Predicting Colombian School Test Performance from Satellite Imagery

Manuel Cárdenas Prieto

Stanford Graduate School of Education

Motivation

- Educational disparities are shaped by infrastructure and neighborhood conditions that are difficult to observe consistently at national scale.
- Repeatable overhead imagery may help prioritize where direct school surveys and support are most urgently needed.[1,2]
- The goal is **screening and hypothesis generation**, not replacing student evaluation or claiming causal effects.

Prediction tasks

We predict whether mean Saber 11 math performance is above:

GLOBAL	WITHIN QUARTILE
the national training mean	peers with similar INSE

Two visual scales

BROAD: ~2.56 km neighborhood

CLOSE: ~300 m school context

Broad imagery captures area context; close imagery resolves school and nearby infrastructure.

Dataset and stratification

11,049 processed schools

11,035 modeling split

2 image scales

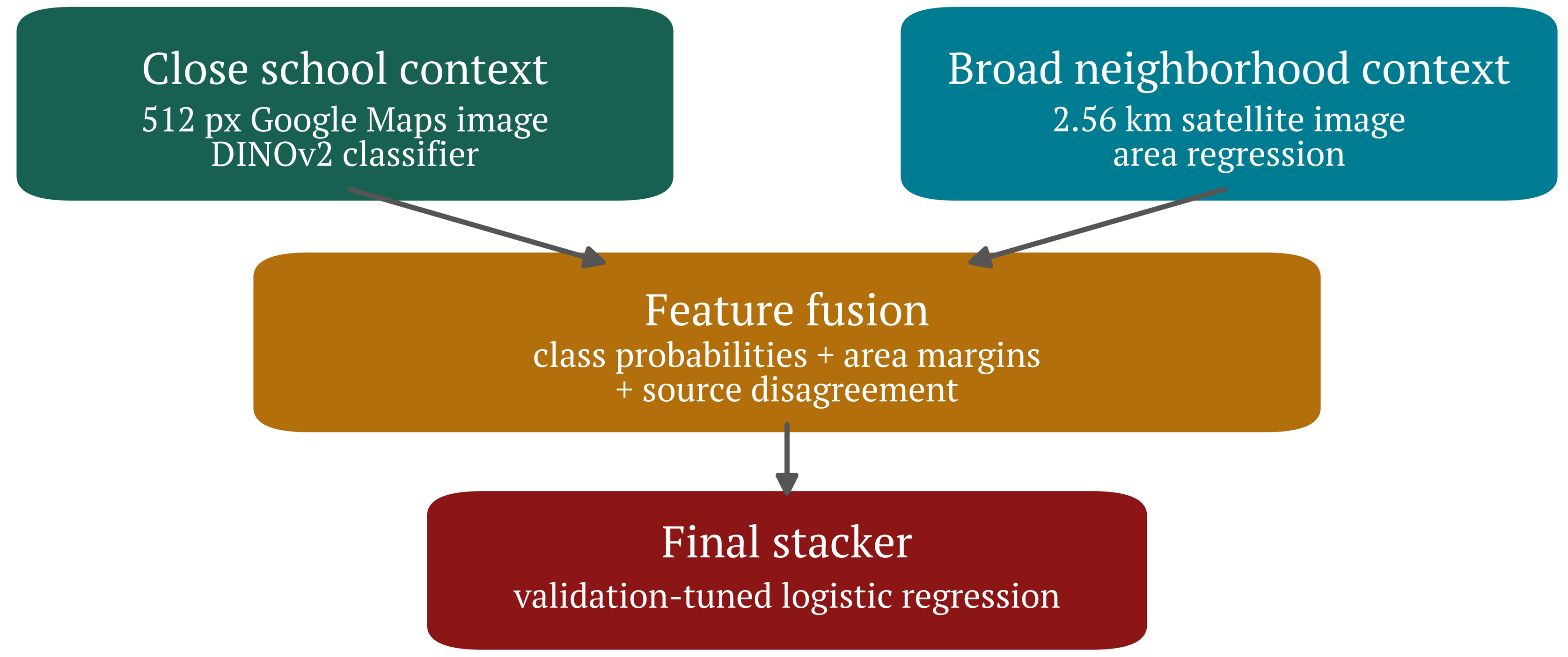
INSE is ICFES's student socioeconomic index, derived from background questions on household goods, services, and parental education.

	Q1	Q2	Q3	Q4	Total
Train	1,931	1,931	1,931	1,931	7,724
Validation	414	414	413	414	1,655
Test	414	414	414	414	1,656

Splitting by INSE quartile preserves socioeconomic composition across train, validation, and test.

The within-quartile task removes much of the neighborhood-wealth shortcut and asks for signal among socioeconomic peers.

Multi-scale modeling pipeline



DINOv2 features[3] are fine-tuned separately at each spatial scale, then fused by a small logistic stacker.

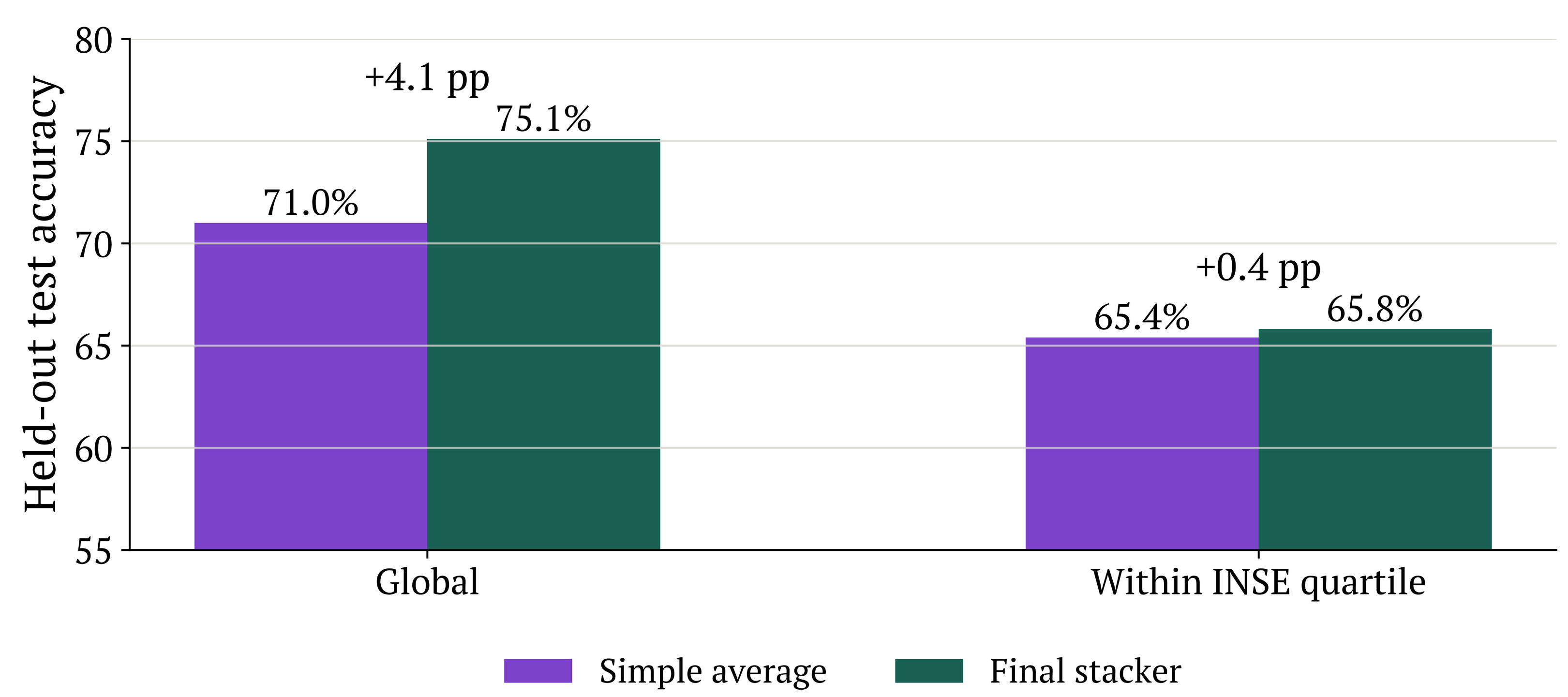
Headline held-out test results

75.1%
Global balanced accuracy
84.1% average precision

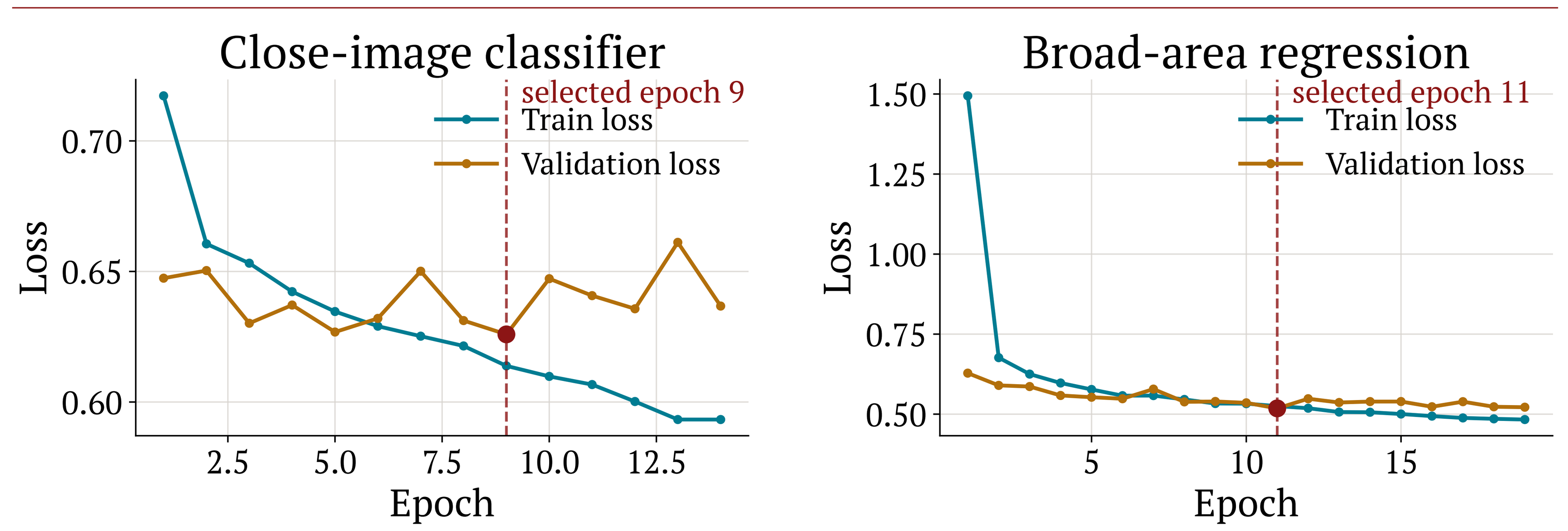
65.8%
Within-quartile balanced accuracy
70.8% average precision

The peer-comparison task remains above chance after much of the visible socioeconomic signal is removed.

Combining scales improves prediction

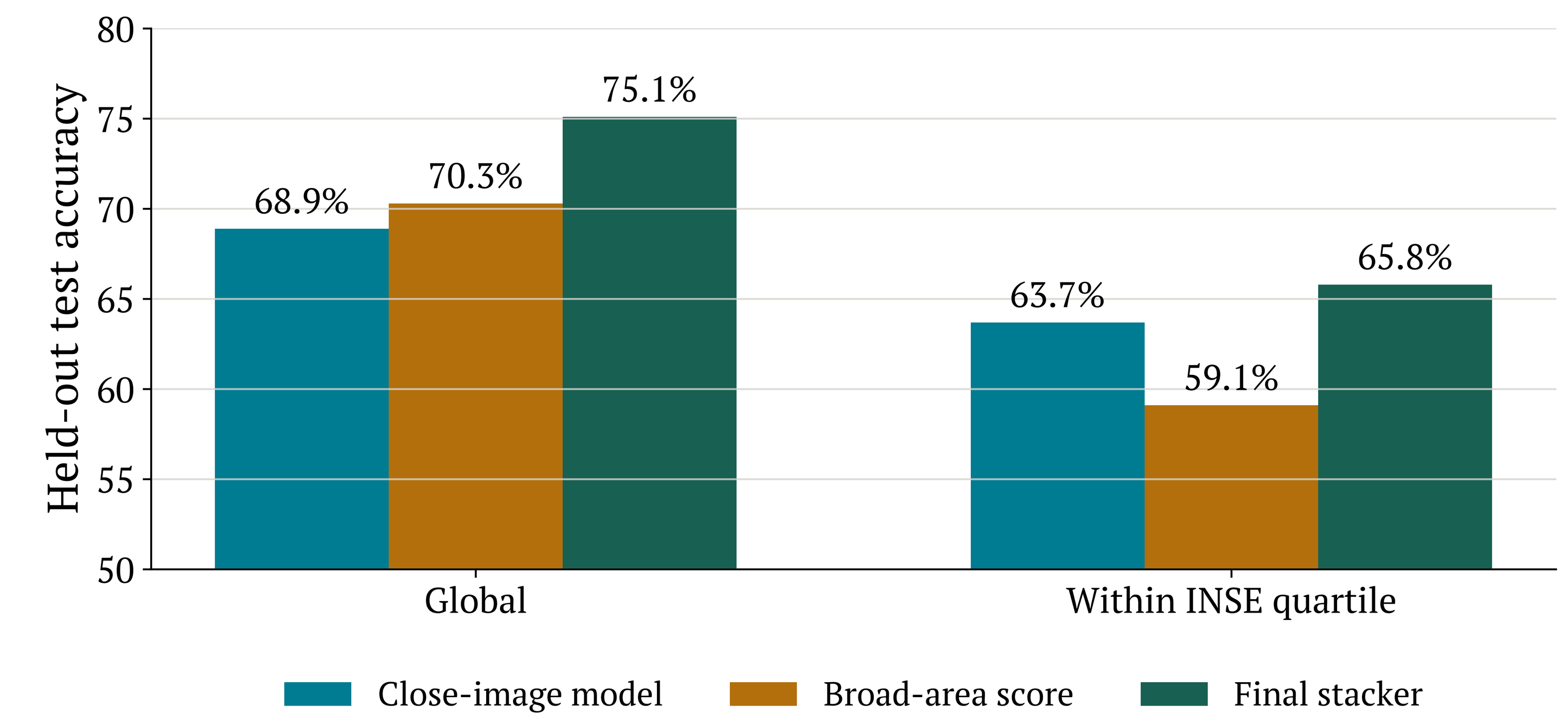


Training dynamics



Checkpoints are selected by validation loss before test reporting.

Performance by visual signal



Broad context is strongest globally; close context carries more signal within INSE quartiles.

Where does the close model look?

Original close image

Integrated Gradients overlay

Integrated Gradients[4] highlights image regions supporting the positive-class prediction.

Interpretation and limitations

- Attributions describe **model behavior**, not causal effects.
- Broad imagery predicts INSE more reliably ($R^2 = 0.64$) than math ($R^2 = 0.35$).
- Imagery misses teachers, curriculum, family support, and student histories.

Main conclusions

- **The within-quartile task is harder:** neighborhood wealth is the dominant signal visible from space, and socioeconomic peer comparison removes much of that shortcut.
- **Some failures reveal unusually successful schools:** imagery predicts low performance, but the focal school outperforms nearby schools facing similar visible conditions.
- These positive outliers merit direct study for leadership, teaching practices, resources, or community programs invisible from space.

References

[1] N. Jean *et al.*, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, 2016.
[2] D. Runfola, A. Stefanidis, and H. Baier, "Using satellite data and deep learning to estimate educational outcomes in data-sparse environments," *Remote Sens. Lett.*, vol. 13, no. 1, pp. 87–97, 2022.
[3] M. Oquab *et al.*, "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.
[4] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, pp. 3319–3328, 2017.

manuelcp@stanford.edu