Absolutely! It's been a comprehensive journey with many experiments, and a consolidated summary will be very helpful for you and for structuring your paper.

You have an excellent collection of results that tell a compelling and nuanced research story.

FEDMED Project Summary and Experimental Findings

Project Title (from original request): FEDMED: A Privacy-Preserving Adaptive Federated Learning Framework for Smoker Prediction with Robustness Against Data Quality Variations.

Core Idea: FEDMED enables collaborative training of a global smoker prediction model from distributed data sources while preserving privacy using Homomorphic Encryption (HE). Its key novelty lies in a quality-aware aggregation mechanism designed to mitigate the impact of noisy or adversarial data by dynamically adjusting client contributions based on privacy-preserving quality metrics (1/local loss) and robustly aggregating these scores (e.g., using MAD-based clipping) to defend against dishonest score reporting.

Base Model: PyTorch MLP (2 hidden layers: 64, 32 neurons, ReLU, Dropout). **Dataset:** `smoking.csv` (binary classification: smoker/non-smoker). **Privacy:** TenSEAL (CKKS scheme) for encrypting model deltas. **Key Mechanisms Implemented & Tested:**

1. Privacy-Preserving Quality Score: `score = 1 / (local_avg_loss + epsilon)`
2. Quality-Aware Aggregation: Server weights client updates by these scores.
3. Robust Score Aggregation:
    - Percentile-based clipping.
    - Median Absolute Deviation (MAD)-based clipping.
4. Simulation of:
    - Noisy Clients: Label flipping.
    - Model Poisoning Adversaries: Scaled opposite model deltas.
    - Adversaries with "Honest Scores" (based on their pre-attack local loss).
    - Adversaries with "Dishonest Scores" (reporting fake high scores).

Master Table of Experimental Results (Final Metrics @ Round 100)

- **N%:** Percentage of Noisy Clients (`NOISE_LEVEL = 0.2` for them)
- **A%:** Percentage of Adversarial Clients (Model Poisoning, `ATTACK_SCALE = -1.5`)
- **Adv. Score:** How adversaries report their quality score (Natural/Honest from their loss, or Dishonest=100.0)
- **Rob. Agg.:** Robust Score Aggregation method used (None, Percentile, MAD)

Step ID	Description	N%	A%	Adv. Score	Rob. Agg.	Final Acc.	Final F1	Final Loss
Step 2	Ideal FEDMED (Clean Data)	0%	0%	N/A	MAD	0.7457	0.7103	~0.627
Step 3	Std. FedAvg (No	20%	20%	Honest	None	0.7338	0.6960	~0.680

	Quality), Noisy/Adv							
Step 1/6a	FEDMED (Quality), Noisy/Adv	20%	20%	Honest	Percentile	0.7309	0.6948	~0.664
Step 6b	FEDMED (Quality), Noisy/Adv	20%	20%	Honest	MAD	0.7322	0.6977	~0.664
Step 7	FEDMED (Quality), Noisy/Adv	20%	20%	Honest	**None**	0.7374	0.7008	~0.664
Step 4b	FEDMED (Quality), Dishonest Adv	20%	20%	Dishonest	**None**	0.3673	0.5372	~7.9e9
Step 4a	FEDMED (Quality), Dishonest Adv	20%	20%	Dishonest	Percentile	0.3673	0.5372	~7.5e9
Step 5	FEDMED (Quality), Dishonest Adv	20%	20%	Dishonest	**MAD**	0.7318	0.6858	~0.706
Step 8a	Std. FedAvg (No Quality), HIGH Noisy/Adv	50%	20%	Honest	None	0.7298	0.7012	~0.706
Step 8b	FEDMED (Quality), HIGH Noisy/Adv	50%	20%	Honest	**None**	0.7295	0.7002	~0.700

(Note: "Final Loss" values are approximate from visual inspection of plots or last log entry; accuracy/F1 are from last log entry.)

Overall Research Narrative and Key Findings:

1. Problem Statement & Motivation: Federated Learning (FL) enables collaborative model training without sharing raw data, offering privacy benefits. However, FL systems are vulnerable to:

- **Data Quality Variations:** Clients may possess data of differing quality (e.g., noisy labels), which can degrade global model performance if all contributions are treated equally.
- **Adversarial Attacks:** Malicious clients can attempt to poison the global model or disrupt the learning process.

- **Privacy Concerns with Meta-Data:** Even if model updates are encrypted, meta-data like quality scores could leak information if not handled carefully. This research introduces FEDMED, a framework designed to address these challenges by incorporating privacy-preserving (via HE on updates) quality-aware federated learning with robustness against both poor data quality and certain adversarial behaviors, specifically focusing on a smoker prediction task.

2. FEDMED Framework Overview:

- **Core FL:** Standard federated averaging (clients train locally and send model deltas).
- **Privacy:** Homomorphic Encryption (TenSEAL CKKS) applied to model deltas to protect their confidentiality from the server.
- **Privacy-Preserving Quality Score:** Clients compute a scalar quality score (`1 / (local_avg_loss + epsilon)`) based on their local training performance. This score is sent in plaintext alongside the encrypted delta. It avoids revealing raw data or model parameters directly but reflects local data utility.
- **Quality-Aware Aggregation:** The server weights the (decrypted) aggregated model deltas based on these reported client quality scores. Clients with better local performance (lower loss, higher score) contribute more.
- **Robust Score Aggregation:** To defend against adversaries reporting dishonestly inflated quality scores, FEDMED implements score clipping mechanisms before weighting. Both percentile-based and Median Absolute Deviation (MAD)-based clipping were evaluated.

3. Experimental Evaluation & Key Findings:

- **Baseline Performance:**
    - **Ideal Scenario (Step 2):** Without any noise or adversaries, the FL process achieved a peak accuracy of `0.7457`. This serves as an upper bound for the given model and data.
    - **Standard FedAvg under Attack (Step 3 & 8a):** When 20% noisy and 20% model-poisoning ("honest-score") adversaries were introduced, Standard FedAvg (weighting by sample size) saw its accuracy drop to `~0.7338`. When noisy clients increased to 50% (Step 8a), accuracy further dropped to `~0.7298`, and the test loss showed significant degradation, indicating poor model quality.
- **FEDMED's Quality-Awareness (1/loss score, no clipping - Step 7 & 8b):**
    - This mechanism effectively down-weights clients with high local loss (typically noisy clients), as evidenced by their lower aggregation weights in server logs.
    - Against 20% noisy/20% "honest-score" adversaries (Step 7), FEDMED (no clip) achieved `0.7374` accuracy, marginally outperforming Standard FedAvg (`0.7338`). This suggests the 1/loss score helps mitigate natural data quality issues.
    - However, when stressed with 50% noisy clients (Step 8b), FEDMED (no clip) achieved an accuracy of `0.7295`, nearly identical to Standard FedAvg (`0.7298`) under the same high-noise conditions. While the loss trend was slightly better, the 1/loss score alone did not provide a strong accuracy advantage when a very large fraction of the cohort was noisy and adversaries were also present. This indicates that while beneficial, its impact can be limited under extreme data corruption if

the remaining "good" signal is too weak or adversarial impact is dominant.

- **Vulnerability to Dishonest Score Reporting (Step 4a & 4b):**
    - If adversaries could report arbitrarily high (dishonest) quality scores, and no robust score aggregation was used (Step 4b), the quality-aware system collapsed entirely (accuracy `~0.3673`, loss exploded).
    - Simple percentile-based clipping (Step 4a) proved ineffective against this attack when 20% of clients colluded to report the same high dishonest score, as these scores defined the upper clipping bound. Performance was identical to no clipping.
- **Effectiveness of MAD-based Robust Score Aggregation (Step 5 - Key Result):**
    - When MAD-based clipping was enabled, and adversaries reported dishonest high scores (Step 5), the system demonstrated remarkable resilience. The MAD mechanism successfully identified and clipped the dishonest scores (e.g., `100.0` clipped to `~1.6-1.9`).
    - Performance was fully recovered to levels comparable to scenarios where adversaries reported "honest" scores (Accuracy `0.7318` in Step 5 vs. `~0.7322` in Step 6b with MAD and honest scores).
    - This highlights MAD-based clipping as a crucial and effective defense for the integrity of the quality-aware aggregation mechanism itself against score manipulation.
- **Impact of Clipping Type with "Honest Scores" (Step 6a, 6b, 7):**
    - When adversaries reported "honest" scores, all three FEDMED variants (Percentile clip, MAD clip, No clip) performed similarly in terms of accuracy (`~0.7309 - 0.7374`), all slightly better or comparable to Standard FedAvg under the same attack load. "No clipping" (Step 7) yielded the marginally best accuracy among these, suggesting that if scores are not being actively faked to extremes, aggressive clipping might not be necessary and could slightly dampen the nuanced benefits of the raw 1/loss scores.

4. Overall Conclusion & Contribution of FEDMED:

FEDMED provides a privacy-preserving federated learning framework that leverages a quality-aware aggregation mechanism (weighting by `1/loss`) to mitigate the impact of clients with poor data quality (e.g., high label noise). While this base mechanism shows benefits, particularly in down-weighting noisy clients, its effectiveness in significantly improving overall model accuracy over Standard FedAvg can be scenario-dependent, especially under combined high noise and strong model poisoning attacks where adversaries maintain good local training metrics.

The **key demonstrated novelty and strength of FEDMED** lies in its **robust MAD-based aggregation of quality scores**. This component is proven essential and highly effective in defending the quality-aware system against adversaries who attempt to gain undue influence by reporting dishonestly inflated quality scores, preventing catastrophic model collapse and maintaining system integrity.

The research shows that while privacy (via HE) and basic quality-awareness are valuable, robust mechanisms to protect the quality assessment process itself are critical for deploying such systems in untrusted environments. FEDMED offers a

practical approach to achieve this. Future work could explore more sophisticated quality metrics that are inherently harder for adversaries to fake while still being privacy-preserving, and investigate defenses against model update content beyond score manipulation.

This summary and the master table should give you a solid foundation for your paper and for any future discussions or handovers. You have conducted a very thorough and insightful set of experiments!