Data Handling

* Missing Data
  + Forward fill vitals (max 7 days): If BP missing on 12th Jan, use BP from 11th Jan.
  + Median impute labs: If HbA1c missing, fill with median HbA1c for males age 50–60 with diabetes.
  + Add missing\_flags: Create “cholesterol\_missing = 1” when cholesterol not measured.
  + Mark irregular gaps: If last HR recorded 20 days ago, add feature “irregular\_gap = 1”.
* Time Windows
  + Acute (7–30 days): Calculate std of BP in past 14 days.
  + Subacute (30–90 days): Compute HbA1c slope over past 60 days.
  + Chronic (90–180 days): Count number of abnormal kidney tests in past 120 days.
  + Rolling stats: Mean HR over last 30 days, time since last HbA1c test.

Feature Engineering

* Biomarkers: BP slope over 90 days (e.g., rising from 120→150 systolic).
* Medication Adherence: PDC = 80% if patient covered 72 out of 90 days.
* Clinical Scores: BP control = 65% days systolic BP < 140 mmHg.
* Temporal: Patient skipped meds on weekends → “weekend\_gap = 1”.

Feature Selection

* Remove non-actionable: Drop hospital ID column.
* Drop correlated: If weight and BMI correlation = 0.98, keep only one.
* Mutual information: Keep cholesterol if it has high MI with heart disease risk.
* SHAP pre-screen: Select top 50 features from LightGBM importance.

Data Validation

* Range checks: Flag HR = 400 bpm as implausible.
* Temporal logic: Drop a glucose test dated “2026-01-01” (future date).
* Cross-feature validation: Height = 170 cm, weight = 300 kg → BMI unrealistic.
* Outliers: Keep BP = 220/130 but flag as “outlier = 1”.

Scaling and Encoding

* Robust scaling: HbA1c values scaled using median and IQR.
* Target encoding: Replace medication “Metformin” with average risk score of patients on it.
* Cyclic encoding: Encode “Sunday = (0,1), Monday = (0.87,0.5)” on a circle.
* Standardize: Convert creatinine to mg/dL before model input.

Modeling Strategy

* RuleFit: “If BP > 140 and HbA1c > 7.0 → high risk” rule generated.
* LightGBM: Learns that HbA1c trend + medication gap best predict diabetes worsening.
* XGBoost: Handles missing cholesterol directly without imputation.

Use a single unified ensemble with age (e.g., 65) and disease type (e.g., diabetes) as features.

Advanced Models

* Temporal Fusion Transformer: Daily glucose over 180 days predicts hypoglycemia risk.
* Bayesian Classifier: If 40% lab data missing, predicts with uncertainty intervals.
* SHAP: Shows HbA1c slope contributed +0.15 risk for patient X.
* LIME: Explains why patient Y with BP=160, age=70 got 85% risk score.

Dashboards

Global View

* Risk tier grid: 200 patients in “High Risk” bucket.
* Condition heatmap: CKD patients show higher risk than diabetes-only patients.
* Care gap bars: 30% of patients missing annual HbA1c test.
* Risk trend lines: Heart failure risk rising steadily past 6 months.

Patient View

* Risk gauge: Patient A risk = 78% (High).
* Biomarker timeline: HbA1c rose from 6.5 → 8.2 in 6 months.
* Factor waterfall: “HbA1c trend (+0.25), missed meds (+0.18), controlled BP (-0.05)”.
* Intervention impact: Simulated adherence improves risk from 78% → 60%.