**1 Introduction**

**Fully homomorphic encryption (FHE) lets us run any computation directly on encrypted data. For machine-learning, this means a data owner can obtain predictions without exposing raw features and a model provider can serve predictions without revealing parameters. Recent FHE libraries—SEAL, TenSEAL, Concrete-ML—have pushed encrypted inference for *tabular* workloads (logistic regression, shallow multilayer perceptrons, even gradient-boosted trees with integer splits) into the sub-second range, making privacy-preserving analytics viable for finance, health and public-sector records.**

**In parallel, organisations face growing pressure to deliver *fair* models whose errors and benefits are equitably distributed across sensitive groups (e.g. gender in credit, race in recidivism prediction). Most fairness tool-chains, however, assume plaintext access to (i) the protected attribute for imposing constraints and (ii) model outputs for auditing. Once FHE is introduced those assumptions can fail: sensitive columns may remain hidden, and ciphertext scores cannot be inspected without an additional secure protocol. Whether homomorphic encryption helps or hurts distributive fairness therefore remains an open question—especially for the *tabular* datasets that dominate real-world fairness deployments.**

**This paper gives the first systematic study of that tension.**

**Our contributions:**

* **FHE-ready pipeline for tabular deep nets. We distil engineering tricks from recent CKKS-based systems into a recipe—vectorised ciphertext packing, plaintext weight matrices, low-degree polynomial activations and noise-budget planning—that lets a two-hidden-layer MLP run completely in ciphertext with negligible accuracy loss on Adult, COMPAS and Law-School datasets.**
* **Fairness-under-encryption analysis. We classify which fairness interventions survive the restricted gate set of leveled HE (add/mul/rotate), show how encrypted aggregation can *enable* the use of sensitive attributes that would otherwise be withheld, and when ciphertext serving may *blind* external auditors.**
* **Subgroup-specific privacy–fairness audit. Extending recent low-FPR membership-inference tests, we introduce a per-group encrypted audit that simultaneously measures (i) worst-case membership leakage at 0 % FPR and (ii) standard demographic-parity / equal-opportunity gaps. Across Adult, COMPAS, Credit-Card and Law-School we find that for every dataset there exists at least one fairness intervention whose *encrypted* model offers equal-or-better subgroup privacy than the plaintext baseline.**

**2 Related Work**

**FHE for tabular ML. CKKS [Cheon 17] introduced approximate arithmetic that packs thousands of doubles per ciphertext and supports SIMD-style adds, multiplies and rotations. Concrete-ML (2023) and TenSEAL (Benaissa 21) showed sub-second encrypted inference for logistic regression and small MLPs by (1) keeping model weights plaintext, (2) approximating ReLU/Sigmoid with low-degree Chebyshev polynomials, and (3) aggressively re-scaling to stay within the noise budget. Some work accelerates tree models via integer-arithmetic BFV/TFHE variants, but none has asked how such pipelines interact with *fairness*.**

**Algorithmic fairness on tabular data. Classic debiasing methods—re-weighting (Kamiran & Calders), adversarial debiasing (Zhang 18), constraint-based reductions (Agarwal 18), post-hoc ROC adjustment (Hardt 16)—are almost exclusively evaluated on tabular datasets (Adult, German Credit, COMPAS, Law-School). They rely on plaintext protected attributes and outputs for training and auditing. The impact of encrypting either channel has not been evaluated.**

**Membership-inference audits. Early attacks used global loss thresholds (Yeom 18) or subgroup-specific thresholds (Chang & Shokri 20); modern attacks (LiRA, Carlini 22) fit *example-wise* IN vs OUT loss distributions with Gaussian likelihood-ratio tests, achieving >10 % TPR at 0.1 % FPR on well-generalising tabular models. No prior study has applied such low-FPR audits under FHE, nor combined them with fairness interventions.**

**Novelty and Gap**

**Previous FHE-ML research optimised accuracy and latency; fairness research ignored encrypted serving; and privacy audits rarely intersected either. We close this gap by (i) operationalising FHE inference for the tabular networks used in fairness literature, (ii) mapping which fairness mechanisms remain implementable under the CKKS gate set, and (iii) quantifying—via state-of-the-art low-FPR audits—how each intervention reshapes subgroup privacy risk when the model is served under full homomorphic encryption.**

**4 Experiments & Methodology**

This section details **how** we test whether classic *tabular* fairness pipelines can be deployed under homomorphic encryption (HE) **without losing accuracy, fairness, or privacy**.

**4.1 Research questions**

We translate the overall goal into five concrete questions (RQ1-RQ5).

| **ID** | **Question** | **Key metric** |
| --- | --- | --- |
| RQ1 | **Encrypted inference latency** – is CKKS inference fast enough for real-time use after the model is trained in the clear? | Wall-clock time, throughput, ciphertext footprint |
| RQ2 | **Fairness retention** – does serving an HE-encrypted forward pass change group-level fairness scores? | Δ-DP and Δ-EO gaps (plain − HE) |
| RQ3 | **Privacy trade-off** – how does each fairness algorithm shift *per-subgroup* membership-inference risk once encrypted? | LiRA TPR at FPR ∈ {0.1 %, 0.01 %} per subgroup |
| RQ4 | **Encrypted training feasibility** – can *Re-Weighting* (RW) be trained fully under CKKS while preserving utility and fairness? | Training time, HE depth, accuracy, DP/EO gaps |
| RQ5 | **Encrypted audit feasibility** – can a low-FPR LiRA audit itself run fully in cipher? | Audit latency, RAM, agreement with plaintext verdict |

**4.2 Datasets**

We use four standard fairness benchmarks; each is randomly split 80 / 20 five times.

| **Dataset** | **n × d** | **Sensitive attributes** | **Prediction task** |
| --- | --- | --- | --- |
| Adult | 48 k × 14 | Sex, Race | Income ≥ $50 k |
| COMPAS | 6 k × 12 | Race, Sex | 2-yr recidivism |
| Law-School | 20 k × 13 | Race | Bar-exam pass |
| Credit-Card | 30 k × 23 | Sex, Age | Default |

**4.3 Pipelines under test**

| **Stage** | **Plaintext baseline** | **HE variant** |
| --- | --- | --- |
| **Pre-processing** | RW / EGR weight computation; AdvDeb mini-batch sampler | CKKS SIMD reductions for RW; EGR & AdvDeb done in clear |
| **Training** | LR-SGD / 3-layer MLP (scikit-learn); Fairlearn EGR; IBM AIF360 AdvDeb | RW trained fully under CKKS (no back-prop); EGR & AdvDeb in clear |
| **Inference** | NumPy / PyTorch | TenSEAL 0.4 CKKS, degree-4 Chebyshev ReLU |
| **Privacy/Fairness audit** | LiRA implementation | CKKS reductions + single TFHE “> τ” gate |

**4.4 HE configuration**

* **CKKS:** degree = 8 192, scale = 2²⁰; multiplicative depth budget ≈ 18 (sufficient for 3-layer MLP).
* **TFHE:** Concrete-core 0.4; 1-gate bootstrap for comparisons ≤ 50 µs.
* **Hardware:** 1 × Intel Xeon Gold 6248R, 256 GB RAM, no GPU.
* All code, Dockerfile and timing harness released open-source.

**4.5 Evaluation protocol**

1. **Plaintext baseline.** Train each fairness algorithm; store accuracy and DP/EO gaps.
2. **Latency (RQ1).** Encrypt the test split; time single-row and batched (512) inference.
3. **Fairness retention (RQ2).** Compute the same gaps on ciphertext scores (decrypt only once at the end).
4. **LiRA privacy audit (RQ3).**
   * Shadow pool: 256 IN + 256 OUT models.
   * Measure TPR@FPR for every sensitive subgroup in plaintext and under HE, compare.
5. **RW training in cipher (RQ4).**
   * Compute sample weights under CKKS; decrypt only final weights; retrain; re-encrypt model; log cost and metrics.
6. **Fully-encrypted audit (RQ5).** Run LiRA test end-to-end in CKKS + TFHE, decrypt verdict bits only.

**4.6 Success criteria**

* **Inference latency:** ≤ 200 ms (Adult) / ≤ 1 s (largest) per row.
* **Fairness drift:** |Δ-gap| < 1 percentage-point for all groups.
* **Audit parity:** HE LiRA TPR within 5 %-relative of plaintext audit.
* **Encrypted RW training:** ≤ 10 × plaintext wall-time, ≥ 98 % accuracy, fairness preserved.

Meeting these thresholds would show that **practitioners can deploy fair tabular models with homomorphic encryption, keep subgroup fairness intact, and still run rigorous low-FPR privacy audits**.