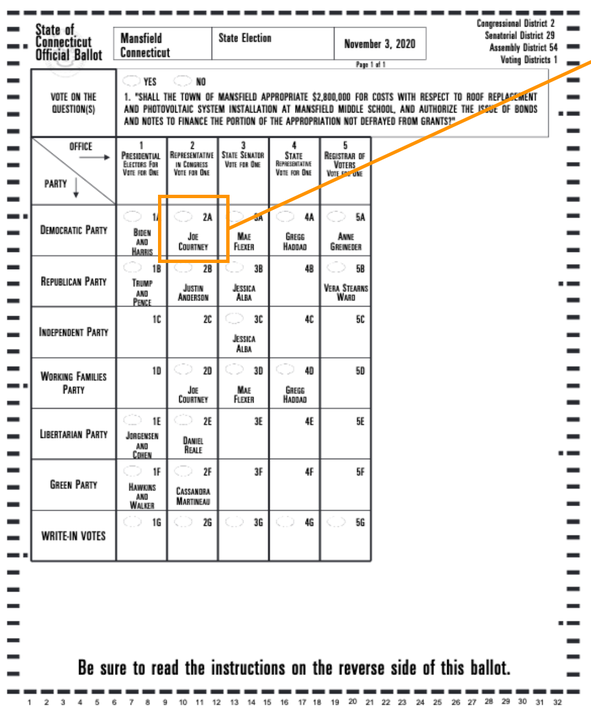
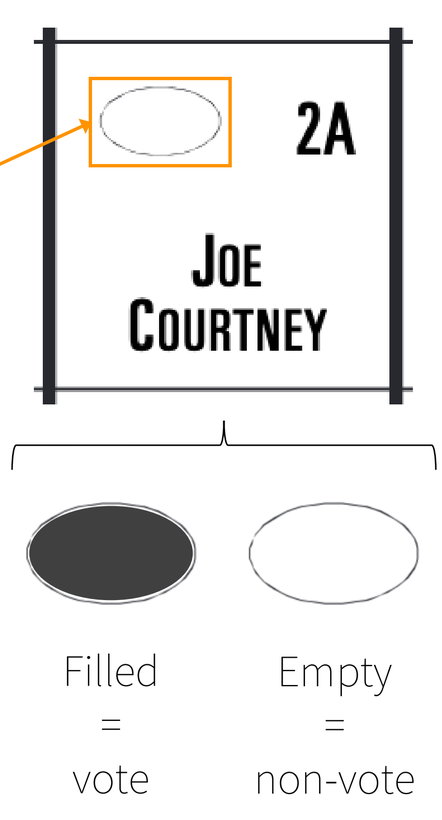
AI can change your vote: can we secure the system?

Background

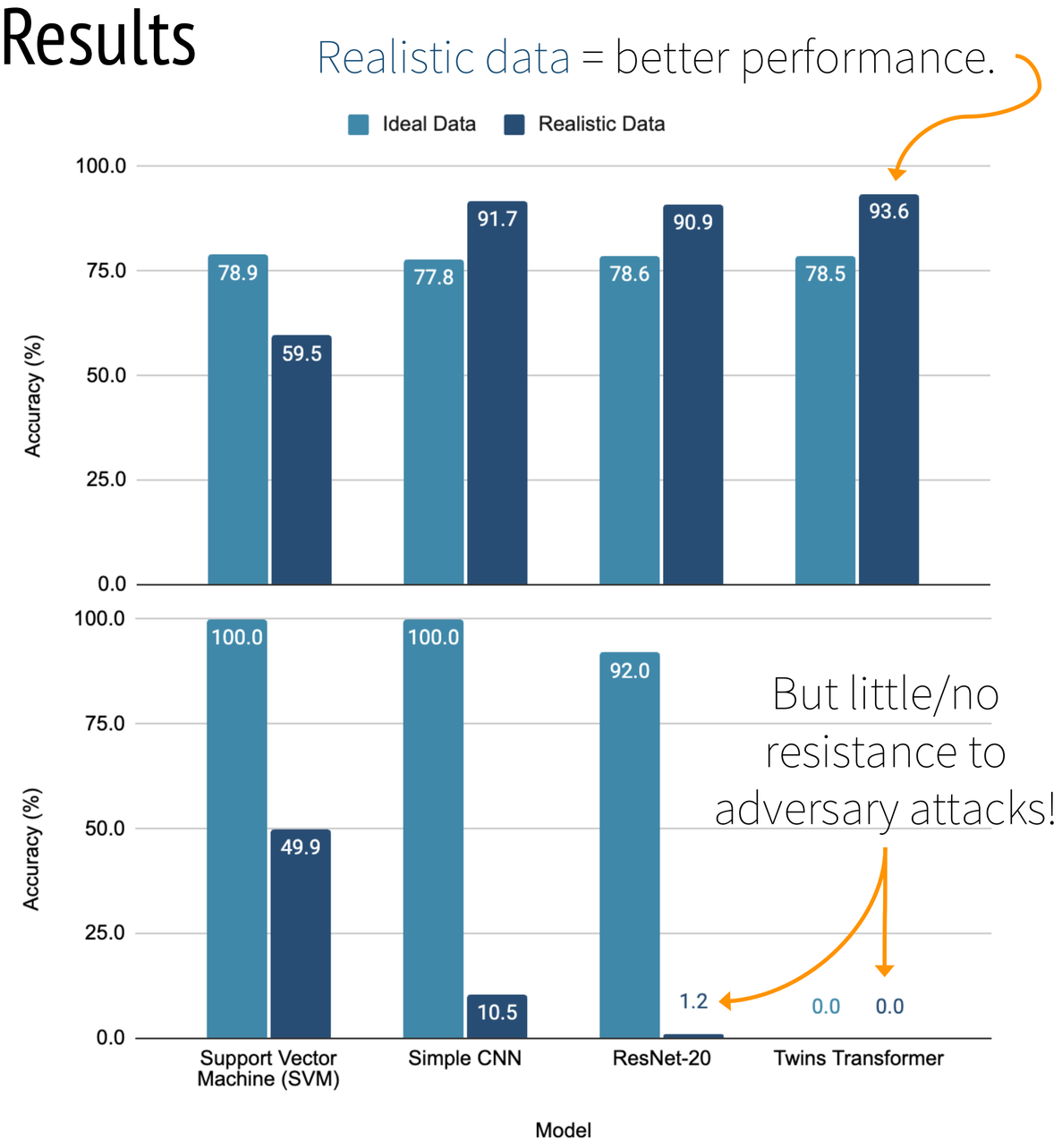




Machine learning models record votes from the bubbles on a ballot.

How vulnerable are these models to: training data, and an adversary attack?

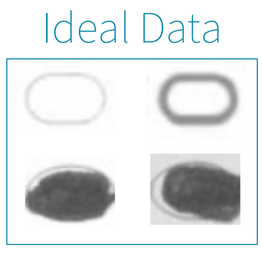
Results



So votes can be misclassified by incorrect model training and/or adversary attacking models!

Methodology

Two datasets: ideal & realistic.
Four models: SVM, Simple-CNN, ResNet-20, Twins Transformer.

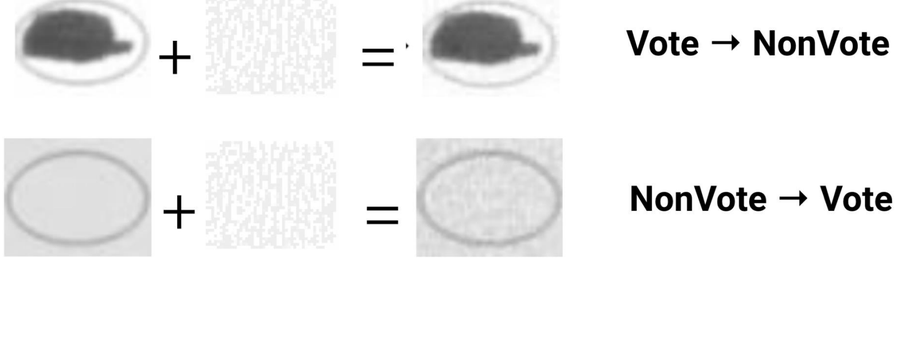


Vulnerability #1 (ambiguity):

- Train each model on each dataset, record accuracy.

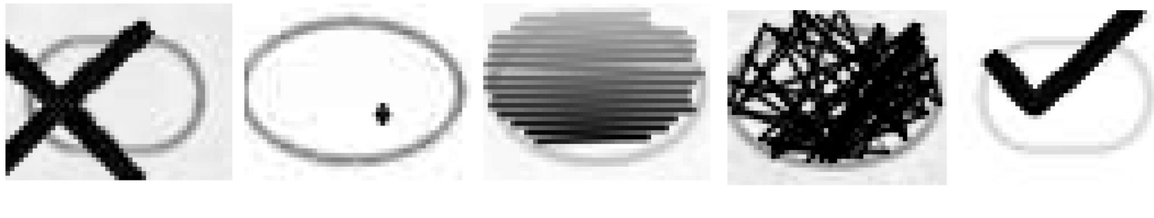
Vulnerability #2 (adversary attacks models):

- Train each model on each dataset.
- Generate adversarial bubbles with APGD.
- Attack each model with APGD-generated images, record accuracy.



Conclusions & Next Steps

- Realistic data → tradeoff between model accuracy and defense against attacks!
- Next steps: defensively train models, address tradeoffs, augment realistic marks.



A cautionary tale of exploitable AI! Need to build secure models!

Disclaimer: The images of ballots in this poster are for illustrative purposes only. They are not intended as endorsements of any candidate, political party, or policy.

Acronym Key. APGD = Adaptive Projected Gradient Descent, a type of adversarial machine learning attack. SVM = Support Vector Machine, CNN = Convolutional Neural Network.

This poster is based on our research: K. Mahmood, C. Manicke, E. Rathbun, A. Verma, S. Ahmed, N. Stamatakis, B. Fuller, and L. Michel, "Busting the paper ballot: Voting meets adversarial machine learning," 2025, in Submission.