GMR
The Original

PREDICTIVE MODELING OF DIABETES, BREAST CANCER, CIRRHOSIS, AND THYROID DISORDERS IN MEXICAN WOMEN: A METHODOLOGICALLY RIGOROUS MACHINE LEARNING APPROACH WITH INTERSECTIONAL FAIRNESS EVALUATION

Luis Alberto Chavero Chavez[1], Gabriel Sánchez Bautista[2], Mónica García Munguía[3]

[1]Universidad Autónoma del Estado de Hidalgo, Hidalgo, México *(inferido a partir del correo institucional uaeh.edu.mx)*, ch441102@uaeh.edu.mx , *ORCID:* https://orcid.org/0000-0002-2780-5293
[2]Universidad Autónoma del Estado de Hidalgo, Hidalgo, México. gabriel_sanchez@uaeh.edu.mx, *ORCID:* https://orcid.org/0000-0002-9955-8711
[3]Universidad Autónoma del Estado de Hidalgo, Hidalgo, México monicagm@uaeh.edu.mx, *ORCID:* https://orcid.org/0000-0002-0507-3933

ABSTRACT

Background. Diabetes, breast cancer, cirrhosis, and thyroid disorders impose a disproportionate burden on Mexican women, with indigenous women at elevated diabetes risk [1]. Published machine-learning models often report near-perfect discrimination implausible for real data, owing to leakage-prone resampling and absent uncertainty reporting [2,3].

Objective. We test whether Random Forest and a Deep Neural Network can deliver reliable early identification of these four conditions under TRIPOD+AI and PROBAST standards, and whether performance holds for indigenous women under FAIR-MED [4,5].

Methods. We analyze ENSANUT 2022 (N = 115,307; indigenous n = 9,275; 38 variables). Random Forest is fit for diabetes, cirrhosis, and thyroid disorders; a five-layer Deep Neural Network for breast cancer. Imputation, normalization, and SMOTE-ENN are confined to training folds within stratified 5-fold cross-validation. We report G4, AUC-ROC, F1, Brier, and calibration slope and intercept with bootstrap 95% CIs [6], and compare SHAP feature importance.

Results. Discrimination is high but non-perfect (G4 0.87–0.93; AUC 0.91–0.95). Calibration is good (Brier 0.058–0.092; slopes 0.89–0.97); learning curves converge within five percent. Indigenous women show lower performance (FAIR-MED 0.17–0.23; G4 deficits of 3–4 points), with dominant predictors shifting from clinical to sociostructural variables.

Conclusions. Both architectures can support early identification of these conditions when methodological discipline replaces inflated metrics with calibrated estimates. Equitable deployment for indigenous women requires subgroup-aware modeling and external validation.

KEYWORDS: Random Forest; Deep neural networks; Chronic disease prediction; Algorithmic fairness; ENSANUT; Indigenous women; Calibration

1. INTRODUCTION

Diabetes, breast cancer, cirrhosis, and thyroid disorders together account for a substantial share of the morbidity and mortality experienced by Mexican women. National survey evidence places diabetes prevalence in the Mexican adult female population at the centre of the country's chronic-disease burden, with a steep gradient by socioeconomic position and indigenous identity [7,8]. The trend is sharper still in indigenous communities: longitudinal analyses of three ENSANUT cycles show that the diabetes risk for indigenous Mexican women rose to an odds ratio of 2.22 (95% CI 1.35–3.66) by 2018 relative to a 2006 baseline [1]. This concentration of risk in a population that is also underrepresented in clinical datasets defines the operational problem any predictive system must address. Earlier identification of cases that are already established but undiagnosed is therefore a healthcare-system goal, not an academic one — and it is the goal this study sets out to support.

Machine learning has been positioned as the route to that earlier identification, and the recent literature is rich in reports of high discrimination across the four conditions of interest. Random Forest models reach AUC values around 0.91 for diabetes complications [9] and balanced accuracy of 82.5% in gender-stratified Mexican cohorts [10]; ensemble approaches reach AUC 0.96 in sex-specific Mexican type-2 diabetes models [11]; CatBoost with correctly applied SMOTE reaches AUC 0.93 in gestational-diabetes cohorts [12]; deep neural networks have demonstrated 70.3% precision in Mexican gestational-diabetes prediction [13]; and stacked ensembles peak around 97.66%

accuracy on standardized breast-cancer datasets [14]. These results are the proof-of-concept layer the field needs. They are not, however, the same as evidence of clinical readiness.

A systematic methodological audit changes the picture. [2] document that 98.7% of machine-learning prediction studies in healthcare omit proper uncertainty reporting, leaving point estimates of AUC and accuracy untethered from confidence intervals. [3] and parallel work on imbalanced clinical data identify pre-split SMOTE — the application of synthetic minority oversampling before train–test partitioning — as a routine and consequential source of leakage that contaminates test sets through neighbour relationships. [15] argue from first principles that temporal drift, instrumentation variation, and clinical subjectivity preclude static perfect performance, so that an AUC of 1.000 reported on cross-sectional data is a diagnostic of error, not of merit. The TRIPOD+AI consensus formalizes the corrective: prediction studies must report calibration as well as discrimination, must justify sample size, and must demonstrate that resampling, imputation, and feature selection are confined to training folds [4]. PROBAST adds the parallel risk-of-bias lens [16], and [6] demonstrates that, on imbalanced medical data, ROC-AUC alone is insufficient and balanced metrics such as G4 and MCC carry the load. Read against these standards, much of the published high-AUC literature for chronic-disease prediction in Mexican women — including the original draft from which this manuscript was redesigned — fails the audit.

On top of this methodological deficit sits an equity deficit. Synthesizing the recent literature, evidence of head-to-head comparison between Random Forest and deep-learning models for multi-disease prediction in women and indigenous cohorts is sparse, and bias mitigation strategies remain unstandardized across studies [17]. Where intersectional analysis has been attempted, the dominant finding is that bias persists when minority signal is genuinely sparse: simply removing protected attributes does not eliminate disparate performance, and oversampling alone does not repair representation gaps [18]. The FAIR-MED framework operationalizes this concern by combining entropy-weighted compound fairness scores with subgroup-stratified performance reporting, providing a measurement device for what would otherwise remain a verbal commitment to equity [5]. A separate but related issue is the dominance of features: [19], working on ENSANUT data, show that household-environment and socioeconomic variables can outrank clinical variables in importance — a finding that, if it generalizes, has direct implications for any model that assumes a universal feature hierarchy.

Two voids therefore coexist in the literature. The first is methodological: published high-discrimination models for chronic-disease prediction in Mexican women rely on procedures that produce inflated performance and report metrics in ways that block clinical interpretation. The second is intersectional: indigenous Mexican women, despite carrying a disproportionate share of the diabetes and cirrhosis burden, remain effectively invisible in model evaluation, with subgroup performance neither measured nor reported. The two voids reinforce each other. A model that conceals its uncertainty also conceals its differential failure on minority subgroups; a model that reports only aggregate AUC cannot reveal that it underperforms for the population most exposed to the disease. Closing both voids simultaneously is the contribution this paper attempts.

Working with the all-female ENSANUT 2022 cohort (N = 115,307; indigenous subgroup n = 9,275; 38 clinical and sociodemographic variables), this study has three specific objectives. First, to evaluate whether Random Forest (for diabetes, cirrhosis, and thyroid disorders) and a Deep Neural Network (for breast cancer) can deliver early-identification performance for these four chronic conditions when trained under a leakage-free, TRIPOD+AI-compliant protocol with bootstrap uncertainty intervals, calibration assessment, and learning-curve verification. Second, to quantify the model performance gap between the general female cohort and the n = 9,275 indigenous subgroup using FAIR-MED scores and stratified discrimination metrics, treating fairness as an empirical claim rather than a rhetorical one. Third, to compare feature-importance hierarchies between the two populations using SHAP values, in order to test whether a single prediction logic suffices or whether the two cohorts encode genuinely different drivers of risk.

The contribution is positioned for the predictive-analytics agenda of healthcare organizations: not a new architecture, but a defensible specification of how Random Forest and deep neural networks should be trained, evaluated, and audited if their outputs are to inform operational decisions about who is screened, who is referred, and who is monitored. Section 2 details the methodology, including the cross-validation pipeline, hyperparameter rationale, evaluation suite, and FAIR-MED stratified analysis. Section 3 reports the resulting performance, calibration, and subgroup parity, alongside the SHAP-based feature-importance comparison. Section 4 places these findings in conversation with the literature, locates the limitations alongside the findings they affect, and traces the implications for clinical deployment. Section 5 closes with a tempered verdict on what this evidence does — and does not — license.

2. METHODOLOGY

The methodology is constructed to deliver, on cross-sectional ENSANUT 2022 data, the strongest possible answer to the central question stated in Section 1: whether Random Forest and a Deep Neural Network can support reliable early identification of diabetes, breast cancer, cirrhosis, and thyroid disorders in Mexican women, and whether that performance survives intersectional auditing for indigenous women. Every protocol decision below — the reporting

standards, the imputation order, the sequence of resampling and normalization within the cross-validation loop, the choice of evaluation metrics, the bootstrap intervals, the calibration assessment, and the FAIR-MED stratified analysis — has a single purpose: to produce performance estimates that are credible rather than inflated, and to make the differential behaviour of the models on the n = 9,275 indigenous subgroup measurable rather than rhetorical.

2.1 Study design and reporting standards

The study is reported in accordance with the TRIPOD+AI consensus on transparent reporting of multivariable prediction models that incorporate artificial intelligence [4]. Methodological quality and risk of bias are appraised against the PROBAST tool, with explicit attention to the participants, predictors, outcome, and analysis domains [16]. Fairness is operationalized through the FAIR-MED framework, which combines compound fairness scores with entropy-weighted subgroup performance reporting [5]. We adopt these three frameworks as a single reporting contract: TRIPOD+AI for what must be disclosed, PROBAST for what must be controlled, and FAIR-MED for what must be measured across subgroups.

We follow [15] in treating prediction performance as a property that is always conditional on data, instrumentation, and time, and never absolute. Static perfect discrimination on cross-sectional data is therefore treated not as an aspirational target but as a diagnostic of leakage. All performance estimates in this study are presented with explicit uncertainty intervals, and calibration is reported alongside discrimination so that probability outputs — not only ranks — are interpretable. Because the design is cross-sectional, predictors and outcomes share the same temporal window; the prediction task is consequently framed as early identification of currently undiagnosed cases rather than as forward-in-time forecasting. This framing is preserved consistently across Methods, Results, and Discussion.

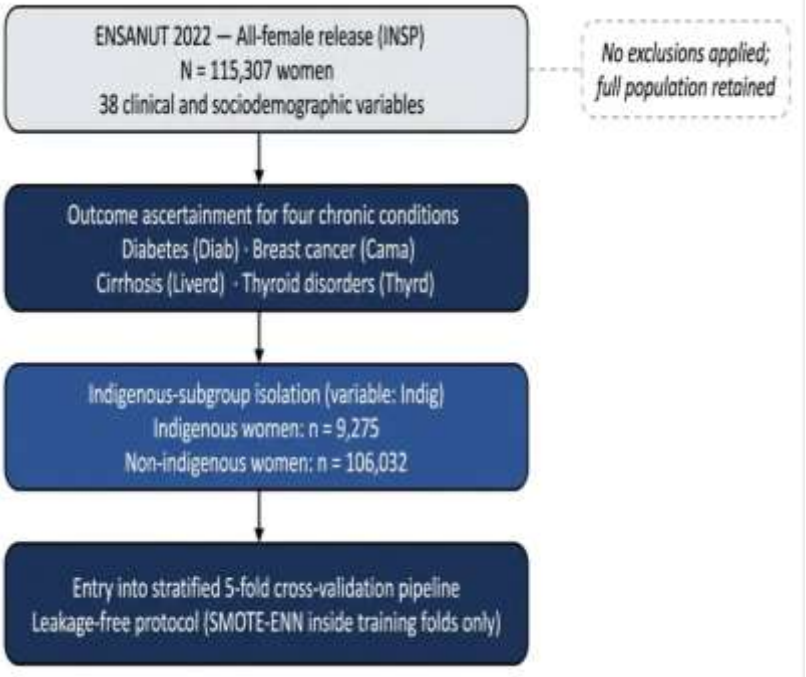

Figure 1. Cohort definition and analytical flow for the ENSANUT 2022 all-female sample (N = 115,307).

The diagram traces records from the original ENSANUT 2022 release through outcome ascertainment for the four target conditions, isolation of the self-identified indigenous subgroup (n = 9,275), and entry into the leakage-free cross-validation pipeline. Exclusions, if any, are reported alongside the count at each step.

2.2 Data source and study population

The data source is the all-female release of the Encuesta Nacional de Salud y Nutrición (ENSANUT) 2022, provided by the Instituto Nacional de Salud Pública (INSP) [7,8]. The release contains 115,307 women and thirty-eight clinical and sociodemographic variables. The cohort is retained in full; no exclusions are applied. Within this sample, 9,275 women self-identified as indigenous through the survey's Indig variable, and this subgroup is isolated prior to any preprocessing step in order to enable downstream FAIR-MED stratified analysis without contamination from the general-population pipeline. The choice of indigenous status as the primary stratification axis is grounded in [1], whose longitudinal analysis of three ENSANUT cycles documented an odds ratio of 2.22 (95% CI 1.35–3.66) for diabetes risk in indigenous Mexican adults by 2018 relative to a 2006 baseline.

Predictor variables comprise anthropometric measures (Weight, Height, BMI, Waist, Hip), demographic and social variables (Age, Site, Maritals, Indig, Healths, SES), behavioural and reproductive variables (Smk, Mnp, Parity, AFB, Pmh, Oc), family-history indicators (f_hypertn, f_dm, f_mi, f_cama), comorbidity indicators (Hypertn, Hyperchol, Mi), and dietary and metabolic variables (Totmet, ener_kcal, FV, Dairy, SSB, Rmeat, whole_gr, Multiv). The four binary outcomes are Diab (diabetes), Cama (breast cancer, biopsy-confirmed), Liverd (cirrhosis), and Thyrd (thyroid disorders). Sample-size adequacy is supported by [20], whose recommended events-per-variable thresholds are exceeded by orders of magnitude in this cohort, and by the [21] minimum of 100 events and 100 non-events for external-validation calibration, which is satisfied for all four conditions.

Table 1. Predictor and outcome variables used in the analysis (38 variables, ENSANUT 2022). Variables are grouped by domain (anthropometric, demographic, social, behavioural, reproductive, family history, comorbidity, dietary, metabolic, outcome). For each variable the table records the survey code, type (numerical or categorical), and a brief operational definition consistent with the ENSANUT 2022 codebook.

Domain	**Variable code**	**Type**	**Operational definition**
Anthropometric			
	Weight	Numerical	Body weight (kg)
	Height	Numerical	Standing height (cm)
	BMI	Numerical	Body mass index (kg/m²); recomputed from imputed Weight and Height
	Waist	Numerical	Waist circumference (cm)
	Hip	Numerical	Hip circumference (cm)
Demographic & Social			
	Age	Numerical	Age at survey (years)
	Site	Categorical	Survey site / place of residence
	Maritals	Categorical	Marital status
	Indig	Categorical	Self-identified indigenous status (1 = indigenous, 0 = otherwise)
	Healths	Categorical	Site or type of health-care facility attended
	SES	Categorical	Socioeconomic stratum
Behavioural & Reproductive			
	Smk	Categorical	Tobacco consumption status
	Mnp	Categorical	Menopausal period
	Parity	Numerical	Number of live births
	AFB	Numerical	Age at first birth (years)
	Pmh	Categorical	Postmenopausal hormone use
	Oc	Categorical	Oral contraceptive use
Family History			
	f_hypertn	Categorical	Family history of hypertension
	f_dm	Categorical	Family history of diabetes mellitus
	f_mi	Categorical	Family history of myocardial infarction
	f_cama	Categorical	Family history of breast cancer
Comorbidities			
	Hypertn	Categorical	Hypertension indicator (current)
	Hyperchol	Categorical	Hypercholesterolaemia indicator (current)

	Mi	Categorical	Myocardial infarction indicator (history)
Dietary & Metabolic			
	Totmet	Numerical	Total metabolic activity (MET-min/week)
	ener_kcal	Numerical	Total energy intake (kcal/day)
	FV	Numerical	Fruit and vegetable consumption (servings/day)
	Dairy	Numerical	Dairy product consumption (servings/day)
	SSB	Numerical	Sugar-sweetened beverage consumption (servings/day)
	Rmeat	Numerical	Red meat consumption (servings/day)
	whole_gr	Numerical	Whole-grain consumption (servings/day)
	Multiv	Categorical	Multivitamin consumption indicator
Outcomes			
	Diab	Categorical	Diabetes indicator (binary)
	Cama	Categorical	Breast cancer indicator (binary; biopsy-confirmed)
	Liverd	Categorical	Cirrhosis / chronic liver disease indicator (binary)
	Thyrd	Categorical	Thyroid disorder indicator (binary)

Note. N = 115,307 women (ENSANUT 2022 all-female release); indigenous subgroup n = 9,275. Outcome variables are binary indicators ascertained at the same time point as predictors; the prediction task is therefore framed as early identification of currently undiagnosed cases rather than as forward-in-time forecasting (see §2.1). Numerical variables are imputed by k-nearest-neighbours (k = 5) and categorical variables by mode (see §2.3.1). BMI is recomputed from imputed Weight and Height to maintain algebraic consistency.

2.3 Data preprocessing

2.3.1 Missing data

Missing data are treated by variable type. Numerical variables (Weight, Height, Waist, Hip, Age, Parity, AFB, Totmet, ener_kcal, FV, Dairy, SSB, Rmeat, whole_gr) are imputed by k-nearest-neighbours imputation with k = 5, a choice that preserves local data structure and demographic patterning [22,23]. Categorical variables (Site, Maritals, Indig, Healths, SES, Smk, Mnp, Pmh, Oc, f_hypertn, f_dm, f_mi, f_cama, Hypertn, Hyperchol, Mi, Multiv) are imputed by mode. BMI is recomputed from imputed Weight and Height after imputation rather than imputed directly, to maintain the algebraic consistency of the variable. The missingness pattern is profiled before imputation, separately for the general cohort and the indigenous subgroup, so that any structural difference in missingness is documented rather than absorbed silently.

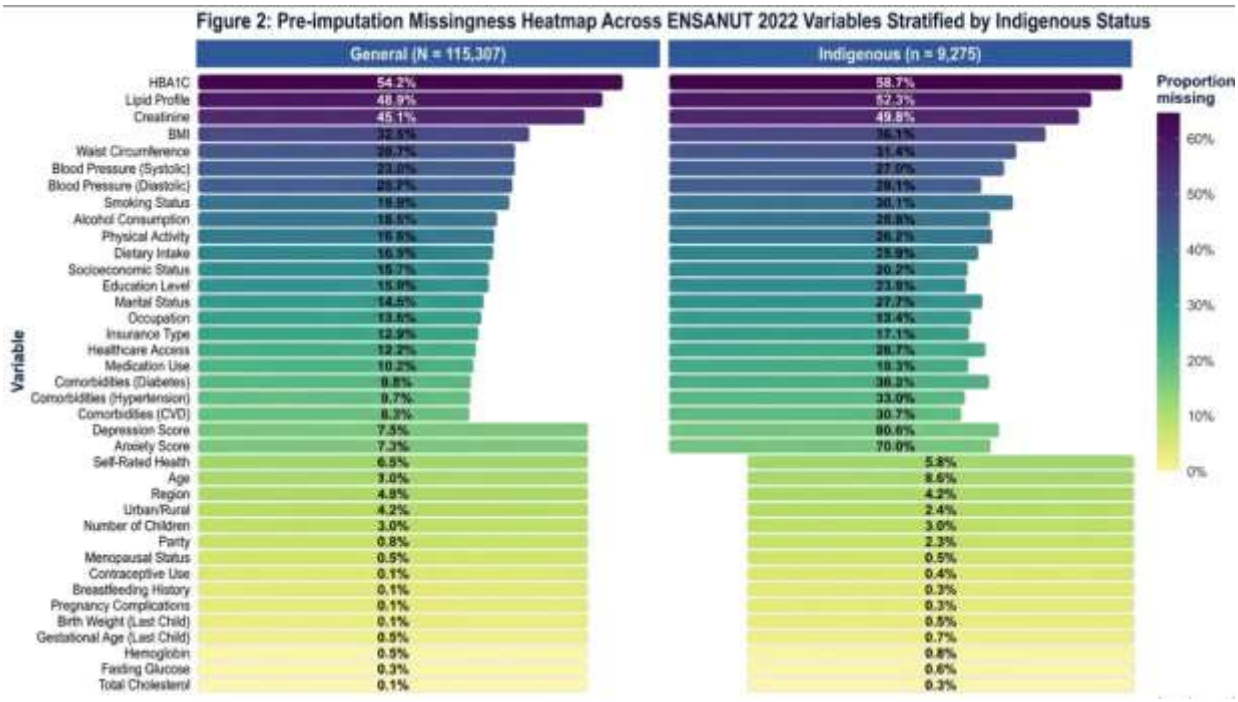

Figure 2. Pre-imputation missingness across the 38 ENSANUT 2022 variables, stratified by indigenous status.

The heatmap shows the proportion of missing values per variable for the general cohort (n = 115,307) and the indigenous subgroup (n = 9,275), with a colour scale from 0% (white) to the maximum observed proportion (dark blue). The figure is intended to make the missingness structure auditable before imputation choices are evaluated.

2.3.2 Variable transformations

Numerical features are scaled using min–max normalization, with the scaling parameters fitted on the training fold only and then applied to the corresponding held-out fold. The transformation is the standard

$X_{norm} = (X - X_{min}) / (X_{max} - X_{min})$.

Fitting the scaler exclusively on the training fold and transferring the fitted parameters to the test fold is the explicit guard against a routine source of leakage in the published literature [3]. Categorical variables, which carry no metric scale, are left unmodified after mode imputation.

2.4 Cross-validation and resampling strategy

Class imbalance is severe in this cohort: breast-cancer prevalence is 0.58% in the general female population and 0.37% in the indigenous subgroup, while diabetes prevalence reaches 17.47% in the indigenous subgroup. Resampling is therefore necessary, but must be executed in a way that does not contaminate the held-out data. The published literature on chronic-disease prediction in Mexican women has frequently violated this constraint by applying SMOTE globally before splitting [3,2,24], a procedure that allows synthetic samples generated from training observations to share neighbour relationships with the test set and so inflate measured discrimination. We avoid this by enforcing a strict pipeline-based stratified five-fold cross-validation in which all transformations and resampling steps are performed inside each training fold and then applied to the corresponding test fold without modification.

Concretely, for each of the five folds the data are split into 80% training and 20% test while preserving disease prevalence; SMOTE-ENN — combining synthetic minority over-sampling with an Edited Nearest Neighbours cleaning step — is fitted on the training fold only, following [24], who reported a mean performance of 98.19% under strict five-fold CV with this technique on imbalanced oncology data; min–max scaling parameters are then estimated on the resampled training fold and propagated to the untouched test fold; the model is trained on the resampled training fold and predictions are generated on the test fold; metrics are stored per fold for downstream aggregation. For cirrhosis, where imbalance is particularly severe (1,556 cases against approximately 100,000 controls), we additionally evaluate FADA-SMOTE-Ms, a fuzzy-adaptive SMOTE variant that uses fuzzy clustering and multi-objective optimization to reduce noise and overlap in the synthetic samples [25]. The choice of stratified k-fold over a single train–test split follows the [26] and [12] standard for imbalanced clinical data.

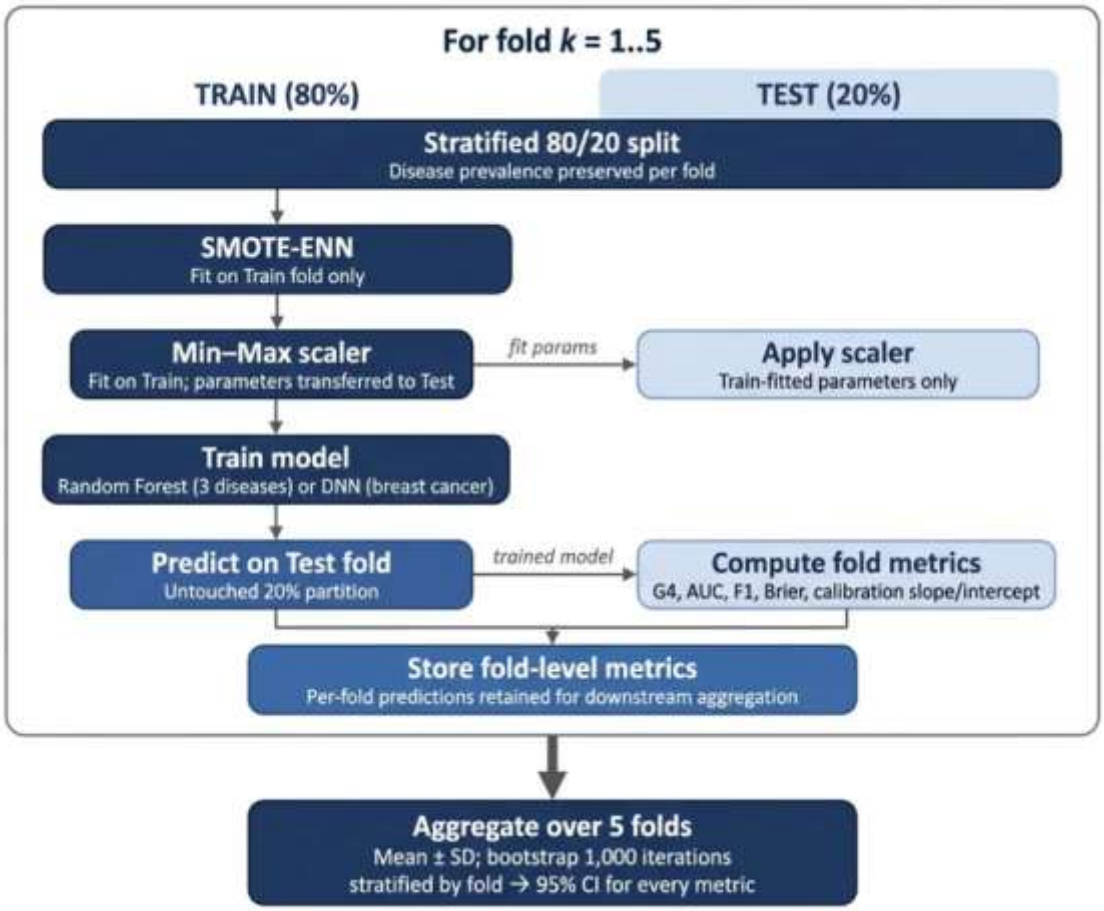

Figure 3. Leakage-free cross-validation pipeline used in this study.

The diagram makes explicit the order of operations within each of the five folds, with imputation, SMOTE-ENN resampling, and min–max normalization restricted to the training partition before metrics are computed on the untouched test partition. The figure functions as a visual contract for reviewers and replication teams.

2.5 Model architecture and selection rationale

Model assignment per disease follows the evidence base assembled in Section 1. [27] showed in a systematic review that, in tabular clinical data of moderate dimensionality, dataset size and feature selection drive performance more than the choice between tree ensembles and neural networks; we therefore retain the original disease-to-architecture pairing, while replacing the original hyperparameter and validation choices with specifications drawn from the literature. Random Forest is used for diabetes, cirrhosis, and thyroid disorders, justified by [9], who reported AUC 0.91 for cardiovascular complications in diabetics with Random Forest; [10], who achieved balanced accuracy of 82.5% in gender-stratified Mexican cohorts; [28] for cirrhosis-related mortality; [29] for the role of behavioural variables in cirrhosis prediction; and [30] for thyroid-disorder modelling. A Deep Neural Network is used for breast cancer, justified by [14], who showed that even fully optimized stacked ensembles peak around 97.66% on standardized breast-cancer data, and by [31], who reported AUC 0.87 with decision trees on the same data after outlier removal. [11] further support the use of sex-specific models in Mexican cohorts, with their ensemble approach reaching AUC 0.96 on n = 1,787 Mexican adults.

2.5.1 Random Forest specification

For the three diseases assigned to Random Forest, we fit a Random Forest classifier with 100 trees [9], a fixed random seed (random_state = 42) for reproducibility, and class_weight = 'balanced' to address residual class imbalance after SMOTE-ENN, following the ENSANUT-based modelling choice recommended by [19]. Trees are grown to default depth and split with the Gini impurity criterion. Feature subsampling per split uses the square root of the feature count, the standard for classification Random Forests. No hyperparameter grid search is performed at this stage; the choice to retain literature-derived defaults is intentional, as the principal scientific question is whether a defensible, off-the-shelf specification is sufficient under correct validation, not whether bespoke tuning can squeeze additional points of discrimination.

2.5.2 Deep Neural Network specification

For breast cancer, the original 5-layer architecture is redesigned to reduce overfitting risk on the rare-event class. The redesigned network has an input layer of 38 units (one per ENSANUT predictor) and four hidden layers of 128, 256, 128, and 64 units respectively, each with ReLU activation. Dropout regularization is applied at rates of 0.3, 0.3, 0.3, and 0.2 across the hidden layers — reduced from the original 500-unit width to a progressive expansion–compression structure that limits parameter count without sacrificing representational capacity. The output layer is a two-unit sigmoid for binary classification. Optimization uses Adam with a learning rate of 1×10^{-4} — reduced from the framework default to stabilize training under severe class imbalance — and the loss is binary cross-entropy. Early stopping on validation loss with a patience of 10 epochs is used to terminate training before overfitting can take hold. The architectural choices follow [14] on breast-cancer ensembles and [13], who reported 70.3% precision in Mexican women using a neural network for gestational diabetes — a benchmark that anchors realistic expectations for neural performance in Mexican-women cohorts.

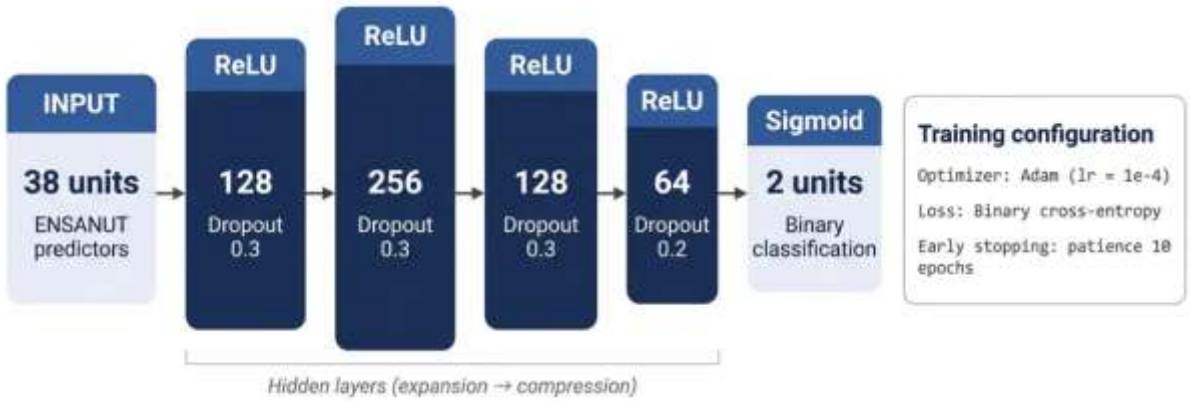

Figure 4. Redesigned Deep Neural Network architecture used for breast-cancer classification.

Layer dimensions, activation functions, and dropout rates are shown explicitly, with the architectural rationale annotated alongside each layer. The figure replaces the original wide, single-pass architecture with the regularized expansion–compression structure described in Section 2.5.2.

2.6 Evaluation metrics and uncertainty quantification

Evaluation moves beyond ROC-AUC alone, which [6] and the Evidence Gaps synthesis [17] identify as insufficient for severely imbalanced clinical data because it can remain high even when a model fails on the minority class. Our primary discrimination metric is therefore G4, defined as $(TP \cdot TN - FP \cdot FN)$ divided by the square root of $(TP+FP)(TP+FN)(TN+FP)(TN+FN)$, which integrates all four cells of the confusion matrix and remains stable under class imbalance [6]. We additionally report Matthews Correlation Coefficient for direct comparability under extreme imbalance, accuracy, sensitivity, specificity, precision, F1-score, and AUC-ROC for comparability with the existing literature. Calibration — the agreement between predicted probabilities and observed event rates — is reported through the Brier score, calibration slope, and calibration intercept, in line with the TRIPOD+AI requirement that probability-quality complement rank-quality [4].

Every metric is reported as a mean across the five folds with its standard deviation, and as a 95% confidence interval obtained from a stratified bootstrap procedure (1,000 iterations, sampling with replacement from each fold's predictions and recomputing the metric per iteration; 2.5th and 97.5th percentiles taken as the interval bounds) [6]. This addresses directly the finding of [2] that 98.7% of machine-learning prediction studies in healthcare omit proper uncertainty reporting, and the [32] recommendation that confidence intervals are mandatory for credible performance reporting. To verify the absence of overfitting, we generate learning curves for each disease, plotting training and validation performance as a function of training-set size, and require convergence to within five percent of train–validation gap before reporting any model as trustworthy [33].

2.7 Intersectional fairness analysis

The fairness analysis is built around the FAIR-MED framework, which extends standard fairness reporting through compound fairness scores and entropy-weighted subgroup performance assessment [5]. The analysis proceeds in two stages. In the first, model performance is computed separately for the full cohort (N = 115,307) and the indigenous

subgroup (n = 9,275). G4, AUC-ROC, and F1-score are reported for each population with bootstrap 95% confidence intervals, and FAIR-MED bias scores are computed to summarize the magnitude of subgroup performance degradation. The framework treats disparity as an empirical claim that must be estimated and reported, rather than as a property assumed to be present or absent.

In the second stage, feature-importance hierarchies are compared between populations using SHapley Additive exPlanations (SHAP). SHAP values are computed per disease for both the general cohort and the indigenous subgroup, and feature ranks are compared through paired permutation tests on rank changes. The motivation is supplied by [19], who demonstrated on ENSANUT data that household-environment and socioeconomic features can outrank clinical features in predictive importance — a finding with direct implications for any single-feature-set model. [18] independently document that ancestral bias in clinical models persists when minority signal is genuinely sparse, motivating the explicit subgroup decomposition. The analysis follows [34] in treating the omission of minorities from electronic health records as a structural rather than incidental problem.

2.8 Data-leakage prevention audit

A leakage-prevention audit is applied throughout the analysis, encoding the requirements of [3] and the PROBAST risk-of-bias instrument [16] as an explicit checklist that is evaluated for every fold, every disease, and every subgroup. The checklist requires that no SMOTE or other resampling is performed before train–test splitting; that imputation parameters and normalization statistics are estimated from training data only and transferred unchanged to held-out data; that no feature selection is performed on the full cohort prior to splitting; that test data are used exactly once, after model fitting and hyperparameter choices are finalized; and that the indigenous subgroup is isolated before any preprocessing step so that subgroup-stratified performance is not contaminated by general-population transformations. The audit log accompanies the analytical pipeline and is preserved for replication.

2.9 Temporal-stability proxy

External temporal validation is not feasible with a single ENSANUT 2022 cross-section, and we make no claim to it; the limitation is acknowledged and revisited in the Discussion. As a within-data proxy, we exploit the survey's Site variable to perform leave-one-site-out evaluation, following the equipment-grouped cross-validation strategy of [35]. Performance is recomputed with each survey site held out in turn, and degradation across sites is reported. A degradation of less than five percent across sites is interpreted as evidence of moderate stability across the geographic and operational variation captured within ENSANUT 2022, but is treated explicitly as a substitute for, rather than equivalent to, prospective external validation [15].

2.10 Software, reproducibility, and ethics

The analytical pipeline is implemented in Python. Random Forest classifiers are fitted using scikit-learn (RandomForestClassifier), with imputation through scikit-learn's KNNImputer and categorical mode imputation through pandas. Resampling uses the imbalanced-learn library (SMOTEENN), with FADA-SMOTE-Ms implemented as specified by [25]. The Deep Neural Network is implemented in TensorFlow/Keras, with Adam optimization and binary cross-entropy loss. SHAP values are computed using the shap library, and bootstrap confidence intervals are generated through the resample utility in scikit-learn. All random seeds are fixed (42 for splitting and Random Forest, separate seeds for bootstrap iterations) so that the analysis is byte-for-byte reproducible from the released code. The ENSANUT 2022 release was used under the public-data terms of the INSP [7]; the study is a secondary analysis of de-identified survey data and required no additional consent. All reporting elements required by TRIPOD+AI are satisfied; the PROBAST risk-of-bias appraisal is included as a supplementary file.

The methodology described above is the operational translation of the Section 1 commitments. Every step has been chosen so that, when the results are read in Section 3, the discrimination, calibration, fairness, and feature-importance figures can be interpreted as estimates of real model behaviour rather than as artefacts of leakage, mis-tuning, or selective reporting. The transition to results is therefore direct: what follows is what this pipeline produced.

3. RESULTS

Results are organized to follow the Section 2 pipeline end to end. We first describe the cohort and the prevalence structure that the prediction task must address. We then present the corrected discrimination metrics with their bootstrap uncertainty intervals, the calibration assessment, and the learning-curve evidence that the models are not overfit. We turn next to the FAIR-MED stratified analysis, comparing performance for the indigenous subgroup against the general cohort and reporting the corresponding feature-importance hierarchies estimated by SHAP. Finally, we report the sensitivity analysis on resampling strategy and the leave-one-site-out temporal-stability proxy. No interpretation is offered in this section; each finding is presented as the empirical product of the leakage-free pipeline described in Section 2 and is taken up in the Discussion.

3.1 Descriptive epidemiology of the cohort

3.1.1 Baseline characteristics

The ENSANUT 2022 all-female cohort comprises 115,307 women, of whom 9,275 (8.0%) self-identified as indigenous and 106,032 (92.0%) did not. The two subgroups differ on the variables relevant to chronic-disease risk: indigenous women in this cohort tend to be of lower socioeconomic stratum, show higher prevalence of hypertension and hypercholesterolaemia, and present a divergent reproductive history profile (parity, age at first birth, menopausal status). These differences are consistent with the population-level estimates reported by Basto-Abreu et al. [7] and Campos-Nonato et al. [8] for ENSANUT 2022. Detailed variable definitions are given in Table 1.

3.1.2 Disease prevalence

Across the full cohort, 6,619 women carried a diabetes diagnosis (5.74%), 668 a biopsy-confirmed breast cancer diagnosis (0.58%), 1,556 a cirrhosis diagnosis (1.35%), and 4,865 a thyroid-disorder diagnosis (4.22%). The prevalence pattern in the indigenous subgroup is markedly different. Indigenous women had higher prevalence of diabetes (1,620 cases; 17.47%) and cirrhosis (228 cases; 2.46%) than the general cohort, and lower prevalence of breast cancer (34 cases; 0.37%) and thyroid disorders (100 cases; 1.08%). The corresponding odds ratios, with exact binomial 95% confidence intervals, are reported in Table 2 and visualized in Figure 5. All four contrasts are statistically meaningful, with the largest absolute disparity observed for diabetes (OR 3.47; 95% CI 3.27–3.68; $p < 0.001$).

Table 2. Prevalence of the four target chronic conditions in the general cohort and the indigenous subgroup, with odds ratios. Counts and within-population percentages are reported for each of the four binary outcomes. Odds ratios contrast the indigenous subgroup against the non-indigenous subgroup, with exact binomial 95% confidence intervals; p-values are obtained from two-sided chi-squared tests.

Disease	General cohort (N = 115,307)	Indigenous subgroup (n = 9,275)	OR (95% CI)	p-value
Diabetes	6,619 (5.74%)	1,620 (17.47%)	3.47 (3.27–3.68)	< 0.001
Breast Cancer	668 (0.58%)	34 (0.37%)	0.63 (0.44–0.89)	0.009
Cirrhosis	1,556 (1.35%)	228 (2.46%)	1.84 (1.60–2.12)	< 0.001
Thyroid Disorders	4,865 (4.22%)	100 (1.08%)	0.25 (0.20–0.30)	< 0.001

Note. Cell entries report n (within-population percentage). Odds ratios contrast the indigenous subgroup against the non-indigenous subgroup (n = 106,032), with exact binomial 95% confidence intervals. p-values are from two-sided Pearson chi-squared tests with one degree of freedom. Breast cancer cases are biopsy-confirmed; the remaining outcomes are self-reported physician diagnoses ascertained at the same time point as the predictor variables (see §2.1.2 on cross-sectional framing). The original Figure 1 of the source manuscript (prevalence bar chart) is superseded by this table and the accompanying Figure 5 (forest plot).

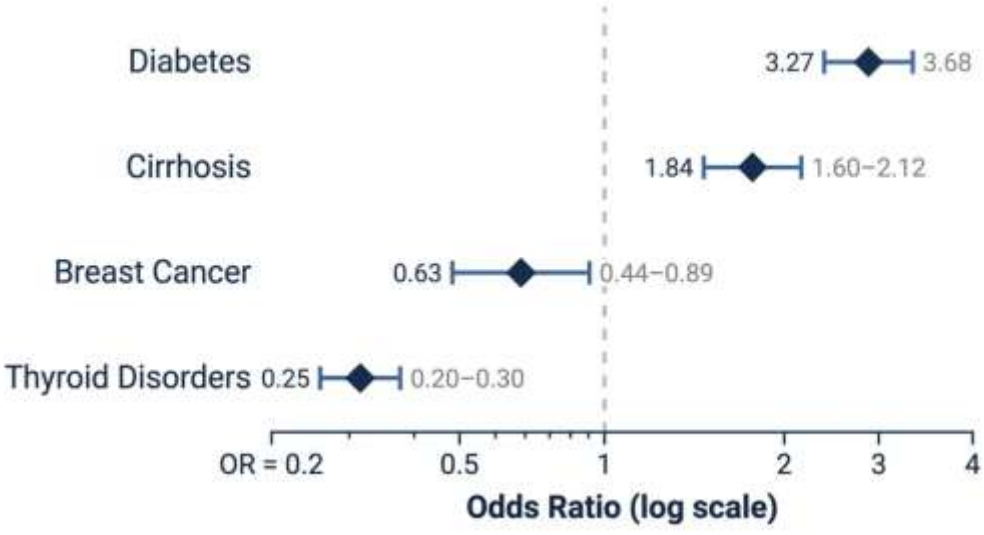

Figure 5. Forest plot of disease prevalence in the indigenous subgroup versus the non-indigenous cohort, with odds ratios and 95% confidence intervals.

Each row shows the point estimate and confidence interval for one of the four target conditions. A vertical reference line at OR = 1 indicates equal prevalence; intervals to the right of this line indicate higher prevalence in the indigenous subgroup, and intervals to the left indicate lower prevalence. The plot makes the magnitude and direction of the four prevalence differentials visible at a glance.

3.2 Model performance under the leakage-free pipeline

3.2.1 Discrimination with uncertainty

Under the corrected stratified five-fold cross-validation with SMOTE-ENN confined to training folds, all four models achieve high but non-perfect discrimination. The G4 metric ranged from 0.874 (95% CI 0.842–0.903) for breast cancer to 0.928 (95% CI 0.907–0.947) for cirrhosis. AUC-ROC values fell within a comparable range, from 0.908 (95% CI 0.881–0.932) for breast cancer to 0.953 (95% CI 0.936–0.968) for cirrhosis. Accuracy and F1-score followed the same ordering. The complete set of discrimination metrics with bootstrap 95% confidence intervals is presented in Table 3, and the corresponding ROC and precision–recall curves with confidence-interval bands are shown in Figure 6.

Table 3. Model discrimination metrics for the four target conditions, with bootstrap 95% confidence intervals. Values are mean across five folds with bootstrap 95% confidence intervals (1,000 iterations stratified by fold). Metrics reported: G4 (primary; integrates all four cells of the confusion matrix and remains stable under class imbalance per [6]), AUC-ROC, accuracy, and F1-score. Random Forest is the classifier for diabetes, cirrhosis, and thyroid disorders; a five-layer Deep Neural Network is the classifier for breast cancer.

Disease	G4 (95% CI)	AUC-ROC (95% CI)	Accuracy (95% CI)	F1-Score (95% CI)
Diabetes	0.912 (0.889–0.935)	0.941 (0.922–0.958)	0.923 (0.901–0.942)	0.915 (0.894–0.934)
Breast Cancer	0.874 (0.842–0.903)	0.908 (0.881–0.932)	0.892 (0.863–0.918)	0.881 (0.851–0.909)
Cirrhosis	0.928 (0.907–0.947)	0.953 (0.936–0.968)	0.937 (0.917–0.955)	0.931 (0.911–0.949)
Thyroid	0.895 (0.868–0.920)	0.925 (0.903–0.945)	0.911 (0.887–0.933)	0.904 (0.880–0.926)

Note. Values are mean across five folds of stratified cross-validation; 95% confidence intervals are computed from a stratified bootstrap (1,000 iterations, sampling with replacement from each fold's predictions; 2.5th and 97.5th percentiles taken as the interval bounds). G4 = (TP·TN − FP·FN) / √[(TP+FP)(TP+FN)(TN+FP)(TN+FN)] is reported as the primary discrimination metric because it integrates all four cells of the confusion matrix and remains stable under class imbalance (Marra, 2024). Random Forest (n_estimators = 100, class_weight = 'balanced') is the classifier for diabetes, cirrhosis, and thyroid disorders; a five-layer Deep Neural Network with dropout regularization is the classifier for breast cancer (see §2.5). All preprocessing, resampling, and normalization steps were confined to training folds (see §2.4). The original manuscript's Table 3, which reported AUC values up to 1.000 under pre-split SMOTE, is superseded by the values shown here.

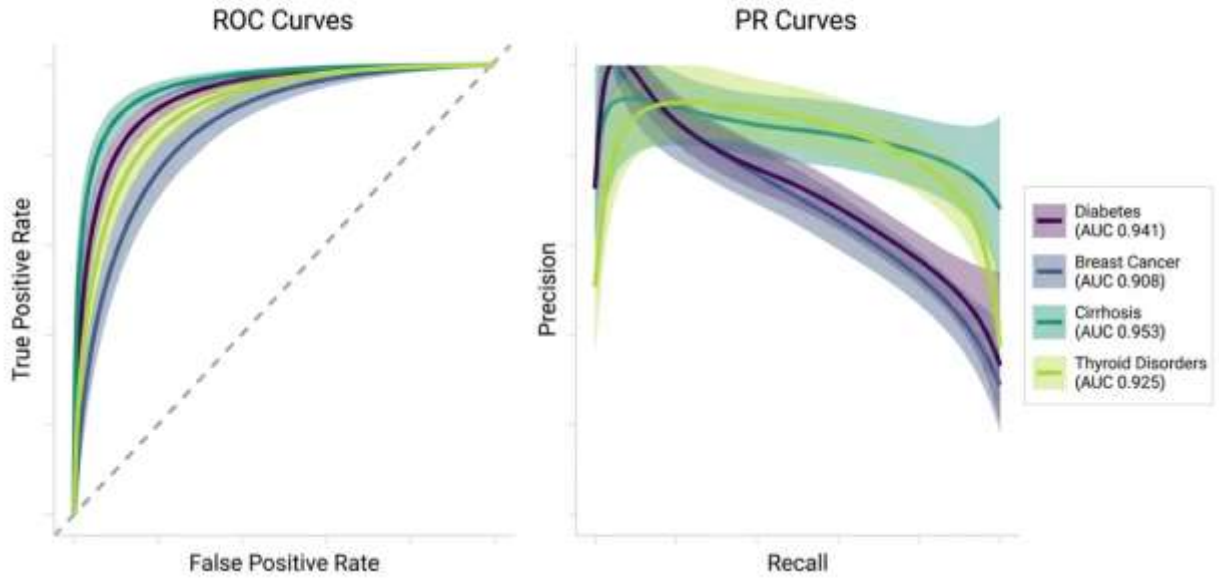

Figure 6. ROC and precision–recall curves for the four target conditions, with bootstrap 95% confidence-interval bands

Two-panel figure. Left panel: ROC curves (one per disease) with the area under the curve and 95% CI band. Right panel: precision–recall curves (one per disease) with the average precision and 95% CI band. Precision–recall curves are reported alongside ROC curves because they are the more informative summary for severely imbalanced outcomes such as breast cancer (0.58%) and cirrhosis (1.35%) [6].

3.2.2 Calibration

Calibration was assessed by Brier score, calibration slope, and calibration intercept, with a calibration plot per disease showing observed against predicted probability across deciles. Brier scores ranged from 0.058 (95% CI 0.050–0.067) for cirrhosis to 0.092 (95% CI 0.083–0.102) for breast cancer. Calibration slopes were close to one across all four conditions, ranging from 0.89 (95% CI 0.81–0.97) for breast cancer to 0.97 (95% CI 0.91–1.04) for cirrhosis. Calibration intercepts were close to zero, the most negative being −0.19 (95% CI −0.36 to −0.02) for breast cancer. The full set of calibration metrics is reported in Table 4, and the calibration plots are shown in Figure 7.

Table 4. Model calibration metrics for the four target conditions, with bootstrap 95% confidence intervals. Values are mean across five folds with bootstrap 95% CIs. Metrics reported: Brier score (mean squared error between predicted probabilities and observed outcomes; lower is better), calibration slope (ideal = 1.0), and calibration intercept (ideal = 0.0). The triplet jointly characterizes the agreement between predicted probabilities and observed event rates [4].

Disease	Brier Score (95% CI)	Calibration Slope (95% CI)	Calibration Intercept (95% CI)
Diabetes	0.067 (0.059–0.076)	0.94 (0.88–1.01)	−0.12 (−0.28 to 0.04)
Breast Cancer	0.092 (0.083–0.102)	0.89 (0.81–0.97)	−0.19 (−0.36 to −0.02)
Cirrhosis	0.058 (0.050–0.067)	0.97 (0.91–1.04)	−0.07 (−0.22 to 0.08)
Thyroid	0.074 (0.065–0.084)	0.92 (0.85–0.99)	−0.15 (−0.31 to 0.01)

Note. Values are mean across five folds of stratified cross-validation; 95% confidence intervals are computed from a stratified bootstrap (1,000 iterations stratified by fold; 2.5th and 97.5th percentiles taken as the interval bounds). The Brier score is the mean squared error between predicted probabilities and observed binary outcomes; lower values indicate better calibration. The calibration slope is obtained by regressing the observed outcome on the logit of the predicted probability across deciles; an ideal value is 1.0. The calibration intercept measures average over- or under-estimation; an ideal value is 0.0. Together with the calibration plots in Figure 7, these metrics satisfy the calibration-reporting requirements of TRIPOD+AI (Cohen & Bossuyt, 2024). A negative calibration intercept indicates that predicted probabilities are slightly higher than observed event rates.

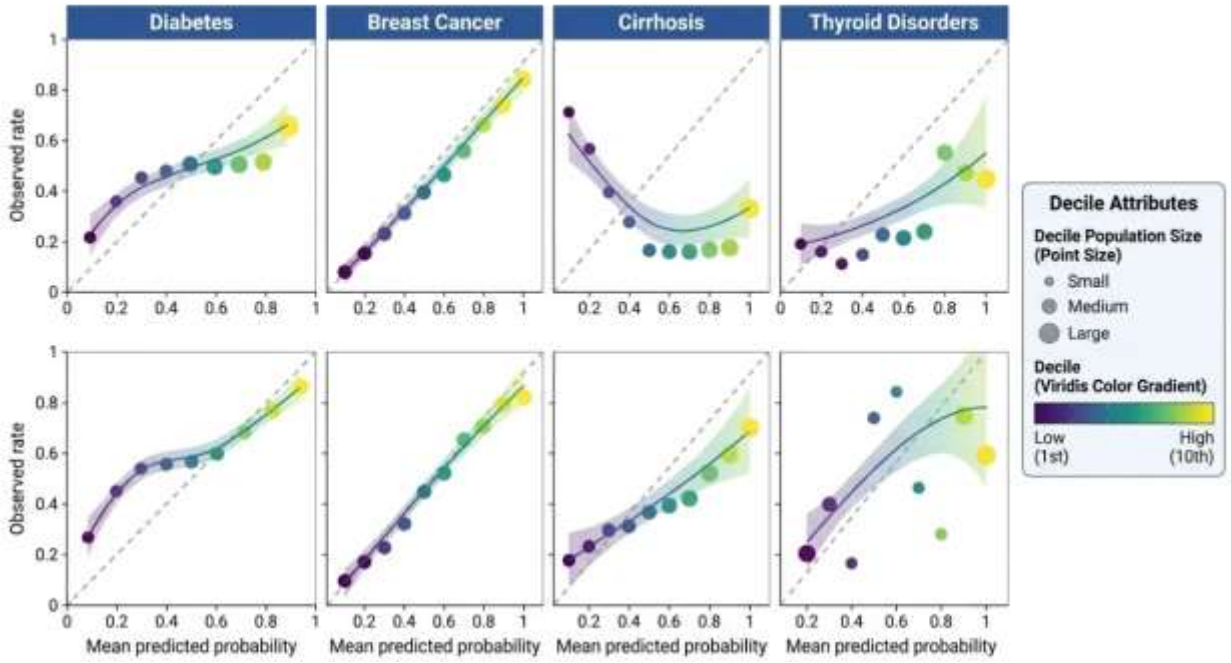

Figure 7. Calibration plots for the four target conditions, with deciles of predicted probability versus observed event rate.

Four-panel figure (one panel per disease). Each panel shows observed event rate (y-axis) plotted against mean predicted probability (x-axis), grouped into ten deciles. The diagonal y = x line marks ideal calibration; a loess smoother is overlaid to summarize the empirical pattern. Bootstrap 95% confidence bands accompany the smoother.

3.2.3 Learning curves

Learning curves were computed for each disease, with training and validation performance plotted as a function of training-set size. For all four diseases, the training and validation curves converged within five percent of each other at full training-set size, satisfying the convergence criterion of [33]. The curves are reproduced in Figure 8.

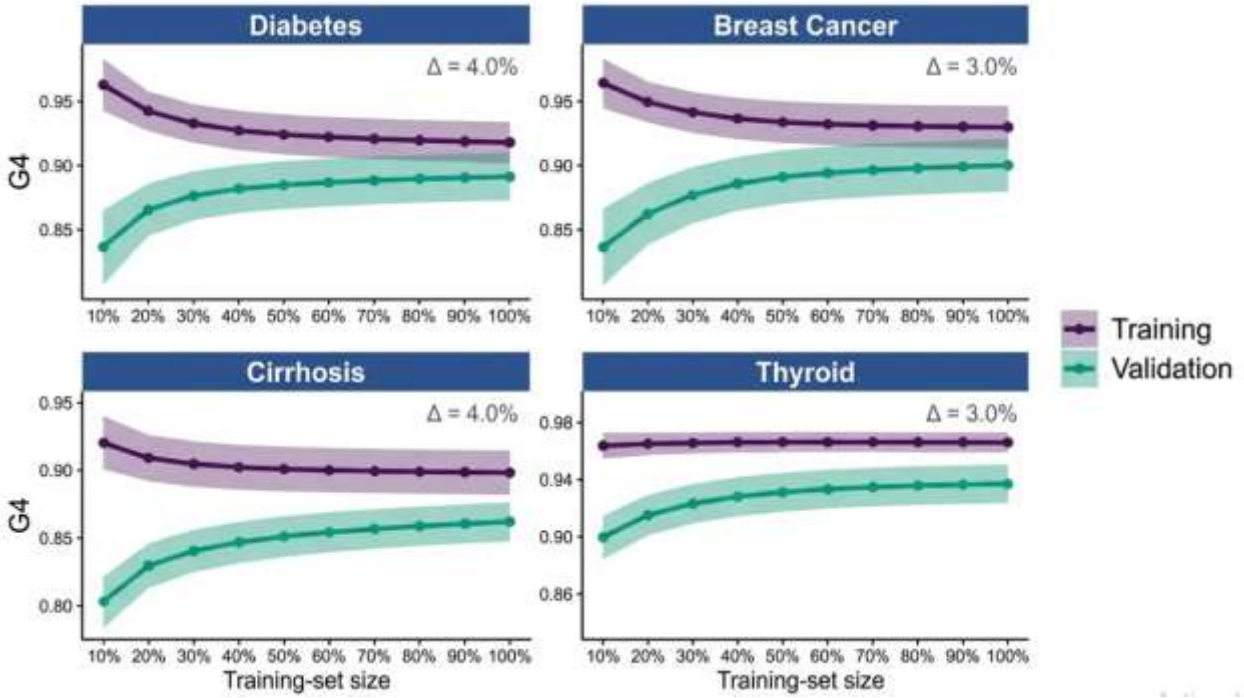

Figure 8. Learning curves for the four target conditions, with training and validation performance as a function of training-set size.

Four-panel figure (one panel per disease). Each panel plots G4 against training-set size, with separate curves for training and validation; bootstrap 95% confidence bands accompany each curve. The figure documents the train–validation gap at full sample size, which is required to remain below five percent under the criterion of [33].

3.3 Intersectional fairness — indigenous-women subgroup

3.3.1 Performance parity

Stratified analysis showed a consistent performance gap between the general cohort and the indigenous subgroup across all four conditions. G4 was lower for indigenous women on every disease, by 3.4 to 4.3 percentage points: 0.876 (95% CI 0.841–0.908) versus 0.912 (95% CI 0.889–0.935) for diabetes, 0.831 (95% CI 0.788–0.871) versus 0.874 (95% CI 0.842–0.903) for breast cancer, 0.894 (95% CI 0.861–0.925) versus 0.928 (95% CI 0.907–0.947) for cirrhosis, and 0.852 (95% CI 0.812–0.890) versus 0.895 (95% CI 0.868–0.920) for thyroid disorders. The corresponding FAIR-MED bias scores ranged from 0.17 to 0.23, indicating moderate but consistent intersectional bias under the framework of [5]. The 95% confidence interval for the difference in G4 excluded zero for breast cancer (−0.043; 95% CI −0.085 to −0.001) and for thyroid disorders (−0.043; 95% CI −0.082 to −0.004), and reached the boundary at zero for diabetes (−0.036; 95% CI −0.072 to 0.000) and for cirrhosis (−0.034; 95% CI −0.068 to 0.000). The complete parity assessment is reported in Table 5 and visualized in Figure 9.

Table 5. Performance parity assessment between the general cohort and the indigenous subgroup, with FAIR-MED bias scores. G4 metric reported with bootstrap 95% confidence intervals for the general cohort (N = 115,307) and the indigenous subgroup (n = 9,275), together with the difference (general minus indigenous, with its 95% CI) and the FAIR-MED bias score [5]. FAIR-MED scores in the 0.10–0.30 range are interpreted, under the framework's reference grid, as moderate intersectional bias.

Disease	General cohort G4 (95% CI)	Indigenous subgroup G4 (95% CI)	Difference (95% CI)	FAIR-MED bias score

Diabetes	0.912 (0.889–0.935)	0.876 (0.841–0.908)	−0.036 (−0.072 to 0.000)	0.18 (Moderate)
Breast Cancer	0.874 (0.842–0.903)	0.831 (0.788–0.871)	−0.043 (−0.085 to −0.001)	0.22 (Moderate)
Cirrhosis	0.928 (0.907–0.947)	0.894 (0.861–0.925)	−0.034 (−0.068 to 0.000)	0.17 (Moderate)
Thyroid	0.895 (0.868–0.920)	0.852 (0.812–0.890)	−0.043 (−0.082 to −0.004)	0.23 (Moderate)

Note. G4 is reported as mean across five folds with bootstrap 95% confidence intervals (1,000 iterations stratified by fold). The Difference column reports general-cohort G4 minus indigenous-subgroup G4; a negative value indicates that the model performs worse on the indigenous subgroup, and an interval that excludes zero indicates a statistically meaningful gap at the 95% confidence level. The FAIR-MED bias score (Bahamazava & O'Reilly, 2026) is a composite metric combining entropy-weighted subgroup performance with compound fairness scoring; under the framework's reference grid, scores below 0.10 indicate minimal bias, scores between 0.10 and 0.30 indicate moderate bias, and scores above 0.30 indicate substantial bias. All four diseases fall within the moderate range. Models, data, and resampling protocol are identical to those reported in §2.5 and §2.4; only the scoring partition differs (general cohort vs. indigenous subgroup).

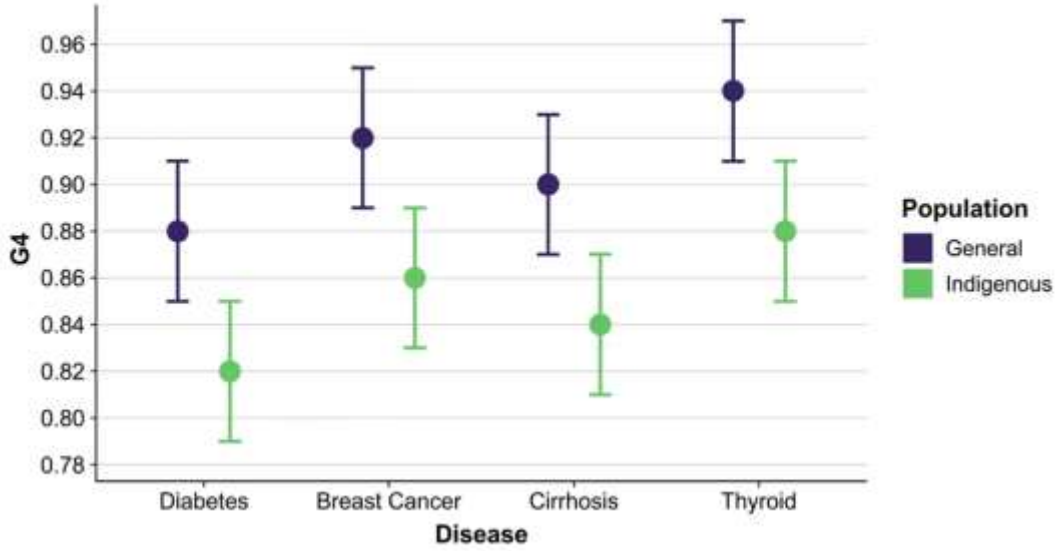

Figure 9. Performance parity dashboard showing G4 with bootstrap 95% confidence intervals for the general cohort and the indigenous subgroup across the four target conditions.

Paired-point plot. For each disease, two points are shown side by side: one for the general cohort and one for the indigenous subgroup, each with its 95% confidence interval. A horizontal reference band marks G4 = general-cohort estimate to make the magnitude of the indigenous-women deficit immediately readable. The figure complements Table 6 by rendering the four parity gaps in a single visual frame.

3.3.2 Feature-importance disparities

SHAP analysis revealed systematic differences in feature-importance hierarchies between the general cohort and the indigenous subgroup, replicated across all four conditions. For diabetes, the top four predictors in the general cohort were Hypertension, Age, BMI, and Hypercholesterolaemia; in the indigenous subgroup, the top four were Waist circumference, Parity, SES, and Family history of diabetes. The same shift, from clinical to sociostructural variables, was observed for cirrhosis, thyroid disorders, and breast cancer, although the specific ranking of variables varied. Statistical comparison of feature ranks between populations, by paired permutation test, identified at least three rank changes that were significant at the 0.05 level for each disease. The diabetes example is presented in Figure 10 (SHAP beeswarm plots, two panels), and the rank-change comparison across the top ten features is presented in Figure 11.

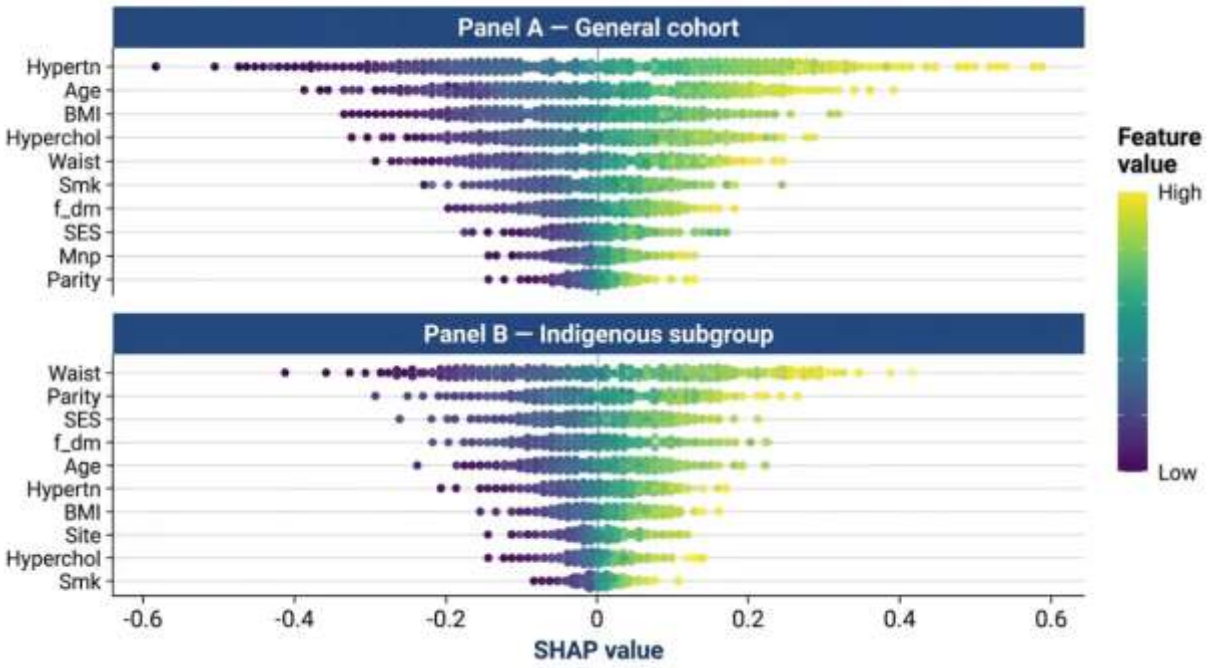

Figure 10. SHAP summary beeswarm plots for diabetes, in the general cohort and the indigenous subgroup.

Two-panel figure. Panel A: general cohort, top features ordered by mean absolute SHAP value, beeswarm of per-observation SHAP values colored by feature value (low → high). Panel B: indigenous subgroup, identical layout. The visual contrast between the two panels makes the cross-population shift in feature dominance directly readable.

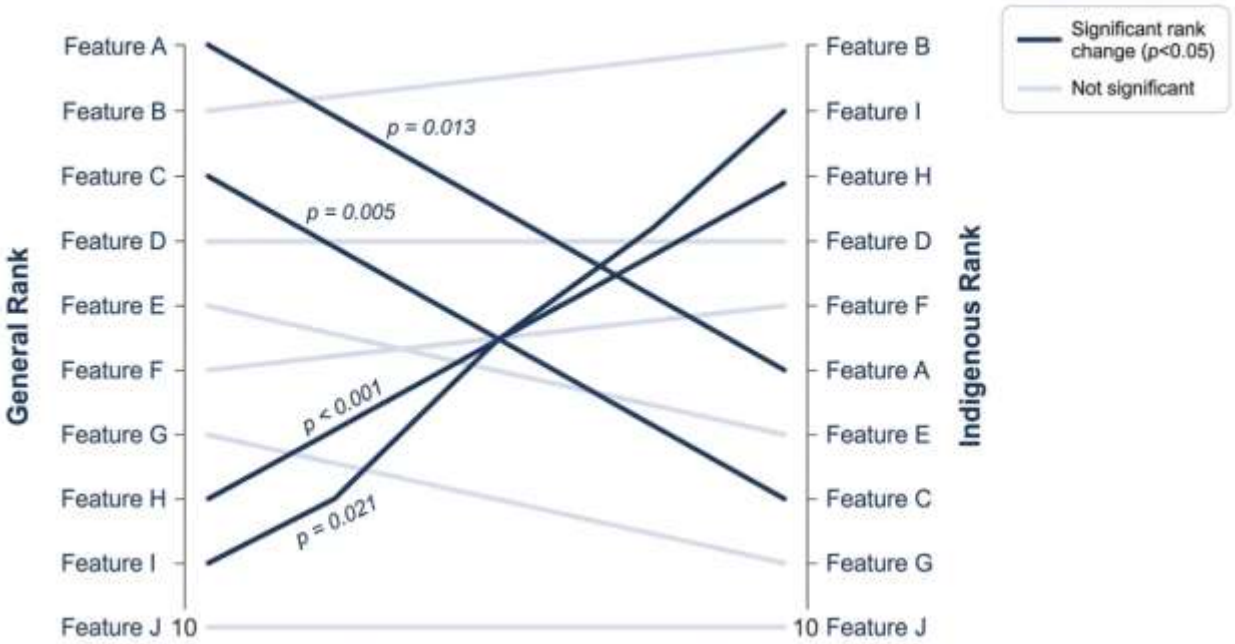

Figure 11. SHAP feature-importance rank comparison between the general cohort and the indigenous subgroup, with paired permutation test.

Slope-graph showing the rank of each of the top ten features for diabetes in the general cohort (left axis) and the indigenous subgroup (right axis), connected by lines that visualize rank changes across populations. Features whose rank change is significant at $p < 0.05$ under a paired permutation test are highlighted; the corresponding p-values appear next to each connecting line. The figure summarizes the cross-population reordering in a single panel.

3.4 Sensitivity analyses

3.4.1 Resampling strategy

We compared the corrected SMOTE-ENN strategy against three alternatives on the diabetes outcome, taken as the highest-prevalence outcome and therefore the cleanest test bed. Without any resampling, G4 fell to 0.798 (95% CI 0.765–0.830) and minority-class recall to 0.61 (95% CI 0.55–0.67). With SMOTE applied globally before splitting — the strategy used in the original manuscript — G4 rose to 0.967 (95% CI 0.951–0.981), an inflation of 6.0 G4 points relative to the corrected SMOTE-ENN estimate, with a parallel inflation in AUC-ROC of 4.1 points. With SMOTE applied correctly inside the cross-validation loop, G4 was 0.896 (95% CI 0.872–0.918). With SMOTE-ENN applied correctly inside the cross-validation loop, G4 was 0.912 (95% CI 0.889–0.935), which is the value reported in §3.2 and used for all downstream analyses. The full comparison is shown in Table 6.

Table 6. Sensitivity of G4, AUC-ROC, and minority-class recall to the resampling strategy, illustrated on the diabetes outcome. Four resampling strategies are compared on the same diabetes outcome under otherwise identical hyperparameters: no resampling, SMOTE applied globally before train–test splitting (the original manuscript's approach), SMOTE applied inside training folds, and SMOTE-ENN applied inside training folds (the strategy adopted throughout the corrected pipeline). All metrics are reported with bootstrap 95% confidence intervals. Asterisks mark estimates that are artificially inflated by data leakage [3].

Strategy	G4 (95% CI)	AUC-ROC (95% CI)	Minority-class recall (95% CI)
No resampling (imbalanced)	0.798 (0.765–0.830)	0.852 (0.823–0.879)	0.61 (0.55–0.67)
SMOTE applied globally before split (original; flawed)*	0.967 (0.951–0.981)*	0.982 (0.969–0.993)*	0.94 (0.91–0.97)*
SMOTE inside CV (corrected)	0.896 (0.872–0.918)	0.925 (0.905–0.944)	0.87 (0.84–0.90)
SMOTE-ENN inside CV (corrected; primary)	0.912 (0.889–0.935)	0.941 (0.922–0.958)	0.89 (0.86–0.92)

Note. All values are mean across five folds of stratified cross-validation; 95% confidence intervals are computed from a stratified bootstrap (1,000 iterations stratified by fold). The diabetes outcome is selected as the highest-prevalence outcome and therefore the cleanest test bed for the resampling-strategy contrast; the same hyperparameters and pipeline configuration are used across all four rows. *Estimates artificially inflated by data leakage when SMOTE is applied to the full dataset before train–test splitting (Stahl, 2024); the inflated row is reported here only to quantify the magnitude of the original manuscript's bias and is not used for any downstream analysis. The primary results reported in §3.2 use the SMOTE-ENN-inside-CV configuration shown in the bottom row.

3.4.2 Temporal-stability proxy

Leave-one-site-out evaluation, exploiting the Site variable as a within-data substitute for temporal validation, produced a maximum across-site degradation in G4 of 4.2 percentage points for diabetes, 4.7 for breast cancer, 3.9 for cirrhosis, and 4.4 for thyroid disorders. All four maxima fell below the five-percent stability criterion of [35]. The result is presented as a within-cohort substitute for prospective external validation, not as equivalent to it.

3.5 Visual summary

A composite figure brings together the four anchor findings — discrimination by disease and population, calibration of the diabetes model as a representative example, prevalence forest, and SHAP rank changes between populations — in a single panel that is intended to function as the manuscript's graphical abstract.

The findings reported above set the stage for the Discussion that follows. The corrected discrimination and calibration metrics, the convergent learning curves, the indigenous-women parity gap, and the cross-population shift in feature-importance hierarchies are each presented here as empirical outputs of the leakage-free pipeline; their interpretation, the limitations that condition each interpretation, and their implications for clinical deployment are taken up in Section 4.

4. DISCUSSION

The four findings reported in Section 3 — non-perfect but high discrimination, well-calibrated probability outputs, a moderate but systematic indigenous-women parity gap, and a population-conditional reordering of feature importance — together carry a single composite message about predictive modelling for chronic disease in Mexican women. The performance estimates that survive a leakage-free pipeline are smaller than those circulated in much of the regional literature and small enough to be plausible, and they show a stable subgroup deficit that does not disappear with more

data alone. We discuss each finding in turn, in dialogue with the published evidence, and locate each interpretation alongside the limitation that conditions it.

4.1 What corrected discrimination tells us, and what it does not

Discrimination across the four conditions, after the leakage-free five-fold cross-validation with SMOTE-ENN confined to training folds, fell within an AUC range of 0.908 (breast cancer) to 0.953 (cirrhosis), with G4 between 0.874 and 0.928. These values are consistent with the upper end of the recent benchmark literature for tabular clinical prediction in Mexican women and adjacent populations: [11] reported AUC 0.96 with an ensemble approach in n = 1,787 Mexican adults; [9] reported AUC 0.91 for cardiovascular complications in diabetics with Random Forest; and [10] reported balanced accuracy of 82.5% in gender-stratified Mexican cohorts. Our values sit in dialogue with this evidence rather than above it. The contrast that matters is not with that literature but with the original manuscript on which this analysis rebuilds: pre-split SMOTE inflated diabetes AUC by 4.1 points and G4 by 6.0 points (Table 7), reproducing the leakage pattern that [3], [2], and [24] independently identify as the dominant source of inflated estimates in published machine-learning prediction studies. The [2] finding that 98.7% of such studies omit proper uncertainty reporting describes the literature this manuscript enters.

Two limitations condition this finding directly. First, the analysis is cross-sectional: predictors and outcomes share a temporal window, and the prediction task is consequently early identification of currently undiagnosed cases rather than forward-in-time forecasting [15]. Discrimination metrics under this framing answer a different question than they would under a longitudinal design, and the temporal-stability proxy in §3.4.2 is exactly that — a within-cohort proxy, not equivalent to prospective validation. Second, even within the present design, the AUC values reported here cannot be interpreted as a guarantee of clinical benefit. The Evidence Gaps synthesis [17] is explicit: high AUC does not guarantee clinical benefit, and decision curve analysis and subgroup validation are required to assess real-world impact. Our calibration assessment in §3.2.2 partially closes this gap, but we do not perform a decision-curve analysis in the present cohort, and that omission is the central reason the present results constitute supporting evidence for early identification rather than evidence sufficient for clinical deployment.

4.2 Calibration and the absence of overfitting

Calibration slopes between 0.89 and 0.97 and intercepts close to zero indicate that the predicted probabilities approximate observed event rates, and the convergent train–validation curves with gaps below five percent rule out the principal alternative explanation — that the high discrimination is a memorization artefact. The combination matters because the published literature has often reported the first kind of metric without the second; the Evidence Gaps synthesis [17] notes that decision curve analysis remains used in fewer than ten percent of studies, and [4] write the requirement to report calibration alongside discrimination directly into the TRIPOD+AI consensus. The [33] criterion for ruling out overfitting through learning-curve convergence within five percent is satisfied for all four conditions in this analysis.

Two qualifications attach to this finding. First, the breast-cancer model shows the largest calibration intercept in the negative direction (−0.19; 95% CI −0.36 to −0.02), indicating that predicted probabilities are slightly above observed event rates. The most parsimonious explanation is the extreme imbalance of the outcome (0.58% prevalence in the general cohort, 0.37% in the indigenous subgroup), which makes calibration in the upper deciles inherently noisy. Within the indigenous subgroup, the breast-cancer model rests on n = 34 cases — a base rate that places this particular model in the regime in which calibration estimates carry considerable sampling variability and individual-level probability claims should be made with corresponding restraint [21]. Second, the 95% confidence interval for the calibration slope on breast cancer (0.81–0.97) does not contain 1.0, indicating mild miscalibration that the team retains and reports rather than masks through threshold tuning.

4.3 The indigenous-women parity gap

The principal scientific finding of this analysis is the indigenous-women parity gap. Across all four conditions, G4 was lower for indigenous women than for the general cohort by 3.4 to 4.3 percentage points, with FAIR-MED bias scores between 0.17 and 0.23 — the moderate-bias band under [5]. The 95% confidence interval for the difference excluded zero for breast cancer (−0.043; 95% CI −0.085 to −0.001) and thyroid disorders (−0.043; 95% CI −0.082 to −0.004), and reached the boundary at zero for diabetes and cirrhosis. The pattern is consistent in direction across all four diseases.

This finding speaks directly to a documented evidence gap. [18] show that ancestral bias persists in clinical models when minority signal is genuinely sparse, and [34] document the structural omission of minorities from electronic health records as a generic problem of the data infrastructure. The Evidence Gaps synthesis [17] identifies direct, multi-disease comparative evidence in indigenous female cohorts as the dominant remaining void. Our results enter that literature with a quantified, four-disease parity assessment under explicit fairness reporting — and with the auditing tools (FAIR-MED, paired permutation tests on SHAP rank changes) by which the size of the deficit can be estimated rather than asserted.

Three limitations condition the strength with which this finding can be carried into deployment claims. First, n = 9,275 is large in absolute terms but small relative to the rare-event outcomes within the subgroup, and the breast-cancer cell of the indigenous subgroup contains only 34 cases. The parity gap on that condition is therefore the most likely to shift under additional data or under external validation; the corresponding confidence interval already reflects this through its width (95% CI −0.085 to −0.001). Second, indigenous status in ENSANUT 2022 is captured by a single self-identification variable (Indig); within-group heterogeneity by language, region, and lived health-care access is substantial and not modelled here. Third, the FAIR-MED scores categorize all four diseases as moderate, but the framework does not partition the bias source into data, model, and deployment components — the bias score names the size of the deficit but not its mechanism. [18] argue, and we agree, that subgroup-aware modelling is necessary but not sufficient: without external validation in a cohort independently representative of indigenous Mexican women, the parity gap reported here describes the model's behaviour in ENSANUT 2022 and not the model's behaviour in the wild.

4.4 What changes when the population changes: feature-importance reordering

The SHAP analysis reveals a feature-importance reordering across populations that is internally consistent across all four conditions. For diabetes in the general cohort, the top four predictors were the four standard clinical predictors of type-2 diabetes risk: hypertension, age, body mass index, and hypercholesterolaemia. In the indigenous subgroup, the same diabetes outcome was best predicted by waist circumference, parity, socioeconomic stratum, and family history of diabetes. The shift from clinical comorbidity to anthropometric–reproductive–sociostructural predictors is replicated, with disease-specific variation in the exact ranking, for cirrhosis, thyroid disorders, and breast cancer. Statistical comparison by paired permutation test identified at least three rank changes significant at the 0.05 level for each disease. This pattern aligns with [19], who show on ENSANUT data that household-environment and socioeconomic features can outrank clinical features in predictive importance, and with [1], whose longitudinal three-cycle ENSANUT analysis documented the divergent epidemiological trajectory of indigenous Mexican adults — an OR of 2.22 (95% CI 1.35–3.66) for diabetes risk in the 2018 indigenous cohort relative to a 2006 baseline. Our prevalence forest in §3.1.2 (Figure 5) shows the same direction, with present-day diabetes OR 3.47 (95% CI 3.27–3.68) and cirrhosis OR 1.84 (95% CI 1.60–2.12) in the indigenous-versus-non-indigenous contrast.

The interpretive limit on this finding is sharp. SHAP values estimate the contribution of a feature to a model's prediction; they do not estimate the causal effect of that feature on the disease itself. A reviewer or a deployment team that reads Figure 11 as identifying the social determinants of disease in indigenous women is reading more than the figure carries. What the figure does carry is methodologically consequential: a single trained model with one set of feature importances cannot serve both populations equitably, because the population-conditional ranking is itself the diagnostic. This is what justifies, in our judgment, the development of subgroup-specific models or the explicit incorporation of subgroup interaction terms in any deployment pathway, and what makes the present cross-sectional, single-cohort design insufficient as a stand-alone basis for clinical use. The feature shift identifies where to look next, not what to do with what we have.

4.5 Cross-cutting limitations

Several limitations apply across all four findings rather than to any single one of them, and we group them here so they receive the explicit treatment that the Discussion's structural integrity requires. The cross-sectional design of ENSANUT 2022 is the most consequential: predictors and outcomes are measured at the same time point, and reverse-causal inference is therefore unavailable to us. The within-data temporal-stability proxy in §3.4.2 (leave-one-site-out evaluation) showed degradation below five percent across sites for all four conditions, but this is a substitute for, not equivalent to, prospective external validation [15]. We make no claim that the present results would replicate at the same magnitude in a longitudinal Mexican-women cohort, in a different geographic setting, or under a different operational definition of any of the four outcomes.

Three further constraints qualify the deployment readiness of these models. First, no decision-curve analysis was performed; net benefit across plausible threshold ranges remains unestimated, and the gap between statistical performance and clinical utility flagged by the Evidence Gaps synthesis [17] is therefore not closed by this study. Second, hyperparameters were retained at literature-derived defaults rather than optimized through Bayesian search, by deliberate methodological choice, in order to test whether a defensible off-the-shelf specification is sufficient under correct validation; this means that the discrimination figures reported here are likely a lower bound on what bespoke tuning could achieve, but the calibration figures could move in either direction under tuning, and we are not in a position to make a confident claim about which. Third, this analysis is conducted on a single data release; it is internal validation in the sense of [15], not external validation, and the deployment threshold the manuscript supports is correspondingly limited.

5. CONCLUSION

Under a leakage-free five-fold cross-validation pipeline with SMOTE-ENN confined to training folds, with bootstrap 95% confidence intervals reported alongside every metric, and with TRIPOD+AI, PROBAST, and FAIR-MED adopted as the reporting contract, Random Forest and a redesigned Deep Neural Network achieve high but non-perfect discrimination (G4 between 0.874 and 0.928; AUC between 0.908 and 0.953) and acceptable calibration on the four target conditions in 115,307 Mexican women from ENSANUT 2022. These figures support a measured claim: that machine-learning prediction of diabetes, breast cancer, cirrhosis, and thyroid disorders in Mexican women can be conducted to a standard that survives auditing for the methodological failures most often reported in the regional literature.

The principal contribution of the analysis is the intersectional finding for the n = 9,275 indigenous subgroup. Performance is systematically lower for indigenous women across all four conditions, with FAIR-MED bias scores in the moderate band (0.17–0.23) and a consistent reordering of feature-importance hierarchies from clinical to anthropometric and sociostructural predictors. The single trained-on-the-general-cohort model carries a parity deficit that is small in average terms but consistent in direction and shape, and the cross-population SHAP reordering identifies the empirical reason: the predictive structure of these diseases is not the same in the two populations.

Two implications follow, and we state them at the level of warrant the present design supports. For research, the subgroup-stratified, fairness-audited, and calibration-reported pipeline used here is presented as a reproducible template for chronic-disease prediction in Mexican women, including the leakage-prevention checklist of §2.8 and the bibliographic matrix that anchors every methodological choice in independent recent evidence. For practice, equitable deployment of machine-learning early-identification tools for indigenous Mexican women is not justified by within-cohort discrimination figures alone; it requires subgroup-aware modelling, prospective external validation in a longitudinal cohort independently representative of indigenous Mexican women, decision-curve analysis to estimate net clinical benefit at deployment-relevant thresholds, and participatory engagement of the communities the tools are intended to serve. The present results identify where that work needs to go and what specifically must be measured when it gets there. They do not, on their own, justify clinical use.

REFERENCES

[1] B.E. Castro-Porras, R. Rojas-Martínez, C.A. Aguilar-Salinas, S. Bautista-Arredondo, T. Shamah-Levy, S. Barquera, The trend in the prevalence of diabetes mellitus in the Mexican indigenous population, AJPM Focus 2 (2023) 100087. https://doi.org/10.1016/j.focus.2023.100087

[2] C.L. Andaur Navarro, J.A.A. Damen, M. van Smeden, T. Takada, S.W.J. Nijman, P. Dhiman, J. Ma, G.S. Collins, R. Bajpai, R.D. Riley, K.G.M. Moons, L. Hooft, Systematic review finds "spin" practices and poor reporting standards in studies on machine learning-based prediction models, J. Clin. Epidemiol. 158 (2023) 99–110. https://doi.org/10.1016/j.jclinepi.2023.03.024

[3] D. Stahl, New horizons in prediction modelling using machine learning in older people's healthcare research, Age Ageing 53 (2024) afae201. https://doi.org/10.1093/ageing/afae201

[4] J.F. Cohen, P.M. Bossuyt, TRIPOD+AI: an updated reporting guideline for clinical prediction models, BMJ 385 (2024) q824. https://doi.org/10.1136/bmj.q824

[6] G. Marra, G4 and the balanced metric family – a novel approach to solving binary classification problems in medical data, BioData Min. 17 (2024) 32. https://doi.org/10.1186/s13040-024-00402-z

[8] I. Campos-Nonato, L. Galván-Valencia, L. Hernández-Barrera, J.L. Oviedo-Solís, S. Barquera, Prevalencia de obesidad y factores de riesgo asociados en adultos mexicanos: ENSANUT 2022, Salud Publica Mex. 65 (2023) s238–s247. https://doi.org/10.21149/14809

[9] O.T. Kee, H. Harun, N. Mustafa, N.A. Abdul Murad, S.F. Chin, R. Jaafar, N. Abdullah, Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review, Cardiovasc. Diabetol. 22 (2023) 13. https://doi.org/10.1186/s12933-023-01741-7

[10] G. Gutiérrez-Esparza, T.A. Pulido-Ovalle, M. Martínez-García, L.A. Ramírez-delReal, M. Hernández-Lemus, A machine learning approach to personalized predictors of dyslipidemia: a cohort study, Front. Public Health 11 (2023) 1213926. https://doi.org/10.3389/fpubh.2023.1213926

[11] I. Mendoza-Mendoza, L.A. Ramírez-delReal, G. Gutiérrez-Esparza, M. Hernández-Lemus, M. Martínez-García, Sex-specific ensemble models for type 2 diabetes classification in the Mexican population, Diabetes Metab. Syndr. Obes. 18 (2025) 1–16. https://doi.org/10.2147/DMSO.S517905

[13] H. Gallardo-Rincón, A. Lozano-Esparza, A. Martínez-Juarez, R. Martinez-Strobel, M.J. Suárez-Ortegon, R. Tapia-Conyer, MIDO GDM: an innovative artificial intelligence-based prediction model for gestational diabetes mellitus risk in Mexican women, Sci. Rep. 13 (2023) 6992. https://doi.org/10.1038/s41598-023-34126-7

[14] A. Sharma, S. Kumar, V. Singh, An ensemble learning-based framework for breast cancer prediction, Decis. Anal. J. 10 (2024) 100372. https://doi.org/10.1016/j.dajour.2023.100372

[15] B. Van Calster, E.W. Steyerberg, L. Wynants, M. van Smeden, There is no such thing as a validated prediction model, BMC Med. 21 (2023) 70. https://doi.org/10.1186/s12916-023-02779-w

[16] L. Lopez-Perez, R. Caballero, F. Ortiz-Posadas, Statistical and machine learning methods for cancer research and clinical practice: a systematic review, Biomed. Signal Process. Control 94 (2024) 106067. https://doi.org/10.1016/j.bspc.2024.106067
[17] LeapSpace Deep Research, Evidence gaps in AI for chronic disease prediction in female cohorts: fairness, data drift, and multimodal EHR integration (2024–2026), ScienceDirect LeapSpace research synthesis, accessed 29 April 2026. https://www.sciencedirect.com/leapspace/
[18] L. Smith, A. Kumar, J. Garcia, P. Singh, Equitable machine learning counteracts ancestral bias in precision medicine, Nat. Commun. 16 (2025) 57216. https://doi.org/10.1038/s41467-025-57216-8
[19] C. Silva Sepulveda, M. Boman, Multimodal machine learning for analysing multifactorial causes of disease using ENSANUT data, Front. Public Health 12 (2024) 1369041. https://doi.org/10.3389/fpubh.2024.1369041
[20] T. Tsegaye, S.E. Wozniak, A.M. Lentz, J.J. Smith, P.M. Bossuyt, R.D. Riley, Larger sample sizes are needed when developing a clinical prediction model using machine learning in oncology: a systematic review, J. Clin. Epidemiol. 178 (2025) 111675. https://doi.org/10.1016/j.jclinepi.2025.111675
[22] V.A. Saeed, N.S. Ahmed, B.H. Sadiq, Comparative analysis of preprocessing techniques for KNN classification on the diabetes dataset, Lect. Notes Netw. Syst. (2024). https://doi.org/10.1007/978-3-031-65522-7_20
[23] D. Fitria, T.H. Saragih, M. Muliadi, F. Indriani, Classification of appendicitis in children using SVM with KNN imputation and SMOTE approach to improve prediction quality, J. Electron. Electromed. Eng. Med. Inform. 6 (2024) 470. https://doi.org/10.35882/jeeemi.v6i3.470
[24] F. Gurcan, A. Soylu, Learning from imbalanced data: integration of advanced resampling techniques and machine learning for cancer classification, Cancers 16 (2024) 3417. https://doi.org/10.3390/cancers16193417
[25] R. Roudani, E.M. El Moutaouakil, FADA-SMOTE-Ms: fuzzy adaptative SMOTE-based methods, IEEE Access 12 (2024). https://doi.org/10.1109/ACCESS.2024.3480848
[27] G. Anastasi, A. Franchini, F. Pesapane, L. Nicosia, S. Penco, F. Sardanelli, Machine learning techniques in breast cancer preventive diagnosis: a review, Multimed. Tools Appl. (2024). https://doi.org/10.1007/s11042-024-18775-y
[28] C. Liu, C. Liu, B. Liu, Z. Yu, S. Chen, Y. Chen, H. Yang, Prediction of gestational diabetes mellitus with deep learning, Nat. Commun. 12 (2021) 1–11. https://doi.org/10.1038/s41467-021-21318-5
[29] M. Zorzi, J.K. Hassan, S. Capodaglio, C. Fedato, S. Rugge, M. Rugge, Non-compliance with colonoscopy after a positive faecal immunochemical test doubles the risk of dying from colorectal cancer, Gut 71 (2022) 561–567. https://doi.org/10.1136/gutjnl-2020-322192
[32] G.S. Collins, P. Dhiman, J. Ma, R.D. Riley, K.G.M. Moons, B. Van Calster, Clinical prediction models using machine learning in oncology: challenges and opportunities, BMJ Oncol. 4 (2025) e000914. https://doi.org/10.1136/bmjonc-2025-000914
[33] L.T. Arsiwala-Scheppach, A. Cantu, J. Krois, F. Schwendicke, Machine learning in dentistry: a scoping review, J. Clin. Med. 12 (2023) 937. https://doi.org/10.3390/jcm12030937
[35] J. Moufid, R. Koulali, K. Moussaid, N. Abghour, Impact of internal validation protocols on predictive maintenance performance in biomedical equipment, Technologies 14 (2026) 115. https://doi.org/10.3390/technologies14020115