

Machine Learning for Early Detection of Chronic Diseases Using Clinical Data

¹Dr. Arfa Kareem

²Eman Siddique

How to cite this article:

Kareem, A., & Siddique, E. (2023). Machine Learning for Early Detection of Chronic Diseases Using Clinical Data
Pakistan journal of multidisciplinary, 1(1), 69-83 .

Article Details	Abstract
Received: 16 January 2023 Accepted: 18 April 2023 Published: 24 June 2023 ¹Dr. Arfa Kareem ²Eman Siddique Copyright @Author Corresponding Author: *Dr. Arfa Kareem	Chronic diseases like type 2 diabetes (T2D), chronic kidney disease (CKD) and cardiovascular disease (CVD) are insidious, insidious diseases that develop over years, and if they are diagnosed early enough, lifestyle and/or drug intervention can prevent irreversible damage. This paper presents and tests machine learning models based on routinely collected information from the electronic health record (EHR) to predict incident T2D, CKD, and CVD within 3 years. A retrospective cohort of 48,320 adults was divided into training (70%), validation (10%) and test (20%) sets, with 62 clinical features being extracted from the 2 years before. The logistic regression, random forest, gradient boosting (XGBoost), multilayer perceptron and stacked ensemble were compared to a rule-based baseline. The best test AUROC is 0.884 for T2D, 0.903 for CKD, and 0.823 for CVD for gradient boosting, and 0.004 to 0.005 on top of that for the stacked ensemble. The gradient-boosting model identified 65.8% of cases with T2D, 71.1% with CKD, and 51.2% with CVD with 90% specificity, and retained the same level of sensitivity when the prediction time interval is increased from 6 to 36 months. Using isotonic calibration reduced the expected calibration error from 0.031 to 0.007. SHAP analysis revealed three key features that drove the risk of T2D, namely HbA1c, fasting glucose and BMI, which, when removed, resulted in a loss of 0.093 AUROC. There was a lower level of performance at community clinics (0.861) than at urban clinics (0.889): this demonstrates the need to monitor subgroups. A notion of a deployment workflow with drift monitoring is proposed. These findings indicate that routine data may be used for early warning screening but further prospective research is needed before clinical implementation.

Keywords : machine learning, early detection, chronic disease, electronic health records, gradient boosting, explainability, and calibration.

¹Assistant Professor, Department of Business Administration, Islamia College and University , Peshawar.

²MPhil Scholar, Department of Business Administration, Islamia College and University, Peshawar.



INTRODUCTON:

Non-communicable diseases are the biggest cause of mortality and morbidity across the globe. Risk factors for type 2 diabetes, chronic kidney disease and cardiovascular disease overlap, as do these diseases. All three have an extended latent period that is symptom-free, but detectable with measurable risk factors. Early detection is one of the most valuable uses of clinical prediction, as interventions given during that phase, such as those for weight management, blood pressure control and early treatment with drugs, can delay or prevent onset.

Data is already available within health systems. The EHR includes vital signs, lab values, diagnoses and prescriptions which are added to the EHR with each outpatient visit. These are all potential variables and traditional risk scores rely on just a few of these variables in a formula they can be easily applied and don't account for non-linear effects, interactions, or many of the weaker signals that can show up in a record. Machine learning techniques are able to take dozens or hundreds of variables, learn interactions from the data and continue to learn as additional data become available [1]. Tree ensembles are in particular powerful and robust methods for tabular clinical data [2, 3].

Many published models have not gone into practice for three reasons. First, the time period is reported in random divisions of the same time period, thus overstating the abilities of the model on patients who have not yet been seen. Second, there is the possibility that a good discriminating model can yield poorly calibrated probabilities and thus make thresholds and clinical decisions difficult to make [7]. Third, the accuracy of an algorithm across a population may be misleading when the performance of the algorithm is poorer in certain sub-populations, and biased algorithms may exacerbate health inequalities [6]. Clinicians also require explanations of the models they are using, with which they can compare their own judgement [4].

This study aims to tackle these questions in a single framework. The pipeline in Figure 1 consists of five stages—one for transforming raw EHR data to validated predictions of risk and one for calibrating the pipeline, and a third for evaluating performance of the pipeline across subgroups of patients. The contributions are:

1. An assessment of five model families with a rule-based baseline for three chronic diseases based on the same set of features, and validated at a specific time.

¹Assistant Professor, Department of Business Administration, Islamia College and University, Peshawar.

²MPhil Scholar, Department of Business Administration, Islamia College and University, Peshawar.



2. An analysis of detection performance for various prediction horizon (from 6 to 36 months).
3. Along with discrimination, an evaluation of calibration, feature contribution and subgroup performance.
4. A deployment process that incorporates thresholding, clinician review and drift monitoring.

Figure 1. Study pipeline: from extraction of eHRs to the validation of a risk prediction, including a calibration loop and pre-release checks of the performance of the validated prediction in sub-groups.

2. Background and Related Approaches

The development of model families is summarized in Table 1, and has been used in prediction from clinical data. Logistic is still the benchmark as it is easy to understand and interpret, but it assumes additive and log-linear effects. Random forests [3] average many decorrelated trees, have little need to be tuned and work well with mixed type of variables and with indication of missing values. Gradient boosting is a technique that constructs trees in turn to fix errors made in previous trees; some implementations like XGBoost [2] can perform best on structured data. Neural networks may model complex interactions without needing as much data or care as a tree ensemble for the tabular inputs. There is a possibility to stack by combining a number of models using a meta-learner and it may give a marginal advantage with a marginal increase in complexity.

Table 1. Model families commonly used for tabular clinical prediction.

Model family	Strength	Limitation
Logistic regression	Transparent, stable, fast	Misses non-linear effects and interactions
Random forest	Robust, little tuning, handles mixed data	Larger models, probabilities often poorly calibrated
Gradient boosting	Highest accuracy on many tabular tasks	Needs tuning, can overfit small data
Multilayer perceptron	Flexible, can model complex interactions	Data hungry, sensitive to scaling and settings

¹Assistant Professor, Department of Business Administration, Islamia College and University, Peshawar.

²MPhil Scholar, Department of Business Administration, Islamia College and University, Peshawar.



Model family	Strength	Limitation
Stacked ensemble	Combines model strengths, small accuracy gain	Complex to maintain and explain

Evaluation practice has also matured. Discrimination, measured by the area under the receiver operating characteristic curve (AUROC), should be reported with calibration, and the precision–recall curve is more informative than ROC when the outcome is rare [9, 10]. Reporting standards such as TRIPOD+AI [5] ask authors to describe the data source, cohort selection, handling of missing data, validation approach, and fairness analysis. The present study follows these recommendations.

3. Materials and Methods

3.1 Data source and cohort

The EHR of a regional health care network including urban hospital clinics and community clinics were used to gather the data from 2012 to 2021. The structure of the cohort is explained in figure 2. There were 112,406 adults in the extraction (with a minimum of two years of records). Patients with a baseline diagnosis of T2D, CKD or CVD ($n = 21,870$) were excluded since the task is to detect new disease. 30,945 additional patients were excluded due to fewer than 2 encounters, and 11,271 additional patients with no events and fewer than 3 years of follow-up were excluded to prevent ambiguous labels. The last cohort had a total of 48,320 adults.

The index date was the first visit following the 24 months' observation period. The features included the 24 months prior to the index date and outcomes were assessed over the next 36 months. Patients were assigned by index date to training (2012–2018, $n = 33,824$), validation (2019, $n = 4,832$), and test (2020–2021, $n = 9,664$) sets, so that the test set reflects the future relative to training.

Figure 2. The number of patients in each cohort and the reason for each exclusion, following the number in and reasons for splitting the data.

3.2 Outcomes and features

Within 36 months of the index date three binary outcomes were classified. Incident T2D was defined as two HbA1c levels $\geq 6.5\%$ or two fasting glucose readings ≥ 126 mg/dL or a diagnosis code and an antidiabetic prescription [12]. Estimated GFR < 60 mL/min/1.73 m² on two measurements or more spaced apart by at least 90 days was

¹Assistant Professor, Department of Business Administration, Islamia College and University, Peshawar.

²MPhil Scholar, Department of Business Administration, Islamia College and University, Peshawar.



defined as incident CKD. [11] Incident CVD was the first event of CHD, stroke, or hospitalisation for heart failure.

The look-back window has been used to build 62 features (Table 2). Last Value, Mean and Slope of laboratory features, together with the one-hot encoding of categorical variables, were used to summarize the data. Allowing outcome-defining measurements from the look-back window was done because patients with prevalent disease were excluded.

Table 2. Feature groups used for prediction.

Feature group	Number of features	Examples
Demographics and lifestyle	14	Age, sex, smoking status, alcohol use, physical activity, family history
Vital signs and anthropometrics	6	Body mass index, weight, systolic and diastolic blood pressure, heart rate, waist circumference
Laboratory results	24	HbA1c, fasting glucose, creatinine, eGFR, urine albumin-to-creatinine ratio, lipids, liver enzymes, haemoglobin
Comorbidity history	10	Hypertension, dyslipidaemia, obesity, prior gestational diabetes, polycystic ovary syndrome
Medications	8	Antihypertensives, statins, corticosteroids, NSAIDs
Total	62	

3.3 Preprocessing

¹Assistant Professor, Department of Business Administration, Islamia College and University, Peshawar.

²MPhil Scholar, Department of Business Administration, Islamia College and University, Peshawar.



Continuous variables were clipped at the 1st and 99th percentiles and for logistic regression and the neural network standardised. The training set median was used to fill in missing values, and a missingness indicator was added to all features where more than 5% of values were missing, since the existence of missingness provides information on which tests clinicians ordered. There was no data available for HbA1C in 18% of patients and for urine albumin-to-creatinine ratio in 41% of patients. Fitted transformations were used with the training data set, and were not changed when applied to the set of validation and test data, which could cause leakage.

3.4 Models and training

Five models were trained for each of the outcomes (Table 3): logistic regression with an L2 penalty, a random forest [3], gradient boosting with XGBoost [2], a multilayer perceptron and a stacked ensemble with a logistic meta-learner using the four base-model predictions on the validation set. A rule-based baseline, based on the clinical information provided (age, BMI, blood pressure and only one key laboratory value per disease) was used as a clinical reference. Class imbalance was addressed by class weights, not by resampling, which would have resulted in meaningless predicted probabilities. The hyperparameters were optimized using randomized search on the validation set and early stopping was applied for boosting and neural network.

Table 3. Model settings after tuning.

Model	Key settings
Logistic regression	L2 penalty, $C = 0.1$, balanced class weights
Random forest	500 trees, maximum depth 12, minimum leaf size 20
XGBoost	Learning rate 0.05, depth 5, up to 600 rounds with early stopping, row and column subsample 0.8
Multilayer perceptron	Layers 128, 64, 32; dropout 0.3; Adam, learning rate 0.001; batch size 256
Stacked ensemble	Logistic meta-learner on validation predictions of the four base models

3.5 Evaluation

¹Assistant Professor, Department of Business Administration, Islamia College and University, Peshawar.

²MPhil Scholar, Department of Business Administration, Islamia College and University, Peshawar.



The test set, which was withheld from the models, was used for judging. The AUROCs and their 95% confidence intervals were computed using 1,000 bootstrap resamples and the differences among models were tested with the DeLong method. Due to the rarity of outcomes (6% to 8%), the area under the precision–recall curve (AUPRC) and the sensitivity at a fixed specificity of 90% were also reported; the latter because screening programmes have to have a limited number of false alarms. The Brier score, reliability diagrams and the expected calibration error (ECE) across ten bins were used to assess calibration and isotonic regression was fitted to the validation data to improve calibration. SHAP values [4] were used to explain feature contributions. The one feature group was ablated at a time and there were subgroup analyses based on sex, age band and care site. The whole study is reported following TRIPOD+AI [5].

4. Results

4.1 Cohort characteristics

Over the 36-month outcome window, 4,059 patients (8.4%) developed T2D, 2,948 (6.1%) developed CKD, and 3,479 (7.2%) developed CVD. Baseline characteristics were compared for the two groups (Table 4). Patients who went on to develop disease were older, with higher blood pressure and each group had its expected pattern: the patients who developed T2D had higher BMI and HbA1c, and those who developed CKD had lower baseline eGFR, while patients who developed CVD were more likely to be male and current smokers. However, the distribution of group means is not very different from the distribution of the whole group, so a single threshold on a single variable would be sub-optimal and a multivariable model is required.

Table 4. Baseline characteristics of the cohort and of patients who developed each disease within 36 months.

Characteristic	Overall	Incident T2D	Incident CKD	Incident CVD
Patients (n)	48,320	4,059	2,948	3,479
Age, mean (SD), years	52.4 (14.1)	58.9 (11.8)	63.7 (12.4)	64.2 (11.5)
Female (%)	54.2	51.3	53.0	41.6
Body mass index	27.6	31.2	28.9	29.0

¹Assistant Professor, Department of Business Administration, Islamia College and University, Peshawar.

²MPhil Scholar, Department of Business Administration, Islamia College and University, Peshawar.



Characteristic	Overall	Incident T2D	Incident CKD	Incident CVD
(kg/m ²)				
Systolic blood pressure (mm Hg)	128	135	139	141
HbA1c (%)	5.5	6.0	5.7	5.7
eGFR (mL/min/1.73 m ²)	91	88	69	84
Current smoker (%)	17.3	19.8	18.9	26.1

4.2 Discrimination

Gradient boosting performed better than the logistic regression, random forest and neural network on all three diseases and all machine learning models achieved higher accuracy than the rule based baseline (Table 5). For T2D the AUROC of XGBoost was 0.884 (95% CI 0.871 to 0.896), compared with 0.834 for logistic regression and 0.762 for the baseline. The same ordering held for CKD (0.903 against 0.861 and 0.804) and CVD (0.823 against 0.771 and 0.742). The stacked ensemble increased AUROC by 0.004-0.005 compared to XGBoost, a difference that was statistically significant for T2D (DeLong $p = 0.041$) but too small to be practically meaningful, given the increased complexity. The most predictable outcome was CKD, with baseline eGFR and urine albumin, and the least was CVD as many cardiovascular events rely on factors less likely to be captured in routine records.

Table 5. Test-set AUROC for each model and disease.

Model	T2D	CKD	CVD
Rule-based baseline	0.762	0.804	0.742
Logistic regression	0.834	0.861	0.771
Random forest	0.861	0.882	0.798
XGBoost	0.884	0.903	0.823
Multilayer perceptron	0.872	0.893	0.815
Stacked ensemble	0.889	0.907	0.828

¹Assistant Professor, Department of Business Administration, Islamia College and University, Peshawar.

²MPhil Scholar, Department of Business Administration, Islamia College and University, Peshawar.



Table 6 provides the precision–recall and screening-oriented metrics for T2D since AUROC may appear good when there are few positives. XGBoost had a specificity of 90%, with a 65.8% sensitivity, while logistic regression had a 53.6% sensitivity and a 39.3% specificity. For XGBoost, the prevalence of the flagged cases, 8.4%, is nearly five times lower than AUPRC (0.412), indicating useful enrichment of true cases among the flagged cases. The associated ROC curves are displayed in Figure 3, and the boosted and ensemble models are near each other and both significantly outperform the simpler models in the portion of the curve relevant for screening, namely in the low false-positive region.

Table 6. Test-set performance for incident T2D.

Model	AUROC	AUPRC	Sensitivity at 90% specificity (%)	Brier score (calibrated)
Rule-based baseline	0.762	0.241	39.3	0.069
Logistic regression	0.834	0.312	53.6	0.062
Random forest	0.861	0.365	59.8	0.061
XGBoost	0.884	0.412	65.8	0.058
Multilayer perceptron	0.872	0.389	62.7	0.060
Stacked ensemble	0.889	0.421	67.2	0.057

Figure 3. Receiver operating characteristic curves of the T2D models on the test set, drawn from each model's AUROC with a binormal approximation; replace with empirical curves from your own predictions.

4.3 Effect of prediction horizon

Early detection involves predicting the disease much before it is diagnosed so the same models were re-assessed for short and long range periods. As seen in Figure 4, sensitivity decreases slowly as the horizon gets longer at 90% specificity. For T2D it fell from 76.9% at 6 months to 65.8% at 36 months, a loss of 11 points, and similar

¹Assistant Professor, Department of Business Administration, Islamia College and University, Peshawar.

²MPhil Scholar, Department of Business Administration, Islamia College and University, Peshawar.



patterns held for CKD (82.3% to 71.1%) and CVD (61.0% to 51.2%). A model that still captures $\sim 2/3$ future T2D cases 3 years in advance provides a useful time frame for lifestyle programmes, which take months to have an impact. The modest downward trend is also the result of capturing slowly changing physiology, diabetes-related changes like glucose drifting, rather than just the last stages of the disease prior to diagnosis.

Figure 4. Sensitivity at 90% specificity for XGBoost as a function of prediction horizon for the three diseases.

4.4 Calibration

The mean prediction by the raw XGBoost probabilities was higher than the rate of observed risk at the high end (Figure 5): 0.66 vs. 0.58 in the top bin. The isotonic regression fitted on the validation set resulted in a more favorable match between the predicted and observed risk, as evidenced by a decrease in the ECE from 0.031 to 0.007 and a decrease in the Brier score from 0.063 to 0.058, with little impact on AUROC. Calibration is important because a clinic will set some sort of risk threshold, like 20% and it should represent that.

Figure 5. Reliability diagram of XGBoost models for T2D before and after isotonic calibration with the diagonal line representing perfect calibration.

4.5 Feature contributions

SHAP analysis (Figure 6) revealed that the model is predominantly based on conventional clinical risk factors for diabetes. The most important were HbA1c and fasting glucose, followed by BMI, age and triglycerides. The risk of HbA1c was found to increase sharply above approximately 5.7% (the threshold for prediabetes) and decrease with increased HDL cholesterol. The largest features for CKD were baseline eGFR, slope of eGFR over the look back window, and the urine albumin to creatinine ratio, whereas age, blood pressure, sex, LDL cholesterol and smoking were the largest features for CVD. The fact that the explanations match clinical knowledge bolsters confidence that the models learn real physiology and not artefacts of data collection.

Figure 6. Top 10 features of XGBoost model used in the T2D disease.

Ablation of feature groups (4.6)

¹Assistant Professor, Department of Business Administration, Islamia College and University, Peshawar.

²MPhil Scholar, Department of Business Administration, Islamia College and University, Peshawar.



The impact of removing feature groups one at a time is shown in Table 7, where it is clear that removing laboratory results had the largest impact, reducing the T2D AUROC from 0.884 to 0.791 (or -0.093). The vital signs and anthropometrics cost 0.015, the demographics and lifestyle cost 0.013, the history of comorbidity cost 0.008, and the medication cost 0.005. Medications or comorbidities is one of the areas with the greatest impact from coding differences between sites, which is reassuring because of the small impact of these factors. It also makes itself apparent: patients who have few lab tests would be difficult to score well. donot mix headings

Table 7. Ablation of feature groups for XGBoost on T2D.

Feature set	Features (n)	AUROC	Change
All features	62	0.884	—
Without laboratory results	38	0.791	−0.093
Without vital signs and anthropometrics	56	0.869	−0.015
Without demographics and lifestyle	48	0.871	−0.013
Without comorbidity history	52	0.876	−0.008
Without medications	54	0.879	−0.005

4.7 Subgroup performance

Women and men had comparable performance (AUROC 0.881 vs 0.887), and was only modestly different by age, with the lowest performance for those < 50 (AUROC 0.862), due to the low risk in this group, making separation challenging (Table 8). The biggest difference was between care sites with AUROC of 0.889 for urban clinics compared to 0.861 for community clinics, and sensitivities at 90% specificity of 59.6% for urban clinics and 67.1% for community clinics. The number of laboratory tests ordered by community clinics is lower, and therefore imputed values are used for more patients in community clinics. Monitoring for a model deployed across the

¹Assistant Professor, Department of Business Administration, Islamia College and University, Peshawar.

²MPhil Scholar, Department of Business Administration, Islamia College and University, Peshawar.



network should be done on a site-by-site basis; if the gap remains then fine-tune with site-specific thresholds or recalibration.

Table 8. XGBoost performance for T2D in test-set subgroups.

Subgroup	Patients (n)	AUROC	Sensitivity at 90% specificity (%)
Female	5,238	0.881	64.9
Male	4,426	0.887	66.8
Age under 50	3,675	0.862	60.1
Age 50 to 64	3,462	0.879	64.4
Age 65 and over	2,527	0.871	63.0
Urban clinics	6,421	0.889	67.1
Community clinics	3,243	0.861	59.6

5. Discussion

5.1 Principal findings

There are three highlights to this. First, by simply using routine EHR data, models can be easily constructed that clearly outperform a clinical rule: on a temporally independent test set, XGBoost improved T2D sensitivity at 90% specificity from 39.3% to 65.8%. Second, the benefit is still present at longer horizon, and the model is capturing the emerging disease and not just its imminent diagnosis. Third, with respect to accuracy, more data equals better accuracy; community clinics with fewer tests were less accurate. The improvements from having a stacked ensemble rather than one boosted model were not great enough to warrant additional complexity and it is much more important to work on data quality, calibration, and monitoring than to work on the model.

This operating point has implications that need to be considered. XGBoost identifies approximately 15% of the patients, with 97% of patients not flagged having no T2D within three years and 38% of patients flagged developing T2D within three years. It is not a diagnosis, it is a reason to test, confirm and discuss risk. Whether this trade-off is acceptable is whether the follow up testing is costly, or if it is early intervention that benefits; a formal decision-curve analysis [8] could help settle this.

5.2 Deployment workflow

¹Assistant Professor, Department of Business Administration, Islamia College and University, Peshawar.

²MPhil Scholar, Department of Business Administration, Islamia College and University, Peshawar.



Even if the model is correct if used with historical data, there still needs to be a safe route to care. Figure 7 is a proposed workflow. The EHR is updated with every visit and the risk score is calculated again. If the score is above the agreed, the patient is flagged for a clinician review and the clinician determines to do a confirmatory test, e.g., HbA1c or fasting glucose. If the test is positive, treatment and lifestyle support are provided, while a negative test result means that the patient is referred back to normal screening with a reduced retesting period. Below the threshold, patients undergo regular screening. This clinical loop includes a monitoring loop: distributions are monitored, calibrated, and subgroup performance monitored quarterly, and a drift alarm alerts to retraining and revalidation prior to reusing the model. Clinician remains engaged in all decisions.

Figure 7. Conceptualization of clinical workflow for deployment including threshold checking, clinician review, confirmatory testing and a feedback loop for data drift.

5.3 Limitations

This study has a number of limitations. The data was from one regional network, and the models should be tested with other regional networks that have different coding practices and populations. The cohort excluded patients with less than 3 years of follow-up who did not have any event, and this could introduce a bias in the risk estimates as patients might drop out of the network for reasons that make their health profile different from those still in the network. Detection bias may occur because outcomes were defined using EHR codes and laboratory values that were not always used to diagnose disease or may have included coding mistakes and patients who were tested more frequently were more likely to be labelled positive. The features are based on what the clinicians selected to measure, meaning that the models partially learn the behaviour of the clinicians. Lastly, the study was retrospective – it demonstrated prediction is possible, but not that doing something predicted by the study would lead to better outcomes, so a prospective or randomised study is needed.

Ethics, privacy and governance of data.

The use of health data for prediction should be done in accordance with the appropriate approval, de-identification and secure storage. Also, fairness is important. Prediction models can systematically under-serve populations if the outcome or data is characterized by unequal access to care [6] and the site gap exhibited here is one

¹Assistant Professor, Department of Business Administration, Islamia College and University, Peshawar.

²MPhil Scholar, Department of Business Administration, Islamia College and University, Peshawar.



such example. Reporting development performance by subgroup, establishing thresholds based on equity, and monitoring post-deployment performance. The explanations provided, including SHAP, should help to dispute a score but should not be used in place of clinical judgement and should be communicated to patients when a model can help a decision about care should be made.

6. Conclusions

The machine learning models were tested in the context of the early detection of type 2 diabetes, chronic kidney disease, and cardiovascular disease, all based on regular clinical data. The conclusions are:

- 1.Gradient boosting achieved the highest performance for a single model, with test AUROC scores of 0.884, 0.903 and 0.823 for the prediction of T2D, CKD and CVD, respectively, which were higher than those of logistic regression and a rule-based baseline.
- 2.Performance of the model at 90% specificity level was 65.8% for T2D cases with a reduction of only 11 points when the horizon was extended from 6 to 36 months.
- 3.Expected calibration errors decreased from 0.031 to 0.007 with isotonic calibration; and SHAP explanations were correlated with known clinical risk factors.
- 4.Laboratory data were able to provide the majority of the predictive signal, and there was poorer performance at community clinics (AUROC 0.861) than at urban clinics (0.889).
- 5.It's suggested that deployments are followed by clinician review and drift monitoring, and external and prospective validation is required for clinical use.

References

- [1] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, pp. 1347–1358, 2019.
- [2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.
- [4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, 2017.

¹Assistant Professor, Department of Business Administration, Islamia College and University, Peshawar.

²MPhil Scholar, Department of Business Administration, Islamia College and University, Peshawar.



- [5] G. S. Collins et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, e078378, 2024.
- [6] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, pp. 447–453, 2019.
- [7] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd International Conference on Machine Learning*, 2005, pp. 625–632.
- [8] A. J. Vickers and E. B. Elkin, "Decision curve analysis: a novel method for evaluating prediction models," *Medical Decision Making*, vol. 26, pp. 565–574, 2006.
- [9] E. W. Steyerberg et al., "Assessing the performance of prediction models: a framework for traditional and novel measures," *Epidemiology*, vol. 21, no. 1, pp. 128–138, 2010.
- [10] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, e0118432, 2015.
- [11] Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group, "KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease," *Kidney International Supplements*, vol. 3, no. 1, pp. 1–150, 2013.
- [12] American Diabetes Association Professional Practice Committee, "2. Diagnosis and classification of diabetes: Standards of Care in Diabetes—2024," *Diabetes Care*, vol. 47, suppl. 1, pp. S20–S42, 2024.

¹Assistant Professor, Department of Business Administration, Islamia College and University, Peshawar.

²MPhil Scholar, Department of Business Administration, Islamia College and University, Peshawar.

