NSS-SocialSec 2023

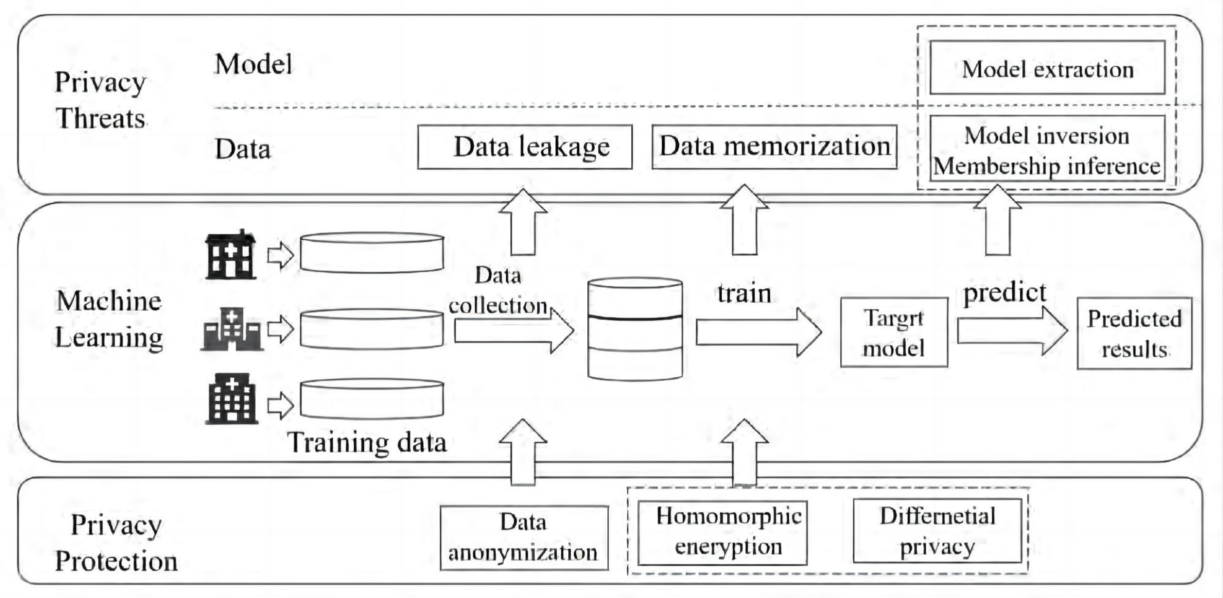
RRML: Privacy Preserving Machine Learning Based on Random Response Technology

Jia Wang[1], Shiqing He[2], and Qiuzhen Lin[1]

[1] Shenzhen University, 3688 Nanhai Avenue, Shenzhen, Guangdong Province, CN
[2] Shangqiu University, No. 66, Beihai East Road, Shangqiu, Henan Province, CN
{jia.wang, qiuzhlin}@szu.edu.cn

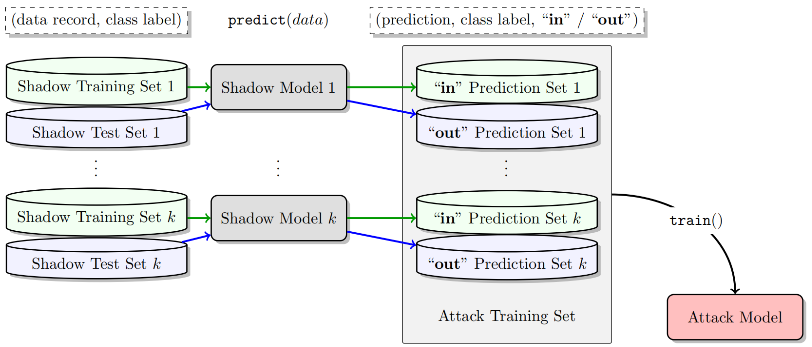
SHENZHEN
UNIVERSITY

A. Background of research field

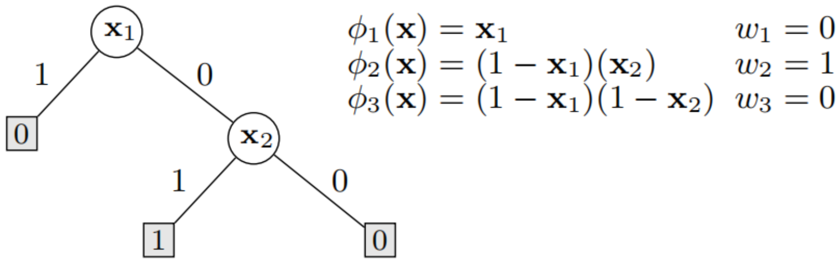


- In recent years, the combination of machine learning models and big data analytics techniques has led to significant advancements in healthcare, clinical decision support, and other fields by harnessing the potential of massive healthcare data for new research insights.
- However, machine learning algorithms also face potential risks of privacy leackage and are vulnerable to various attacks

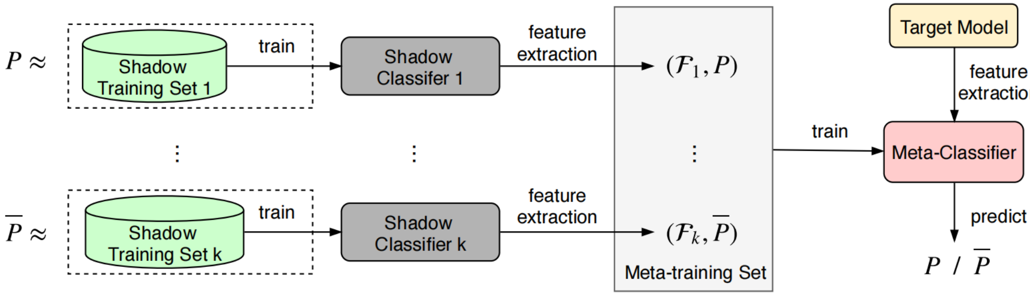
B. The main attack methods in multi-party data sharing



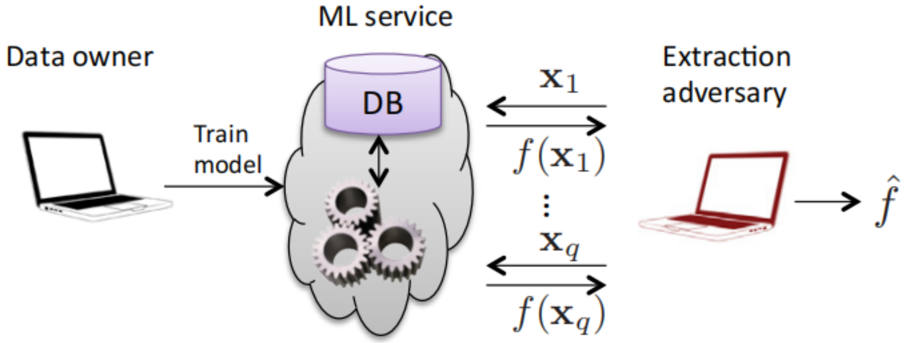
[1]


$\mathbf{x}_1$
1
0
0
$\mathbf{x}_2$
1
0
1
0
$\phi_1(\mathbf{x}) = \mathbf{x}_1$ $w_1 = 0$
$\phi_2(\mathbf{x}) = (1 - \mathbf{x}_1)(\mathbf{x}_2)$ $w_2 = 1$
$\phi_3(\mathbf{x}) = (1 - \mathbf{x}_1)(1 - \mathbf{x}_2)$ $w_3 = 0$

[2]



[3]



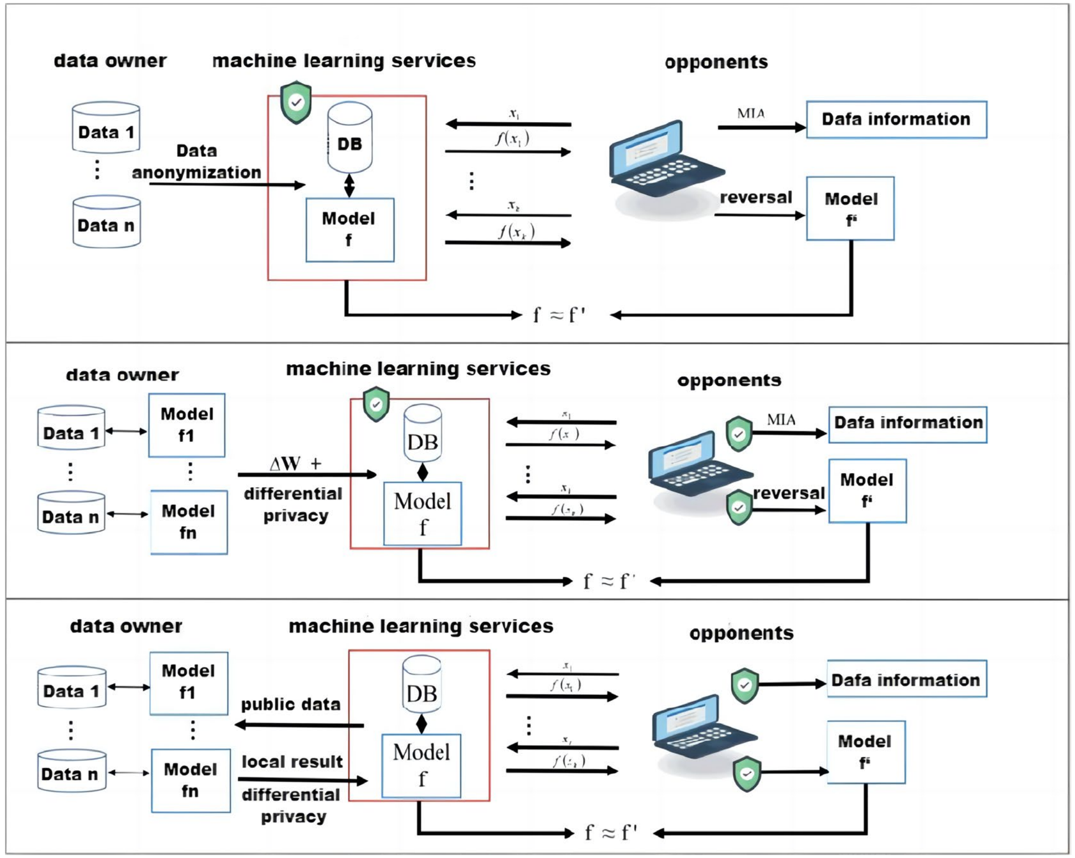
[4]

- Four methods:

  Membership Inference Attack(MIA), Model Inversion, Attribute Inference and Model Extraction
- These attacks primarily target training data and model parameters with the goal of extracting sensitive information. These attacks can be conducted individually or in combination, posing varying degrees of threats to privacy.

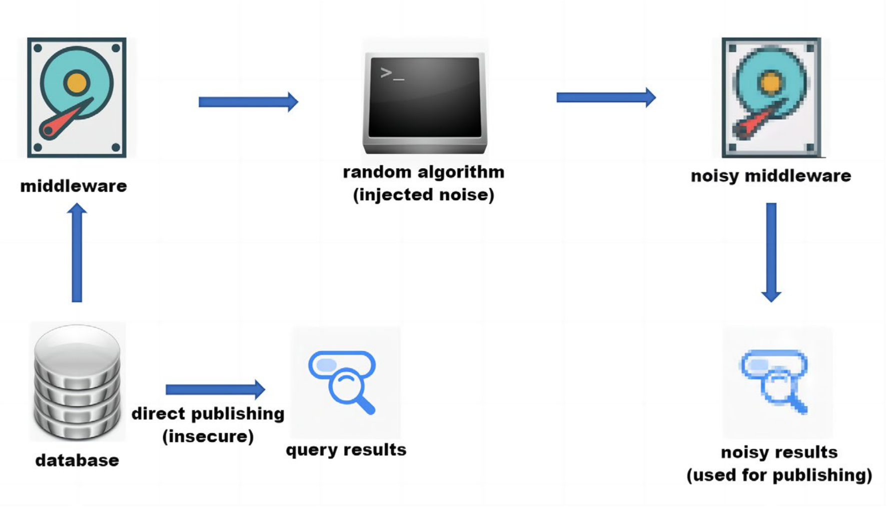
[1] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In IEEE S&P, 2017.
[2] M. Fredrikson, S. Jha, and T. Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In ACM CCS, 2015.
[3] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov. Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations. In CCS, 2018.
[4] F. Tramer, F. Zhang, A. Juels, M. K. Reiter, and T. Risten- ` part. Stealing machine learning models via prediction APIs. In USENIX Security, 2016.

B. Methods for privacy protection in multi-party data sharing



- these methods: Data anonymization techniques, Adding differential privacy to model parameters and Third-party differential privacy(such as PATE).
- These methods provide means for privacy protection, but there are still issues that need further exploration in achieving the optimal balance between security and utility.

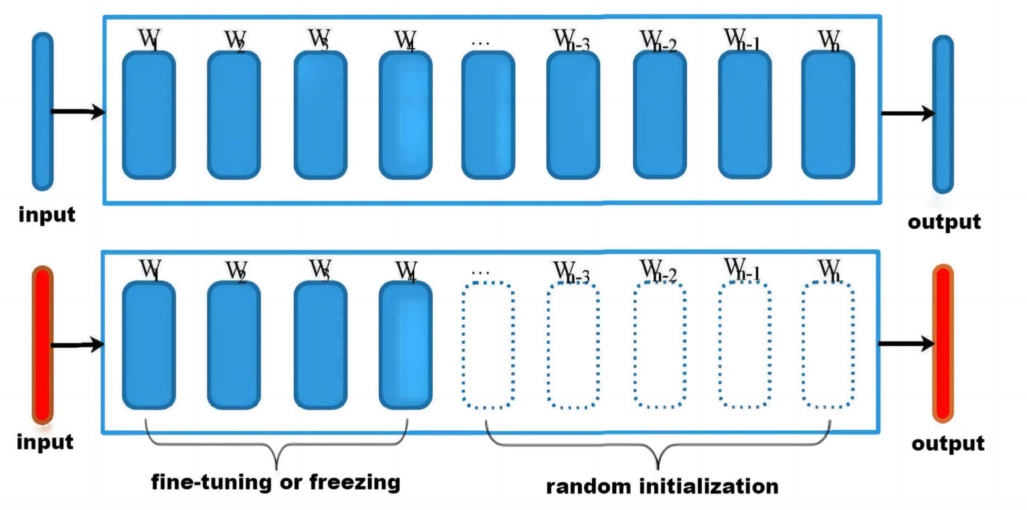
C. Research Objectives: some specific algorithms



[1]

$$c(o;A,aux,d,d') \overset{\Delta}{=} \log \frac{\Pr\left[A(aux,d)=o\right]}{\Pr\left[A(aux,d')=o\right]}$$

[2]



[4]

$$P(X_i =\text{" }yes\text{ "}) = \pi p + (1-\pi)(1-p)$$
$$P(X_i =\text{" }no\text{ "}) = (1-\pi)p + \pi(1-p)$$

[3]

- *differential privacy, the mathematical definition and computation methods of moments accountant*, and *the privacy computation methods using random response for binary and multi-class classification*.
- *the specific implementation method of parameter-based transfer learning*.

[1] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, Automata, Languages and Programming, pages 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[3] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pages 308–318. ACM, 2016.

[4] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10):1345–1359, 2009.

C. RRML: Teacher Consensus Aggregation Learning Mechanism based on Randomized Response Differential Privacy

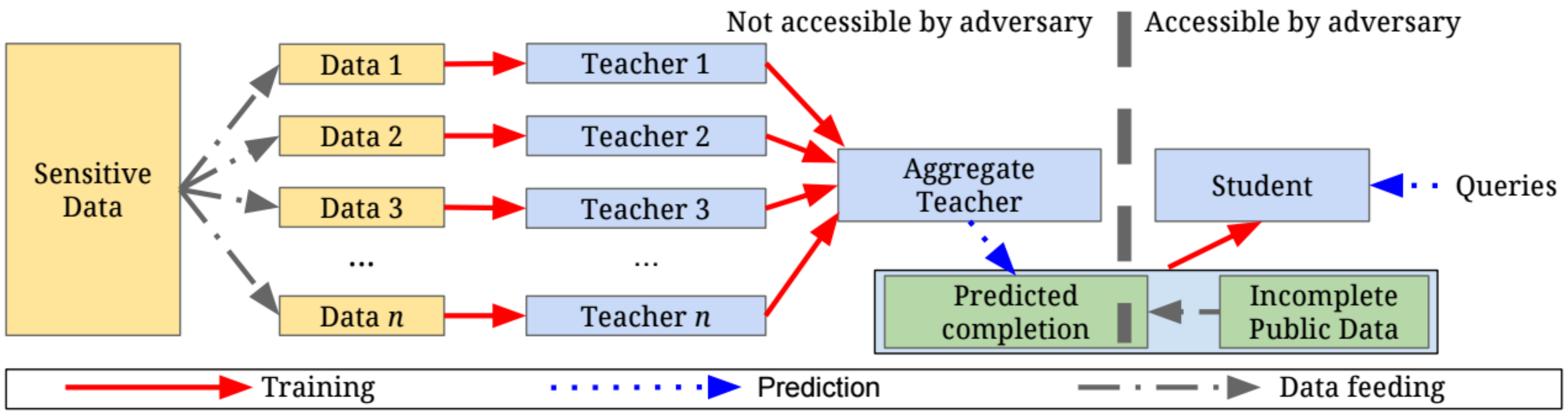
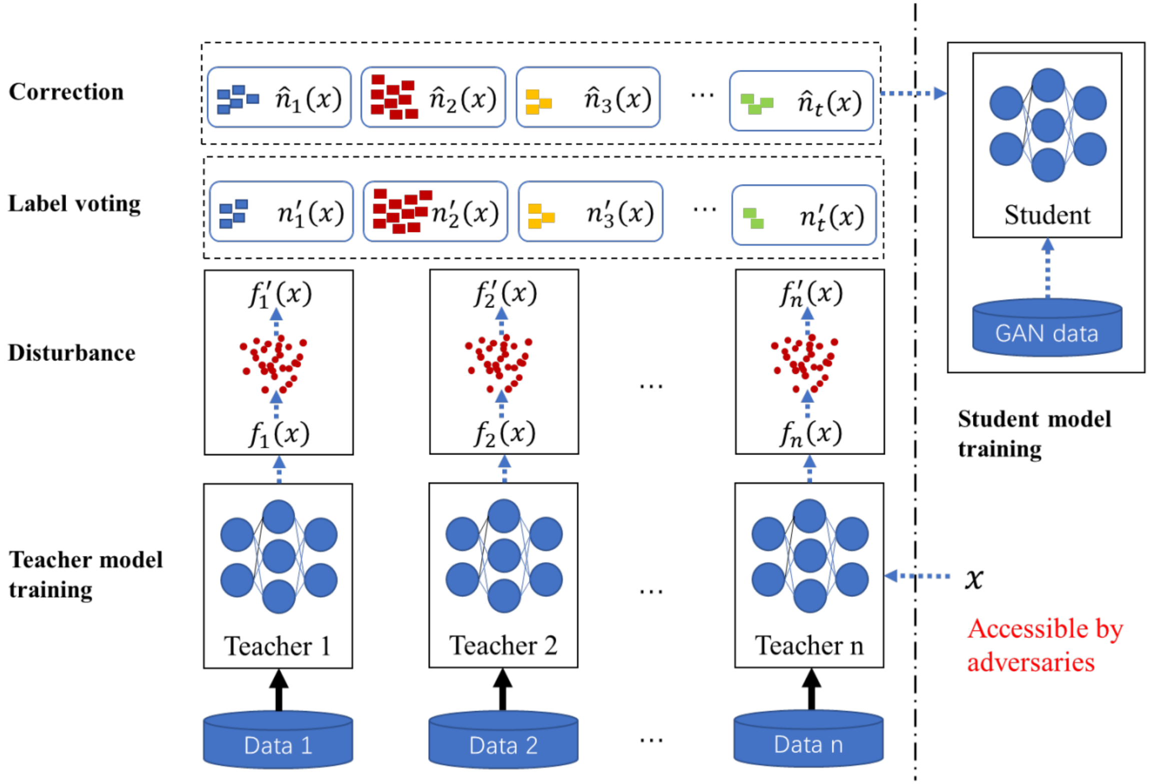


Figure : Overview of the approach: (1) an ensemble of teachers is trained on disjoint subsets of the sensitive data, (2) a student model is trained on public data labeled using the ensemble. [1]

- Challeges: The main problem with the PATE framework is: (1) Trusted third party requirement. (2)Decreased accuracy with more teacher models.
- Potential solutions: the RRML algorithm: (1) Localized noise addition using randomized response differential privacy. (2)Improved performance under low budget constraints.

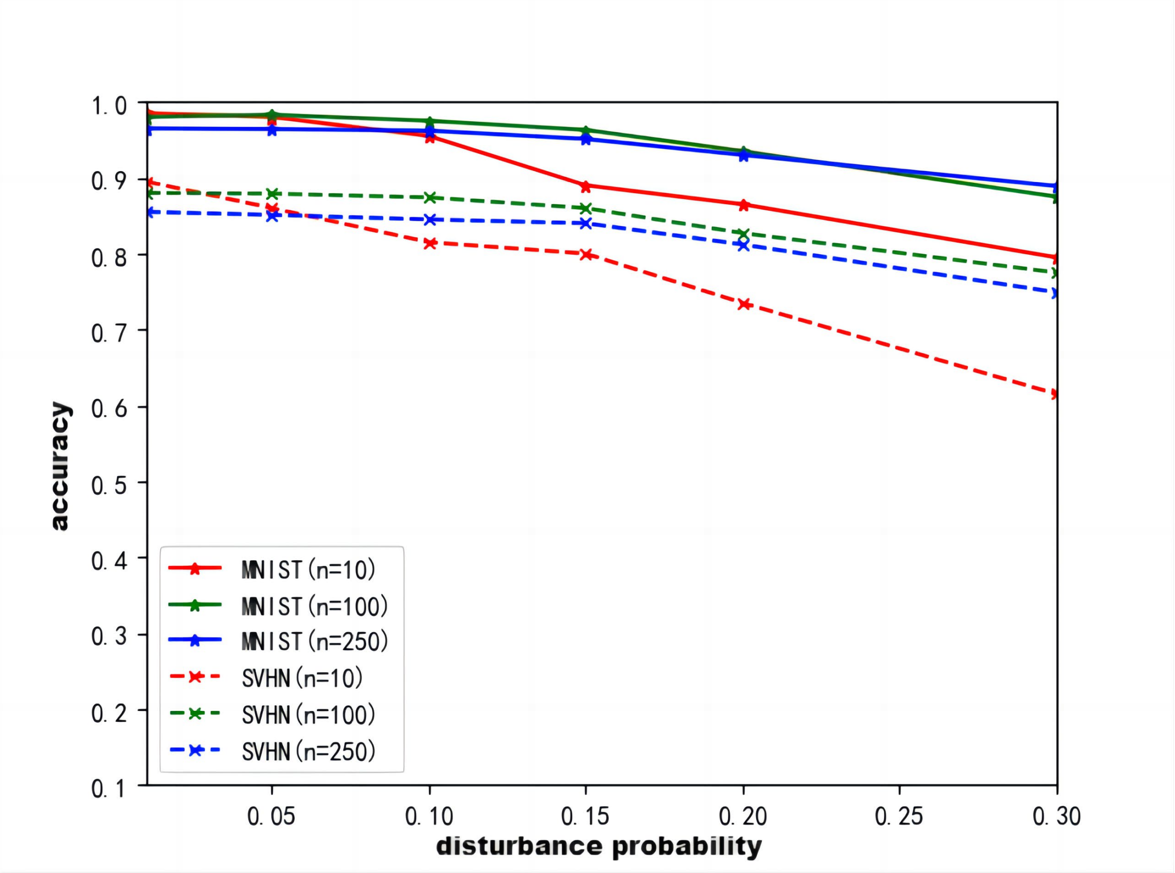
[1] N. Papernot, I. Goodfellow, M. Abadi, K. Talwar, and Ú. Erlingsson: Semi-supervised knowledge transfer for deep learning from private training data. ICLR, 2017.

C. Steps of the RRML algorithm



- Partition the data and train multiple teacher models.
- Apply local random response perturbation to the teacher models' outputs.
- Calculate the unbiased estimator of the perturbed labels.
- Determine the estimated labels corresponding to queries.
- Train a student model to achieve privacy-preserving machine learning.

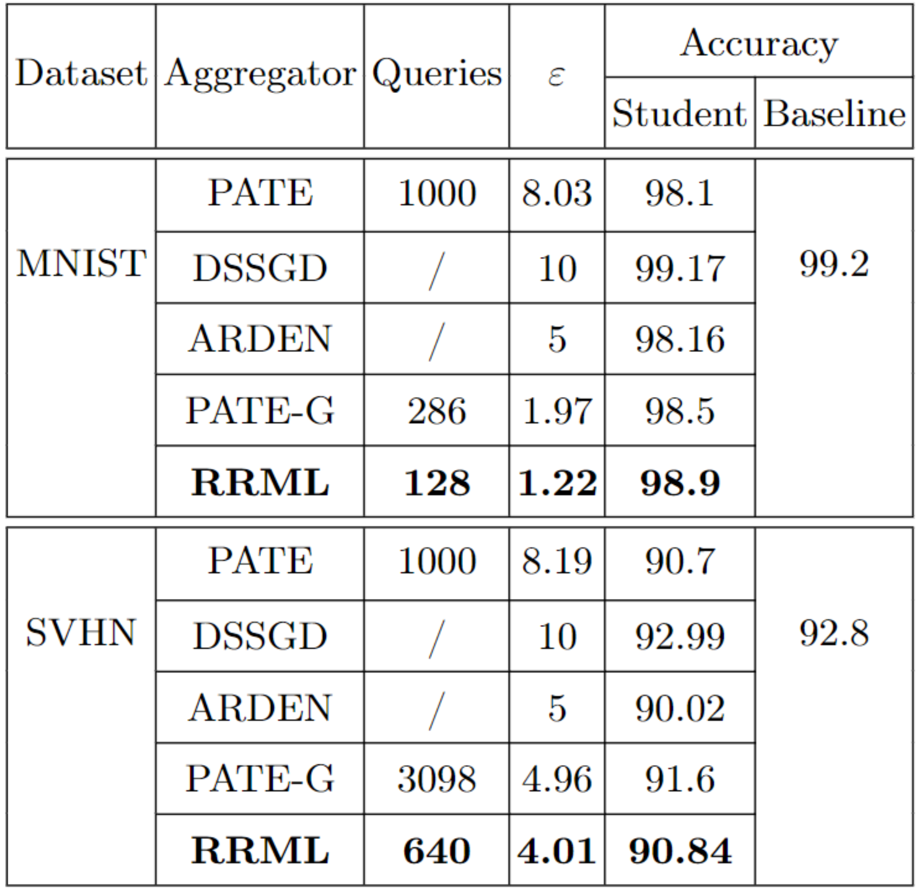
C. RRML: "Utility of the RRML algorithm on MNIST and SVHN datasets"



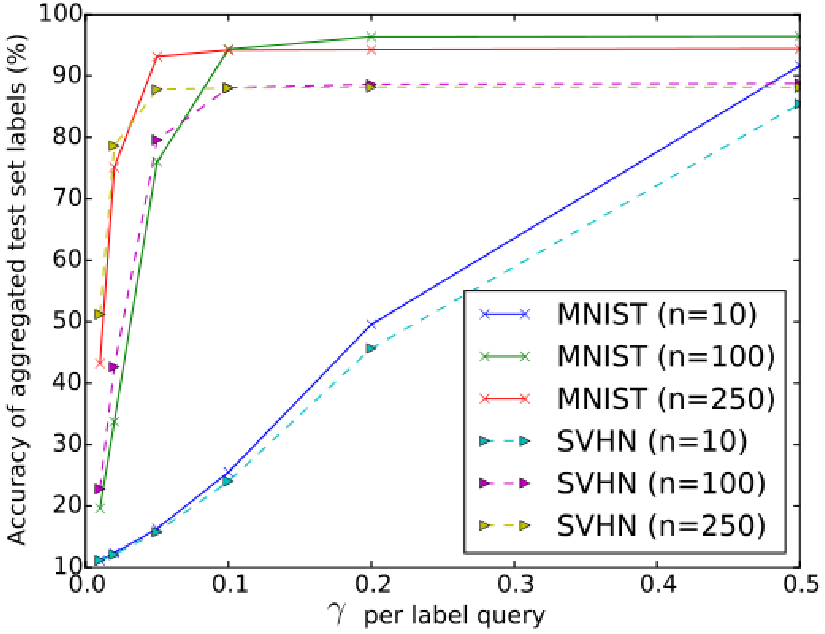
Utility of the RRML algorithm on MNIST and SVHN datasets:

- The RRML framework achieves privacy protection while minimizing accuracy loss.
- Increasing the perturbation probability reduces the privacy budget and enhances privacy protection.

C. RRML: “Performance comparison between RRML and other privacy frameworks under equivalent conditions”



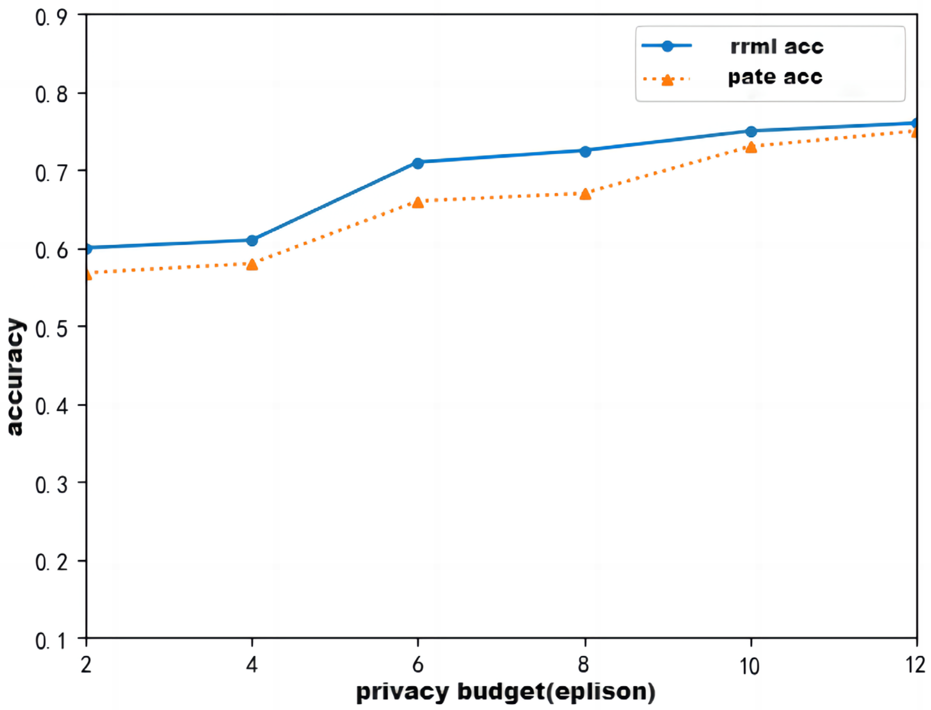
Dataset	Aggregator	Queries	ε	Accuracy	
				Student	Baseline
MNIST	PATE	1000	8.03	98.1	99.2
	DSSGD	/	10	99.17	
	ARDEN	/	5	98.16	
	PATE-G	286	1.97	98.5	
	RRML	**128**	**1.22**	**98.9**	
SVHN	PATE	1000	8.19	90.7	92.8
	DSSGD	/	10	92.99	
	ARDEN	/	5	90.02	
	PATE-G	3098	4.96	91.6	
	RRML	**640**	**4.01**	**90.84**	



Performance comparison between RRML and other privacy frameworks under equivalent conditions:

- The proposed RRML framework shows significantly better privacy-utility trade-off compared to other frameworks.

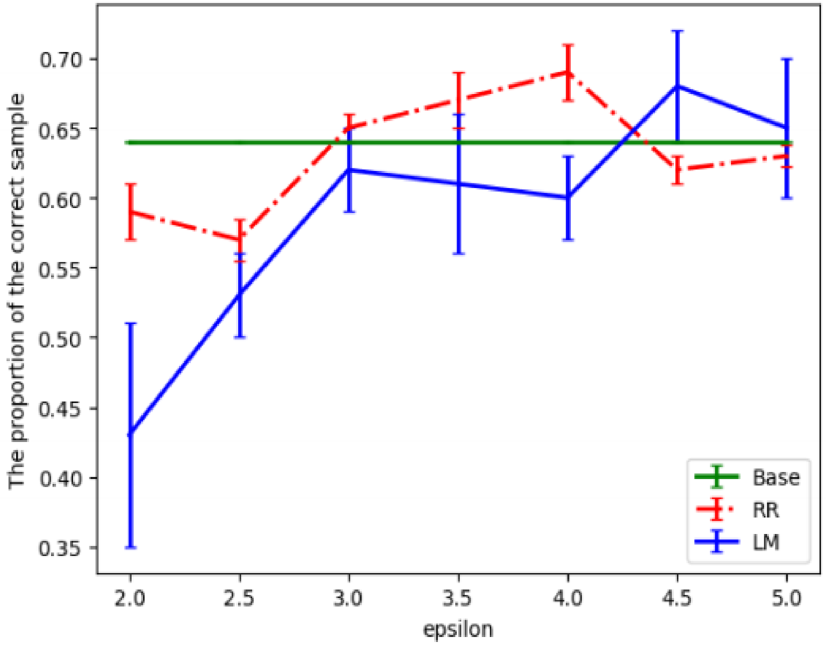
C. RRML: "Comparison of model accuracy loss based on RRML and PATE on the COVID-CT dataset"

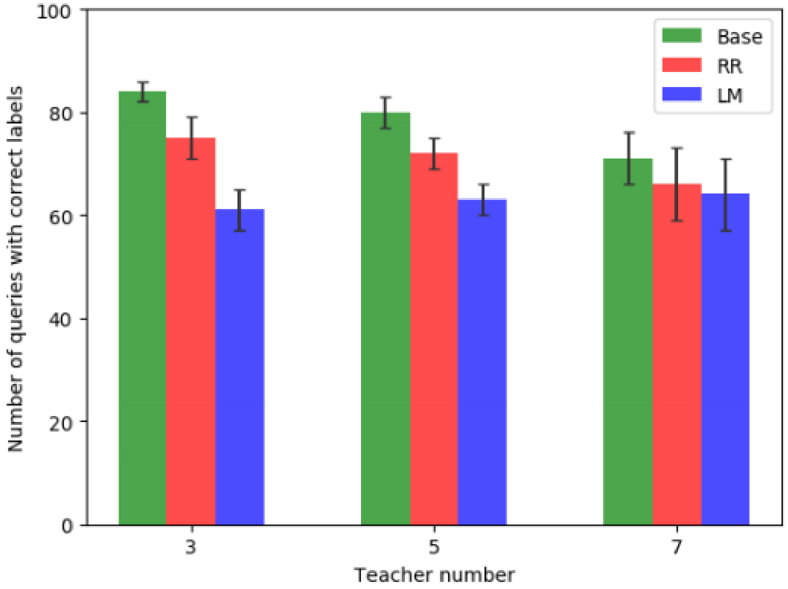


Comparison of model accuracy loss based on RRML and PATE on the COVID-CT dataset:

- The RRML framework exhibits smaller accuracy loss than PATE, particularly when the epsilon value is small.
- RRML is more suitable for cases with stricter privacy requirements.

C. RRML: "Statistical utility of adding Laplace noise and random response perturbation to the COVID-CT dataset"

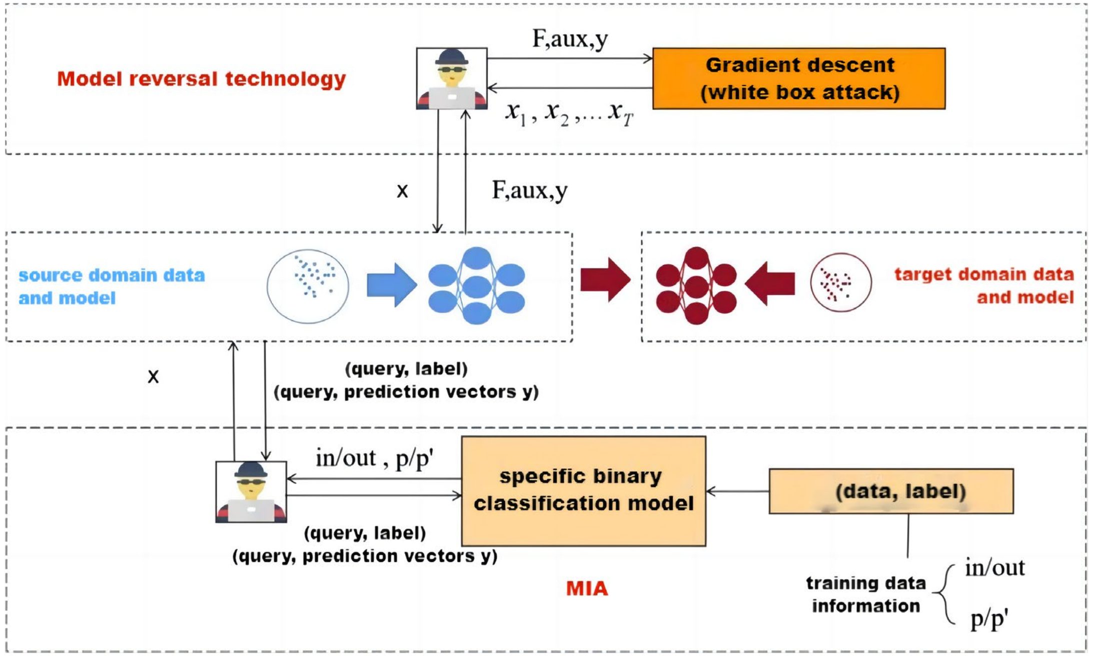




Statistical utility of adding Laplace noise and random response perturbation to the COVID-CT dataset:

- As the number of teacher models increases, correct prediction count gradually decreases due to limited training data.
- Random response perturbation performs better than noise methods, especially with a small number of teacher models.
- Noise methods require a sufficient number of teacher models for strict privacy protection, and inadequate training data may significantly reduce overall utility.

C. RRMTL: Application of RRML in Transfer Learning


Model reversal technology
F,aux,y
Gradient descent (white box attack)
$x_1, x_2, \dots x_T$
x
F,aux,y
source domain data and model
target domain data and model
x
(query, label)
(query, prediction vectors y)
in/out , p/p'
specific binary classification model
(data, label)
(query, label)
(query, prediction vectors y)
MIA
training data information
in/out
p/p'

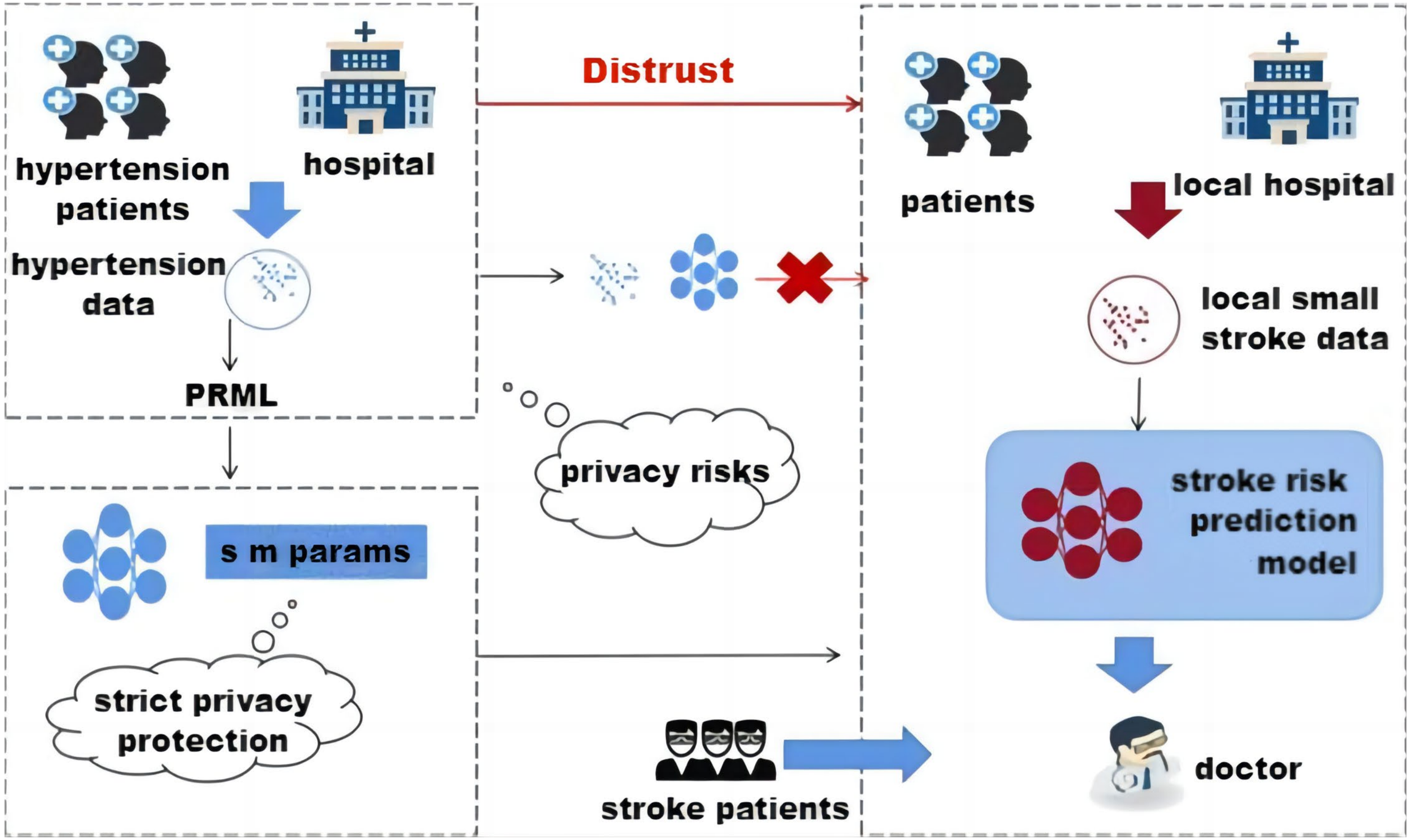
Traditional transfer learning vulnerabilities:

- Challenge: (1) Model inversion attacks and membership inference attacks exploit the sharing of source domain model parameters to reverse-recover training data information.

Introduction of differential privacy protection in RRMTL:

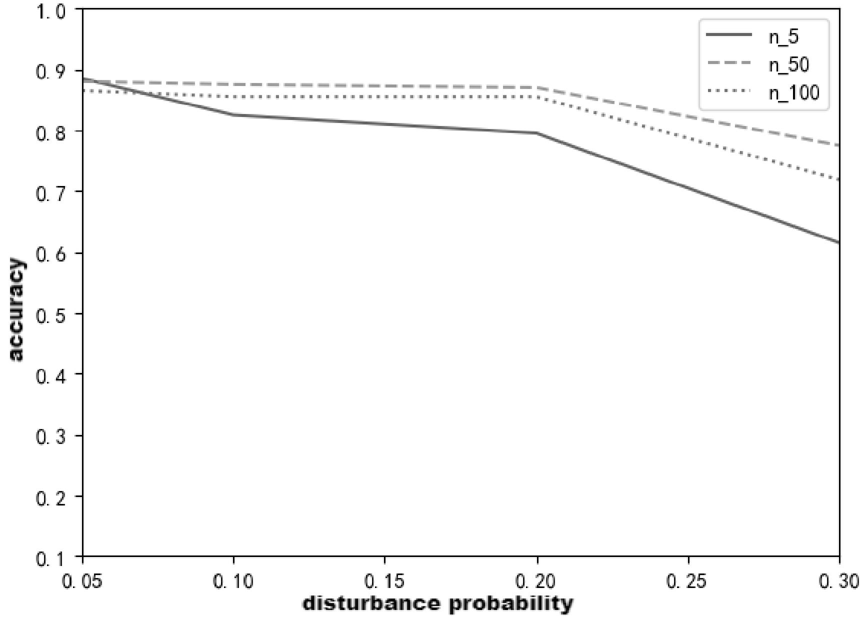
- Solution: RRMTL incorporates the differential privacy protection mechanism from RRML to effectively prevent privacy leakage.
- Improving transfer learning effectiveness with FOA: FOA (feature optimization algorithm) is proposed to enhance transfer learning by selecting more similar features based on the similarity between the source and target domain features.

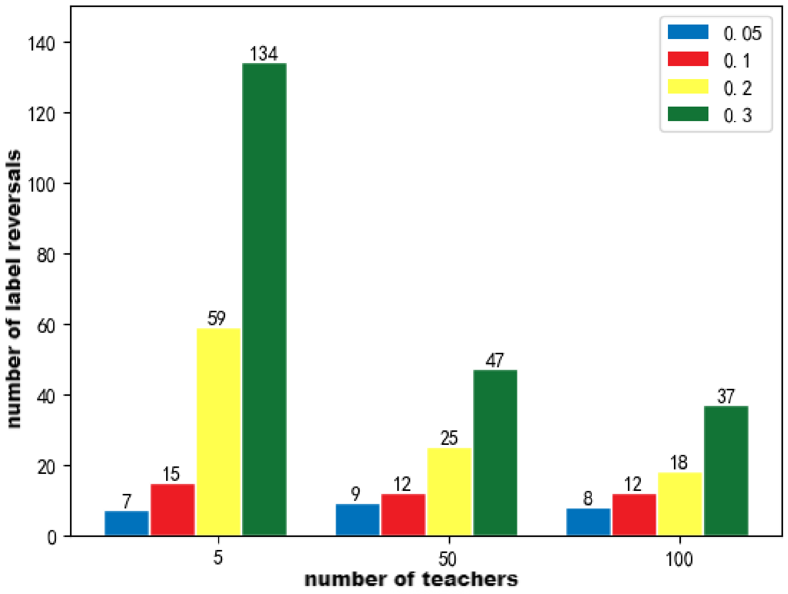
C. The basic architecture of RRMTL



- Train a privacy-protected student model using RRML in the source domain, then transfer the student model as a pre-trained model to the target domain.
- This avoids direct sharing of source domain data and provides privacy protection.

C. RRMTL: “Testing accuracy of RRML algorithm on hypertension data”

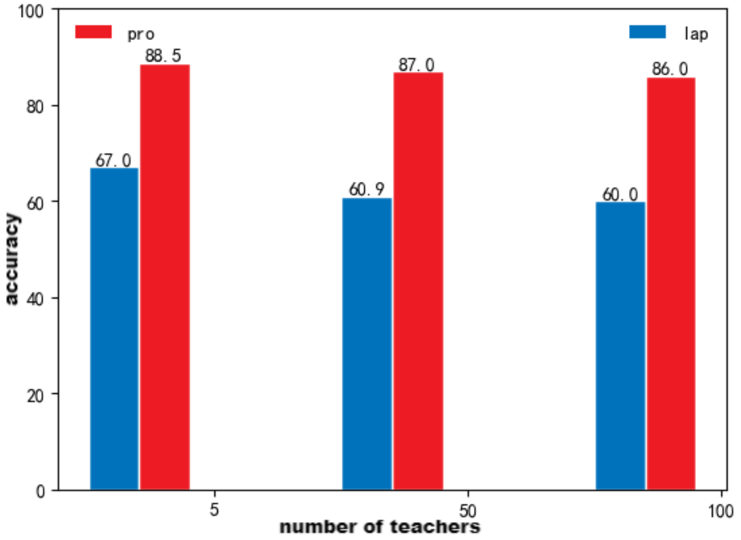


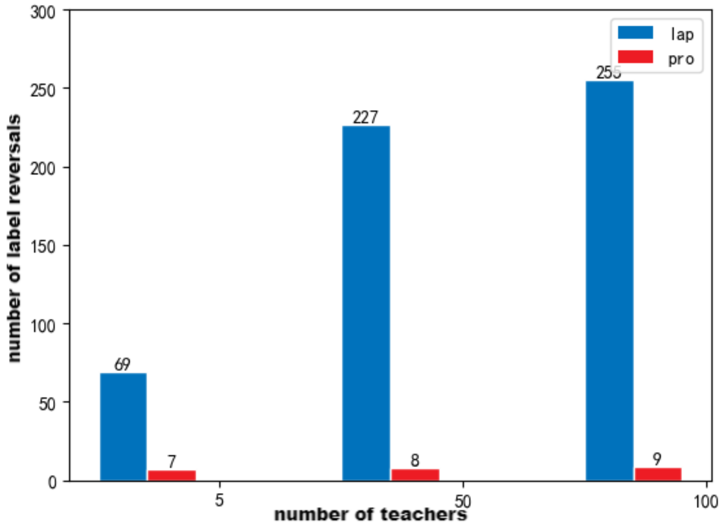


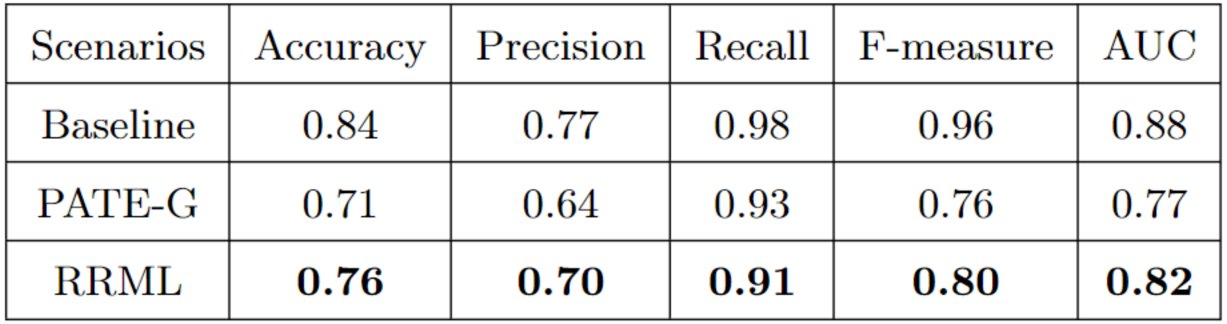
Testing accuracy of RRML algorithm on hypertension data:

- The accuracy of the RRML algorithm on hypertension data improves when the number of teacher models is appropriately set.

C. RRMTL: "Performance comparison between RRML and PATE-G"







Scenarios	Accuracy	Precision	Recall	F-measure	AUC
Baseline	0.84	0.77	0.98	0.96	0.88
PATE-G	0.71	0.64	0.93	0.76	0.77
RRML	**0.76**	**0.70**	**0.91**	**0.80**	**0.82**

Performance comparison between RRML and PATE-G:

- RRML outperforms PATE-G in terms of high-level data utility, especially with a small privacy budget.
- PATE-G uses Laplace noise for perturbing voting results, while RRML employs the random response technique as the perturbation mechanism.

C. RRMTL: “Effectiveness of FOA in transfer learning”

Performance Metrics	Accuracy	Precision	Recall	F-Score	AUC
Non-transfer	0.732	0.695	0.804	0.762	0.759
Direct transfer	0.723	0.709	0.758	0.732	0.774
Non-transfer (FOA)	0.728	0.69	0.83	0.753	0.738
Transfer after FOA	**0.805**	**0.75**	**0.842**	**0.797**	**0.812**

Effectiveness of FOA in transfer learning:

- Applying FOA to transfer learning leads to significant performance improvements in accuracy, precision, recall, and other metrics compared to direct transfer.

C. RRMTL: “Performance of stroke risk prediction model based on RRMTL”

Scenario	Accuracy	Precision	Recall	F-Score	AUC
Non-transfer	0.728	0.69	0.83	0.753	0.738
RRMTL（n=5)	0.749	0.722	0.81	0.763	0.797
RRMTL（n=50)	**0.788**	**0.745**	**0.876**	**0.805**	**0.818**
RRMTL（n=100)	0.77	0.74	0.822	0.781	0.794

Performance of stroke risk prediction model based on RRMTL:

- The RRMTL-based stroke risk prediction model achieves a better balance between data privacy and utility.

D. Conclusion

RRML and RRMTL mechanisms for privacy protection:

- Proposed to address privacy leakage in machine learning and transfer learning.
- RRML uses random response mechanism for better classification performance with privacy preservation.
- RRMTL applies RRML to transfer learning, ensuring local differential privacy in data fusion and knowledge transfer.
- Successfully validated the effectiveness of RRMTL in stroke risk prediction tasks.

THANK YOU