ADR Modeling Pipeline : Research and Implementation Report

Abstract

We present an integrated pipeline for modeling adverse drug reactions (ADRs) under limited labels and high-dimensional side-effect profiles. Key components include semi-supervised learning to handle label scarcity, GAN-based synthetic data generation for class balancing, dimensionality reduction for both modeling and visualization, and explainability tools to interpret predictions. Our experiments on the SIDER dataset (1,430 drugs × 5,868 side effects) demonstrate measurable improvements at each stage and yield actionable insights into drivers of “abdominal pain” ADRs.

1. Semi-Supervised Learning

1.1 Background

When only a small fraction of ADR labels are known, semi-supervised methods exploit unlabelled data to improve classification. Two approaches:

- Masked Autoencoders randomly hide features and train a neural network to reconstruct them, capturing latent correlations.
- Label Propagation builds a similarity graph among drugs (using k-nearest neighbours) and diffuses the few known labels across that graph.

1.2 Application & Findings

- We masked 95% of the “abdominal pain” labels and trained Label Propagation on the full 5,868-dimensional drug profiles.
- **Baseline performance**: Accuracy ≈ 0.50, positive-class recall ≈ 0.07.
- After reducing dimensionality via PCA to 50 components, Label Propagation achieved accuracy ≈ 0.58 and recall ≈ 0.23.

Conclusion: PCA markedly improves label diffusion, boosting sensitivity to rare ADR signals.

2. Synthetic Data Generation

2.1 Background

GANs for tabular data address class imbalance by generating realistic synthetic samples:

- CTGAN conditions on feature statistics to model mixed data types, producing synthetic drug profiles.

- Alternative SDV Synthesizers (TVAE, CopulaGAN) can be compared for fidelity, especially in biomedical contexts.

2.2 Application & Findings

- We focused on drugs labeled positive for abdominal pain, sampling 300 drugs and selecting the top 100 frequent side effects as features.
- CTGAN was trained on the positive subset (≈ 143 samples) and generated an equal number of synthetic positives.
- **Quality checks**:
  - Histograms for the top five side effects showed that synthetic samples captured marginal probabilities but underestimated rare patterns.
  - A PCA scatter of combined real vs. synthetic profiles revealed that synthetic points cluster tightly around the mean, indicating limited variance.

Conclusion: Synthetic augmentation successfully balances classes but requires deeper GAN training (more epochs or alternative architectures) to match real-data diversity.

3. Dimensionality Reduction

3.1 Background

High-dimensional side-effect matrices hamper both learning and interpretation. Common tools:

- PCA is a linear method that captures maximal variance in orthogonal components.
- UMAP nonlinearly embeds data while preserving both local neighborhood and global structure.
- MOFA+ (for multi-view factor analysis) can jointly decompose heterogeneous data sources if available.

3.2 Application & Findings

- PCA was used both to improve Label Propagation (Section 1) and to visualize real vs. synthetic distributions (Section 2).
- UMAP on the full drug-side-effect matrix produced a two-dimensional embedding in which:
  - Drugs with abdominal-pain labels concentrate in specific regions.

- KMeans (k=4) on the UMAP embedding delineated clusters, one of which was enriched for the target ADR.

Conclusion: UMAP combined with clustering provides a powerful exploratory tool to detect ADR-related drug subgroups

4. Explainability Tools

4.1 Background

Understanding model decisions is crucial for clinical trust:

- SHAP (Shapley values) attributes each prediction to feature contributions in a theoretically consistent manner.
- LIME uses local surrogate models to approximate behavior around individual predictions.
- ELI5 inspects weights of linear and tree-based models and integrates with SHAP/LIME for easy visualization.

4.2 Application & Findings

- We trained an XGBoost classifier on the balanced and augmented dataset.
- SHAP analysis identified the top ten side effects driving “abdominal pain” predictions. These included clinically plausible gastrointestinal and systemic ADRs, reinforcing model validity.

Conclusion: SHAP explanations not only validate model behavior against domain knowledge but also highlight potential comorbid ADR patterns worthy of further biological investigation.

5. Final Evaluation

To avoid overfitting and data leakage, we:

1. Split the 300-drug subset into a training set (200 drugs) and an unseen test set (100 drugs) with stratified sampling.
2. Augmented only the training positives via CTGAN, then balanced against negatives.
3. Trained XGBoost on this balanced+synthetic training set.
4. Evaluated on the untouched 100-drug test set.

Result: The hold-out test performance (precision, recall, F1) reflects realistic model generalization, confirming the benefit of PCA preprocessing and synthetic augmentation.

6. Implementation Notes

- All code for data preprocessing, model training, augmentation, visualization, and explainability is contained in the accompanying Jupyter notebook adr_modeling.ipynb.
- Key parameters (e.g., PCA components, CTGAN epochs, UMAP neighbors) are documented inline for reproducibility.
- Quality-check plots and classification reports are displayed sequentially under each methodological section.

7. Conclusions & Next Steps

This pipeline demonstrates a robust framework for ADR modeling under challenging conditions:

- Semi-supervised learning with PCA boosts scarce-label inference.
- CTGAN augmentation balances classes but requires further tuning for diversity.
- UMAP visualizations uncover ADR-enriched clusters.
- SHAP explanations provide actionable insights into side-effect drivers.

Future work includes exploring MOFA+ factors, comparing SDV’s TVAE/CopulaGAN to CTGAN, and integrating drug-chemical features to further enrich the model.