DP conceptually relevant to feature-level protection.

Federated learning changes the learning architecture by keeping data distributed, but the literature shows that shared updates may still leak sensitive information [2], [10], [11]. Therefore, FL is often combined with secure aggregation, differential privacy, or encryption. This makes it a hybrid direction rather than a complete privacy solution on its own.

B. Privacy-Utility Trade-Off

A recurring issue across PPML methods is the privacy-utility trade-off. In differential privacy, stronger privacy usually requires more noise, which may reduce accuracy or slow convergence [7], [8]. In cryptographic approaches, accuracy may remain close to plaintext performance, but the cost appears in computation time, memory consumption, communication, and architectural constraints [5], [6]. In federated learning, data locality improves privacy compared with centralized training, but update leakage remains a concern if no additional protection is applied [10], [11].

Selective protection is motivated by these trade-offs. Instead of applying the same protection level to every feature, label, chunk, or update, selective methods focus protection where it is most useful. Feature selection before encrypted learning can reduce the number of features processed under FHE, while chunk-based encrypted image classification can reduce the amount of encrypted visual data [13], [14]. These examples suggest that selective protection can reduce cost, but only if the selection process does not overlook privacy leakage through correlated or unprotected features.

C. Computational Cost and Scalability

Scalability remains one of the main challenges in PPML. Fully encrypted computation is expensive because ciphertext operations are slower than plaintext operations and encrypted data representations are larger [5], [12]. Deep learning under HE is especially challenging because common neural network operations such as ReLU, softmax, and pooling are not naturally efficient under encrypted computation [6]. As a result, researchers often use approximations or modified model architectures.

Feature selection and selective encryption directly target this scalability problem. If fewer features or components require strong protection, the computational burden may be reduced. Lo et al. show that feature selection can reduce the number of encrypted features in practical FHE-based machine learning experiments [13]. Similarly, chunk-based encryption reduces encrypted image processing by focusing on informative regions instead of encrypting the whole image [14].

However, efficiency should not be improved by ignoring privacy risks. A feature that appears less important for prediction may still reveal sensitive information through correlation

TABLE I
COMPARISON OF MAJOR PRIVACY-PRESERVING MACHINE LEARNING APPROACHES

Approach	Main Idea	Privacy Target	Main Strength	Main Limitation
Secure Multi-Party Computation	Jointly computes over private inputs without directly revealing them.	Training data and intermediate computation.	Strong privacy during collaborative computation.	High communication and computation cost; depends on system and adversarial assumptions [4].
Homomorphic Encryption	Performs computation directly on encrypted data.	Input data and inference process.	Strong confidentiality for outsourced computation.	High computational overhead, large ciphertexts, and model adaptation requirements [5], [6], [12].
Differential Privacy	Adds controlled noise to limit the influence of individual records.	Training samples, gradients, labels, or outputs.	Formal privacy guarantee and practical deployment potential.	Privacy-utility trade-off; stronger privacy can reduce accuracy [7]–[9].
Federated Learning	Keeps raw data local and shares model updates instead of datasets.	Raw local data in distributed learning.	Reduces centralized data collection.	Shared gradients and updates can still leak private information [2], [10], [11].
Selective Protection	Applies stronger protection only to selected components, labels, chunks, or features.	Important or sensitive parts of the data or model pipeline.	Can reduce unnecessary overhead and preserve utility.	Requires careful selection criteria and may miss leakage through correlated features [9], [13]–[15].

with protected attributes. Alishahi et al. show that privacy-preserving mechanisms can affect feature ranking, feature relationships, and model performance, which means that feature selection and privacy protection should be analyzed together rather than independently [15]. Therefore, feature-importance-based selective protection should consider both predictive contribution and privacy sensitivity.

Overall, current PPML methods offer valuable but incomplete solutions. Cryptographic approaches provide strong confidentiality but are expensive. Differential privacy provides formal guarantees but may reduce utility. Federated learning reduces data centralization but remains vulnerable to update leakage. Selective and feature-level protection offers a promising direction for balancing these limitations, but it requires careful threat modeling, feature analysis, and evaluation.

IX. RESEARCH GAP AND FUTURE DIRECTION

The reviewed literature shows that PPML has developed through several major directions, including cryptographic computation, differential privacy, federated learning, and selective protection mechanisms. However, a key gap remains in how privacy protection is assigned across different parts of the data. Many existing methods apply privacy mechanisms uniformly, such as encrypting all features, adding noise throughout training, or protecting all model updates. While this can improve privacy, it may also introduce unnecessary computational cost, communication overhead, or utility loss [4]–[7].

The limitations discussed in the previous sections suggest that future PPML systems should not only ask how data can be protected, but also which parts of the data or learning process require stronger protection. Cryptographic approaches can provide strong confidentiality but become costly with complex models and large data [4], [5], [12]. Differential privacy provides formal guarantees, but stronger privacy usually requires more noise, which can reduce model performance [7], [8]. Federated learning reduces data centralization, but gradients and shared updates can still leak sensitive information [2], [10], [11].

Selective protection has started to address this problem, but the literature is still limited. Label differential privacy focuses protection on labels when labels are the sensitive component [9]. Chunk-based encrypted image classification applies encryption to selected image regions instead of the entire image [14]. Practical FHE-based studies also show that feature selection can reduce the number of encrypted features and improve efficiency [13]. In addition, work on the relationship between feature selection and privacy-preserving mechanisms shows that privacy protection can affect feature ranking, feature relationships, and model utility [15]. These studies support selective privacy protection, but they do not yet provide a general framework for feature-importance-based selective protection across different data types and models.

Based on this gap, feature-importance-based selective protection can be considered a promising future research direction. The core idea is to combine feature importance with privacy sensitivity when deciding which features require stronger protection. Instead of encrypting or perturbing all features equally, a system could rank features by predictive contribution, identify sensitive or privacy-critical attributes, and then apply stronger protection to features that are both important and sensitive. Fig. 1 illustrates this conceptual workflow.

In this workflow, feature importance analysis identifies variables that contribute most to model performance, while sensitivity assessment evaluates whether these variables may reveal private or confidential information. The selected features can then be protected using an appropriate mechanism, such as encryption, perturbation, masking, or a hybrid method. The final model should be evaluated not only by accuracy, but also by privacy risk and computational cost, since selective protection may reduce overhead while still introducing risks if unprotected features are correlated with sensitive information.

The research question guiding this future direction is: how can feature importance and feature sensitivity be used to guide selective privacy protection in machine learning while balancing privacy, model utility, and computational efficiency? This question narrows the broader PPML topic into a feature-

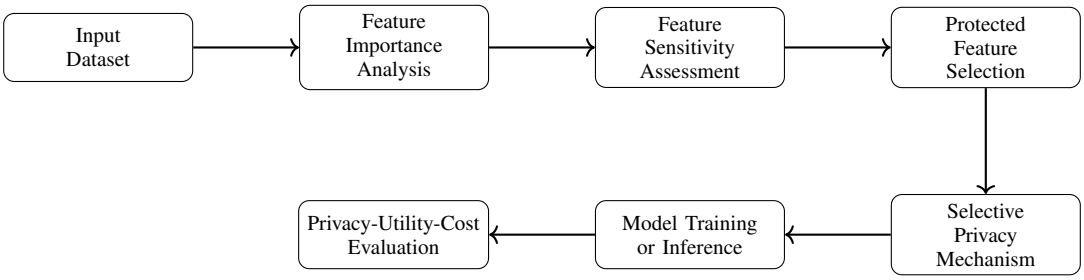

Fig. 1. Conceptual workflow for feature-importance-based selective protection in privacy-preserving machine learning.

level perspective and directly reflects the limitations found in full cryptographic protection, noise-based privacy mechanisms, and distributed learning systems.

Future research can investigate this direction in several ways. First, different feature importance methods, such as tree-based importance, permutation importance, or explainability-based methods, can be compared for selecting features to protect. Second, different protection mechanisms can be applied selectively, including homomorphic encryption for important features, differential privacy for sensitive attributes, or hybrid combinations. Third, evaluation should include prediction accuracy, privacy leakage resistance, execution time, memory usage, and communication cost. These directions connect the proposed idea to the practical limitations identified in existing PPML studies [13]–[15].

Overall, the main gap identified in this survey is the lack of a general and systematic approach for connecting feature importance, feature sensitivity, and selective privacy protection. Existing studies show that selective protection is possible, but they are often limited to labels, image chunks, or reduced feature sets. A feature-importance-based selective protection framework could provide a more flexible direction for future PPML research by protecting the most relevant and sensitive parts of the data while reducing unnecessary overhead.

X. Conclusion

This survey reviewed privacy-preserving machine learning from three main perspectives: privacy risks, protection mechanisms, and selective feature-level protection. The literature shows that machine learning models can expose sensitive information through membership inference, reconstruction, model inversion, gradient leakage, and white-box attacks [1]–[3]. These risks show that privacy protection must address not only raw datasets, but also trained models, outputs, gradients, and distributed learning updates.

Existing PPML methods provide different forms of protection, each with its own limitations. Cryptographic approaches such as secure multi-party computation and homomorphic encryption protect data during computation, but often introduce high computational and communication overhead [4]–[6]. Differential privacy provides formal guarantees through controlled noise, but may reduce model utility under stronger privacy requirements [7]–[9]. Federated learning reduces raw data centralization, but shared gradients and updates may still leak private information without additional protection mechanisms [2], [10], [11].

The reviewed studies suggest that selective protection is a promising direction for reducing the limitations of uniform privacy mechanisms. Feature selection before encrypted learning, chunk-based encrypted image classification, label differential privacy, and studies on the relationship between feature selection and privacy-preserving mechanisms all indicate that privacy can be applied more strategically to selected data components [9], [13]–[15].

The main gap identified in this survey is the lack of a general approach that connects feature importance, feature sensitivity, and selective privacy protection. Future research can build on this gap by developing methods that identify important and sensitive features, apply suitable privacy mechanisms selectively, and evaluate the resulting trade-offs among privacy, utility, and computational cost. Therefore, feature-importance-based selective protection represents a focused and meaningful direction for future PPML research.

References

[1] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 3–18.

[2] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE Symposium on Security and Privacy (SP)*, 2019, pp. 739–753.

[3] M. Rigaki and S. Garcia, "A survey of privacy attacks in machine learning," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–34, 2023.

[4] P. Mohassel and Y. Zhang, "SecureML: A system for scalable privacy-preserving machine learning," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 19–38.

[5] J.-W. Lee, H. Kang, Y. Lee, W. Choi, J. Eom, M. Deryabin, E. Lee, J. Lee, D. Yoo, Y.-S. Kim, and J.-S. No, "Privacy-preserving machine learning with fully homomorphic encryption for deep neural network," *IEEE Access*, vol. 10, pp. 30 039–30 054, 2022.

[6] R. Podschwadt, D. Takabi, P. Hu, M. H. Rafiei, and Z. Cai, "A survey of deep learning architectures for privacy-preserving machine learning with fully homomorphic encryption," *IEEE Access*, vol. 10, pp. 117 477–117 500, 2022.

[7] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318.

[8] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," in *International Conference on Learning Representations*, 2017.

[9] B. Ghazi, N. Golowich, R. Kumar, P. Manurangsi, and C. Zhang, "Deep learning with label differential privacy," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 27 131–27 145.

[10] N. Truong, K. Sun, S. Wang, F. Guitton, and Y. K. Guo, "Privacy preservation in federated learning: An insightful survey from the GDPR perspective," *Computers & Security*, vol. 110, p. 102402, 2021.
[11] L. Ge, H. Li, X. Wang, and Z. Wang, "A review of secure federated learning: Privacy leakage threats, protection technologies, challenges and future directions," *Neurocomputing*, vol. 561, p. 126897, 2023.
[12] A. Falcetta and M. Roveri, "Privacy-preserving deep learning with homomorphic encryption: An introduction," *IEEE Computational Intelligence Magazine*, vol. 17, no. 3, pp. 14–25, 2022.
[13] D. C.-T. Lo, Y. Shi, H. Shahriar, B. Deng, X. Zhang, and M.-L. Chen, "Practical considerations of fully homomorphic encryption in privacy-preserving machine learning," in *2024 IEEE International Conference on Big Data (BigData)*, 2024, pp. 6330–6335.
[14] H. Jia, D. Cai, J. Yang, W. Qian, C. Wang, X. Li, and S. Yang, "Efficient and privacy-preserving image classification using homomorphic encryption and chunk-based convolutional neural network," *Journal of Cloud Computing*, vol. 12, no. 175, 2023.
[15] M. Alishahi, V. Moghtadaiee, A. Fathalizadeh, and M. Rabiei, "Mutual impact of feature selection and privacy-preserving mechanisms," *International Journal of Machine Learning and Cybernetics*, vol. 17, no. 3, p. 106, 2026.
[16] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–36, 2021.
[17] J. Yuan, W. Liu, J. Shi, and Q. Li, "Approximate homomorphic encryption based privacy-preserving machine learning: A survey," *Artificial Intelligence Review*, vol. 58, no. 82, 2025.
[18] D.-E. Petrean and R. Potolea, "Random forest evaluation using multi-key homomorphic encryption and lookup tables," *International Journal of Information Security*, vol. 23, pp. 2023–2041, 2024.