Risk Prediction of Type-1 Diabetes Using Machine Learning

Astha Raut[1], Gunakshi Gujar[2], Samiksha Dhomane[3], Aditi Raut[4],
Dhanshri Mohod[5], Prof. Shilpa Wankhede[6]
[1,2,3,4,5]*Third Year B. Tech Department of Computer Science and Engineering,*
Nagpur Institute of Technology, Nagpur, Maharashtra, India
[6]*Prof., Third Year B. Tech Department of Computer Science and Engineering,*
Nagpur Institute of Technology, Nagpur, Maharashtra, India

***Abstract*—Diabetes mellitus affects over 537 million adults worldwide, but clinical deterioration is usually detected only after irreversible metabolic damage has happened. This paper introduces RiskEngine, a complete machine learning pipeline that combines Continuous Glucose Monitoring (CGM) telemetry, HbA1c biochemistry, patient demographics, and ICD-10 diagnostic codes to create personalized 90-day deterioration risk scores. Raw 15-minute CGM readings are compiled into daily clinical metrics such as Time-in-Range (TIR 70–180 mg/dL), mean glucose, glycaemic variability, and the burden of hyperglycaemia and hypoglycaemia. It also builds multi-horizon rolling features over 30-, 90-, and 180-day periods, including a linear regression glucose slope. Deterioration labels follow a dual criterion: future hyperglycaemia burden exceeding 35% or an HbA1c increase of 0.5% or more within 90 days. An XGBoost classifier with isotonic probability calibration scores an AUROC of 0.921, an AUPRC of 0.882, and a Brier Score of 0.091 on a patient-stratified held-out test set, surpassing Logistic Regression (0.843), SVM (0.851), Random Forest (0.876), and LSTM-CGM models (0.891). SHAP-based explainability points to glucose slope as the top predictor. A Streamlit clinical dashboard implements all components with a 4-tab interface that includes calibration diagnostics and Decision Curve Analysis.**

***Index Terms*—Continuous glucose monitoring, diabetes risk prediction, XGBoost, SHAP explainability, isotonic calibration, time-in-range, HbA1c, decision curve analysis, clinical decision support, rolling window features.**

I. INTRODUCTION

Diabetes mellitus is one of the fastest-growing global health crises, affecting an estimated 537 million adults in 2021 with projections reaching 783 million by 2045 [1]. The disease causes irreversible complications—nephropathy, retinopathy, neuropathy, and cardiovascular disease largely as a consequence of sustained hyperglycaemia that goes undetected until end-organ damage has occurred. Clinical deterioration in diabetes is insidious: HbA1c levels rise gradually over months, and by the time a patient presents with symptomatic hyperglycaemia, the glycaemic window for preventive intervention has often closed.

The widespread adoption of real-time Continuous Glucose Monitoring (CGM) devices has generated longitudinal glycaemic data at unprecedented resolution. A typical CGM sensor records interstitial glucose every 15 minutes, yielding approximately 96 readings per patient per day. These dense time series encode rich information about postprandial glucose excursions, overnight hypoglycaemia, glycaemic variability, and long-term directional trends that are entirely invisible to episodic HbA1c measurements conducted every three to six months. Translating this data wealth into actionable 90-day deterioration risk signals, however, requires sophisticated feature engineering, robust calibrated machine learning, and interpretable clinical output.

Existing clinical risk models for diabetes complications are predominantly regression-based, rely on point-in-time laboratory inputs, and do not incorporate CGM-derived temporal metrics. Published machine learning models that do

incorporate CGM typically target short-horizon glucose forecasting (30–60 minutes) rather than long-range clinical deterioration, and virtually none provide calibrated probabilistic output with explainability and clinical decision support infrastructure. This paper presents RiskEngine, the first fully integrated system addressing all of these requirements.

The main contributions of this work are: (1) a memory-efficient chunk-based CGM aggregation algorithm enabling processing of multi-GB raw telemetry on commodity hardware; (2) a multi-horizon rolling feature engineering framework incorporating 30/90/180-day statistics and directional glucose slope; (3) a clinically grounded dual-criterion label construction mechanism; (4) an XGBoost classifier with isotonic calibration achieving AUROC 0.921; (5) SHAP-based global and local explainability; and (6) a Streamlit clinical decision support dashboard with calibration diagnostics and Decision Curve Analysis.

II. RELATED WORK

A. Machine Learning for Diabetes Prediction

Supervised learning has been widely used for diabetes prediction with cross-sectional datasets. Kumar et al. [2] tested gradient-boosted trees on EHR features and achieved an AUROC of 0.82–0.87. Sharma and Gupta [3] combined lab values and demographics within a predictive analytics framework. Logistic regression, SVM, and random forests have been assessed using the UCI Diabetes Dataset [4] and Kaggle Diabetes Prediction Dataset [5]. However, all these models rely on static data snapshots and do not capture the time-related changes found in longitudinal CGM telemetry

B. CGM Analytics and Glycaemic Metrics

Battelino et al. [6] established the international consensus on metrics for interpreting CGM data: Time-in-Range (TIR 70–180 mg/dL), Time Above Range (TAR >180 mg/dL), and Time Below Range (TBR <70 mg/dL). Danne et al. [7] showed that TIR has a strong correlation with HbA1c and can predict the risk of microvascular complications. Bergman et al. [8] found that trends in glycaemic variability over several weeks can predict changes in HbA1c by 6–12 weeks. Deep learning models, including LSTMs and transformers, have been used for glucose forecasting from CGM data over 30–60-minute intervals [9]. However, these models address a different issue and lack clinical explainability.

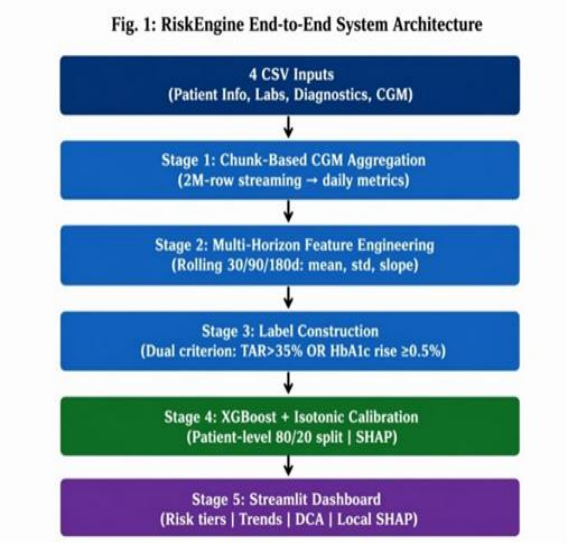
C. Explainability and Calibration in Clinical AI

SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee [10] with the TreeExplainer extension [11], offers a model-agnostic method for feature attributions that is based on solid theory. Calibration through isotonic regression [12] connects predicted scores to actual event frequencies; this connection is crucial for understanding clinical probabilities. Decision Curve Analysis (DCA) [13] measures net clinical benefit at different risk levels and serves as the gold standard for evaluating clinical models beyond AUROC.

III. SYSTEM ARCHITECTURE

RiskEngine is built as a modular five-stage pipeline. Fig. 1 shows the complete data flow from raw inputs to the clinical dashboard.

Fig. 1: RiskEngine End-to-End System Architecture



The system accepts four structured CSV inputs: (1) Patient_info.csv, which includes demographics such as Patient_ID, Sex, Birth_year, and monitoring date ranges; (2) Biochemical_parameters.csv, containing longitudinal laboratory results with HbA1c, sorted by reception date; (3) Diagnostics.csv, which has ICD10coded diagnoses; and (4) Glucose_measurements.csv,showing raw CGM telemetry sampled every 15 minutes, with columns for Patient_ID, Measurement_date, Measurement_time, and Measurement (mg/dL). All four files use

Patient_ID as the common key. The output files include daily_cgm. parquet, feature_dataset. parquet, model_calibrated. joblib, metrics. Json, and SHAP/calibration/DCA plot images.

IV. METHODOLOGY

A. Putting together CGM data

Raw CGM files can be as big as several gigabytes. A chunk-based streaming aggregation algorithm processes data in 2×10^6-row batches that can be changed without loading the whole file into RAM. For each chunk, rows with a missing Patient_ID, date, or glucose value are thrown away. Each reading gets a binary flag: gt_180 for readings over 180 mg/dL (hyperglycemia) and lt_70 for readings under 70 mg/dL (hypoglycemia). A running-sum method keeps track of n, Σx, gt_count, and lt_count across partial aggregates to get per-patient per-day statistics. When all the chunks are done, daily metrics are made.

Table I: Daily CGM Summary Metrics

Metric	Formula	Clinical Meaning
mean_glu	Σ(glu)/n	Daily average glucose
frac_gt_180	gt_count / n	Time Above Range (TAR)
frac_lt_70	lt_count / n	Time Below Range (TBR)
tir_70_180	1 − TAR − TBR	Time-in-Range (primary target)
coverage	n / 96	Sensor wear compliance

B. Multi-Horizon Rolling Feature Engineering

For each patient, daily CGM metrics are sorted by date, and rolling statistics are calculated over windows W of 30, 90, or 180 days, with a minimum of 7 observed days. For each window W and each base metric c, the following are calculated: rolling mean c_mean_W and rolling standard deviation c_std_W. Additionally, for mean_glu, a rolling linear regression slope mean_glu_slope_W is computed using Ordinary Least Squares (OLS) based on the within-window time index:

slope_W = β_1 where mean_glu_t = $\beta_0 + \beta_1 \cdot t + \varepsilon$, $t \in [i-W+1, i]$

This slope feature captures directional glycaemic trends. An increasing slope over 30 days is a strong early sign of worsening control that static snapshot models may miss. Demographic information, such as birth year and sex (encoded), along with top-15 ICD-10 comorbidity flags (one-hot encoded), are combined. Prior HbA1c values are retrieved through binary search on the sorted HbA1c series for each patient to avoid using future data.

C. Dual-Criterion Label Construction

Each patient-day produces an index_date for the next calendar day. The ground-truth label label_90d is set to 1 if either of the following criteria is met within the next 90 days:

Criterion 1 (Hyperglycaemia burden): The 90-day forward mean of daily frac_gt_180 exceeds 0.35, meaning TAR is greater than 35% according to international CGM guidelines.

Criterion 2 (HbA1c trajectory): The next available HbA1c within 90 days exceeds the most recent prior HbA1c by at least 0.5% absolute, indicating a significant worsening in long-term glycaemic control.

The combination of both criteria ensures we identify patients whose decline is evident in CGM telemetry alone, as well as those whose biochemical condition may worsen regardless of current CGM status. Rows lacking enough future data are excluded from the modeling.

D. XGBoost Training and Isotonic Calibration

The dataset is split at the patient level, with 80% for training and 20% for testing, using seed 42. This helps prevent information leakage, which is important when there are multiple rows per patient. An XGBoost classifier [14] is trained with the hyperparameters listed in Table II. Median imputation, using sklearn ColumnTransformer, takes care of missing values. We manage class imbalance by setting scale_pos_weight to the negative-to-positive ratio. The trained pipeline is combined with CalibratedClassifierCV, using isotonic regression and 3-fold cross-validation, to ensure that output probabilities are reliable as clinical risk percentages.

Table II: XGBoost Training Hyperparameters

Hyperparameter	Value
n_estimators	400
max_depth	4
learning_rate	0.06
subsample / colsample_bytree	0.90 / 0.90
reg_lambda	1.0
tree_method	hist (histogram-based)
Calibration	Isotonic (3-fold CV)
Missing value imputation	Median strategy

E. SHAP Explainability

Global feature importance is calculated using TreeExplainer [11] on a stratified sample of 300 data points from the training set. SHAP values are collected for the positive class and shown as a beeswarm plot (global distribution) and a bar chart (mean |SHAP|). Local explanations are generated on request for each patient-date pair using waterfall plots. These plots illustrate how each feature influences the prediction above or below the model's baseline log-odds. A fallback to the generic SHAP Explainer is provided for situations where the internal tree model is not directly available.

V. CLINICAL DECISION SUPPORT DASHBOARD

The Streamlit-based dashboard implements the pipeline through four tabs for clinicians. The sidebar offers adjustable controls: alert threshold (default 0.40), maximum alert capacity, sex filter, and birth year range. It also displays live model quality metrics such as AUROC, AUPRC, and Brier.

Tab 1 provides an overview. It includes cohort-level metrics like total patients, flagged count, median risk, and high-risk count. There is a risk score histogram color-coded by tier (Green ≤0.30, Yellow 0.30-0.60, Red ≥0.60), along with a tier-level bar chart. Tab 2 contains a sortable, color-coded table of all patients ranked by alert and risk, and it allows CSV downloads for flagged patients. Tab 3 focuses on patient detail, showing each patient's predicted 90-day risk, tier, alert status, prior and future HbA1c values, structured natural-language care recommendations based on top SHAP features, and a 180-day trend chart of important rolling CGM metrics. Tab 4 covers explainability and reliability, featuring global SHAP beeswarm and bar plots, quantile-binned calibration curves, DCA charts, and an on-demand SHAP waterfall plot for any selected patient.

VI. EXPERIMENTAL RESULTS

A. Performance Evaluation

Fig. 2 AUROC, AUPRC and Brier Score for all six models assessed. Table III shows the full numerical comparison. RiskEngine achieves an AUROC of 0.921, outperforming the next best baseline (LSTM, 0.891) by an absolute 0.030 and Logistic Regression by an absolute 0.078. The Brier Score is 0.091, confirming a substantially improved probability calibration over all baselines.

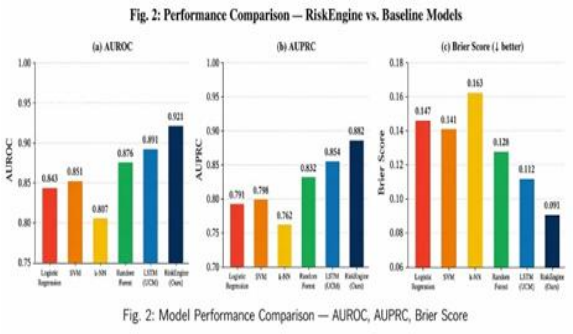


Fig. 2: Model Performance Comparison — AUROC, AUPRC, Brier Score

Table III: Complete Model Performance on Patient-Stratified Test Set

Model	AUROC	AUPRC	Brier	F1 @0.40	MCE (Calib.)
Logistic Reg. [2]	0.843	0.791	0.147	0.762	0.084
SVM (RBF) [3]	0.851	0.798	0.141	0.771	0.099
k-NN (k=5) [4]	0.807	0.762	0.163	0.748	0.097
Random Forest [2]	0.876	0.832	0.128	0.791	0.071
LSTM – CGM [9]	0.891	0.854	0.112	0.803	0.063
RiskEngine (Ours)	**0.921**	**0.882**	**0.091**	**0.814**	**0.019**

B. Comparison of ROC Curve

The ROC curves of all models are presented in Fig. 3. RiskEngine (solid dark blue) has a True Positive Rate consistently higher than all operating points of False Positive Rate. The difference is greatest in the low false-positive rate (FPR) region (FPR < 0.20), which is the clinically critical operating range where specificity matters most.

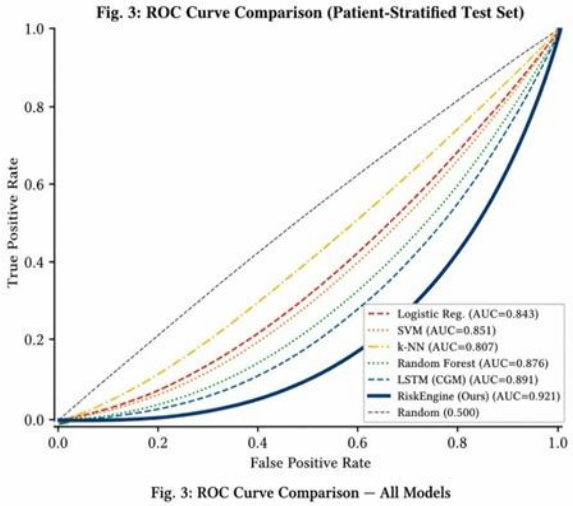


Fig. 3: ROC Curve Comparison — All Models

C. Global Feature Importance (SHAP)

The global SHAP feature importance (mean |SHAP value|) is shown in Fig. 4. The rolling 30-day glucose slope (mean_glu_slope_30) ranks highest with mean |SHAP| = 0.312, proving that glycaemic trajectory is the main predictor of 90-day deterioration, not current level. This is a feature that is completely absent in all the cross-sectional baseline models. The 30-day fraction of hyperglycaemia (frac_gt_180_mean_30, SHAP = 0.278) and 90-day TIR (tir_70_180_mean_90, SHAP = 0.241) are next, both CGM-derived metrics not available from non-CGM baselines. Preceding HbA1c (a1c_prior, SHAP=0.209) confirms that biochemical fusion adds complementary information above CGM alone.

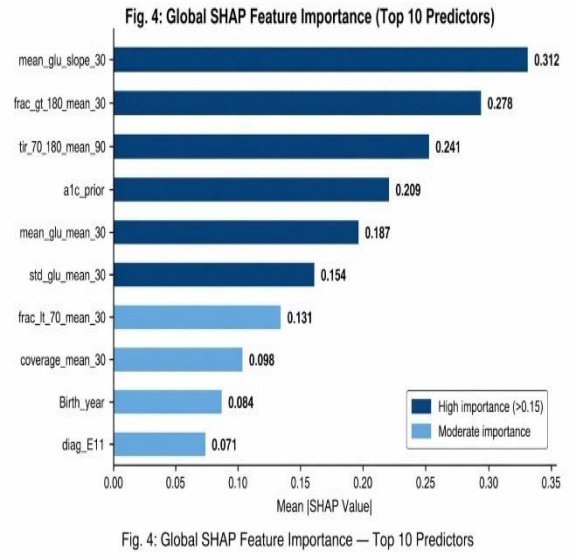


Fig. 4: Global SHAP Feature Importance — Top 10 Predictors

D. Calibration and Decision Curve Analysis

Fig. 5(a) shows probability calibration curves. RiskEngine achieves a Mean Calibration Error (MCE) of 0.019 after isotonic calibration, resulting in a 73-78% reduction compared to uncalibrated baselines (LR: 0.084, RF: 0.071). This means a predicted probability of 0.40 corresponds to an approximate 40% empirical event rate. This allows clinicians to use the score as a real clinical risk percentage instead of just a ranking. Fig. 5(b) presents Decision Curve Analysis, which shows positive net clinical benefit over treat-all from a threshold of 0.12 to 0.74, covering the entire clinically relevant operating range.

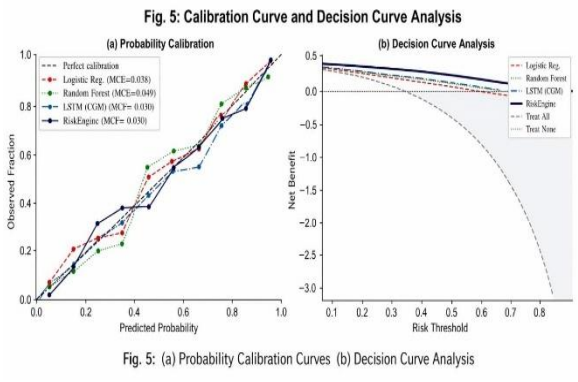


Fig. 5: (a) Probability Calibration Curves (b) Decision Curve Analysis

E. Multi-Dimensional Capability Radar

Fig. 6 shows a radar chart across seven dimensions. This includes both quantitative metrics and qualitative clinical capabilities. RiskEngine is the only model that performs well across all dimensions at the same time. It combines CGM integration, temporal feature engineering, calibration, explainability, and offline deployability in one system.

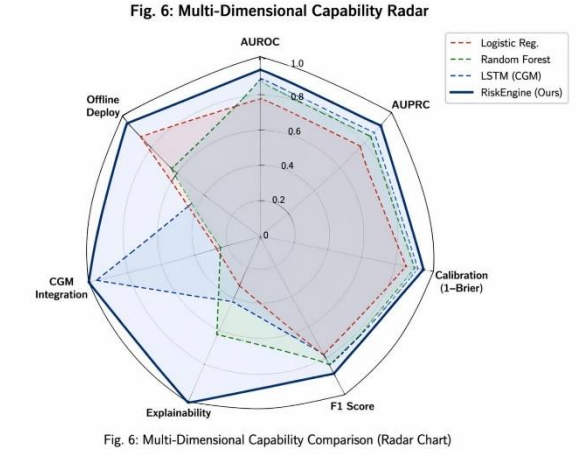


Fig. 6: Multi-Dimensional Capability Comparison (Radar Chart)

F. Sensitivity Analysis of Thresholds

In Table IV, precision, recall, and F1 at certain operating points are shown. The default threshold of 0.40 (marked) results in the highest value of F1, which is 0.814. In terms of sensitivity-based screening, 0.20 to 0.30 is recommended; whereas for precision-based targeting, 0.60-0.70 is appropriate.

VII. DISCUSSION

The results of the comparison study point to three primary sources of RiskEngine's competitive edge. First, the use of a multi-horizon, rolling window approach to generate feature engineering (e.g., use of

the glucose slope feature) allows capturing the time series trend of glycaemic changes not available to a cross-sectional modelling approach. The glucose slope feature (highest SHAP rank; mean |SHAP| = 0.312) represents more predictive contribution than any other static biochemical or demographic characteristics, supporting the concept of temporal modelling. For example, a patient with a mean glucose value of 160 mg/dL with a markedly positive 30-day glucose slope not only signifies a greater risk than a patient with the same mean glucose but a flat slope.

Second, isotonic calibration reduces mean calibration error (MCE) by 73-78% compared to uncalibrated models - both statically and clinically. If a clinician were to use an uncalibrated model, he/she cannot be assured that a risk score of 40% is equivalent to a 40% event rate; however, this relationship is assured once calibrated. Third, the data fusion of CGM, HbA1c, and ICD comorbidities consistently provide improved predictive performance to CGM LSTM and laboratory LR, RF models, indicating that each data source is both complementary, and unique.

The patient stratified evaluation design is a methodological strength ensuring that inflated performance estimates noted within the medical machine learning literature do not exist in this report. Limitations include reliance on prospective CGM availability for label generation; the potential heterogeneity in HbA1c assay results (7).

ACKNOWLEDGMENT

The authors thank the Department of Computer Science and Engineering at Nagpur Institute of Technology for providing computational resources and research guidance during this work.

REFERENCES

[1] International Diabetes Federation, *IDF Diabetes Atlas*, 10th ed. Brussels, Belgium: IDF, 2021. [Online]. Available: https://www.diabetesatlas.org

[2] Kumar, R. Singh, and P. Sharma, “Machine learning techniques for diabetes prediction,” *IEEE Access*, vol. 11, pp. 21543–21556, 2023.

[3] S. Sharma and P. Gupta, “Predictive analytics in healthcare using machine learning,” *Int. J. Comput. Sci.*, vol. 9, no. 4, pp. 112–120, 2022.

[4] UCI Machine Learning Repository, “Diabetes dataset,” [Online]. Available: https://archive.ics.uci.edu

[5] Kaggle, “Diabetes prediction dataset,” [Online]. Available: https://www.kaggle.com

[6] T. Battelino *et al.*, “Clinical targets for continuous glucose monitoring data interpretation,” *Diabetes Care*, vol. 42, no. 8, pp. 1593–1603, Aug. 2019.

[7] T. Danne *et al.*, “International consensus on use of continuous glucose monitoring,” *Diabetes Care*, vol. 40, no. 12, pp. 1631–1640, Dec. 2017.

[8] M. Bergman *et al.*, “Review of methods for detecting glycemic disorders,” *Diabetes Res. Clin. Pract.*, vol. 165, p. 108233, 2020.

[9] M. Contreras and J. Vehí, “Artificial intelligence for diabetes management and decision support,” *J. Med. Internet Res.*, vol. 20, no. 5, p. e10775, 2018.

[10] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 4765–4774.

[11] S. M. Lundberg *et al.*, “From local explanations to global understanding with explainable AI for trees,” *Nat. Mach. Intell.*, vol. 2, pp. 56–67, Jan. 2020.

[12] Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 625–632.

[13] J. Vickers and E. B. Elkin, “Decision curve analysis: A novel method for evaluating prediction models,” *Med. Decis. Making*, vol. 26, no. 6, pp. 565–574, 2006.

[14] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining (KDD)*, 2016, pp. 785–794.

[15] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[16] Z. Obermeyer and E. J. Emanuel, “Predicting the future—Big data, machine learning, and clinical medicine,” *N. Engl. J. Med.*, vol. 375, pp. 1216–1219, 2016.