Early Detection of Hypertensive Disorders in Pregnancy

Executive Summary

This report presents a comprehensive analysis of clinical data aimed at developing a predictive model for early detection of hypertensive disorders in pregnancy (HDP) at week 15 of gestation. The analysis leverages multimodal data including structured clinical measurements, laboratory results, demographic information, and unstructured clinical notes to identify high-risk pregnancies for targeted intervention.

Key Findings:

- **Dataset**: 10,000 low-to-moderate risk pregnancies with 4.32% positive cases (432 patients with HDP)
- **Best Model**: XGBoost achieved 58.1% recall and 83.3% precision with AUC of 0.974
- **Budget Optimization**: Top 100 patients capture 70.9% of true cases with 61% precision
- **Critical Features**: Blood pressure trends, diagnosis history, and lab ratios are key predictors

Business Impact: The model enables cost-effective screening by prioritizing high-risk patients for expensive laboratory testing, potentially reducing maternal and fetal morbidity through early intervention.

1. Methodology

Data Overview

- **Cohort**: Low and moderate-risk pregnancies only (high-risk patients excluded)
- **Prediction Point**: Week 15 of gestation
- **Target Variable**: Development of hypertensive complications (preeclampsia, gestational hypertension, eclampsia)

Feature Categories

1. **Demographics**: Age, socioeconomic status (capitation coefficient)
2. **Laboratory Tests**: CBC, biochemical markers, urine analysis
3. **Blood Pressure**: Systolic/diastolic measurements with trend analysis
4. **Diagnosis History**: ICD-9 codes from 4 and 24-month windows

5. **Clinical Text**: Physician notes and clinical documentation
6. **Smoking Status**: Self-reported and text-extracted smoking information

Data Preprocessing Strategy:

- **Missing Value Handling**:
    - Lab values: Median imputation with missing indicators values (most had <1% missing)
    - Blood pressure: -1 imputation with presence indicators
    - Demographics: Row removal for critical missing values - only 9 missing values, none of them were positive cases (Y=1)
    - Smoking: Default to non-smoker with missing indicators
    - Diagnosis: Transformed into aggregate features capturing the number and recency of diagnoses within 4- and 24-month windows. Missing values were filled with a default of 999 to indicate absence of a diagnosis and were included in engineered recency scores
    - Clinical text: Filled with empty string ""

- **Feature Engineering**:
    - Diagnosis aggregation (count, recency, recency scores)
    - Blood pressure trends and ranges
    - Lab ratios and clinical flags
    - Text features via TF-IDF and SVD

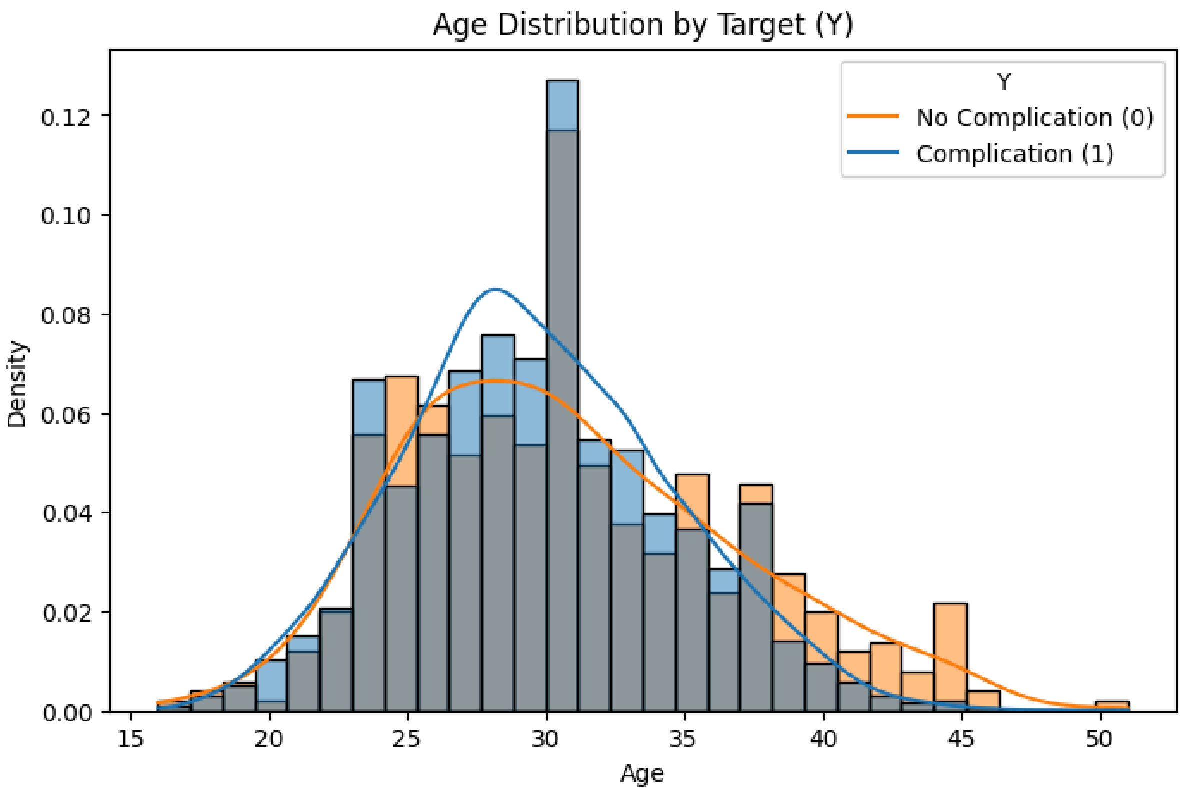
- **Data Leakage Prevention**: Excluded post-week-15 features and outcome-derived variables

2. Data Exploration Results

Target Distribution:

- **Class Imbalance**: 95.68% negative cases (9,568), 4.32% positive cases (432)
- **Complication Types**:
    - Preeclampsia: 128 cases (29.6% of positive cases)
    - Gestational hypertension: 111 cases (25.7%)
    - Essential hypertension: 100 cases (23.1%)
    - Eclampsia: 17 cases (3.9%)

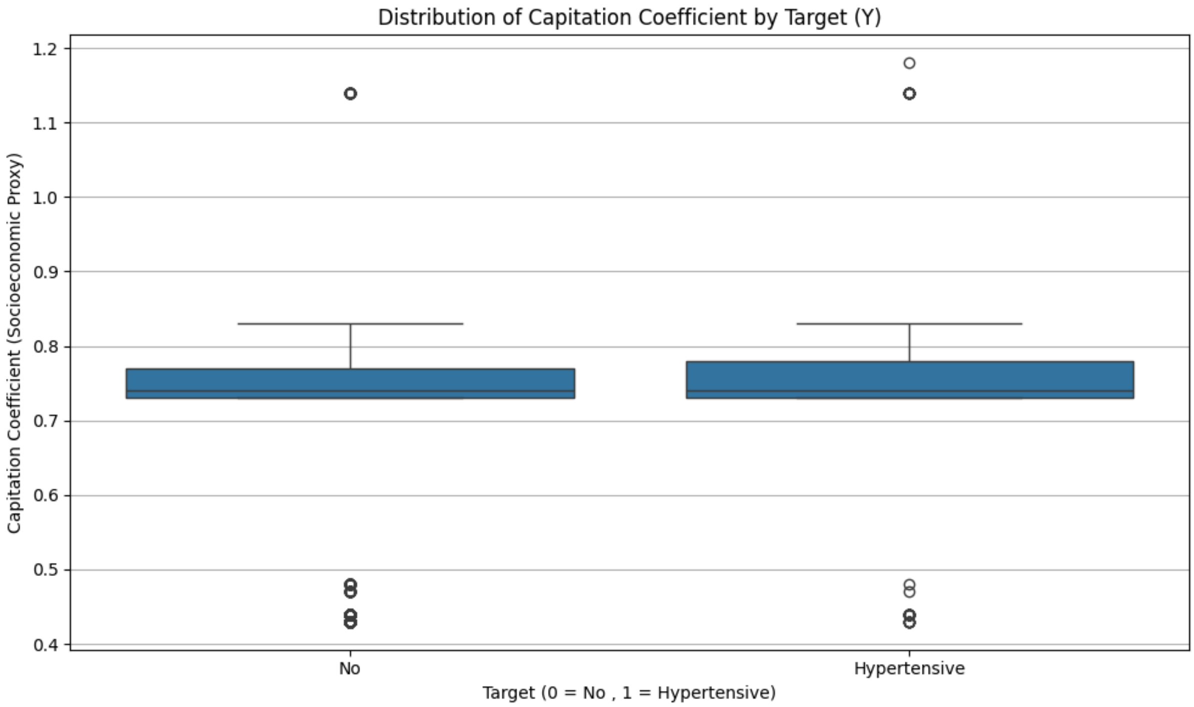
Key Demographic Insights:



Group	Mean Age	Median	Std
Y=0	29.7	29	4.88
Y=1	30.8	30	5.95

T-test confirms the difference is statistically significant: T = -3.80, p = 0.00016

- **Age Effect**: Patients with complications had higher mean age (30.8 vs 29.7 years, p<0.001). Older maternal age is modestly associated with increased risk of hypertensive disorders. While the effect size is small, age remains a useful predictive feature.



Group	Mean	Median	Std
Y=0	0.72	0.74	0.099
Y=1	0.74	0.74	0.117

t-test confirmed that the difference in means is statistically significant: **t = -3.83, p = 0.00015**

- The **mean capitation coefficient** was slightly higher among patients who developed HDP (Y=1: 0.742) compared to those who did not (Y=0: 0.720).

- Both groups had the same median (0.74), but variance was higher in the HDP group.
- t-test confirmed that the difference in means is **statistically significant**:

  t = -3.83, **p = 0.00015**

Conclusion:
Socioeconomic status shows a weak but statistically significant association with hypertensive outcomes. While not a strong standalone predictor, it may contribute useful signal when combined with clinical features.

Source Flag Analysis:

Among positive cases, diagnosis sources were:

- **ICD-9 codes**: 54% (main administrative channel)
- **Aspirin prescription**: 22% (clinical suspicion)
- **Hospital notes**: 21% (severe cases)
- **Medical registry**: 3% (limited documentation)
- **BP measurements**: 0% (not used in current labeling)

Conclusion:

The label (Y) heavily depends on post-week-15 diagnosis codes and clinical documentation. Therefore, early signals such as blood pressure, urine protein, or early lab results are likely underrepresented in Y = 1. This should be taken into account when training a model for early prediction.

3. Feature Engineering

3.1 Diagnosis History Features

- **Aggregation Strategy**: Converted 40+ individual diagnosis columns into 6 engineered features
- **Key Features**:

  `num_recent_diags_4m/24m`: Count of recent diagnoses

  `min_days_since_diag_4m/24m`: Recency of most recent diagnosis

  `recency_score_4m/24m`: Exponential decay of diagnosis recency

3.2 Blood Pressure Engineering

- **Trend Analysis**: First-to-last value differences
- **Range Features**: Max-min differences
- **Clinical Flags**: High BP indicators (>130/85)
- **Interaction Terms**: Age × BP interactions

3.3 Laboratory Features

- **Ratio Features**: Neutrophil-to-lymphocyte ratio
- **Clinical Flags**: Low hemoglobin indicators (<11 g/dL)
- **Missing Indicators**: Binary flags for missing lab values

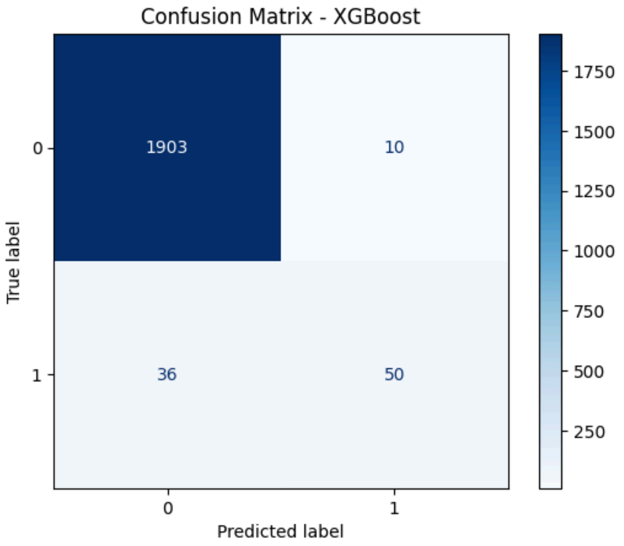
3.4 Text Feature Extraction

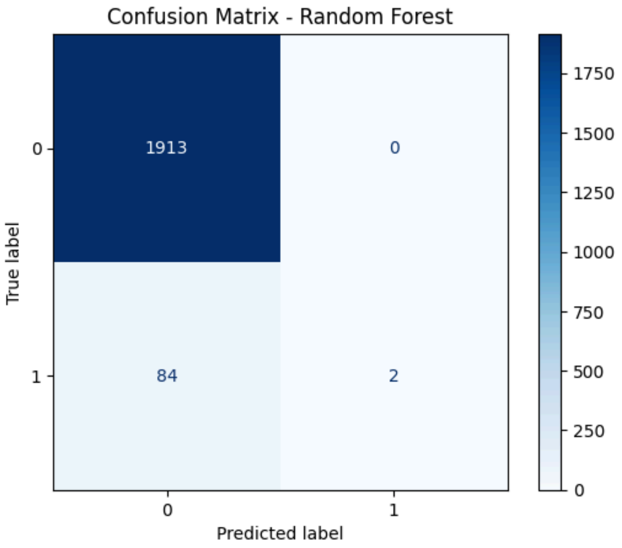
- **Preprocessing**: Removed numeric entities, punctuation, and Hebrew text cleaning
- **Vectorization**: TF-IDF with 500 features
- **Dimensionality Reduction**: SVD to 50 components
- **Result**: 50 text-derived features (txt_svd_0 to txt_svd_49)

4. Modeling Results

Model Comparison:

Metric	Random Forest	XGBoost
AUC	**0.85**	**0.974**
Recall	**2.3%**	**58.1%**
Precision	**100%**	**83.3%**
Accuracy	**95.8%**	**97.7%**


Confusion Matrix - XGBoost


Confusion Matrix - Random Forest

XGBoost Performance Analysis

- **True Positives**: 50 out of 86 cases (58.1% recall)
- **False Positives**: 10 cases (83.3% precision)
- **False Negatives**: 36 cases (missed cases)
- **True Negatives**: 1,903 cases (high specificity)

Both Random Forest and XGBoost were trained and evaluated for predicting hypertensive complications. Random Forest achieved high precision but extremely low recall, indicating it missed the majority of true cases. XGBoost, on the other hand, provided a better balance between recall (58.1%) and precision (83.3%), along with a significantly higher AUC (0.974 vs. 0.85).

XGBoost's advantages include better handling of imbalanced data, superior performance on complex feature interactions, and built-in regularization to reduce overfitting. Therefore, XGBoost was selected as the final model for deployment and further analysis.

Budget-Constrained Evaluation

Top-K Patients	Precision@K	Recall@K	Efficiency Gain
50	88%	51.2%	20.5x
100	61%	70.9%	14.2x
200	38.5%	89.5%	9.0x
300	26.3%	91.9%	6.1x
500	16.8%	97.7%	3.9x

Efficiency Gain = Precision@K / Baseline Positive Rate (4.3%)

Observations:

- Screening the top 100–200 patients captures the majority of true positive cases while limiting the number of costly lab tests.
- This ranking approach is suitable for screening prioritization when resources are limited and early detection is critical.

5. Feature Importance Analysis

Top Predictive Features:

1. **Blood Pressure Trends**: Systolic/diastolic means, trends, and ranges
2. **Diagnosis Recency**: Recent diagnosis history (4-24 months)
3. **Laboratory Ratios**: Neutrophil-to-lymphocyte ratio
4. **Clinical Flags**: High BP indicators and lab abnormalities
5. **Demographics**: Age and socioeconomic status
6. **Text Features**: Clinical note-derived semantic features

Clinical Interpretation

- **Early Warning Signs**: Blood pressure trends provide early signals before clinical thresholds
- **Risk Accumulation**: Multiple recent diagnoses indicate elevated risk
- **Inflammatory Markers**: Lab ratios suggest underlying inflammatory processes
- **Socioeconomic Factors**: Modest but significant contribution to risk assessment

6. Recommendations

Model Improvements & Next Steps:

- **Feature Importance Analysis**: Use gain-based importance to focus on high-impact predictors and simplify the model.
- **Hyperparameter Tuning**: Apply grid search or Bayesian optimization to fine-tune learning rate, tree depth, and regularization.
- **Enhanced Feature Engineering**: Incorporate additional interaction terms, lab trend features, and richer text embeddings.
- **Threshold Calibration**: Use precision-recall curves to optimize decision thresholds under different clinical budget constraints.
- **Data Augmentation**: Expand the dataset with more positive (Y=1) cases to improve balance and model learning.
- **Diagnosis Source Analysis**: Use match_*_after fields to evaluate how the model performs across different diagnosis types and explore multi-label classification.

7. Limitations and Considerations

Data Limitations:

- **Class Imbalance**: 4.32% positive rate may limit model sensitivity
- **Missing Data**: High missingness in some features reduces signal
- **Label Quality**: Dependence on administrative coding for outcomes

Clinical Considerations:

- **False Negatives**: 36 missed cases require careful clinical oversight
- **False Positives**: 10 false alarms may increase patient anxiety

8. Conclusion

The analysis demonstrates that early prediction of hypertensive disorders is feasible using multimodal clinical data. XGBoost outperformed Random Forest by achieving strong recall and precision, making it suitable for prioritizing high-risk patients. Key predictors include blood pressure patterns, diagnosis history, and lab markers. The model supports more targeted and cost-effective screening strategies, with clear paths for future optimization and deployment.