IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154298

A Comprehensive Study on Privacy-Preserving Machine Learning (PPML) in Modern AI Systems

Naveen Kumar[1], Shiva Kumar[2], Paras Kaushik[3], Ms. Usha Kumari[4], Mr. Satish Kumar Soni[5], Mr. Uruj Jaleel[6]

Student, MCA, Meerut Institute of Engineering and Technology, U.P. India[1-3]

Assistant Professor, MCA, Meerut Institute of Engineering and Technology, India[4]

Associate Professor, MCA,Meerut Institute of Engineering and Technology, India[5]

Professor, MCA, Meerut Institute of Engineering and Technology, India[6]

Abstract: We are living in a time when AI is no longer a futuristic concept but a daily reality. From health apps that monitor our vitals to banking systems that flag fraud, machine learning is making decisions that directly affect people's lives. But this progress comes with a serious trade-off: to make AI smarter, we feed it enormous amounts of personal data. This raises a question that researchers and policymakers are urgently trying to answer - how do we build intelligent systems without putting people's privacy at risk? Privacy-Preserving Machine Learning (PPML) is the answer the research community has been working toward. It is a growing subfield of AI that explores how machine learning models can be trained and deployed without ever needing access to raw personal data. This paper takes a comprehensive look at PPML - what it is, how it works, where it is being used, and what still stands in the way of its widespread adoption. We cover the four main privacy-enhancing techniques: Differential Privacy, Federated Learning, Homomorphic Encryption, and Secure Multi-Party Computation. We support our analysis with comparative diagrams and performance charts, and we discuss both the progress made between 2021 and 2026 and the challenges that researchers are still working to solve. Our goal is to give readers - whether students, engineers, or policy professionals - a clear and honest picture of where PPML stands today and where it needs to go.

Keywords: Privacy-Preserving Machine Learning, Differential Privacy, Federated Learning, Homomorphic Encryption, SMPC, Data Security, AI Ethics, GDPR, Deep Learning.

1. INTRODUCTION

Think about the last time you typed something into a search bar, asked your phone for directions, or had a medical test analyzed by an AI system. In each of these moments, your personal data played a role - and in most cases, that data was processed by systems you have little visibility into. AI has become deeply woven into our daily routines, and that is not necessarily a bad thing. But it does raise some uncomfortable questions about who controls our data and how it is being used.

The core problem is straightforward: machine learning models need data to learn, and the more data they have, the better they get. But that data often belongs to real people - patients, customers, students - who reasonably expect it to stay private. The traditional approach of collecting everything into a central database and running a model on it feels increasingly out of step with both public expectations and legal requirements. Laws like GDPR in Europe, CCPA in the United States, and India's Digital Personal Data Protection Act have made it clear that organizations cannot simply treat user data as a free resource [7].

Privacy-Preserving Machine Learning emerged as a response to this tension. The goal is not to choose between privacy and performance - it is to find smarter ways to achieve both. Instead of shipping raw data to a server, what if the model came to the data? Instead of exposing individual records, what if we could learn the patterns without the specifics? These are the kinds of questions PPML is designed to answer.

This paper looks at PPML from multiple angles. We start by examining the privacy problems that make PPML necessary in the first place. We then walk through the main techniques in detail - what they are, how they work, and where they fall

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154298

short. We look at how PPML is being used in healthcare, finance, consumer technology, and autonomous driving. And we take an honest look at the challenges that remain unsolved and the research directions that seem most promising. Our aim is not just to survey the field, but to make it accessible to anyone trying to understand why PPML matters and where it is headed.

2. EXISTING PRIVACY CHALLENGES IN AI SYSTEMS

Before diving into solutions, it is worth spending some time on the problem itself. Privacy risks in AI are not abstract concerns - they show up in very concrete and sometimes alarming ways. Understanding these challenges is what motivates the entire field of PPML.

2.1 The Problem with Centralizing Data

The standard machine learning workflow involves gathering as much data as possible, moving it to a central server, and training a model on it. It is simple, effective, and deeply problematic from a privacy standpoint. When all the data lives in one place, a single security failure can expose millions of people at once. The 2021 Facebook breach, which affected over 533 million accounts, is a stark reminder of what happens when massive datasets are poorly protected. Healthcare breaches are even more damaging - in 2023, the average cost of a medical data breach in the US reached USD 10.9 million [10], not counting the human harm to patients whose most sensitive information was exposed.

Beyond breaches, centralization creates structural privacy problems. Data that was collected for one purpose gets used for another. Third parties gain access. Employees with legitimate system access can browse records they have no business seeing. Centralization is, in many ways, the root cause of most AI privacy issues.

2.2 Models Can Remember Too Much

One of the more surprising privacy risks in AI is that a trained model can effectively memorize parts of its training data. Carlini et al. [2] showed this in striking detail: by prompting a large language model like GPT-2 with certain inputs, they were able to extract verbatim sequences from its training set - including names, phone numbers, and email addresses. The model had not been asked to reveal this information; it simply had it stored in its weights, accessible to anyone who knew how to ask.

This is troubling because it means that even after a company deletes someone's data from its database, the information may still be embedded in any model that was trained on it. Data deletion requirements under GDPR become almost meaningless if a model can reproduce what was deleted.

2.3 Federated Learning Is Not Automatically Private

Federated Learning was designed with privacy in mind - instead of sending raw data to a server, devices send only gradient updates. This sounded like a clean solution. But Huang et al. [4] demonstrated that gradient updates are not as opaque as they seem. By carefully analyzing the gradients a device sends, attackers can often reconstruct the original training samples with surprising accuracy - including images at batch sizes of up to 48. The assumption that "we are only sharing gradients, not data" turns out to be much weaker than it appears.

2.4 Proving Someone Was in the Dataset

Membership inference attacks take a different angle. Instead of recovering actual data, the goal is simply to determine whether a specific person's record was part of the training set. This might sound like a narrow concern, but the implications are serious. If a model was trained on hospital records of cancer patients, and an attacker can confirm that a particular individual's data was included, that effectively reveals a sensitive medical fact about that person. Carlini et al. [8] showed that these attacks are far more powerful than the research community had previously assumed, and that standard DP implementations may not fully protect against them.

2.5 Data Silos Block Collaboration

Some of the most valuable AI applications require data from multiple organizations. A model that could detect rare diseases would benefit from medical records across hundreds of hospitals. A fraud detection system would be stronger if it could learn from transaction patterns across all banks, not just one. But hospitals cannot share patient records. Banks cannot share customer data. Researchers cannot pool genomic datasets without navigating a maze of consent and regulatory requirements. The result is that models end up being trained on whatever data a single institution happens to have, which is rarely enough to see the full picture [6].

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154298

3. LITERATURE REVIEW

Research on PPML has moved quickly over the past few years, driven both by the urgency of the privacy problem and by the rapid growth of AI itself. Here we review the key works from 2021 to 2026 that have most directly shaped our understanding of the field.

The most ambitious attempt to map the full scope of Federated Learning came from Kairouz et al. [1], whose survey covered over 200 papers and identified the core open problems: how to handle data that is not evenly distributed across devices, how to reduce the cost of communicating model updates, and how to integrate differential privacy into FL without destroying model accuracy. This work remains the definitive reference for anyone building production FL systems.

The memorization problem in language models was brought into sharp focus by Carlini et al. [2], who demonstrated that GPT-2 could be prompted to spit out verbatim training data - names, addresses, even unique text strings. Their work forced the research community to take model memorization seriously as a practical threat rather than a theoretical curiosity.

On the solutions side, Andrew et al. [3] made a meaningful improvement to DP-SGD by introducing adaptive gradient clipping. The original DP-SGD algorithm required researchers to manually set a clipping threshold, which had a big effect on model accuracy and was hard to tune correctly. Adaptive clipping estimates this threshold automatically during training, using the data in a differentially private way. The result was 2 to 3 percent better model accuracy at the same privacy level - a modest-sounding improvement that matters a lot in practice.

Huang et al. [4] took a rigorous look at gradient inversion attacks and the defenses against them. They tested five different attack methods against six different defenses and found that most defenses provide weaker protection than claimed. Only differential privacy at the gradient level consistently blocked the attacks - but always at some cost to model performance. Their work established a more honest picture of what FL alone can and cannot guarantee.

Perhaps the most encouraging result from this period came from De et al. [9], who showed that the accuracy penalty of DP training shrinks dramatically when model and dataset size are large enough. Their experiments on ImageNet demonstrated that a large ViT model trained with strong privacy guarantees (epsilon = 10) achieved 86.7 percent accuracy, compared to 90 percent without privacy. That is a meaningful gap, but far smaller than earlier research had suggested. The key insight was that large models are more robust to the noise that DP injects.

For practitioners trying to actually deploy DP in real systems, Ponomareva et al. [11] published what amounts to a manual for doing it right. Drawing on Google's internal deployments, they document the privacy accounting methods, clipping strategies, and tuning approaches that work in production - alongside the pitfalls that are easy to fall into. This kind of applied guidance has been missing from the academic literature for a long time.

The most compelling real-world validation of PPML came from Dayan et al. [10], who trained a COVID-19 outcome prediction model across 20 hospitals on three continents without any patient data ever leaving those hospitals. Not only did the federated model match the performance of a centrally trained model, it outperformed every single-hospital model. This result matters because it directly answers the skeptics who argue that privacy-preserving methods are interesting in theory but too costly in practice.

4. PPML FUNDAMENTALS AND PRIVACY-ENHANCING TECHNIQUES

There is no single technique that solves all privacy problems in machine learning. Different approaches make different trade-offs, and the right choice depends on the specific application, the threat model, the available computational resources, and the accuracy requirements. This section introduces the four main privacy-enhancing techniques and explains how each one works.

4.1 Differential Privacy (DP)

Differential Privacy is the most mathematically rigorous approach to privacy in ML. The core idea is elegant: if you add enough carefully calibrated noise to the information that leaves your system, no outside observer should be able to tell whether any specific individual's data was included. Formally, a mechanism M satisfies (epsilon, delta)-DP if the probability of any output changes by at most a factor of e^epsilon when one person's record is added or removed from the dataset. A smaller epsilon means stronger privacy but also more noise - and more noise means lower model accuracy.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154298

In practice, DP is implemented in ML through an algorithm called DP-SGD. During training, instead of computing gradients over a whole batch, the system computes gradients per individual sample, clips each one to prevent any single record from having too much influence, adds Gaussian noise, and then averages. The privacy cost accumulates over training steps and is tracked using a privacy accountant. Andrew et al. [3] improved this process significantly by automating the clipping threshold, removing one of the trickiest hyperparameters to tune.

4.2 Federated Learning (FL)

Federated Learning flips the standard ML workflow on its head. Instead of bringing data to the model, you bring the model to the data. A central server sends the current model to a set of devices - phones, hospitals, banks - each device trains locally on its own data and sends back only the weight updates, never the raw data itself. The server aggregates these updates and sends a new model out. Repeat until the model is good enough.

This architecture is genuinely useful in situations where data cannot move. Your keyboard app can learn your typing habits without those habits ever leaving your phone. A hospital can contribute to a shared diagnostic model without sharing patient records. Kairouz et al. [1] identified several real engineering challenges in making this work at scale: data is distributed unevenly across devices, network connections are unreliable, and some participants may try to manipulate the model by sending malicious updates.

4.3 Homomorphic Encryption (HE)

Homomorphic Encryption sounds like science fiction: it lets you perform mathematical calculations on encrypted data, and when you decrypt the result, you get the same answer you would have gotten by calculating on the original unencrypted data. In an ML context, this means a hospital could encrypt a patient's scan, send it to an AI service, receive an encrypted diagnosis, and decrypt it locally - without the service ever seeing the actual scan or the actual result. Chillotti et al. [5] made this significantly more practical by introducing programmable bootstrapping for the TFHE scheme, enabling more complex neural network operations on encrypted data. The catch is that HE is still very slow - often 100 to 1000 times slower than unencrypted computation - though specialized hardware is beginning to change that.

4.4 Secure Multi-Party Computation (SMPC)

SMPC is a cryptographic technique that lets multiple parties compute something together without any party learning what the others contributed. Imagine three banks that want to know the average fraud rate across all of them, but none wants to reveal its own numbers. With SMPC, they can compute the average collaboratively, and each bank learns only the final result - not the inputs from the others. In ML, SMPC is used for cross-institutional model training and secure aggregation in federated settings. It provides stronger guarantees than FL alone, but at a higher computational cost.

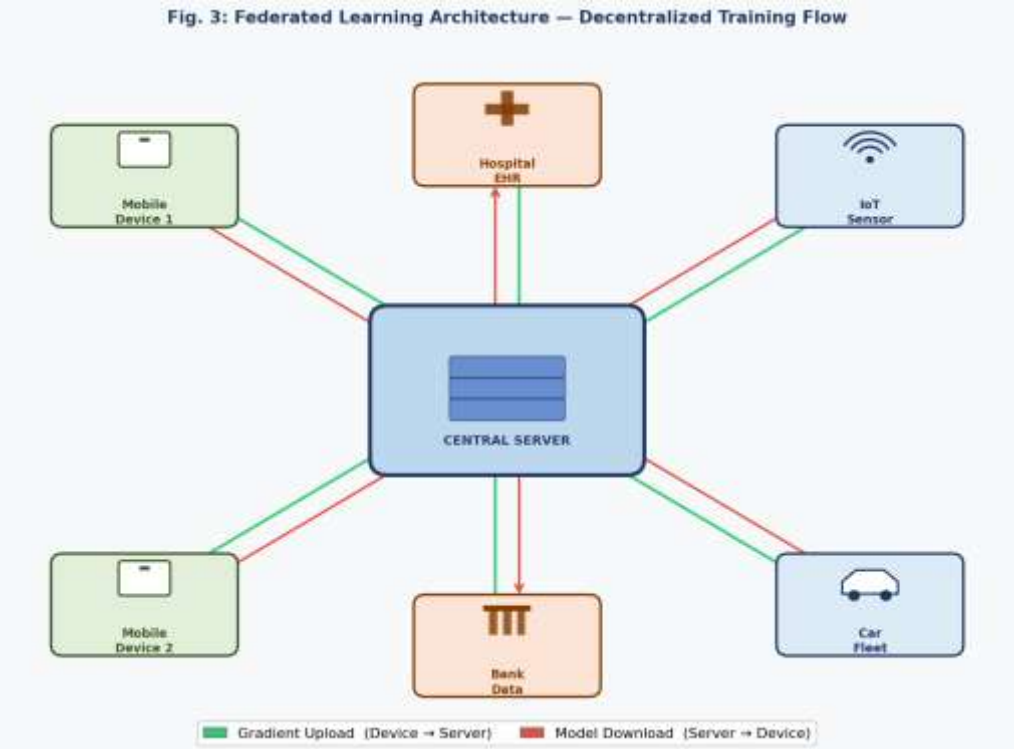

Fig. 3: Federated Learning Architecture — Decentralized Training Flow

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154298

Table 1. Comparative Overview of the Four Main Privacy-Enhancing Techniques

Technique	Privacy Level	Speed	Accuracy	Best Use Case
Differential Privacy	High (tunable)	Very Fast	85-95%	Cloud ML training
Federated Learning	Medium-High	Fast	88-95%	Mobile / Edge AI
Homomorphic Enc.	Highest	Very Slow	~95%	Secure inference
SMPC	Very High	Slow	90-95%	Cross-institutional
Hybrid (DP + FL)	High	Moderate	87-93%	Production systems

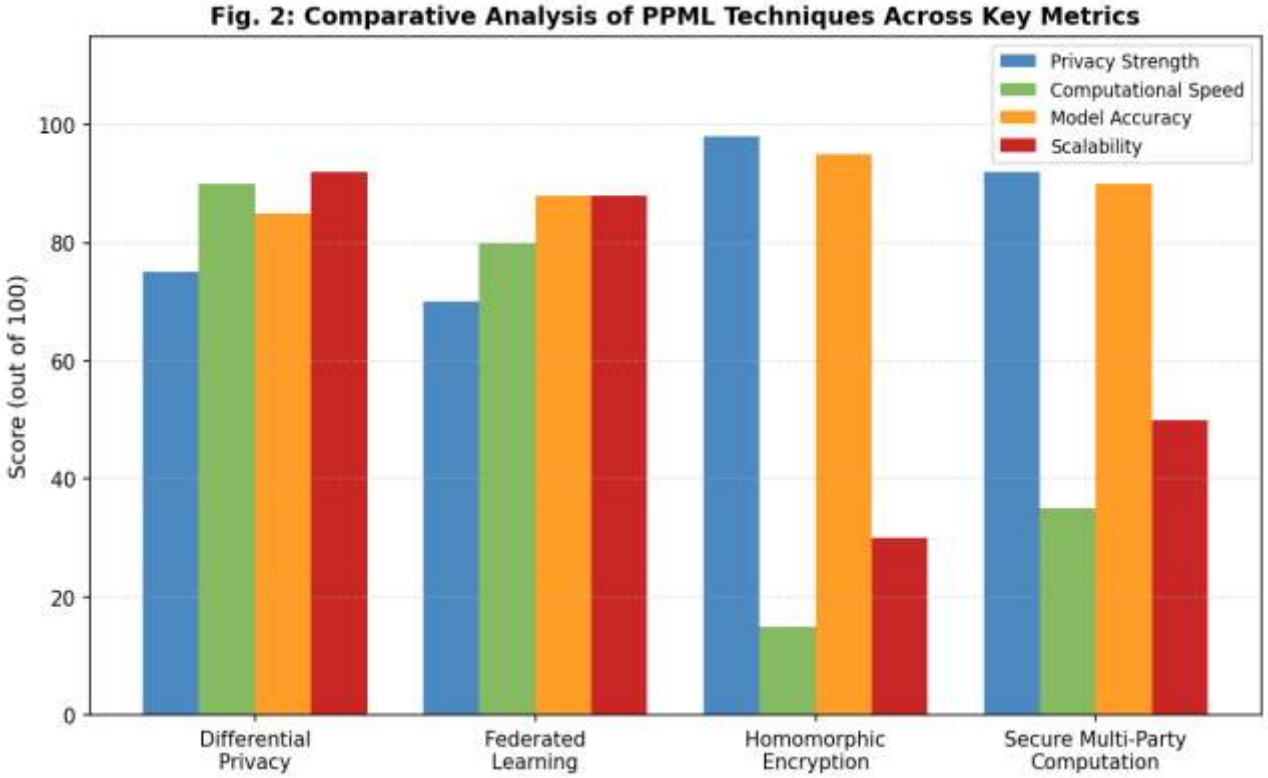

Fig. 2: Comparative Analysis of PPML Techniques Across Key Performance Metrics

5. STRUCTURAL PIPELINES OF A PPML SYSTEM

A PPML system is not just a model with some noise added on top. It requires privacy to be designed in at every stage of the pipeline, from the moment data is collected to the moment a prediction is returned. The five stages below describe how a well-designed PPML system is structured.

5.1 Stage 1: Secure Data Acquisition

Privacy protection has to start before the data even leaves the user. In a properly designed system, data is either kept entirely on the user's device or encrypted at the point of collection before being transmitted anywhere. Local Differential Privacy can be applied here as well - adding randomization to individual data points before they are ever seen by a server. This provides the strongest possible privacy guarantee but comes at the cost of needing more data to achieve the same model accuracy, since each data point carries more noise.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154298

5.2 Stage 2: Privacy-Aware Training

This is where the main privacy mechanism - whether DP-SGD, FL, or HE - is applied. In the case of DP-SGD, the system clips individual gradient contributions, adds noise, and tracks the accumulating privacy budget using an accountant such as the Renyi DP method. The training process has to be designed so that the total privacy cost over all steps stays within the agreed budget. Ponomareva et al. [11] provide detailed guidance on how to set these parameters correctly and what mistakes to avoid - their work is required reading for any team attempting a production deployment.

5.3 Stage 3: Encrypted Model Storage

A trained model is not just a black box - its weights contain information about the training data. This is what makes model inversion attacks possible. To prevent this, production systems store model weights in encrypted form, protected by hardware security modules or secure enclaves like Intel SGX. Even if someone gains access to the storage system, they should not be able to extract meaningful information from the encrypted weights.

5.4 Stage 4: Secure Inference

When a user submits a query to a deployed model, the privacy protections should not stop there. In HE-based inference using the TFHE scheme [5], the user encrypts their input before sending it. The model server performs its computations on the ciphertext and returns an encrypted result. The user decrypts it on their own device. The server sees neither the input nor the output in any usable form. This matters most in sensitive domains like medical diagnosis, where the query itself - a description of symptoms, an image of a scan - may be as sensitive as the answer.

5.5 Stage 5: Continuous Privacy Auditing

Privacy budgets are not infinite. Each query, each training step, each model update consumes some portion of the total allowed privacy cost. A responsible PPML system tracks this continuously and enforces limits. When the budget is running low, the system can add more noise, stop accepting queries, or retrain from a fresh budget. Without this kind of ongoing monitoring, it is easy for the cumulative privacy cost to drift far beyond what was originally intended.

6. DETAILED REVIEW OF TECHNICAL APPROACHES

6.1 Where the Noise Goes: Perturbation Strategies

One of the most important design decisions in a PPML system is where to inject the noise that protects individual privacy. There are three main options, and the choice involves real trade-offs:

* Input Perturbation: Noise is added to the raw data before it ever enters the training pipeline. This approach - called Local Differential Privacy - gives the strongest individual guarantee because no one, not even the model owner, ever sees the true data. The downside is efficiency: because each data point is independently noised without any aggregation benefit, you typically need far more data to achieve the same model quality [6].
* Gradient Perturbation (DP-SGD): Noise is added to the gradient updates during backpropagation. This is the most widely used approach in practice because it offers a good balance - stronger accuracy than local DP, and provable privacy guarantees. Andrew et al. [3] made this more practical by showing that the clipping threshold, which previously required careful manual tuning, can be set adaptively during training.
* Objective Function Perturbation: The loss function itself is modified to discourage the model from fitting too closely to individual examples. This works best for simpler models like logistic regression and SVMs, where strong theoretical guarantees on both privacy and convergence are available.

6.2 When to Use Cryptography and When Not To

Cryptographic methods like HE and SMPC provide the strongest possible privacy - their guarantees hold regardless of what the adversary knows or how much compute they have. But they are expensive. Training a model on HE-encrypted data can easily be 100 to 1000 times slower than training on plaintext. For many applications, that cost is simply not acceptable.

Non-cryptographic methods like DP and FL are much faster and have been successfully deployed at massive scale - billions of devices, hundreds of millions of users. Their privacy guarantees are statistical rather than absolute, but for many practical threat models, that is good enough. Most serious real-world deployments today use a hybrid approach: FL handles the architectural problem of keeping data on-device, while DP handles the statistical problem of ensuring gradient updates do not reveal too much. Together they offer a practical sweet spot between security and performance [11].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154298

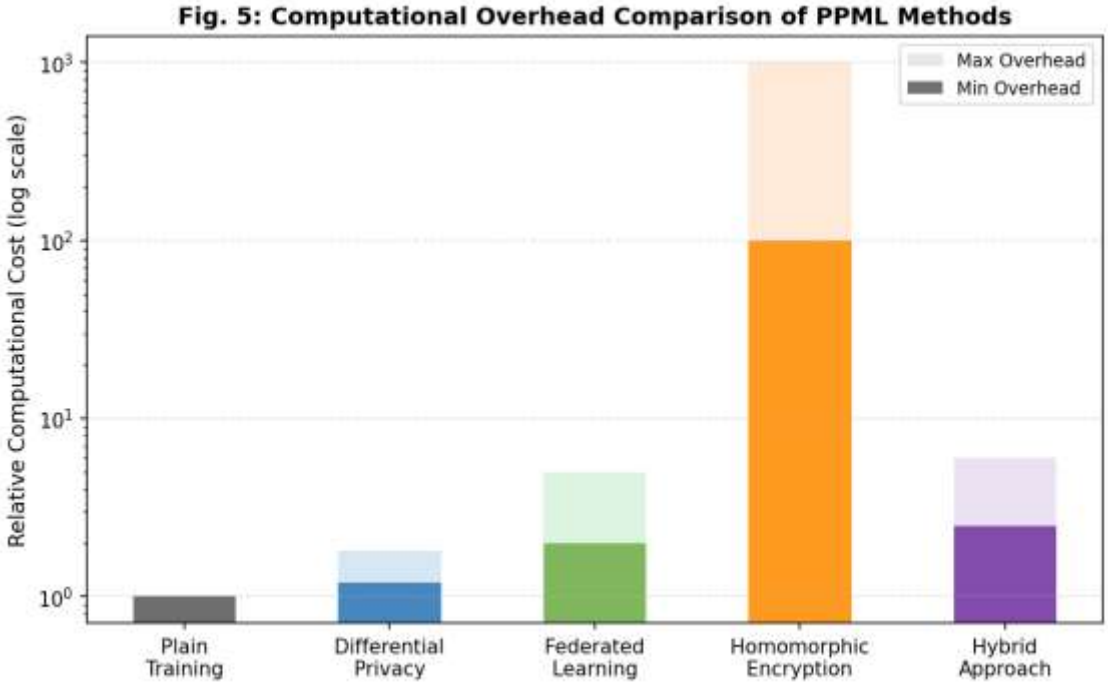

Fig. 5: Computational Overhead Comparison of PPML Methods (Log Scale)

7. REAL-WORLD INDUSTRY APPLICATIONS

It is one thing to describe PPML techniques in the abstract. It is another to see them solving real problems for real organizations. The following examples come from deployments that have been documented in the research literature and show that PPML is not just a research concept but a practical engineering discipline.

7.1 Healthcare: Learning Without Sharing Patient Records

Medicine is perhaps the most compelling use case for PPML. Medical data is sensitive by definition, sharing it across institutions is heavily regulated, and yet the most important medical discoveries come from studying large, diverse patient populations. The COVID-19 pandemic put this tension in sharp relief. Dayan et al. [10] responded by building a federated learning system that trained a COVID-19 outcome prediction model across 20 hospitals in the US, Europe, and Asia - without a single patient record leaving any hospital. The resulting model outperformed every single-hospital model and came close to matching what could have been achieved with fully centralized data. This is the kind of result that makes hospital CIOs pay attention.

7.2 Finance: Finding Fraud Across Institutional Boundaries

Fraud detection is a classic case where collaboration would help enormously but data sharing is off the table. A fraudster who bounces transactions across multiple banks is invisible to any single institution, but their pattern would be obvious if the banks could share data. SMPC provides a way out: the banks can collectively train a fraud model without any bank seeing another's transactions. European banking pilots using this approach have reported significant improvements in fraud detection rates [10] - enough to make the computational overhead worthwhile.

7.3 Consumer Technology: Keeping Your Typing Private

When you use your phone's keyboard app, every keystroke is potentially sensitive. Google's Gboard uses Federated Learning to improve next-word prediction across over a billion Android devices - learning from how people actually type without any keystrokes being sent to Google's servers. Apple takes a similar approach for Siri and autocorrect. These are not experimental prototypes; they are production systems running on devices in people's pockets right now. They work because FL allows the model to improve from real-world usage without collecting the usage data centrally.

7.4 Autonomous Vehicles: Smarter Driving Without Location Tracking

Autonomous vehicles generate an enormous amount of data - camera feeds, sensor readings, GPS coordinates, near-miss events. This data is invaluable for training better driving models, but it is also deeply revealing: location data can expose where someone lives, works, and worships. Federated learning allows car manufacturers to train on real driving data

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154298

collected by their entire fleet without that location and camera data ever leaving the vehicles. Kairouz et al. [1] point to this as one of the clearest examples of cross-silo FL in action.

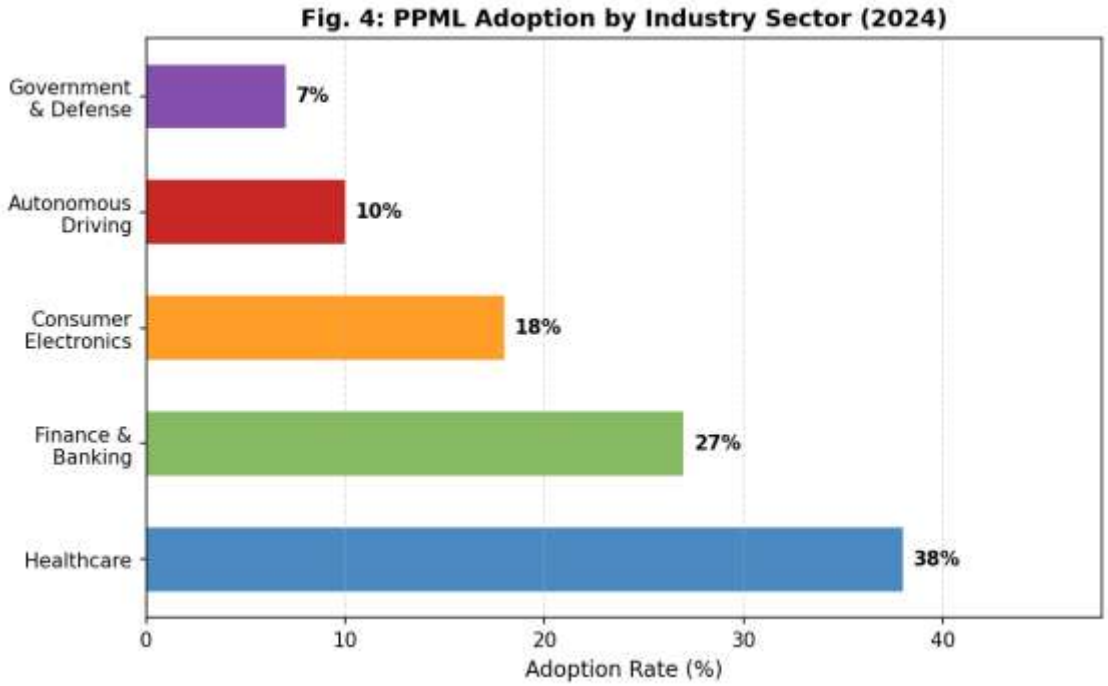

Fig. 4: PPML Adoption Rate by Industry Sector (2024)

Table 2. PPML Deployments by Industry - Technique, Regulation, and Key Benefit

Industry	Primary Technique	Regulation	Key Benefit
Healthcare	FL + DP-SGD	HIPAA / GDPR	Cross-hospital training without data sharing
Finance	SMPC	FINRA / PCI-DSS	Cross-bank fraud detection
Consumer IoT	FL + Local DP	GDPR / CCPA	On-device personalization
Autonomous Driving	FL + DP	ISO 26262	Fleet-wide safety learning
Government	HE + SMPC	NIST / FISMA	Secure census and statistical analytics

8. PERFORMANCE AND PRIVACY EVALUATION

Evaluating a PPML system is more nuanced than evaluating a standard ML model. Accuracy alone does not tell you whether the system is actually private, and privacy metrics alone do not tell you whether the system is actually useful. You need both, and you need to understand the trade-off between them.

8.1 The Privacy Budget: Epsilon

Epsilon is the core metric in differential privacy - it quantifies how much information leaks through the mechanism. A smaller epsilon means stronger privacy and more noise, which typically means lower accuracy. A larger epsilon means the model is more accurate but individuals are less protected. In practice, epsilon values below 1.0 are considered strong privacy, while values above 10.0 provide limited formal guarantees. De et al. [9] made an important contribution by showing that at epsilon = 10, a large vision model can still achieve 86.7 percent accuracy on ImageNet - demonstrating that "private" and "accurate" are not mutually exclusive when the model is large enough. The chart below shows how accuracy degrades as epsilon decreases, illustrating the trade-off clearly.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154298

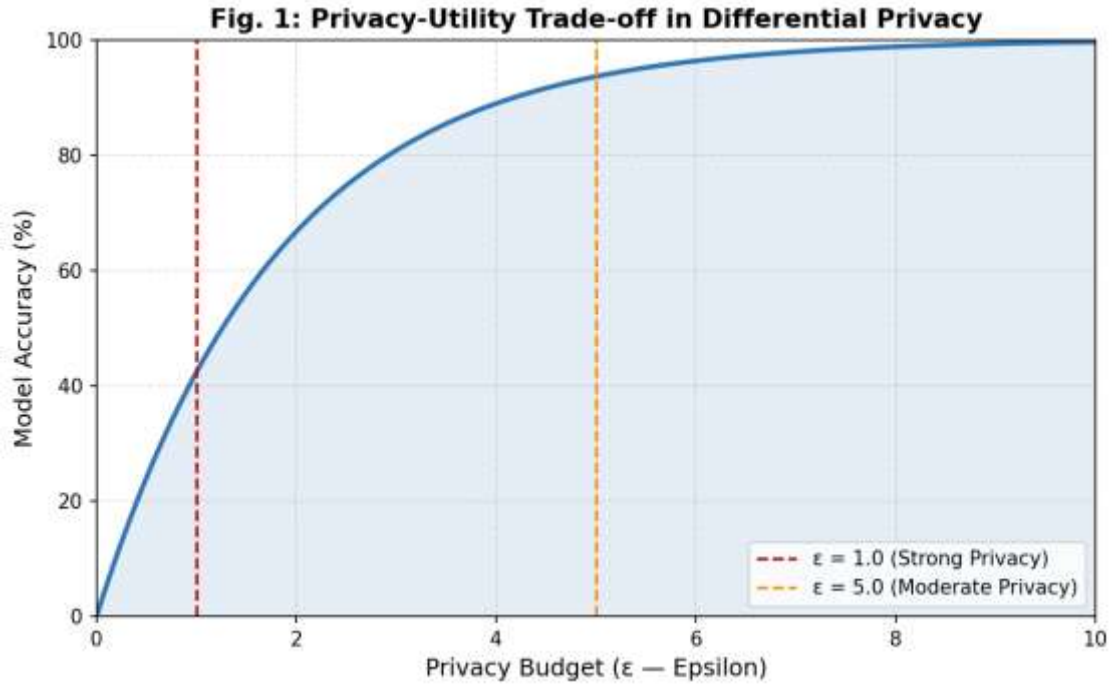

Fig. 1: Privacy-Utility Trade-off: Epsilon vs. Model Accuracy in Differential Privacy

8.2 Model Fidelity

Model fidelity asks a simple question: how much accuracy do we lose by adding privacy? The answer depends heavily on the task, the dataset size, and the model architecture. For large language model tasks, research from this period suggests the gap is shrinking. At epsilon = 8, differentially private BERT loses less than 1.5 percent accuracy compared to non-private BERT on standard benchmarks - a difference that would be acceptable in most production settings. For smaller datasets or more specialized tasks, the gap can be much larger.

8.3 Computational and Communication Overhead

Privacy comes with a performance cost, and for systems deployed on mobile devices or in low-bandwidth environments, this matters. DP-SGD adds roughly 2 to 5 times the computation of standard training. FL adds communication overhead that scales with the number of training rounds and the size of model updates. HE adds the most - currently 100 to 1000 times the compute of plaintext operations, though hardware accelerators are beginning to bring this down. Any serious PPML deployment needs to measure and budget for these overheads alongside accuracy metrics.

Table 3. Key Evaluation Metrics for PPML Systems

Metric	What It Measures	Target Range	How to Measure
Epsilon (e)	Total privacy loss over training	e < 1.0 (Strong)	Renyi DP Accountant
Model Fidelity	Accuracy vs. non-private baseline	> 90%	Cross-validation
MIA Resistance	How hard is membership inference	AUC < 0.6	Shadow model attack
Communication Cost	Gradient transmission overhead in FL	< 5x baseline	Network profiling
Training Overhead	Extra time vs. standard training	< 10x (DP-SGD)	Wall-clock timing

9. CRITICAL CHALLENGES AND LIMITATIONS

PPML has come a long way, but it would be misleading to suggest the problem is solved. Several challenges remain that limit how far and how fast these techniques can be deployed. Being honest about these limitations is important both for researchers deciding where to focus their efforts and for practitioners deciding what they can realistically guarantee.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154298

9.1 The Accuracy Cost Is Real, Even If Shrinking

The fundamental trade-off between privacy and utility has not been eliminated - it has just been pushed further out. De et al. [9] showed that large models lose only a few percent accuracy at moderate epsilon values. But many real-world applications do not have access to large models or large datasets. A rural clinic trying to train a diagnostic model on a few hundred patient records will experience a much steeper accuracy penalty than a tech giant fine-tuning GPT on billions of examples. The privacy-utility trade-off is not evenly distributed, and the organizations with the least data often face the hardest choices.

9.2 Encryption Is Still Too Slow for Many Use Cases

Homomorphic Encryption offers the strongest privacy guarantees of any approach in the PPML toolkit, but it remains computationally expensive to a degree that limits its practical applications. Training a neural network on HE-encrypted data is currently 100 to 1000 times slower than training on plaintext. Hardware accelerators are making progress - specialized chips have demonstrated dramatic speedups in research settings - but these solutions are not yet available in standard cloud infrastructure. Until they are, HE will remain confined to inference tasks and relatively simple models.

9.3 Attacks Keep Getting Better

One of the uncomfortable realities of PPML research is that the attackers and defenders are in an arms race, and the attackers have not run out of ideas. Carlini et al. [8] showed that membership inference attacks are much more powerful than previously thought. Huang et al. [4] showed that gradient inversion works at larger batch sizes than previously demonstrated. Byzantine attacks on federated learning - where malicious participants submit crafted gradient updates to manipulate the global model - remain an open problem. Implementing PPML correctly is not enough; systems need to be continuously re-evaluated against new attack methods.

9.4 Privacy Can Hurt Fairness

One of the less-discussed side effects of differential privacy is its uneven impact across demographic groups. Research has shown that DP-SGD tends to degrade accuracy more for minority groups and underrepresented populations than for majority groups. This creates a troubling situation where privacy protection and fairness protection can work against each other: the same noise that shields a minority individual from re-identification may also reduce the quality of predictions they receive. This is not a theoretical concern - it has practical implications for any PPML deployment in healthcare, lending, or hiring.

9.5 There Is No Agreed Standard for Reporting Privacy

Ask five different research papers what epsilon they used and you may get five answers that look comparable but are not. Pure DP, approximate DP, Renyi DP, and zero-concentrated DP all measure privacy in subtly different ways, and a reported epsilon of 1.0 under one framework may correspond to a very different privacy level under another. Without standardization, it is difficult for practitioners to compare different systems honestly or for regulators to set meaningful requirements. NIST's AI Risk Management Framework [12] is a step toward addressing this, but comprehensive technical standards for PPML are still being developed.

10. FUTURE RESEARCH AND REGULATORY DIRECTIONS

Despite the challenges above, there is genuine reason for optimism about where PPML is headed. The pace of progress over the past five years has been remarkable, and several research directions look particularly promising.

10.1 Letting Users Choose Their Own Privacy Level

Right now, privacy decisions in PPML systems are made entirely by the organizations that deploy them. A hospital or a tech company sets the epsilon value and decides what technique to use, with no input from the individuals whose data is at stake. There is growing interest in developing interfaces that let users meaningfully participate in these decisions - choosing, for example, whether they prefer stronger privacy at the cost of a less personalized experience. This is harder than it sounds: most people have no intuition for what epsilon = 1.0 means. Making privacy choices accessible to non-technical users is an open and important research problem.

10.2 Making Private Models Explainable

GDPR gives individuals the right to an explanation when an automated system makes a decision that affects them. This creates a tension with PPML: the noise and obfuscation that protect privacy also tend to make models harder to interpret. Researchers are beginning to explore how to satisfy both requirements at once - building models that are private by design and explainable by design. This is challenging technically, but it is the right problem to be working on.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154298

10.3 Hardware That Can Handle Encrypted Computation

The computational overhead of Homomorphic Encryption is not a fundamental limit - it is an engineering problem. Dedicated hardware designed specifically for HE operations has demonstrated speedups of 4000x or more over general-purpose software implementations. As this hardware matures and becomes commercially available in cloud data centers, it will open up use cases that are currently off the table. This is one of the areas where investment from both industry and government is likely to pay off significantly over the next decade.

10.4 Private Fine-Tuning of Foundation Models

Foundation models - large pre-trained models like GPT - have changed how AI development works. Instead of training from scratch, organizations fine-tune a foundation model on their own domain data. Recent work has shown that this two-stage approach (public pre-training, private fine-tuning) dramatically reduces the accuracy cost of DP training. The intuition is that the model already knows a lot from the public pre-training; the fine-tuning only needs to adjust a little, so there is less noise needed to protect each update. This is a practical pathway to privacy-preserving AI for organizations that cannot afford to train large models from scratch.

10.5 Clearer Regulations and Technical Standards

Technical solutions can only go so far without clear regulatory frameworks to tell organizations what is actually required of them. GDPR set an important precedent by making privacy a legal requirement, but it says relatively little about how to achieve it technically. NIST's AI RMF [12] and the proposed EU AI Act are moving in the right direction - establishing risk-based requirements that scale with the sensitivity of the application. International coordination on technical standards for PPML, through bodies like ISO/IEC, would help organizations operating across multiple jurisdictions avoid contradictory compliance requirements.

Table 4. Emerging Research Directions in PPML (2025-2030)

Direction	What It Aims to Solve	Timeline
User-Controlled Privacy	Let individuals set their own privacy preferences	2025-2027
Explainable PPML	Private and interpretable models for regulated industries	2025-2028
FHE Hardware Chips	Specialized silicon to make HE fast enough for production	2026-2030
Private Foundation Models	DP fine-tuning of LLMs using public pre-training	2025-2027
PPML Standards	ISO/IEC and NIST compliance frameworks for PPML	2025-2027
Fair Private ML	Preventing DP from amplifying bias in minority groups	2025-2028

11. CONCLUSION

Privacy and intelligence are not opposites. That is perhaps the central lesson of this field. For a long time, the assumption in machine learning was that you had to choose: either you collected all the data you needed and built the best possible model, or you protected privacy and settled for something worse. PPML has challenged that assumption, and the results from the past five years suggest the challenge is justified.

We have seen federated learning deployed across a billion devices and twenty hospitals. We have seen differential privacy move from a theoretical framework into production systems at Google, Apple, and healthcare institutions around the world. We have seen encryption-based inference become fast enough for real applications, and we have seen researchers demonstrate that large models can be trained with strong privacy guarantees while losing only a few percent accuracy.
At the same time, we tried to be honest about what has not been solved. The privacy-utility trade-off is real and falls hardest on those with the least data. Cryptographic methods are still too slow for many use cases. Attacks keep getting

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154298

more sophisticated. Privacy protections can inadvertently worsen fairness for already underrepresented groups. And without standardization, comparing and regulating PPML systems remains difficult.

What gives us confidence in the direction of travel is not just the technical progress – it is the alignment of incentives. Regulations are pushing organizations toward privacy-preserving architectures. Users are becoming more aware of and concerned about how their data is used. And the research community is producing increasingly practical solutions. PPML is no longer a niche subfield of cryptography – it is becoming a core discipline for anyone building AI systems that interact with real people and real data. That is where it belongs.

REFERENCES

[1]. Kairouz, P., et al. (2021). Advances and Open Problems in Federated Learning. Foundations and Trends in Machine Learning, 14(1-2), pp. 1-210.
[2]. Carlini, N., et al. (2021). Extracting Training Data from Large Language Models. Proc. 30th USENIX Security Symposium, pp. 2633-2650.
[3]. Andrew, G., et al. (2021). Differentially Private Learning with Adaptive Clipping. Advances in Neural Information Processing Systems (NeurIPS), Vol. 34, pp. 17455-17466.
[4]. Huang, Y., et al. (2021). Evaluating Gradient Inversion Attacks and Defenses in Federated Learning. NeurIPS, Vol. 34, pp. 7232-7241.
[5]. Chillotti, I., et al. (2021). Programmable Bootstrapping Enables Efficient Homomorphic Inference of Deep Neural Networks. Proc. CSCML 2021, LNCS Vol. 12716, Springer, pp. 1-19.
[6]. Li, T., et al. (2021). Federated Learning: Challenges, Methods, and Future Directions. IEEE Signal Processing Magazine, 37(3), pp. 50-60.
[7]. European Data Protection Board. (2022). EDPB Annual Report 2022: GDPR Enforcement and Compliance Overview. EDPB, Brussels.
[8]. Carlini, N., et al. (2022). Membership Inference Attacks From First Principles. Proc. IEEE Symposium on Security and Privacy (SP), pp. 1897-1914.
[9]. De, S., et al. (2022). Unlocking High-Accuracy Differentially Private Image Classification through Scale. NeurIPS, Vol. 35, pp. 21358-21371.
[10]. Dayan, I., et al. (2021). Federated Learning for Predicting Clinical Outcomes in Patients with COVID-19. Nature Medicine, 27, pp. 1735-1743.
[11]. Ponomareva, N., et al. (2023). How to DP-fy ML: A Practical Guide to Machine Learning with Differential Privacy. Journal of Machine Learning Research, 24(226), pp. 1-77.
[12]. NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, Gaithersburg, MD.

