Project G6: Deep Learning for Predicting Chronic Wasting Disease Spread in U.S. Deer Populations

- **Start with logistic regression first**: Before MLP, train logistic regression on county-year features (historical detections, neighbor counts, environmental variables)—if this achieves AUC>0.75, justify MLP lift. I suggest gradient-boosted trees with class weighting; and calibrate probabilities (isotonic/Platt); This sets a very hard baseline that a small MLP must beat.

- sounds like a graph-aware neural model is a better alternative for this problem.

- **Define neighbor adjacency precisely**: Queen contiguity (shared border/corner) is mentioned, but clarify time-lag (t-1 only or rolling 3-year window?) and whether you compute neighbor detection counts or binary flags.

Recommended actions before Milestone-2

- **Dataset verification**: Download USGS CWD Distribution, verify county-year pairs, check detection counts per state, visualize spatial spread pattern (map), report class imbalance (positive detection rate likely low, e.g. <5%).

- **Define environmental data sources precisely**: List exact datasets for land cover (NLCD?), temperature (PRISM?), elevation (USGS NED?), deer density proxy (state harvest data?), human population (Census)—verify spatial resolution matches county level.

- **Implement spatial-temporal join**: Test joining county-year records to environmental features, validate no missing values (>10% missingness requires imputation strategy), document join success rate.

- **Create validation split**: Use temporal split (train ≤2022, val 2023, test 2024+), verify no temporal leakage (neighbor features use only t-1 information), report train/val/test sizes.

- **Logistic regression baseline**: Train L2-regularized logistic regression on all features, report AUC, precision/recall for positive class (new detection), log training time—this establishes non-DL ceiling before MLP.

Ablations

- **Feature groups**: Compare historical-only (prior detections, neighbor counts) vs environmental-only (land cover, temperature, elevation) vs combined—hypothesis is historical features dominate (expect +15–20% AUC over environmental-only).

- **Neighbor features on/off**: Remove neighbor detection counts and flags—expect -10% AUC drop, testing whether spatial contagion signal is critical for spread prediction.

- **MLP depth**: Test 2-layer (shallow) vs 4-layer (deep) MLP with same total parameters—hypothesis is shallow network sufficient for tabular data, deep network may overfit without regularization (dropout >0.3).

Risks & mitigations

- **Risk**: Severe class imbalance (new detections likely <5% of county-years)—**Mitigation**: Use class-weighted cross-entropy loss (weight=10–20 for positive class), tune threshold on validation PR-AUC, report precision@90% recall.

- **Risk**: Surveillance bias (detections depend on sampling effort, human population density is weak proxy)—**Mitigation**: Add state-level fixed effects or stratify evaluation by state, report false positive rate in high-effort vs low-effort counties.

- **Risk**: Neighbor features may leak future information if not properly lagged—**Mitigation**: Validate that neighbor counts use only t-1 data (not concurrent t), test temporal robustness by shifting train/val splits backward 1 year.

Open questions

- What is the positive class rate (new detections / total county-years) in the USGS dataset, and how will you handle imbalance if rate <5%?
- What specific environmental datasets will you use (NLCD for land cover, PRISM for temperature?)—have you verified these are publicly available at county resolution?
- What is your MLP architecture exactly (e.g., 2 hidden layers of 128 units each, ReLU activation, dropout=0.3, final sigmoid output)—can you commit to this before coding?