International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Balancing Privacy, Fairness, and Predictive Performance in AI-Based Credit Risk Prediction

M. Devi Harshini
BSCS , Vishwa Vishwani Institutes of Systems and Management
Hyderabad, Telangana, India
Email:deviharshini21@gmail.com

A. Sahasra Sri
AIML, BVRIT Hyderabad College of Engineering for Women
Hyderabad , Telangana, India
Email:alavalapatisahasrasrii@gmail.com

A. Vaishnavi
AIML , BVRIT Hyderabad College of Engineering for Women
Hyderabad , Telangana,India
Email:alavalapativaishnavii@gmail.com

ABSTRACT

AI-based credit scoring can help lenders assess loan applications, but it also raises questions about how much personal information is collected, whether applicants are treated fairly, and how reliable the predictions are. This paper sets out a plan to evaluate two privacy safeguards in credit risk prediction: data minimization and differential privacy, with the Indian lending context in mind. Data minimization limits a model to features that are relevant and justified for the task. Differential privacy is designed to limit what a trained model can reveal about any one record, often by clipping and adding noise to gradients during training. The proposed study compares four approaches: a baseline model, a model trained on a reduced set of features, a differentially private model, and a model that combines both safeguards. It assesses fairness by comparing applicants with limited credit histories with those who have more extensive histories. Classification measures will be used to assess predictive performance, while group-level error rates and calibration will help identify possible differences in outcomes. Membership-inference attacks will provide an additional check for privacy leakage. The paper also describes a reproducible evaluation process, including data checks, consistent train–test splits, repeated runs, and clear reporting of privacy settings. No dataset or experimental results were provided for this draft, so it makes no claims about measured improvements or which safeguard works best. Instead, it offers a method for examining the trade-offs and emphasizes the need for suitable data, interpretation of fairness measures, and human review in consequential lending decisions.

Keywords: Artificial Intelligence, Credit Risk Prediction, Data Minimization, Differential Privacy, Algorithmic Fairness, Membership-Inference Attacks.

1. INTRODUCTION

Artificial intelligence (AI) and machine learning are now used to help with decisions involving large amounts of information. Credit risk prediction is one example: a model estimates how likely an applicant is to default, giving a lender information when reviewing a loan application. A well-designed model can process data consistently and quickly. But credit decisions affect people's access to financial services, so an inaccurate or unfair prediction can affect their economic opportunities. Credit models may use information about income, employment, repayment behaviour, existing debt, and credit-bureau records. Detailed information may improve predictions, but it can also increase privacy risks. For example, a model may reveal clues about the records used to train it. Applicants also differ in how much credit history they have. Someone new to borrowing may have only a few records, while someone who has used credit for years may have many. If a model depends heavily on past credit activity, it may assess people with limited histories less reliably. This paper looks at two safeguards that address different parts of the problem. Data minimization limits the model to information that is relevant and justified. Differential privacy is intended to limit how much any one record can influence the trained model. The main question is whether either safeguard, or both together, can reduce privacy risks while preserving useful predictions and avoiding larger differences in outcomes for applicants with limited and extensive credit histories. The discussion focuses on lending in India, where responsible data use and fair access to credit are important concerns. India's Digital Personal Data Protection Act, 2023 offers relevant context for handling digital personal data. Determining legal compliance, however, requires a separate review of the applicable provisions and commencement notifications (Government of India, 2023).

2. REVIEW OF LITERATURE

2.1 Differential privacy in machine learning. Differential privacy gives a formal way to describe how much a computation's output may change if one person's record is added to or removed from the data. Dwork and Roth (2014) established the foundations of this framework. For machine learning, Abadi et al. (2016) developed a practical method for training deep neural networks with differential privacy. Their approach clips individual gradients and adds noise during training. These methods provide a starting point for evaluating privacy-preserving training. The privacy guarantee comes with trade-offs: the amount of noise, the number of training steps, and the size of the available dataset can all affect model usefulness. Researchers should therefore report the privacy budget, usually denoted by epsilon (ε), alongside model

performance. Under otherwise comparable conditions, a smaller ε generally means a stronger formal privacy guarantee.

2.2 Membership-inference attacks. A model can sometimes give away clues about whether a particular record was part of its training data. Shokri et al. (2017) demonstrated attacks that use a model's outputs to make this inference. This is one reason to test privacy protections rather than assume that removing direct identifiers makes a model safe. Membership-inference testing covers only one kind of privacy risk, so it cannot serve as a complete privacy audit. It can, however, offer a useful measure of one possible form of information leakage.

2.3 Fairness in automated decisions. There is no single measure that captures every aspect of fairness. Hardt et al. (2016), for example, proposed equality of opportunity, which compares true-positive rates across groups. Barocas et al. (2023) explain how algorithmic decisions can reproduce or introduce harm even when a system performs well technically. In credit prediction, a strong overall accuracy score may hide differences in error rates. One group, for instance, might face more false rejections or false approvals than another. A responsible evaluation should report overall results as well as outcomes for each group, rather than relying on a single score.

2.4 Data minimization and the research opportunity. Data minimization means limiting the personal information used for a task. This can reduce exposure and make data governance easier, although removing features may also discard information that helps with prediction. Data minimization and differential privacy are connected, but they do different things: minimization restricts what information is collected or kept for the task, while differential privacy limits the effect of individual records during computation. Their combined effect on credit prediction, especially when applicants have different amounts of credit history, needs to be measured rather than assumed.

Earlier research provides a foundation for privacy-preserving training, membership-inference testing, and group fairness. This paper brings those strands together in one evaluation plan and asks whether privacy safeguards affect applicants with limited credit histories differently from those with extensive histories.

3. RESEARCH GAP

Studies of privacy, fairness, and predictive performance often examine these goals separately or focus on one technique at a time. A model can perform well overall without showing whether its mistakes are distributed fairly across groups. Similarly, a privacy method may offer a formal

guarantee while reducing the model's practical usefulness. Data minimization can prevent unnecessary data use, but it may affect applicants differently depending on how much credit history they have. Evaluating these concerns together would make the trade-offs easier to identify and report clearly.

The research also needs to address context and generalizability. This study is motivated by credit assessment in India, but a public dataset from another country cannot, on its own, show how a model would perform for Indian applicants. Any empirical work should describe where the dataset came from, whom it represents, how its target outcome is defined, and what its limitations are. The proposed design treats India as the application context while requiring the completed study to use appropriate, lawfully accessible data. If it uses a dataset from another market, it should identify the work as a methodological pilot rather than evidence about Indian borrowers.

4. OBJECTIVES OF THE STUDY

This study aims to:

• Compare a baseline credit risk model with models that use data minimization, differential privacy, or both safeguards together.

• Assess predictive performance with suitable classification measures, such as ROC-AUC, precision, recall, and F1-score where appropriate.

• Compare error patterns for applicants with limited credit histories and applicants with more extensive histories.

• Test for privacy leakage using a controlled membership-inference attack and explain the test's limitations.

• Describe the practical trade-offs among privacy, fairness, and predictive performance, and identify responsible evaluation practices.

5. RESEARCH METHODOLOGY

5.1 Research design. The study uses a quantitative comparative experiment. It evaluates four model configurations under the same data split and comparable training conditions: (A) a baseline, (B) a model using a reduced feature set, (C) a differentially private model, and (D) a model using both data minimization and differential privacy. Comparing them this way helps distinguish the effects of each safeguard from the effects of using them together.

5.2 Data source and suitability. The dataset needs to include a credit-risk outcome and enough information to define, or reasonably estimate, the depth of each applicant's credit history. Access to the data must be lawful, and direct identifiers should not be used as predictive features. The Home Credit Default Risk dataset is one possible public benchmark for a methodological pilot; its documentation describes loan applications and related prior-credit and repayment tables (Home Credit Group, n.d.). It does not represent India specifically, however. Results from this dataset should not be presented as evidence about Indian borrowers. A study that makes claims about India should use suitable Indian data with appropriate permissions or clearly limit its conclusions to the benchmark it used.

5.3 Data preparation. Before modeling, the dataset should be checked for missing values, duplicates, class imbalance, leakage, and inconsistent records. Features that would not have been available when the loan decision was made should be removed to prevent information from the future leaking into the model. Preprocessing, feature selection, and model tuning should use training data only. The training, validation, and test sets should be separated before fitting the model. If the dataset has several records for the same applicant, all of that applicant's records should remain in the same split so that the person does not appear in both the training and test data.

5.4 Defining credit-history groups. Applicants should be placed into limited-history and extensive-history groups using a rule based on the number or duration of their previous credit records. The threshold should be justified in advance and reported clearly. If the dataset does not contain a reliable measure of credit-history depth, the study should not create one without evidence. It should instead choose another dataset or state that the planned group comparison cannot be completed. Group sizes and missing data should also be reported, since estimates based on very small groups may be unstable.

5.5 Model configurations. The study could begin with a simple, interpretable model such as logistic regression and add a tree-based model if the data justify it. Where possible, each configuration should use the same model type and comparable tuning procedures. For data minimization, features should be retained because they are relevant to the task, necessary, and supported by a documented selection process—not simply because they improve test-set performance. For differential privacy, the training method should support formal privacy accounting. The study should report the clipping norm, noise multiplier, number of training steps, sampling method, and resulting privacy budget. It should claim differential privacy only if the implementation and accounting method support that claim.

5.6 Evaluation metrics. Predictive performance should include ROC-AUC and, when useful for an imbalanced dataset, precision, recall, F1-score, and a confusion matrix. Calibration should also be checked if the predicted probabilities are meant to represent default risk. For fairness, the study should report false-positive and false-negative rates, true-positive rates, and relevant differences between the two credit-history groups. Equality of opportunity can be assessed by comparing true-positive rates, provided the study clearly defines which outcome counts as positive (Hardt et al., 2016). No single measure should be treated as proof that a model is completely fair.

5.7 Privacy evaluation. The study should run a membership-inference attack against each model using a clearly described attacker setup and held-out data. Depending on the implementation, attack performance can be summarized with ROC-AUC, precision, recall, or advantage. The results should be compared with a meaningful baseline and considered alongside the formal differential-privacy settings. A weak attack does not prove that no information can leak, just as a strong attack does not capture every possible privacy risk.

5.8 Reliability and reporting. The study should document its data splits, use repeated random seeds where feasible, and report confidence intervals or variation across runs. It should also provide group sample sizes, preprocessing choices, feature sets, model settings, privacy parameters, and evaluation measures. Failed runs and study limitations should be disclosed. The test set should be reserved for final evaluation rather than repeatedly used to select the model.

5.9 Statistical analysis plan

Before examining test results, the researchers should write down the comparisons they intend to make and the measures they will use. The primary comparison can be the change in ROC-AUC between the baseline and each safeguarded configuration. Secondary comparisons can cover precision, recall, calibration, group-specific error rates, and attack performance. Defining these comparisons in advance reduces the chance of highlighting only the measures that happen to look favorable.

Each configuration should be evaluated on the same held-out test cases. This paired design makes the comparison more informative because differences are less likely to reflect different test populations. Results across repeated runs should be summarized with a central estimate and a measure of variation. Where sample size supports it, confidence intervals can describe uncertainty around performance and group differences. A difference in a point estimate alone should not be treated as meaningful evidence if it is small relative to the uncertainty. The

analysis should also report denominators, since equal percentages can represent very different numbers of applicants in large and small groups.

Threshold-based measures depend on the decision threshold used to convert scores into accept or reject outcomes. The study should state how that threshold is selected and keep its selection separate from final test evaluation. It may report results at a threshold chosen on validation data and include a sensitivity analysis over a reasonable range of thresholds. This helps show whether a group difference is stable or appears only at one operating point. The paper should avoid presenting a single threshold as universally correct because lenders may face different decision costs and legal obligations.

5.10 Reproducibility and implementation controls

A reproducible study needs more than a list of model names. The authors should record the software libraries and versions, preprocessing steps, feature definitions, random seeds, and hardware or training settings that could affect the result. If the data provider changes a public benchmark over time, the dataset version or retrieval date should be documented. Any exclusions should be counted and explained. These details help another researcher determine whether a different result comes from a changed method, a changed dataset, or ordinary variation between runs.

The four configurations should share the same evaluation pipeline wherever possible. Feature selection for the minimized model should occur inside the training process rather than using information from the validation or test set. Hyperparameter tuning should use the same strategy and comparable effort across models. If the private training method limits the available optimization choices, that constraint should be stated rather than hidden. The study should also keep a record of failed experiments and deviations from the planned protocol, since selective reporting can make a method appear more reliable than it was.

5.11 Data governance and participant protection

Credit data are sensitive because they can reveal financial circumstances and may affect access to essential services. Before using a dataset, researchers should establish who is authorized to access it, what purpose the data may be used for, and how long copies will be retained. Access should be limited to people who need it for the study, and files should be protected during storage and transfer. The manuscript should describe these safeguards at a level that readers can assess without exposing private information or weakening security.

6. CONCEPTUAL BACKGROUND

Data minimization means using only the personal information needed for a defined purpose. In a credit model, that could involve removing redundant, irrelevant, or unjustified attributes while keeping the information needed for a sound assessment. It does not mean deleting as many columns as possible: removing a useful feature could weaken predictions or affect groups in different ways.

Differential privacy is a mathematical framework for protecting privacy. In model training, it is used to limit how strongly the final model can depend on any one training record. Common methods clip individual gradients and add carefully calibrated noise. The privacy budget ε and, where applicable, the parameter δ describe the guarantee under stated assumptions. The paper should report the parameters and accounting method; simply calling a model "privacy-preserving" does not explain what protection it provides.

Credit risk prediction estimates the likelihood of an outcome such as loan default. It can inform a lender's decision, but it should not automatically replace the broader review process. A limited credit history means that relatively few prior credit records are available; an extensive history means that more records are available. These categories describe the data on hand, not a person's reliability or ability to repay.

Algorithmic fairness considers whether a model's outcomes or errors create unjustified differences between groups. Some fairness criteria can conflict, particularly when groups have different base rates. The study should explain why each selected measure matters for lending and avoid treating any one measure as a complete ethical judgment.

6.1 Interpreting group comparisons

Group comparisons must account for missing or uncertain history information. Missing records should not automatically mean an applicant has a short history. The study should report how such cases are handled and acknowledge when the data cannot distinguish a genuinely short history from incomplete reporting.

Differences in error rates do not by themselves explain why they occurred or prove that a feature caused them. Interpretation should consider data quality, group size, model choices, and decision thresholds. If fairness measures point in different directions, report the conflict rather than assigning a single fairness ranking.

7. ANALYSIS AND DISCUSSION FRAMEWORK

The analysis should compare all four configurations across the same three areas. The baseline provides a reference for predictive performance, group-level errors, and vulnerability to membership-inference attacks. The data-minimized model shows what changes when the

feature set is restricted. The differentially private model tests privacy-protected training while retaining the baseline features. Finally, the combined model shows whether using both safeguards introduces additional benefits or costs.

The discussion should weigh trade-offs rather than select a model using accuracy alone. A small performance drop may be acceptable if privacy risk falls and group errors remain stable. However, a large increase in errors for applicants with limited histories needs further investigation. These are examples, not findings.

The safeguards may interact in different ways. Feature reduction could remove useful information or reduce reliance on noisy proxy variables, while privacy noise may affect estimates unevenly when groups are small. Testing is needed to establish the direction and size of these effects, including for the combined model.

A benchmark from another country can test the method but cannot represent all Indian lending practices or applicant populations. Real-world use requires local validation, lawful data handling, security, suitable explanations, and a way for applicants to request human review or challenge decisions.

7.1 How results should be interpreted

Keep three types of evidence separate: predictive metrics measure model performance, group metrics describe differences in errors, and privacy metrics assess formal guarantees or a particular attack. None replaces the others. Report them together because a model may perform well on one dimension and poorly on another.

If results are mixed, state that directly. Feature reduction may reduce data use but lower performance, and private training may change overall or group-specific errors. Conclusions should reflect measured trade-offs, privacy settings, uncertainty, and the dataset. Avoid calling one configuration universally best when results depend on the data or threshold.

7.2 From evaluation to responsible use

A favorable benchmark result supports further review, not automatic deployment. Lending use requires validation on representative data, review of applicable legal requirements, confirmation that the data suit the purpose, and clear accountability for decisions supported by the model.

Human reviewers need the authority and information to investigate questionable outcomes, correct inaccurate records, and explain how applicants can seek reconsideration. The process should be monitored. This paper recommends these safeguards but does not evaluate a lender's review process.

8. EXPECTED CONTRIBUTIONS

The draft does not include experimental results. It therefore reports no measured findings, numerical improvements, or best-performing model. The points below describe possible contributions and questions for the completed study; they are not confirmed results:

• A shared comparison process could make it easier to assess the effects of data minimization and differential privacy.

• Looking only at overall predictive performance could miss differences in error rates between applicants with limited and extensive credit histories.

• Reporting the formal privacy guarantee alongside the results of a membership-inference attack would show two related but distinct aspects of privacy protection.

• The effects of either safeguard on predictive performance and fairness need to be measured on the chosen dataset rather than assumed in advance.

• Reporting the data source, group definitions, privacy settings, and study limitations would support reproducibility and more careful interpretation.

9. LIMITATIONS AND FUTURE WORK

The current paper presents a research plan rather than completed empirical work. It includes no dataset-specific results, so it cannot establish whether data minimization or differential privacy improves performance, reduces observed attack success, or changes group disparities. Those questions remain open until the protocol is implemented and the results are analyzed. The proposed comparisons also depend on finding data with a suitable outcome and a defensible measure of credit-history depth. If no such data are available, the planned fairness comparison may need to be revised.

A public benchmark can support development and testing, but it may not reflect the population or decision process for which a model is eventually intended. The Home Credit dataset, if selected, would provide a methodological test rather than direct evidence about Indian lending. Differences in data collection, credit access, outcome definitions, and institutional practices can limit transfer to another setting. Future work should therefore validate the protocol with context-appropriate data and report the limits of any comparison across datasets.

The proposed fairness groups capture only one dimension of difference between applicants. They do not cover every relevant social or economic group, and a dataset may not contain lawful, reliable information for evaluating all such groups. The privacy evaluation is also limited to the chosen membership-inference attack and does not test every possible disclosure pathway. Formal differential-privacy guarantees depend on the implementation, privacy

accountant, and stated assumptions. Future work could broaden the threat model and examine intersections between credit-history depth and other legally and ethically appropriate group definitions.

The study may also face practical limits in statistical power, particularly for small groups or rare outcomes. Wide confidence intervals would make it difficult to distinguish a real difference from sampling variation. Repeated runs and careful uncertainty reporting can make this limitation visible, but they cannot replace adequate and appropriate data. Researchers should resist making strong claims when group counts are small, records are incomplete, or the target outcome is an imperfect proxy for the concept of credit risk.

Future research can build on this framework by testing alternative definitions of data minimization, privacy budgets, and model classes, provided the comparisons are planned and reported transparently. It can also examine whether explanations, applicant recourse, and human review work as intended in real decision settings. Such extensions should be treated as separate questions with their own evidence requirements. The present study can provide a starting protocol, but it cannot answer those implementation questions without additional data and evaluation.

10. RECOMMENDATIONS

• Collect and retain only information that is relevant, justified, and lawfully processed for the stated credit-assessment purpose.

• Test privacy safeguards before deployment and report the differential-privacy parameters and accounting method.

• Report model performance and error rates separately for applicants with limited and extensive credit histories, as well as overall.

• Do not treat a limited credit history as proof that someone is a high-risk borrower. When information is insufficient, consider whether additional lawful evidence or human review is appropriate.

• Use data from the intended deployment context whenever possible. If the study uses an international public dataset, describe it as a benchmark and avoid applying its results directly to India.

• Document feature selection, model versions, evaluation splits, and assumptions about the attack so that others can reproduce and audit the study.

• Set up ongoing checks for model drift, a process for reviewing complaints and correcting data errors, and meaningful human oversight for consequential decisions.

11. CONCLUSION

A credit risk model should be judged not only by how accurately it predicts default, but also by how it uses personal information and how its mistakes affect applicants with different amounts of credit history. Data minimization and differential privacy address separate concerns: one limits unnecessary data use, while the other limits the influence of individual records during training. Used together, they may provide complementary protections, but they could also change predictive performance or the distribution of errors across groups.

This paper proposes a way to compare a baseline model, data minimization, differential privacy, and the combination of both safeguards. It also includes group-level fairness measures and membership-inference testing. Because no experimental results were provided, the paper does not claim that either safeguard improves privacy or fairness in this application. The next step is to carry out the evaluation using a suitable dataset, explain where the data came from and what they can and cannot show, and report the measured results. Any conclusions intended for lending in India should be checked with relevant local data and supported by responsible data governance, transparency, and human review.

REFERENCES

Alcarva, P. (2026). Algorithmic fairness as a risk-management problem in banking and insurance: Regulatory frameworks, model governance, and fairness-aware credit scoring. *Risks, 14*(9), Article 205. https://doi.org/10.3390/risks14090205

Amed, S., Sen, T., & Banerjee, S. (2026). FSL-BDP: Federated survival learning with Bayesian differential privacy for credit risk modeling. *arXiv preprint arXiv:2601.11134*. https://arxiv.org/abs/2601.11134

Carvalho, T., Moniz, N., & Antunes, L. (2023). A three-way knot: Privacy, fairness, and predictive performance dynamics. In N. Moniz, Z. Vale, J. Cascalho, C. Silva, & R. Sebastião (Eds.), *Progress in artificial intelligence: 22nd EPIA conference on artificial intelligence* (pp. 55–66). Springer. https://doi.org/10.1007/978-3-031-49008-8_5

Chang, H., & Shokri, R. (2021). On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)* (pp. 292–303). IEEE. https://doi.org/10.1109/EuroSP51992.2021.00028

Chaudhuri, K., Monteleoni, C., & Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research, 12*(29), 1069–1109. https://www.jmlr.org/papers/volume12/chaudhuri11a/chaudhuri11a.html

Fioretto, F., Tran, C., Van Hentenryck, P., & Zhu, K. (2022). Differential privacy and fairness in decisions and learning tasks: A survey. *arXiv preprint arXiv:2202.08187*. https://doi.org/10.48550/arXiv.2202.08187

Gonzalez-Zelaya, V., Salas, J., Megías, D., & Missier, P. (2024). Fair and private data preprocessing through microaggregation. *ACM Transactions on Knowledge Discovery from Data*. https://doi.org/10.1145/3617377

Hansen, V., Neerkaje, A., Sawhney, R., Flek, L., & Søgaard, A. (2024). The impact of differential privacy on group disparity mitigation. *Findings of the Association for Computational Linguistics: NAACL 2024*, 3952–3965. https://doi.org/10.18653/v1/2024.findings-naacl.249

Hurlin, C., Pérignon, C., & Saurin, S. (2026). The fairness of credit scoring models. *Management Science, 72*(1), 406–425. https://doi.org/10.1287/mnsc.2022.03888

Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., & Ullman, J. (2019). Differentially private fair learning. *Proceedings of the 36th International Conference on Machine Learning, 97*, 3000–3008. https://proceedings.mlr.press/v97/jagielski19a.html

Kaarat, A., Batthula, V. J. R., & Segall, R. S. (2026). Unified federated AI framework for credit scoring: For privacy, fairness, and scalability. *International Journal of Artificial Intelligence (AI) in Business and Management, 2*(1), 1–31. https://doi.org/10.4018/IJAIBM.398859

Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research,*

297(3), 1083–1094. https://www.sciencedirect.com/science/article/abs/pii/S0377221721005385

Kpatcha, E. (2025). Balancing fairness and accuracy in machine learning-based probability of default modeling via threshold optimization. *Journal of Risk and Financial Management, 18*(12), 724. https://www.mdpi.com/1911-8074/18/12/724

Makhlouf, K., Stefanović, T., Arcolezi, H. H., & Palamidessi, C. (2024). A systematic and formal study of the impact of local differential privacy on fairness: Preliminary results. *2024 IEEE 37th Computer Security Foundations Symposium (CSF)*, 1–16. https://ieeexplore.ieee.org/abstract/document/10664426

Maniar, T., Akkinepally, A., & Sharma, A. (2021). Differential privacy for credit risk model. *arXiv preprint arXiv:2106.15343*. https://doi.org/10.48550/arXiv.2106.15343

Say, L., Denis, C., & Pinot, R. (2025). Fairness meets privacy: Integrating differential privacy and demographic parity in multi-class classification. *arXiv preprint arXiv:2511.18876*. https://doi.org/10.48550/arXiv.2511.18876

Szepannek, G., & Lübke, K. (2021). Facing the challenges of developing fair risk scoring models. *Frontiers in Artificial Intelligence, 4*, Article 681915. https://doi.org/10.3389/frai.2021.681915

Tian, H., Zhang, G., Liu, B., Zhu, T., Ding, M., & Zhou, W. (2025). Do fairness interventions come at the cost of privacy: Evaluations for binary classifiers. *IEEE Transactions on Dependable and Secure Computing*. https://ieeexplore.ieee.org/abstract/document/10925515

Tran, C., Dinh, M., & Fioretto, F. (2021). Differentially private empirical risk minimization under the fairness lens. *Advances in Neural Information Processing Systems, 34*, 27555–27565.https://proceedings.neurips.cc/paper_files/paper/2021/hash/e7e8f8e5982b3298c8addedf6811d500-Abstract.html

Zhang, H., Xu, J., He, J., & Fang, X. (2026). Privacy-preserving credit risk prediction with alternative data. *arXiv preprint arXiv:2606.10333*. https://doi.org/10.48550/arXiv.2606.10333