Tabular Deep Learning vs Classical Machine Learning for Urban Land Cover Classification

Muntasir Tabasum [1] Tanpia Tasnim [2] Md. Ekramul islam [3] Al Zadid Sultan Bin Habib [1]

[1]West Virginia University, Morgantown, WV 26506, USA [2]Green University of Bangladesh, Narayanganj-1461, Bangladesh [3]Stamford University Bangladesh, Dhaka-1217, Bangladesh

Introduction & Motivation

Urban Land Cover (ULC) maps are crucial for planning, flood and heat-risk assessment, and environmental monitoring. In many workflows, decisions are made from tabular features derived from aerial/satellite imagery rather than raw images.

The ULC dataset we use provides such descriptors for nine land-cover classes (e.g., roads, trees, grass, water) and exhibits common challenges:

- **High-dimensional, heterogeneous features** (spectral, spatial, contextual).
- **Class imbalance** between majority and minority classes.
- **Nonlinear interactions** that simple linear models struggle to capture.

Classical ML (Logistic Regression, SVM, tree ensembles, GBDTs) is still the standard for tabular ULC, while Tabular Deep Learning (TDL) models promise to better exploit complex interactions. It remains unclear when the extra complexity of TDL is actually beneficial in practice.

Research Questions & Contributions

Research Questions

- **RQ1:** How do strong classical ML methods compare to TDL models on the ULC dataset?
- **RQ2:** How do class-imbalance handling and non-linear feature interactions affect this comparison?
- **RQ3:** What guidelines can we offer for choosing models in real-world ULC pipelines?

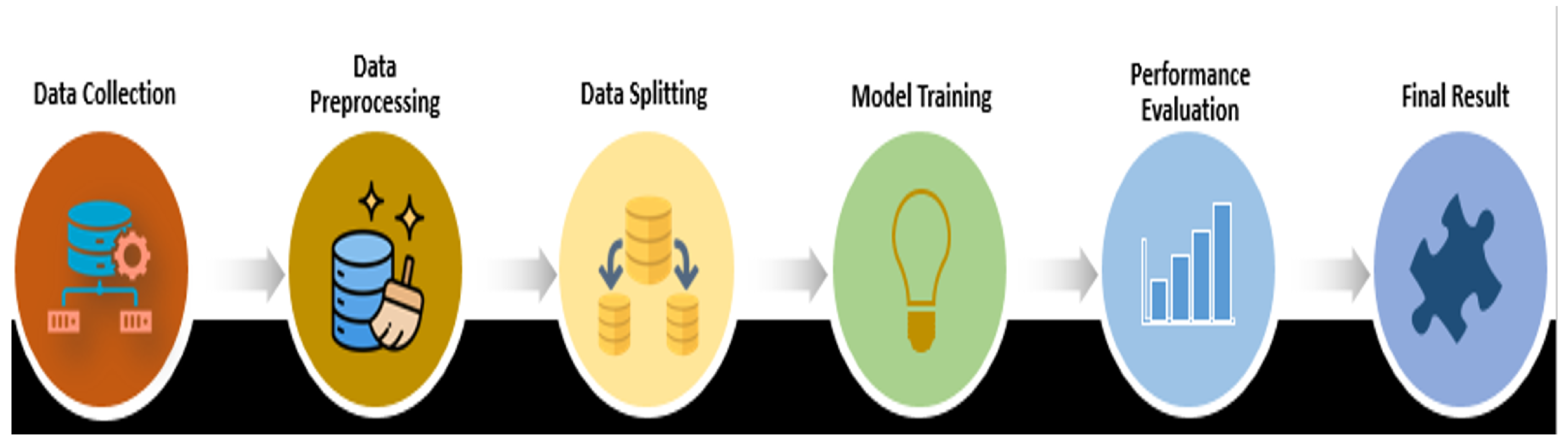
Contributions

- Unified, reproducible benchmarking of classical ML and several TDL models on ULC.
- Evaluation with accuracy and macro metrics to reveal per-class strengths/weaknesses.
- Analysis of imbalance-aware TDL (e.g., class-weighted loss) leading to practical recommendations for ULC practitioners.

Data & Experimental Setup

- **Dataset:** UCI ULC with tabular descriptors derived from aerial imagery; nine land-cover classes with mixed class frequencies.
- **Preprocessing:** Standardization of continuous features; stratified train–validation–test splits with repeated runs; class weights in the loss for TDL, standard settings for tree-based models.
- **Evaluation:** Overall accuracy as the main metric, with macro precision/recall/F1 and confusion matrices to inspect per-class behaviour.

Methods Overview



Classical ML vs TDL Models

We train all models on the same ULC tabular features and splits to enable a fair comparison between classical ML and TDL.

- **Classical ML:** Logistic Regression, SVM, Decision Tree, kNN, Naive Bayes, Random Forest, Gradient Boosting (GBM), AdaBoost, XGBoost, and CatBoost.
- **TDL:** MLP, 1D CNN, and tabular-specific architectures (TabNet, FT-Transformer, TabTransformer, TabSeq) trained with class-weighted cross-entropy.

Results: Overall Performance

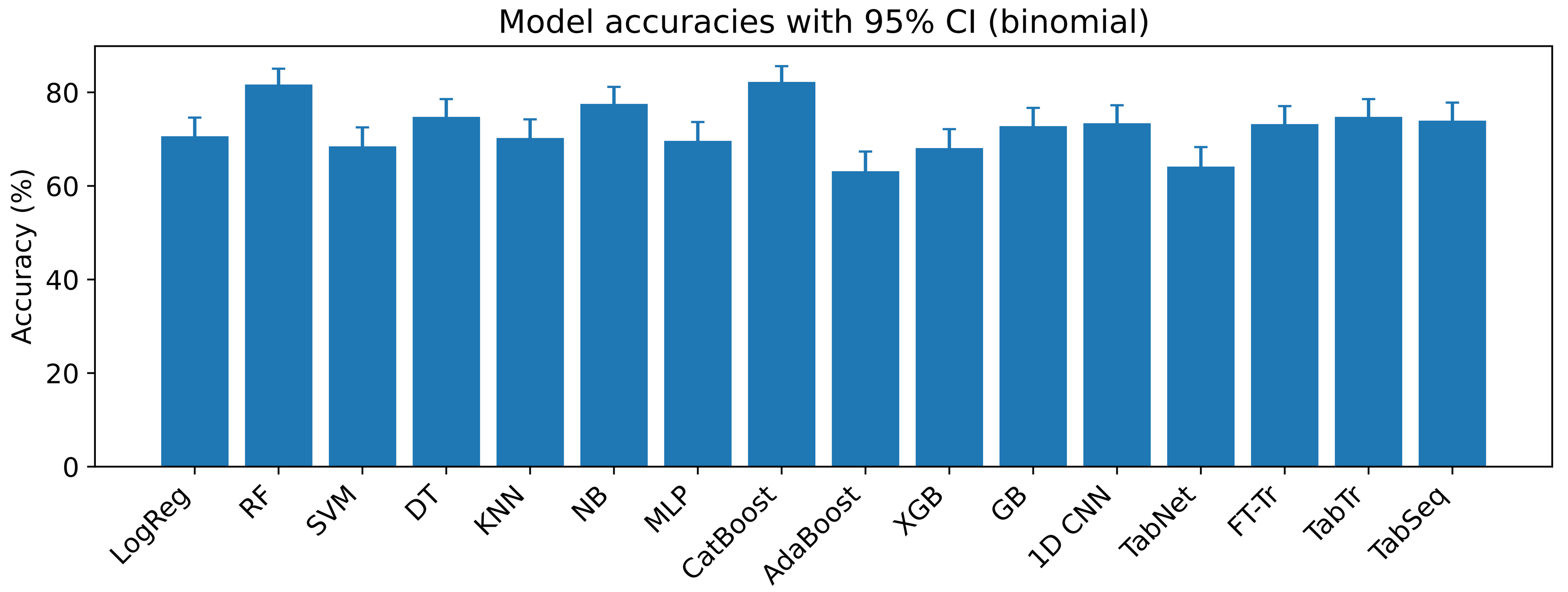


Figure 1. Model accuracies on the ULC test set with 95% binomial confidence intervals.

Table 1. Top 8 models on ULC (test set), sorted by test accuracy.

Model	Test Acc. (%)	Precision	Recall	F1	AUC
CatBoost	**82.25**	0.81	0.82	0.82	0.98
Random Forest	81.66	0.81	0.83	0.81	0.97
Naive Bayes	77.51	0.76	0.78	0.76	0.96
Decision Tree	74.75	0.72	0.75	0.73	0.96
TabTransformer	74.75	0.71	0.72	0.72	0.96
TabSeq	73.96	0.70	0.74	0.74	0.95
1D CNN	73.37	0.73	0.75	0.75	0.97
FT-Transformer	73.18	0.70	0.74	0.70	0.93

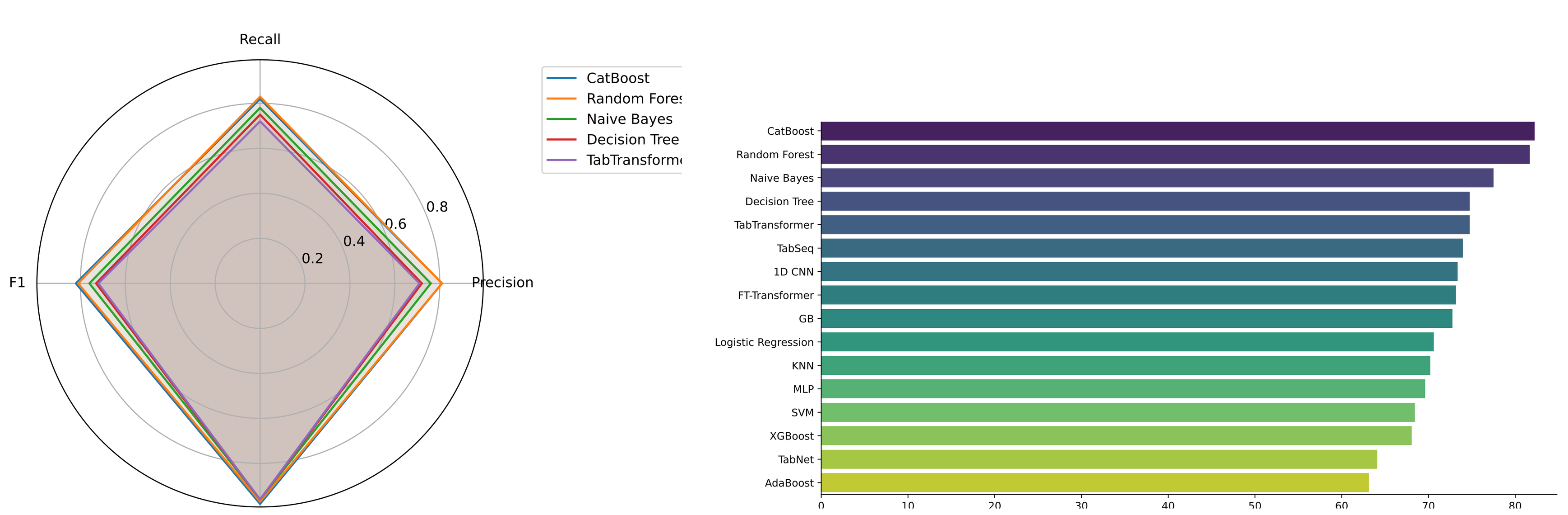


Figure 2. (a) Top models: Precision/Recall/F1/AUC (b) Comparison of test accuracy across models.

Results: Deep Models & Class Imbalance

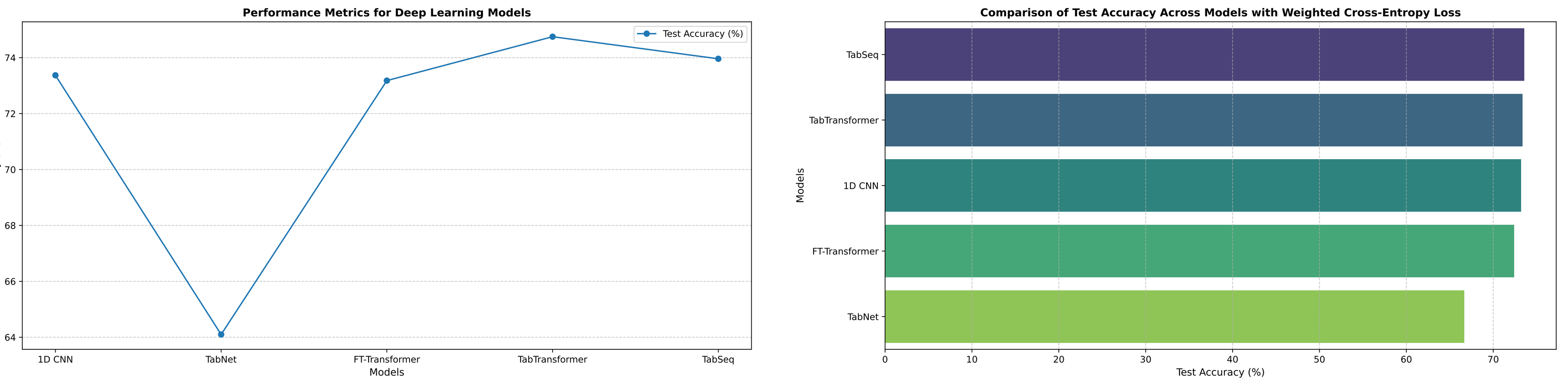


Figure 3. Test Accuracy of TDL models with (a) standard cross-entropy (b) class-weighted cross-entropy.

Table 2. TDL models trained with class-weighted cross-entropy on ULC.

Model	Test Acc.(%)	Precision	Recall	F1	AUC
1D CNN	73.18	0.70	0.74	0.71	0.94
TabNet	66.67	0.65	0.70	0.66	0.93
FT-Transformer	72.39	0.66	0.72	0.68	0.93
TabTransformer	73.37	0.69	0.72	0.71	0.96
TabSeq	73.57	0.70	0.74	0.71	0.95

Discussion & Practical Takeaways

Summary. On this ULC dataset, strong tree ensembles (especially CatBoost and Random Forest) achieve the best overall accuracy, while modern TDL models are competitive but do not consistently surpass classical methods.

Practical takeaways for ULC practitioners

- **Start with strong ensembles.** Gradient boosting and related tree-based models remain robust baselines for tabular remote sensing data, especially when data are limited and moderately imbalanced.
- **Use TDL selectively.** TDL models (TabTransformer, TabSeq, 1D CNN) can match the best classical models while providing flexible feature representations, but require careful tuning and regularization.
- **Handle imbalance explicitly.** Class-weighted cross-entropy improves deep models on minority classes but does not fully close the gap to the best tree ensembles, suggesting that imbalance-aware design is still necessary.

Future work

- Develop **tabular transformer–style, remote sensing–specific TDL model** that encodes spectral–spatial structure, class hierarchies, and imbalance directly into the architecture.
- Extend the study to larger multi-city ULC or related datasets and additional evaluation criteria (e.g., calibration, fairness across land-cover types, and deployment robustness).

Key References

[1] L Prokhorenkova et al. *CatBoost: Unbiased Boosting with Categorical Features.* NeurIPS, 2018.
[2] SÖ Arik and T Pfister. *TabNet: Attentive Interpretable Tabular Learning.* AAAI, 2021.
[3] X Huang et al. *TabTransformer: Tabular Data Modeling Using Contextual Embeddings.* arXiv, 2020.
[4] Y Gorishniy et al. *Revisiting Deep Learning Models for Tabular Data.* NeurIPS, 2021.
[5] AZSB Habib et al. *TabSeq: A Framework for Deep Learning on Tabular Data via Sequential Ordering.* ICPR, 2024.

See code: https://github.com/mtesha/tdl-vs-ml-urbanlandcover

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 5th Muslims In ML (MusIML)

Corresponding author: ah00069@mix.wvu.edu