Introduction/Motivation

Identifying newborns at high risk for infant mortality, a rare yet significant public health outcome, can direct early intervention and monitoring efforts. Our research question assessed this. Which factors are the strongest predictors of infant mortality within the first year of life and which machine learning model best predicts these outcomes? At the start of our project, we discovered that the raw dataset contained many variables directly tied to the death outcome. Keeping these would create label leakage, so we had to modify our dataset before modeling. After removing outcome-dependent variables, dropping values with high-missing columns, and restructuring features to avoid leakage, we built a complete predictive modeling pipeline to evaluate how well different algorithms handle a heavily imbalanced classification problem.

Methods

Our analysis used a cleaned subset of the 2013 NVSS Linked Birth–Infant Death dataset. The original raw files contained nearly four million birth records and over 240 variables, many of which were administrative, had extremely high missing values/columns, or directly revealed the infant's death and causation. To prevent label leakage and ensure modeling integrity, we constructed a cleaned dataset (nvss_aggregated.csv) containing approximately 500,000 records and 40 medically relevant predictors. These variables include maternal characteristics, prenatal care indicators, birth outcomes, congenital anomalies, and neonatal interventions. All categorical features were encoded numerically, and variables with more than 80% missing values or known post outcome information (e.g., cause-of-death and details) were removed. We cleaned the dataset by dropping rows with missing values rather than imputing, since imputation could distort patterns in a rare event outcome like infant mortality (<0.5%). We created a binary target variable, infant_death, indicating whether the infant died before age one. We then used the cleaned dataset to create train, validation, and test splits. Our final modeling pipeline included three files (nvss_train.csv, nvss_val.csv, and nvss_test.csv) each containing the same feature structure and encoding as the aggregated dataset. This helped the consistency with all of our descriptive analysis, cross-validation, and final model evaluation.

We trained four predictive models on the split NVSS infant mortality data. These were penalized logistic regression, a linear Support Vector Machine, an XGBoost ensemble, and a PyTorch neural network. Before modeling, we noted the extreme class imbalance (only 0.6% infant deaths). For logistic regression, we created a balanced training subset by keeping all positive cases and randomly sampling 10% of survivors, then applied StandardScaler and class_weight="balanced" within a pipeline. We trained Ridge (L2), Lasso (L1), and Elastic Net models using GridSearchCV with stratified 3-fold cross-validation, tuning the regularization strength and penalty parameters based on F1 score. The SVM model used the full dataset, scaling all features and tuning the C parameter and class weights through the same cross-validation setup. For the XGBoost model, which does not require scaling, we used the full numeric dataset and handled imbalance using scale_pos_weight (negatives/positives). RandomizedSearchCV tuned the main hyperparameters such as tree depth, learning rate, number of estimators, and subsampling. The neural network was built with three hidden layers (64-32-16), ReLU activation, BatchNorm, and Dropout for regularization. The inputs were standardized, converted to GPU tensors, and trained using optimization and a weighted BCEWithLogitsLoss to handle rare positive class. We performed a small grid search over learning rates and applied early stopping based on validation loss to prevent overfitting. All models, except neural networks, were evaluated using stratified K-fold cross-validation and selected based on F1 score, which better reflects performance on rare outcomes than accuracy. The neural networks instead used a fixed validation set with early stopping instead of full cross validation. This is because running k fold cross validation would multiply the training time by the number of fold. It is computationally expensive and we run out of GPU computers. Additionally we have also included ROC-AUC on the validation and test sets to better understand each model performance under such class imbalance.

Results

Using cross-validation, we compared the performance of the four models (penalized logistic regression, linear SVM, XGBoost, and a neural network) using F1 score as the primary selection metric because of extreme class imbalance. Across the validation folds, XGBoost and the neural network consistently achieved the strongest balance between recall and precision for the positive class, while the linear SVM displayed very high accuracy driven influenced by predicting the majority class. The penalized logistic models (Ridge, Lasso, Elastic Net) performed similarly to one another, providing a moderate recall of infant deaths. Based on cross-validated F1 scores, the neural network and XGBoost emerged as the top two models, with both models substantially outperforming linear SVM in correctly identifying rare death outcomes.

We then evaluated each best model on the untouched test set of 712,884 births. XGBoost achieved the highest overall accuracy (93.34%) and recalled 76.8% of true infant deaths. The neural network demonstrated similar overall accuracy (91.93%) but achieved the **highest recall of death cases** (78.15%), along with strong ranking metrics (ROC-AUC = 0.922, PR-AUC = 0.495). In contrast, the linear SVM reached 99.55% overall accuracy yet detected only 25.8% of infant deaths, which shows that accuracy alone is misleading under extreme imbalance. The penalized logistic models performed similarly to XGBoost in positive-class recall but with slightly lower overall accuracy. Overall, the neural networks performed best for our goal of identifying high risk infants. It achieved the strongest recall on the rare positive cases while still maintaining solid overall accuracy, making it the most reliable model for detecting infant mortality risk.

	CV F1 score	Best Hyperparameter
Penalized Linear Regression (Lasso)	0.50191	Penalty = ‘L1”, C= 0.01
Support Vector Machine	Val accuracy: 0.99	C = 1, class_weight = None
Ensemble (XGBoost)	0.5416	colsample_bytree = 0.8, subsample = 0.8
Neural Network	– (no K-fold CV)	Best lr = 3e-03

Discussion

Although almost every model achieved high overall accuracy due to the extreme class imbalance (over 99% survivors), XGBoost and the neural network were much better at identifying the rare positive cases of infant death. On the test set, XGBoost correctly flagged around 77% of deaths, while the neural network slightly surpassed it at around 78–79% recall. In contrast, models such as Ridge, Lasso, and Elastic Net all performed similarly and hovered around 76% recall for the positive class, showing they could pick up basic linear patterns but struggled to capture the more complex interactions present in the data. Notably, the SVM, which initially looked extremely strong based on its overall accuracy, performed poorly on death cases (only about 26% recall), which showed that accuracy alone is misleading in heavily imbalanced datasets.

These differences largely reflect how flexible each model is in capturing nonlinear and high-dimensional interactions. XGBoost benefits from tree-based boosting, which naturally handles feature interactions and rare cases better than linear decision boundaries. The neural network also benefits from its multilayer (deeper) structure and class-weighting. This enables it to pick up on subtle patterns in the data and assign higher risk scores to infants with important but less obvious risk factors. On the other hand, the penalized logistic regression models were more constrained. This is because they rely on linear relationships and can’t capture complex patterns so it’s hard to model the data’s subtle interactions effectively. Several limitations impacted our project as well. First, infant death is extremely rare, so even small mistakes cause big drops in recall for the positive class. Second, our models only use the variables included in the dataset; important factors like prenatal care quality, hospital conditions, or environmental stressors aren’t measured, which limits how accurate the models can be. Finally, the neural network was trained on a single train/validation split instead of full cross-validation, so its performance may vary more across splits. Overall, while the results are strong, especially for XGBoost and the neural network, they should be interpreted with the understanding that rare-event prediction is difficult and would require careful clinical validation before real-world use.