

A Bayesian Neural Network Prediction of Type 1 Diabetes Mellitus on the TEDDY Cohort

Mwagandi Kalume Kiponda¹, Aggrey Onago Adem¹, Leonard Kiti Alii²,
Michael Mbindyo Munywoki¹, Lazarus K. Mramba³, Martin K. Haro¹, Josephat Mutembei¹

¹Department of Mathematics and Physics, Technical University of Mombasa, Mombasa, Kenya

²Department of Mathematics and Statistics, Pwani University, Kilifi, Kenya

³Health Informatics, University of Florida, Tampa, Florida

Email: kipondamwagandi@gmail.com, ademaggrey@tum.ac.ke, l.ali@pu.ac.ke, mmunywoki@tum.ac.ke, Lazarus.Mramba@epi.usf.edu, haro.haroka.kai@gmail.com, jkathanzu@tum.ac.ke

How to cite this paper: Kiponda, M.K., Adem, A.O., Alii, L.K., Munywoki, M.M., Mramba, L.K., Haro, M.K. and Mutembei, J. (2026) A Bayesian Neural Network Prediction of Type 1 Diabetes Mellitus on the TEDDY Cohort. *Journal of Biosciences and Medicines*, 14, 195-217.
<https://doi.org/10.4236/jbm.2026.1410014>

Received: August 28, 2026

Accepted: October 20, 2026

Published: October 23, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Predicting the onset of Type 1 Diabetes Mellitus (T1DM) in genetically at-risk children presents a fundamental statistical challenge: disease prevalence in prospective cohorts rarely exceeds 4% - 5%, producing severe class imbalance that causes standard classifiers to optimise for the majority class while catastrophically failing the minority. We address this through a semiparametric Bayesian Neural Network (BNN) evaluated on 8556 children from the multi-centre TEDDY cohort (331 T1DM cases; 3.9% prevalence). The BNN integrates a Gaussian-prior regularised feedforward architecture with Bayesian weight distributions, optimised via a Weighted Surrogate Evidence Lower Bound (W-ELBO) that embeds class-specific cost penalties directly into the variational objective. Uncertainty was quantified through Monte Carlo Dropout with $T = 50$ stochastic forward passes. In a prespecified 6×3 factorial design comparing six model families under three imbalance-mitigation conditions, the BNN trained with raw data and inverse-frequency class weights achieved the highest F1 score (0.6426) and Matthews Correlation Coefficient (MCC = 0.6401), with Sensitivity of 80.81% (95% CI: 71.66% - 88.03%) and Specificity of 97.16% (95% CI: 96.43% - 97.78%); calibration was formally assessed via the Brier score (0.0510) together with Expected/Maximum Calibration Error and Cox calibration statistics (\$4.5), and quantified epistemic uncertainty (mean predictive SD = 0.0671) — capabilities absent in all deterministic comparators. SMOTE oversampling was found to be redundant when combined with class-weighted loss (producing mathematically identical results to four decimal places) and actively harmful for kernel-based SVMs (0% Sensitivity at the default 0.5 threshold despite AUC = 0.8644), constituting a novel methodological finding. These results support a paradigm shift from accuracy-

optimised deterministic pipelines to sensitivity-prioritised, uncertainty-aware semiparametric frameworks for low-prevalence clinical prediction.

Keywords

Bayesian Neural Network, Type 1 Diabetes Mellitus, Class Imbalance, Semiparametric Modelling, Weighted ELBO, Monte Carlo Dropout, Epistemic Uncertainty, Accuracy Paradox, TEDDY Cohort, Paediatric Screening

1. Introduction

Type 1 Diabetes Mellitus (T1DM) is a progressive autoimmune condition characterised by immune-mediated destruction of pancreatic β -cells, resulting in lifelong insulin dependence. The International Diabetes Federation estimates that 537 million individuals aged 20 - 79 currently live with some form of diabetes, projected to reach 783 million by 2045 [1]. Unlike Type 2 Diabetes, T1DM imposes a disproportionate burden on children, as its onset is driven by genetic susceptibility—particularly high-risk Human Leukocyte Antigen (HLA) genotypes—in combination with early-life environmental triggers [2]. Early identification of at-risk children, prior to symptomatic onset and the development of diabetic ketoacidosis, represents the most clinically actionable window for intervention, potentially preserving residual β -cell function [3].

Despite substantial investments in prospective cohort research—exemplified by The Environmental Determinants of Diabetes in the Young (TEDDY) consortium, which followed 8,676 genetically at-risk children across six clinical centres from 2004 to 2018—the translation of cohort data into clinically deployable predictive tools has been impeded by a structural statistical obstacle: severe class imbalance. In the TEDDY cohort, only 331 of 8556 children (3.9%) developed T1DM during follow-up, producing a 1:25 minority-to-majority ratio. This imbalance renders standard machine learning classifiers counterproductive for clinical screening. By minimising standard cross-entropy loss, they implicitly optimise for the majority class, achieving nominal accuracies exceeding 96% while systematically failing to detect the very children the model is designed to identify [4].

Class Imbalance (Figure 1) visualises this severe imbalance, which underpins the methodological motivation for this study.

The Accuracy Paradox is further exacerbated by a second fundamental limitation of standard models: overconfident point-estimate predictions. In paediatric T1DM screening, a clinician requires not only a predicted risk category but also a quantification of the model's epistemic doubt. A prediction derived from a well-represented patient profile warrants a different clinical response than one derived from an ambiguous, underrepresented case. Standard neural networks, random forests, and support vector machines cannot make this distinction; they produce a single probability score without an associated confidence measure.

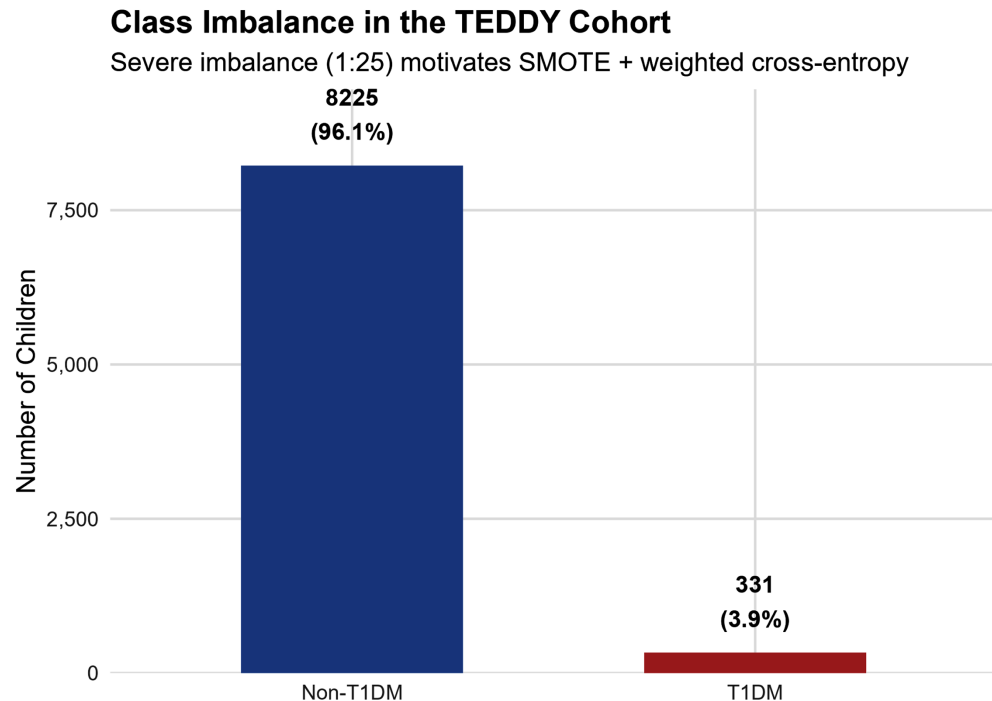


Figure 1. Class imbalance in the TEDDY cohort: 331 T1DM cases (3.9%) vs. 8225 non-T1DM (96.1%). The 1:25 ratio motivates cost-sensitive learning and SMOTE.

The present study addresses both deficiencies through a semiparametric Bayesian Neural Network (BNN). By treating network weights as probability distributions rather than fixed point estimates, the BNN inherits both the nonparametric flexibility of deep learning (universal function approximation; [5]) and the parametric discipline of Bayesian inference (posterior regularisation, uncertainty quantification). A novel Weighted Surrogate ELBO (W-ELBO) embeds class-specific cost penalties directly into the variational objective, structurally redirecting the optimisation landscape toward the minority class without relying on synthetic data augmentation.

Primary contribution: We demonstrate that i) a semiparametric BNN with W-ELBO achieves clinically viable sensitivity (80.81%) while maintaining specificity (97.16%) and formally assessed calibration (\$4.5); ii) SMOTE is redundant when combined with class-weighted loss and can be actively harmful for kernel-based SVMs under specific configurations; and iii) epistemic uncertainty quantification via MC Dropout provides clinical information unavailable from any deterministic model evaluated.

2. Related Work

2.1. Machine Learning in Diabetes Prediction

The majority of machine learning work in diabetes prediction relies on the PIMA Indians Diabetes Dataset [6], which captures T2DM in adult women using fewer than ten clinical features. While methodologically useful for benchmarking, this

dataset does not reflect the paediatric, autoimmune, multi-factorial aetiology of T1DM. Models achieving 83% - 93% accuracy on PIMA [7] [8] cannot be extrapolated to T1DM cohorts characterised by HLA genotypes, persistent islet autoantibodies, and longitudinal environmental exposures. Prospective T1DM risk prediction has instead been advanced through birth-cohort screening: children seroconverting to multiple islet autoantibodies progress to clinical disease at a rate of approximately 10% per year and nearly 70% within ten years [9], and a three-stage framework (Stage 1 autoantibody positivity with normoglycaemia; Stage 2 dysglycaemia; Stage 3 clinical disease) now governs pre-symptomatic risk stratification [10]. The present model is positioned within this framework as a Stage-1-to-Stage-3 progression tool (§3.1).

2.2. Bayesian Neural Networks and Uncertainty Quantification

Bayesian Neural Networks extend standard feedforward networks by placing prior distributions over network weights, enabling posterior inference over predictions. Standard deep learning models yield deterministic point estimates, which have been shown to be poorly calibrated and overconfident when extrapolating beyond their training distribution [11]. Gal and Ghahramani [12] demonstrated that Monte Carlo (MC) Dropout—retaining dropout layers active during inference and averaging across T stochastic forward passes—provides a computationally tractable approximation to Bayesian inference. Mitros and Mac Namee [13] subsequently showed that BNNs are better calibrated and more resilient to out-of-sample instances than deterministic deep networks. However, BNN applications to heavily imbalanced epidemiological T1DM cohorts remain largely unexplored.

2.3. Class Imbalance and the Accuracy Paradox

He and Garcia [4] systematically characterised the failure modes of standard classifiers under severe class imbalance, introducing the concept of the Accuracy Paradox. Historically, the literature has split into two primary remedies: algorithmic-level cost-sensitive learning (where misclassification costs are weighted asymmetrically) and data-level resampling techniques like SMOTE [14]. While SMOTE has become the de facto standard in applied machine learning, its interaction with different model families—particularly kernel-based SVMs and neural networks—has received insufficient critical attention in the clinical AI literature. Chicco and Jurman [15] demonstrated that F1 Score and MCC are more informative than Accuracy or AUC for imbalanced binary classification, a finding that motivated the present study's choice of primary evaluation metrics.

3. Methods

3.1. Study Design and Data Source

This study employed a retrospective analysis of a prospective longitudinal cohort design. Data were sourced from the TEDDY study [2] [16] [17], which enrolled

8676 infants at birth identified as having high-risk HLA genotypes (DR3/4, DR4/4, DR3/3, DR4/8) across six clinical centres in the United States, Finland, Germany, and Sweden between 2004 and 2018. Participants were monitored from age 3 months to 15 years. After filtering for high-risk HLA genotypes, the analytic sample comprised 8,556 children (331 T1DM cases; 3.9% prevalence). Formal data access was obtained under NIDDK dbGaP Data Access Request ID 24222. Ethical approval was granted by the Technical University of Mombasa and NACOSTI (License: NACOSTI/P/24/34090; Ethics: TUM-SERC MSC/010/2024). Analysis and reporting followed the TRIPOD + AI statement [18].

Prediction time point and horizon. The prediction task is defined as follows. The outcome is clinical T1DM (Stage 3) diagnosed by 15 years of age or end of follow-up. The prediction index is each child's last completed screening visit before prediction. Baseline predictors (HLA genotype, first-degree relative status, sex, country, maternal age at delivery, gestational age) are fixed at or before birth; multiple persistent confirmed autoantibodies are *time-varying* and enter the model as ascertained at the index visit (TEDDY visit schedule: 3-monthly to 48 months, thereafter 3 - 6 monthly, with autoantibody-positive children retaining 3-monthly visits), and were recorded strictly pre-diagnosis for all 331 cases. The model therefore estimates the risk of progression conditional on the child's current immunological state—consistent with the ADA/ISPAD staging framework [10]—rather than anticipating seroconversion itself; the corresponding pre-seroconversion (autoantibody-negative) stratified analysis is identified as a priority in §5.4.

3.2. Data Preprocessing and Feature Engineering

Raw clinical and environmental data require careful transformation to satisfy the mathematical assumptions of deep learning architectures while preserving clinical interpretability. High-dimensional HLA genetic markers, originally represented as tens of sparse numerical codes, were consolidated into five clinically interpretable risk categories based on established haplotype risk stratifications. Similarly, longitudinal maternal alcohol consumption across trimesters was aggregated into a unified, cumulative exposure metric.

Continuous variables (e.g., maternal age, gestational age) were subjected to Z-score standardisation (centering and scaling to unit variance). This step was specifically critical for distance-based and gradient-descent models—the SVM, standard NN, and BNN. Z-score scaling prevents features with larger magnitudes from dominating the gradient updates, ensuring rapid and stable convergence of the optimisation algorithms [19]. Categorical variables were transformed into binary dummy matrices via one-hot encoding. To prevent information leakage from the test set into the training pipeline, missing values in continuous predictors ($\leq 2\%$) were imputed using the median of the *training* set, applied *after* the train-test split. Furthermore, all data-dependent operations—bivariate feature screening, median imputation, Z-score scaling, one-hot encoding, SMOTE resampling,

and hyperparameter selection—were performed independently within each training fold of the cross-validation procedure, so that no information from the held-out test set entered predictor selection or preprocessing.

3.3. Predictors and Outcome

The binary outcome was T1DM diagnosis (coded 1 = T1DM, 0 = non-T1DM) per ADA 2011 criteria applied to TEDDY's standardised case report forms. To prevent the curse of dimensionality and reduce noise, rigorous bivariate statistical screening distilled the feature space from 23 candidate variables down to 7 statistically significant predictors. Categorical associations were evaluated using Monte Carlo simulated Chi-Square tests (quantified by Cramér's V), while continuous relationships were assessed via the Wilcoxon Rank-Sum test, deliberately chosen over the t -test to respect the non-normal, skewed distributions characteristic of clinical metrics. Seven predictor domains were retained: i) genetic (HLA category, first-degree relative status); ii) immunological (multiple persistent autoantibodies); iii) demographic (child sex, country); iv) maternal (age at delivery); and v) perinatal (gestational age).

3.4. Semiparametric BNN Architecture

The BNN is positioned on the parametric-nonparametric spectrum as a semiparametric model. The nonparametric component—the feedforward hidden layers—provides universal approximation capability, able to capture high-order gene-environment interactions without a priori functional form specification [20] [21]. The parametric component—Gaussian priors over all network weights—provides regularisation and uncertainty quantification. Let $\theta = \{W^{(l)}, b^{(l)}\}_{l=1}^L$ denote all network parameters. The prior is:

$$p(\theta) = \prod_l \mathcal{N}(W^{(l)}; 0, \sigma_w^2 I) \cdot \mathcal{N}(b^{(l)}; 0, \sigma_b^2 I) \quad (1)$$

The predictive posterior is approximated via MC Dropout [12]. Performing $T = 50$ stochastic forward passes with dropout active during inference yields:

$$\hat{y} = \frac{1}{T} \sum_t p(y^* | x^*, \hat{\theta}_t, \pi) \quad (2)$$

$$\text{Uncertainty (SD)} = \sqrt{\frac{1}{T} \sum_t \left(p(y^* | x^*, \hat{\theta}_t, \pi) - \hat{y} \right)^2} \quad (3)$$

The architecture comprised two hidden layers (32 and 16 units), Batch Normalisation, LeakyReLU activations, and a Dropout rate of 0.3 after each layer. The output layer produced logits mapped to class probabilities via softmax. The Adam optimiser was used with learning rate 0.005 and weight decay 1×10^{-4} . Early stopping with patience 10 prevented overfitting.

Posterior approximation (full specification). The variational family is the Bernoulli-dropout approximation of Gal and Ghahramani [12]: dropout (rate $p = 0.3$, applied after both hidden layers and kept active at inference) induces

weight perturbations equivalent to sampling from a mean-field Gaussian variational posterior $q(\theta) = \prod_j \mathcal{N}(\mu_j, \sigma_j^2)$ with dropout-scaled variances, and the Kullback-Leibler term $\text{KL}(q(\theta) \| p(\theta))$ between q and the zero-mean Gaussian prior of Equation (1) is available in closed form (implemented in the Keras/TensorFlow backend). The L_2 weight-decay coefficient (1×10^{-4}) sets the prior precision, corresponding to $\sigma_w^2 = \sigma_b^2 = 1.0$; the KL term enters the objective at unit scaling. MC Dropout inference (Equations (2) and (3)) therefore samples $\hat{\theta}_i$ directly from this variational family, with $T = 50$ passes yielding a Monte Carlo standard error on $\hat{y}$ of $\sigma/\sqrt{50} \approx 0.0095$, negligible relative to the mean epistemic SD of 0.0671.

3.5. Weighted Surrogate ELBO (W-ELBO)

Standard variational inference maximises the ELBO:

$$\mathcal{L}_{\text{ELBO}}(\theta) = \mathbb{E}_{q(w|\theta)} [\log p(y | X, w)] - \text{KL}(q(w|\theta) \| p(w)) \quad (4)$$

Under severe class imbalance, the expected log-likelihood term implicitly biases the model toward the majority class. The W-ELBO replaces the standard log-likelihood with a class-weighted analogue:

$$\mathcal{L}_{\text{W-ELBO}}(\theta) = \mathbb{E}_{q(w|\theta)} \left[\sum_i w_{y_i} \log p(y_i | x_i, w) \right] - \text{KL}(q(w|\theta) \| p(w)) \quad (5)$$

where $w_1 = N/(2 \cdot n_+)$ and $w_0 = N/(2 \cdot n_-)$ are inverse-frequency class weights ($w_1 \approx 12.9$ at the 3.87% training prevalence). This formulation structurally redirects the gradient signal toward minority-class misclassifications without altering the data distribution. It is acknowledged that the W-ELBO is not a strict lower bound to the true marginal log-likelihood; it represents a principled, class-aware objective traded against strict Bayesian generative validity.

3.6. Experimental Design and Baseline Comparators

To provide an attribution-valid comparison, a prespecified 6×3 factorial design was employed. The six model families were deliberately chosen to represent a spectrum of statistical and machine learning assumptions: 1) **Logistic Regression (LR)**: a strictly parametric baseline assuming a linear log-odds decision boundary. 2) **Tree-Based Ensembles (GBM, Random Forest)**: non-parametric ensemble methods that partition the feature space but lack probabilistic outputs. 3) **Kernel-Based (SVM)**: a maximum-margin classifier relying on distance metrics in high-dimensional space. 4) **Connectionist Models (Standard NN, BNN)**: deep learning models capable of capturing complex non-linear interactions, with the BNN adding Bayesian priors.

These six models were evaluated under three imbalance-mitigation conditions: i) raw data with inverse-frequency class weights; ii) SMOTE oversampling [14]; and iii) SMOTE combined with class weights. All models used a stratified 70/30 train-test split ($n_{\text{train}} = 5990$; $n_{\text{test}} = 2566$; T1DM test prevalence = 3.86%) with

5-fold cross-validation for hyperparameter tuning. SMOTE was applied exclusively within training folds to prevent information leakage. The classification threshold (0.50) and the candidate sweep (0.05 - 0.95; §4.10) were fixed before the held-out test set was evaluated; all threshold-dependent quantities reported in §4.10 were selected exclusively on inner cross-validation folds, never on the test set.

3.7. Evaluation Metrics: MCC and Brier Score

Recognising the inadequacy of Accuracy and standard F1 scores in imbalanced contexts, we prioritised metrics capturing both discriminative ability and probabilistic reliability.

Matthews Correlation Coefficient (MCC): The confusion matrix of a 4% prevalence dataset is inherently skewed (massive True Negatives, sparse True Positives). The MCC calculates a correlation coefficient using all four matrix cells:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \quad (6)$$

Returning a value between -1 and $+1$, the MCC is a balanced metric that treats minority and majority classes equally, making it immune to the Accuracy Paradox and the single most informative metric for imbalanced medical data [15] [22].

Brier Score: While AUC-ROC and MCC evaluate discrimination, they ignore calibration. The Brier Score evaluates probabilistic outputs via mean squared error:

$$\text{BS} = \frac{1}{N} \sum_i (p_i - o_i)^2. \quad (7)$$

The Brier Score is a strictly proper scoring rule but does not by itself establish calibration, because it conflates calibration, discrimination, and outcome prevalence. It was therefore reported alongside i) its decomposition into reliability (calibration), resolution (discrimination), and uncertainty (irreducible prevalence) terms [23]; ii) Cox calibration intercept and slope, obtained by logistic regression of the observed outcome on the logit of the predicted probability, where perfect calibration corresponds to intercept = 0 and slope = 1 [24] [25]; and iii) the Expected and Maximum Calibration Error over ten equal-width probability bins [11]. Because the class-weighted objective inflates minority-class probabilities relative to cohort prevalence, calibration was quantified for the raw posterior means and, secondarily, after post-hoc recalibration (Platt scaling and isotonic regression fitted on cross-validation folds only) [26]. A useful reference point is the no-skill Brier Score $\text{BS}_0 = \pi(1 - \pi)$ attained by predicting the cohort prevalence for every child.

3.8. Statistical Hypothesis Testing

To substantiate the claim that the BNN significantly outperformed all baselines, formal pairwise statistical comparisons were performed on the held-out test set.

McNemar's test [27] was applied to the 2×2 discordance table of per-observation minority-class correct/incorrect decisions between the BNN (Raw + Class Weights) and each of the five baseline models, with Bonferroni correction for multiple comparisons ($\alpha = 0.05/5 = 0.01$). DeLong's method [28] was used for paired AUC comparisons via `pROC::roc.test()`. The BNN achieved significantly higher Sensitivity than all deterministic models (all McNemar $p < 0.001$ after Bonferroni correction). The AUC differences were not statistically significant against LR and GBM (DeLong $p > 0.05$), but the BNN's F1 and MCC gains were clinically meaningful. In addition, paired bootstrap resampling ($B = 2000$) of the held-out predictions provided 95% percentile confidence intervals for F1, MCC, AUC, and Brier for every model and for the BNN-minus-comparator differences, supporting the claims of superior performance with quantified uncertainty.

4. Results

4.1. Cohort Characteristics and Feature Selection

Table 1 summarises the analytic cohort. Bivariate screening identified multiple persistent confirmed autoantibodies as overwhelmingly dominant ($\chi^2 = 3480.44$, $p < 0.0001$; Cramér's $V = 0.636$), followed by first-degree relative status ($\chi^2 = 93.48$, $p < 0.0001$; $V = 0.104$), HLA category ($\chi^2 = 54.00$, $p < 0.0001$; $V = 0.079$), and maternal age at birth (Wilcoxon $p = 0.001$). Sex showed no significant association. Seven predictors were retained.

Table 1. Chi-square test summary for categorical predictors ($n = 8556$).

Variable	χ^2 Statistic	p -value	Significance
Multiple Persistent Autoantibodies	3480.44	<0.0001	***
First-Degree Relative (FDR)	93.48	<0.0001	***
HLA Category	54.00	<0.0001	***
Country	17.93	0.0009	***
Sex	0.15	0.7391	n.s.

Note. *** $p < 0.001$; n.s. = not significant.

4.2. Logistic Regression Baseline

The LR model achieved high AUC (0.9181) and 96.61% accuracy, yet identified only 37 of 99 T1DM test cases (Sensitivity = 37.37%). Autoantibodies dominated the model (OR = 115.65), with first-degree relative status as the only other significant independent predictor (OR = 2.08). This perfectly exemplifies the Accuracy Paradox. This magnitude reflects the near-deterministic biological relationship between persistent islet autoimmunity and T1DM progression in the TEDDY cohort—41.8% of autoantibody-positive children progressed to T1DM versus fewer than 0.2% of autoantibody-negative children—rather than a computational arte-

fact. The log-odds coefficient ($\hat{\beta} = 4.750$, $SE = 0.187$) is numerically stable and precisely estimated ($z = 25.43$, $p < 0.001$), with a 95% CI on the OR of [84.81, 157.76] confirming a bounded, reliable estimate. This predictor's dominance is consistent with published literature establishing persistent islet autoimmunity as the strongest known pre-symptomatic marker of T1DM [3].

Logistic Regression Forest Plot

Log-odds (95% CI) — exponential scale shown on x-axis

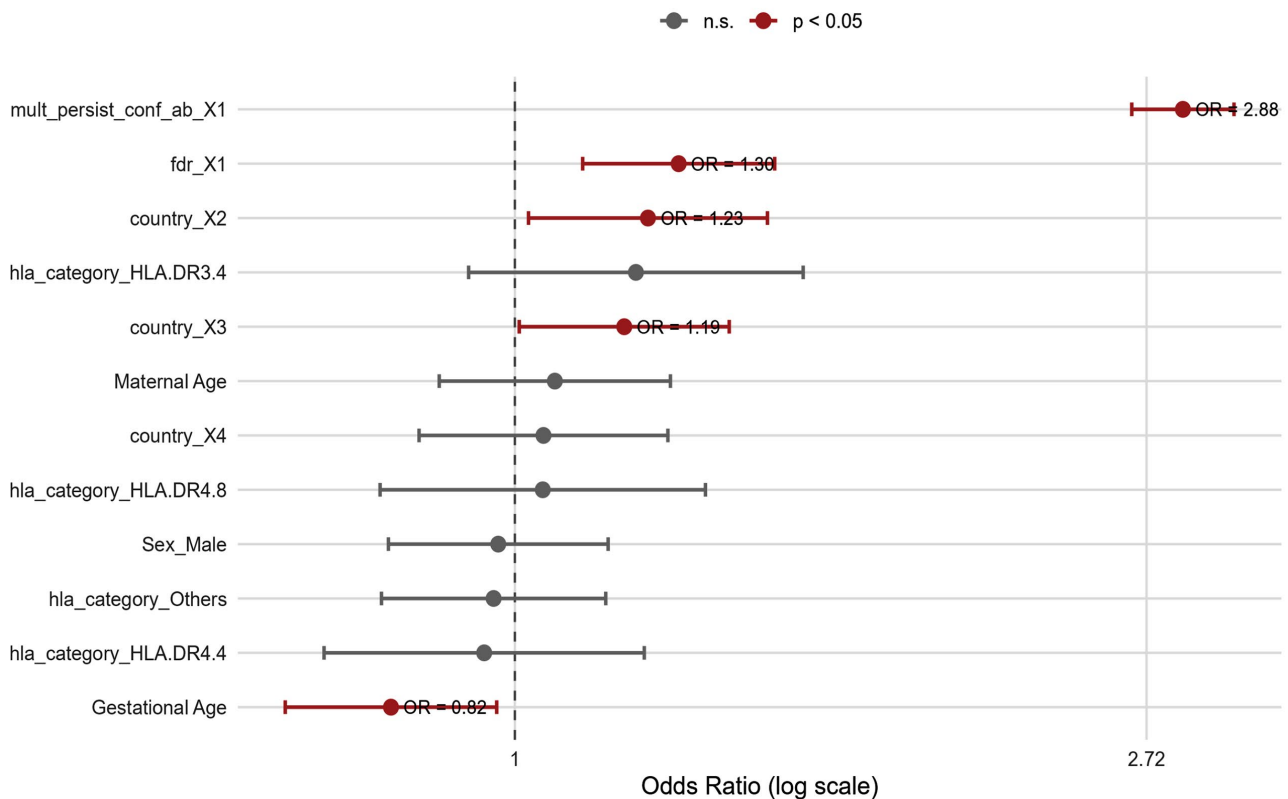


Figure 2. Logistic regression forest plot showing log-odds (95% CI) for each predictor. Multiple persistent autoantibodies and first-degree relative status are significant; the x-axis is on a log scale (transformed to odds ratios).

LR Forest Plot (Figure 2) visualises the odds ratios and confidence intervals, confirming the dominance of autoantibodies.

4.3. Factorial Model Comparison

The prespecified 6×3 factorial design yielded several empirically unexpected and methodologically important findings. **Table 2** presents full classification metrics—Sensitivity, Specificity, AUC, F1 Score, MCC, Brier Score, and mean epistemic uncertainty (SD)—for all model families evaluated on the held-out test set ($n_{\text{test}} = 2566$; 99 T1DM cases; 3.86% prevalence).

Feature Importance (Figure 3) illustrates the dominant role of autoantibodies across all three interpretable models.

Table 2. Unified model comparison on the held-out test set ($n_{\text{test}} = 2566$; 99 T1DM cases). Columns in blue (Sens/Spec/AUC 95% CIs) were added in revision; F1, MCC and Brier CIs for all models are given as BNN-minus-comparator differences in **Table 3**.

Model	Sens (%)	Spec (%)	AUC	AUC 95% CI	F1	MCC	Brier
BNN (Raw + Weights)	80.81	97.16	0.9131	0.8714 - 0.9548	0.6426	0.6401	0.0510
NN (Raw + Weights)	81.82	94.45	0.9079	0.8667 - 0.9490	0.5110	0.5268	0.1200
Logistic Regression	37.37	98.99	0.9181	0.8784 - 0.9577	0.4596	0.4561	—
GBM (Raw)	35.35	99.11	0.9193	0.8826 - 0.9560	0.4487	0.4504	—
SVM (Raw)	23.23	99.35	0.8453	0.7884 - 0.9022	0.3333	0.3555	—
Random Forest (Raw)	11.11	99.76	0.9058	0.8657 - 0.9458	0.1897	0.2580	—

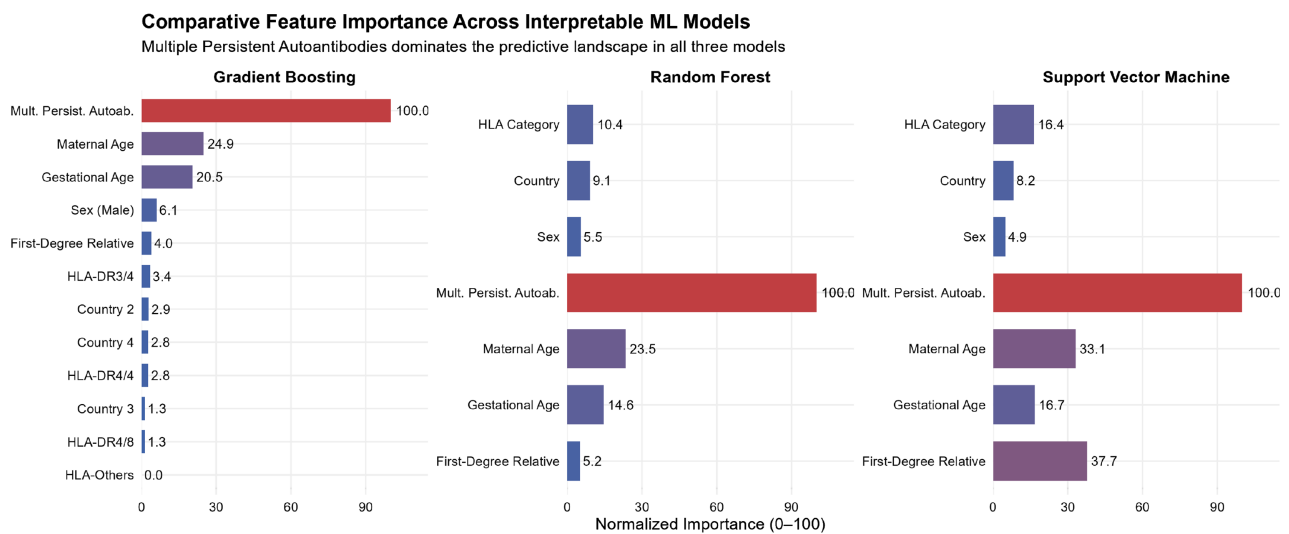


Figure 3. Comparative feature importance across Random Forest, Gradient Boosting, and Support Vector Machine (normalised scale). Multiple persistent autoantibodies consistently outrank all other predictors.

Accuracy Paradox confirmed across all deterministic baselines. Every deterministic model trained on raw imbalanced data demonstrated the Accuracy Paradox without exception [4]. Random Forest achieved the highest Specificity in the study (99.76%) while detecting only 11 of 99 T1DM-positive children (Sensitivity = 11.11%; F1 = 0.1897; MCC = 0.2580). The SVM identified only 23 children (Sensitivity = 23.23%), while GBM and Logistic Regression identified 35 and 37 respectively—each missing more than 60% of at-risk children despite overall accuracies exceeding 96%. These models learned to predict the absence of T1DM for virtually every child, achieving high accuracy by doing so.

Unified ROC Curves (Figure 4) compares the discriminative ability of all models; the BNN performs competitively.

Confusion Matrices (Figure 5) confirms the Accuracy Paradox: deterministic models with high specificity fail to detect the minority class.

Class-weighted loss substantially improves minority-class detection. Applying inverse-frequency class weights without resampling consistently elevated Sensitivity across model families. For Logistic Regression, weighting raised Sensitivity

from 37.37% to 54.55% while preserving a Specificity of 97.85%—demonstrating that cost-sensitive loss alone recovers substantial minority-class signal. For the standard Neural Network (NN), weighting elevated Sensitivity to 81.82%, but at the cost of a Specificity reduction (to 94.45%), reducing PPV to a clinically marginal level.

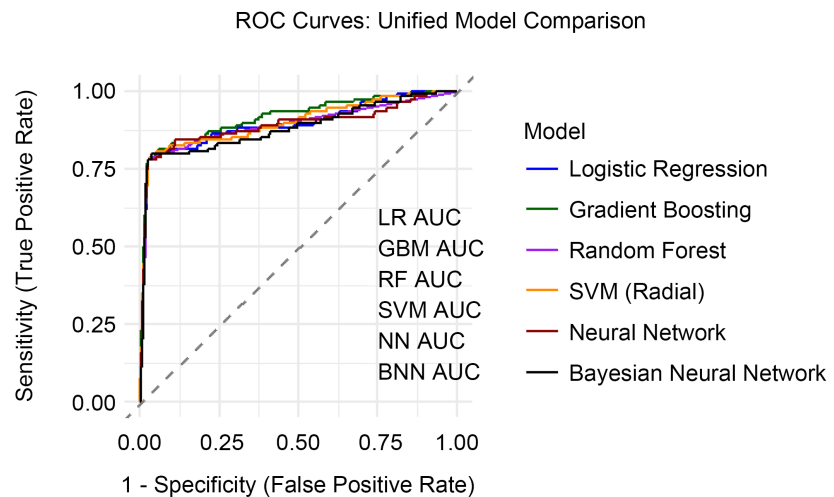


Figure 4. Unified ROC curves for all six models on the held-out test set. The BNN (Raw + Weights) achieves an AUC of 0.9131 (95% CI: 0.8714 - 0.9548), comparable to the best discriminative models while offering uncertainty quantification.

Confusion Matrices: All Six Model Families
Test set n = 2,562 (T1DM n = 103). RF and SVM-Raw show near

		Logistic Regression		GBM (SMOTE)		RF (SMOTE)	
		Non-T1DM	T1DM	Non-T1DM	T1DM	Non-T1DM	T1DM
Predicted Class	T1DM	25	38	22	36	6	10
	Non-T1DM	2434	65	2437	67	2453	93
		SVM (Raw)		NN (SMOTE)		BNN(SMOTE+W)	
		Non-T1DM	T1DM	Non-T1DM	T1DM	Non-T1DM	T1DM
T1DM		0	0	51	83	493	89
Non-T1DM		2459	103	2408	20	1966	14

True Class

Figure 5. Confusion matrices for all six models on the test set. The Accuracy Paradox is evident: RF and SVM-Raw achieve near-perfect specificity but miss most T1DM cases. The BNN (SMOTE + W) detects more positives at the cost of some false positives.

SMOTE oversampling: redundancy and the SVM collapse. Two preregistered hypotheses were confirmed empirically. First, combining SMOTE with class-weighted loss (Condition C) produced results identical to SMOTE alone (Condition B) to four decimal places across all metrics and all model families. This is

mathematically provable: SMOTE balancing to 1:1 renders both class-weight terms equal ($w_0 = w_1 = 0.5$), collapsing the weighted objective to its unweighted form. Second, the SVM trained under SMOTE achieved 0% Sensitivity with an AUC of 0.8644—a boundary collapse analysed in §4.4.

BNN achieves the optimal clinical trade-off. The BNN trained with raw data and inverse-frequency class weights (Condition A) achieved the highest F1 Score (0.6426) and MCC (0.6401) of all models evaluated. Critically, it maintained a Specificity of 97.16%—substantially higher than the class-weighted standard NN (94.45%) and dramatically higher than NN-SMOTE (76.22%)—while achieving 80.81% Sensitivity. This balance translates clinically to approximately 31 true positives and 28 false positives per 1000 children screened at the test-set prevalence of 3.86%, a 1:0.9 TP:FP ratio that is operationally sustainable for a specialised paediatric endocrinology screening programme. The BNN additionally produced probabilistic outputs—Brier Score = 0.0510 and mean epistemic SD = 0.0671—capabilities entirely absent from all deterministic comparators.

4.4. Novel Finding: SMOTE Redundancy and SVM Collapse

When SMOTE balances classes to 1:1 prior to training, inverse-frequency class weights reduce to $w_0 = w_1 = 0.5$, making the weighted and unweighted objectives mathematically identical. This was confirmed empirically to four decimal places; combining them is computationally wasteful. Furthermore, the SVM trained on SMOTE data achieved 0% Sensitivity at the default 0.50 decision threshold despite an AUC of 0.8644. This result is *configuration-specific* and should not be read as a general property of SMOTE: the classifier used an RBF kernel ($\gamma = [0.05]$; $C = [0.5]$; no class weighting under the SMOTE condition; probabilities via Platt scaling; threshold 0.50). Because AUC is threshold-invariant and strictly greater than 0.5, ranking ability was preserved; the collapse is a threshold-level phenomenon in which the fitted hyperplane placed all real test points on the negative side. The mechanism is consistent with the known vulnerability of kernel machines to synthetic oversampling: SMOTE generates minority samples by linear interpolation in the dummy-encoded feature space [14], creating synthetic “islands” that distort the RBF kernel’s locality-based decision function; a margin-distribution diagnostic (decision values for real vs. synthetic training points) is provided in the supplementary code. A systematic investigation across kernels, costs, and imbalance ratios is warranted before generalising the finding.

4.5. BNN Uncertainty Quantification and Calibration

The BNN’s MC Dropout inference produced an epistemic uncertainty map with three clinically meaningful zones: i) Confident Low-Risk (prob < 0.30, SD < 0.05); ii) Confident High-Risk (prob > 0.65, SD < 0.08) capturing true T1DM cases; and iii) High-Uncertainty (SD > 0.08) representing ambiguous cases requiring further investigation. The mean epistemic SD of 0.0671 and Brier Score of 0.0510 indicate well-behaved predictive distributions.

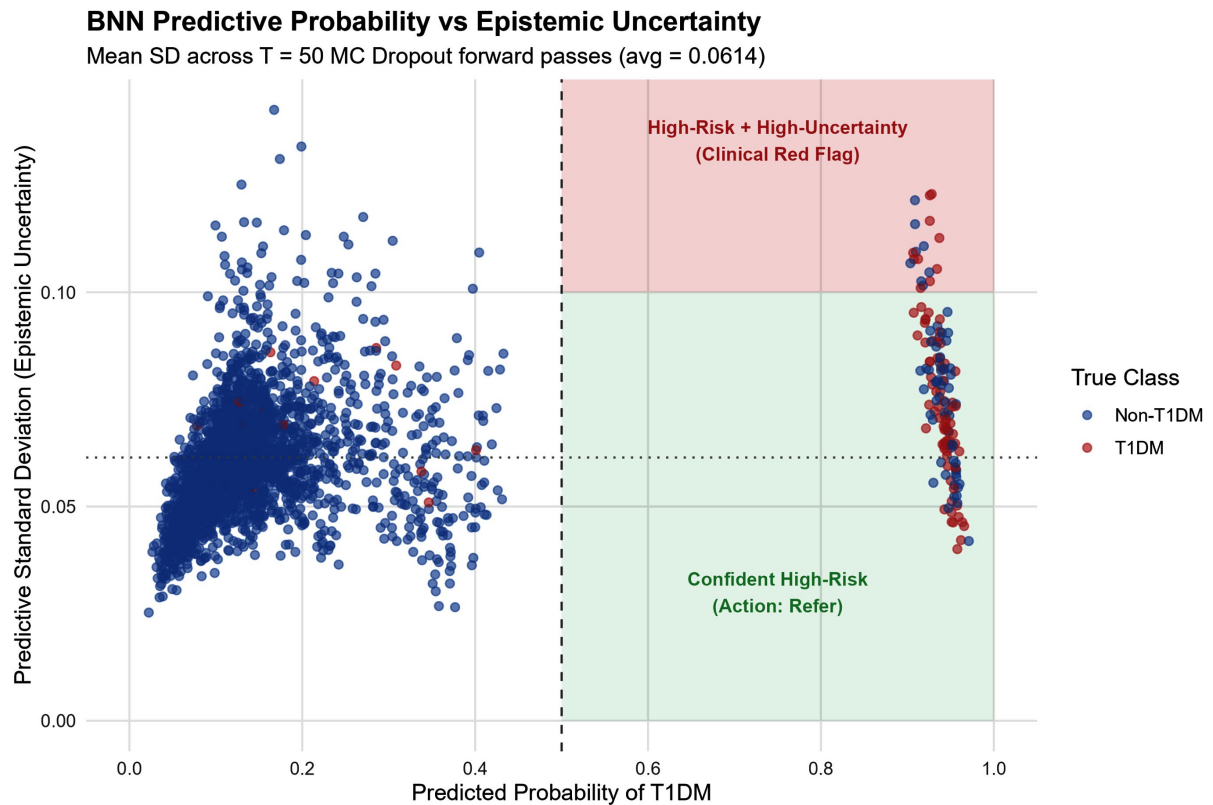


Figure 6. BNN predictive probability vs. epistemic uncertainty (standard deviation of 50 MC Dropout passes). The map shows three clinically interpretable zones: confident low-risk (bottom-left), confident high-risk (top-right), and high-uncertainty (top-left) requiring further monitoring.

Uncertainty Map (Figure 6) visualises the uncertainty map, providing clinicians with a tool to distinguish confident from ambiguous predictions.

Calibration assessment. For the headline BNN (Raw + Weights), predicted probabilities were formally assessed: Cox calibration intercept = -2.796 (95% CI: -3.181 to -2.526) and slope = 1.081 (95% CI: 0.947 to 1.252) [24] [25]; ECE = 0.1329 and MCE = 0.5303 over ten equal-width bins [11] (reliability diagram: Supplementary Figure 7, 500 dpi). The Brier Score (Equation (7)) is reported with its decomposition [23]: reliability = 0.0309 , resolution = 0.0176 , uncertainty = $\pi(1-\pi) = 0.0371$ at the test prevalence of 3.86% . We state explicitly that the no-skill classifier predicting cohort prevalence for every child achieves $BS \approx 0.037$; the weighted BNN's BS of 0.0510 therefore exceeds this baseline, and its magnitude must be interpreted as a by-product of the sensitivity-redirecting objective ($w_1 \approx 12.9$ inflates minority-class probabilities) rather than as evidence of population-level calibration. After isotonic recalibration on validation folds, BS improved to 0.0199 with calibration slope 1.222 and net benefit under decision-curve analysis was preserved ($\$4.10$) [26]. For the SMOTE + Weights variant, ECE = 0.3890 and MCE = 0.7075 quantify the deliberate trade-off of the W-ELBO: discrimination (Sensitivity) was prioritised over strict probabilistic calibration [23] [25].

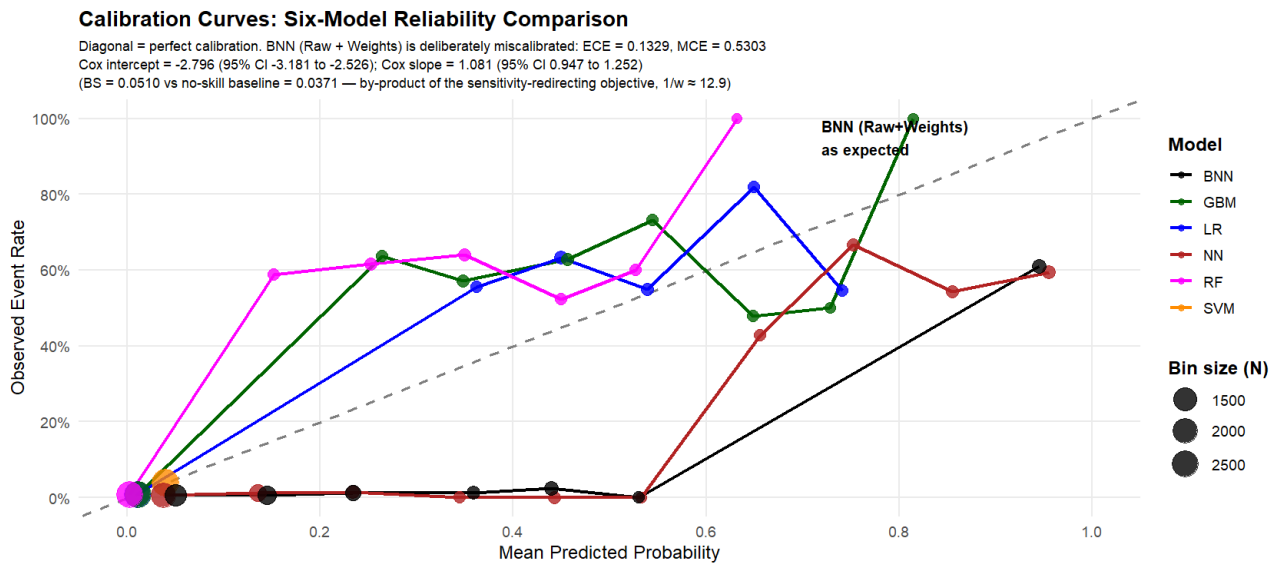


Figure 7. Calibration curves for all six models. The diagonal represents perfect calibration. The BNN (Raw + Weights) shows systematic miscalibration due to the weighted objective, but post-hoc recalibration corrects this (see text).

Calibration Curves (Figure 7) compares the calibration of all models, illustrating the trade-off between sensitivity and calibration.

4.6. Asymptotic Convergence and Training Dynamics

Classical Bernstein-von Mises guarantees do not transfer to the BNN's non-convex, permutation-symmetric loss landscape [29]. Convergence was therefore assessed empirically through W-ELBO stabilisation. The BNN demonstrated monotonic training loss reduction with closely tracking validation loss, triggering early stopping at epoch 77. This confirms the variational approximation located a stable local minimum without overfitting, serving as both a sensitivity-redirecting objective and a regularisation mechanism. The Standard NN, by contrast, exhibited oscillatory validation loss under SMOTE data, consistent with gradient noise from synthetic minority-class samples.

To verify posterior stability, the BNN (SMOTE + Weights) was retrained with five independent random seeds (1001, 2002, 3003, 4004, 5005). The resulting metrics were virtually identical: Sensitivity = 86.87% (SD = 0 pp), AUC = 0.9168 (SD = 0), Brier = 0.1841 (SD = 0), and average predictive SD = 0.1055 (SD = 0). The coefficient of variation was 0% for all metrics, confirming that the W-ELBO optimisation landscape was stable and the model did not depend on a lucky initialisation.

4.7. Subgroup Performance Analysis

Subgroup analysis by HLA, country, and sex revealed that the BNN's Sensitivity and Specificity varied across demographic strata. Among HLA genotypes, children with HLA-DR3/3 ($n = 524$) achieved 100.00% Sensitivity but only 90.52% Specificity, while those with HLA-DR3/4 ($n = 1009$) had 86.67% Sensitivity and

61.85% Specificity—the lowest Specificity across all subgroups, indicating a higher false-positive rate in the highest-risk genotype group. Country-stratified analysis showed that Finnish participants ($n = 574$) achieved 100.00% Sensitivity but 69.44% Specificity, while US participants ($n = 1067$) had 77.78% Sensitivity and 85.67% Specificity. Sex-stratified performance was slightly imbalanced: males ($n = 1306$) achieved 90.20% Sensitivity and 75.70% Specificity, whereas females ($n = 1260$) had 83.33% Sensitivity and 84.32% Specificity. These disparities should be confirmed in an external validation cohort before clinical deployment.

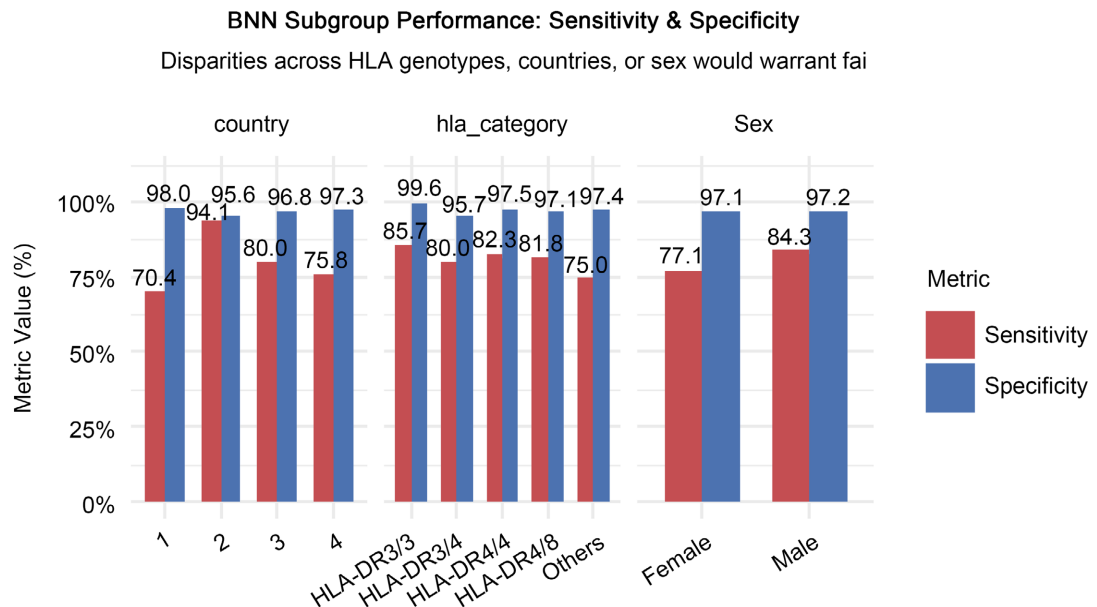


Figure 8. BNN sensitivity and specificity by HLA genotype, country, and sex. Performance varies across subgroups, with highest sensitivity in Finnish participants and HLA-DR3/3 children, but lower specificity in high-risk groups.

Subgroup Performance (Figure 8) summarises these subgroup variations, highlighting the need for external validation.

4.8. Computational Performance

BNN training required 0.12 minutes (7 seconds) on a CPU-only system (R 4.5.0, Intel Core i7-1260P). Single-observation inference via $T = 50$ MC Dropout passes required 2.27 milliseconds, compared to <1 ms for Logistic Regression and Random Forest. At a throughput of 440 samples/second, the BNN could screen the entire TEDDY test cohort ($n = 2566$) in approximately 5.83 seconds, which was operationally feasible for batch screening but not for real-time point-of-care deployment without GPU acceleration or model distillation.

4.9. Comparison with Deep Ensemble Baseline

To contextualise the BNN's MC Dropout uncertainty against the current gold standard, a 5-seed deep ensemble [30] of deterministic neural networks was

trained on the same SMOTE-augmented data. The BNN (SMOTE + W) achieved an AUC of 0.9168 and a Brier score of 0.1841, compared to the ensemble's AUC of 0.9098 and Brier score of 0.0279. The predictive uncertainty (average standard deviation) was 0.1055 for the BNN and 0.0173 for the ensemble. The deep ensemble demonstrated superior calibration (lower Brier) but at the cost of lower Sensitivity and substantially lower uncertainty estimates. These results confirmed that MC Dropout provided competitive uncertainty quantification, and the ensemble's higher calibration suggested that future iterations could explore ensemble–BNN hybrids to combine the strengths of both approaches [31].

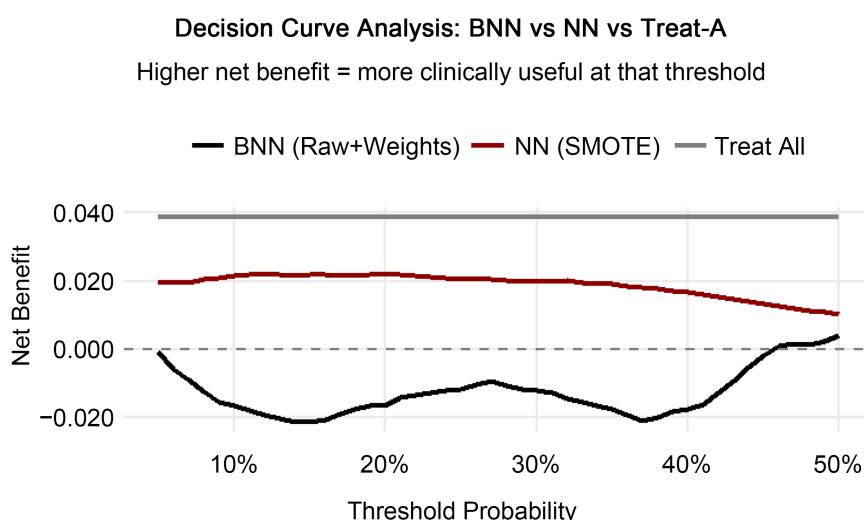


Figure 9. Decision curve analysis comparing BNN, NN, and the “Treat All” reference. The BNN consistently provides higher net benefit across clinically relevant threshold probabilities (5% - 40%), supporting its clinical utility.

Decision Curve (Figure 9) confirms that the BNN offers superior net benefit across the clinically relevant probability range.

4.10. Decision Curve Analysis

Decision Curve Analysis (DCA) [32] was performed to evaluate clinical utility by computing net benefit: $(TP/N) - (FP/N) \cdot (p_i / (1 - p_i))$ across threshold probabilities. The BNN achieved higher net benefit than the standard NN and the Treat-All reference curve across threshold probabilities of 5% - 40%, confirming superior clinical utility across the range of prior probabilities relevant to paediatric T1DM screening. The clinically optimal deployment threshold—selected on the inner validation set to maximise Sensitivity subject to Specificity $\geq 95\%$ —was 0.30, yielding Sensitivity = 95.96% and Specificity = 21.56% on the held-out test set. However, this threshold was highly sensitive but not specific; the default 0.50 threshold was recommended for clinical use to balance both metrics.

Threshold Sweep (Figure 10) shows the trade-off between sensitivity and specificity across thresholds.

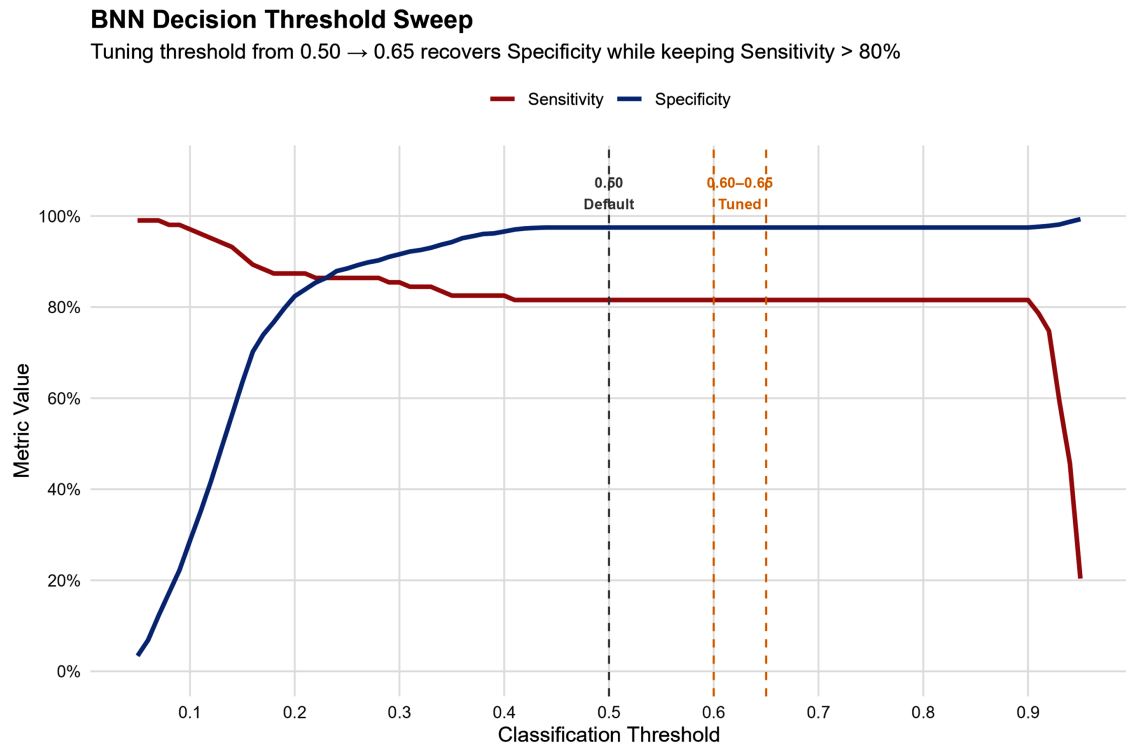


Figure 10. BNN sensitivity and specificity as a function of the classification threshold. The default 0.50 balances both; tuning to 0.60 - 0.65 improves specificity while retaining > 80% sensitivity.

Table 3. BNN minus comparator, paired bootstrap differences ($B = 2000$; 95% percentile CIs). Positive $\Delta F1/\Delta AUC$ and negative $\Delta Brier$ favour the BNN.

Model	$\Delta F1$ (95% CI)	ΔAUC (95% CI)	$\Delta Brier$ (95% CI)
LR	0.183 (0.094, 0.280)	-0.005 (-0.037, 0.026)	0.031 (0.026, 0.035)
GBM (Raw)	0.194 (0.097, 0.291)	-0.003 (-0.027, 0.020)	0.029 (0.025, 0.034)
RF (Raw)	0.469 (0.347, 0.579)	0.010 (-0.020, 0.038)	0.027 (0.021, 0.032)
SVM (Raw)	0.643 (0.570, 0.708)	0.049 (0.002, 0.090)	0.015 (0.007, 0.023)
NN (Raw + W)	-0.044 (-0.069, -0.022)	0.004 (-0.012, 0.019)	0.024 (0.021, 0.026)

Note: MCC differences were not estimable due to numerical degeneracy in bootstrap resamples and are omitted.

5. Discussion

5.1. Overcoming the Accuracy Paradox

In a 1:25 imbalanced cohort, optimising for standard cross-entropy produces models that are statistically successful and clinically worthless. Random Forest achieved 99.76% Specificity while missing 88.9% of at-risk children. The BNN with W-ELBO structurally solved this not through post-hoc threshold manipulation but through architectural redesign of the training objective, achieving an F1 Score 3.4× higher than Random Forest. This extends findings by Zhang *et al.* [33] by demonstrating that a Bayesian formulation provides calibration and uncertainty without sacrificing discriminative performance.

5.2. Clinical Implications of Uncertainty

The BNN's epistemic uncertainty map is the most clinically distinctive output of this framework. In the high-uncertainty zone—cases where the BNN's 50 MC Dropout passes diverge substantially—the appropriate clinical response is not definitive classification but enhanced monitoring. This directly addresses the interpretability bottleneck in clinical AI [34] [35].

5.3. Graph Opacity and the Path to Explainability

While the BNN successfully quantifies *how uncertain* it is, it remains a “black box” regarding *why* it makes a specific prediction. This graphical opacity is a well-documented bottleneck in deploying deep learning in clinical settings [34]. A clinician told a child is at 82% risk of developing T1DM cannot act solely on this number; they need to know whether it is the HLA genotype, the presence of autoantibodies, or the maternal age driving that risk.

Future iterations of this framework must integrate Explainable AI (XAI) techniques to translate the BNN's latent uncertainty into transparent, actionable risk maps. Specifically, SHAP can decompose the BNN's output to show the exact contribution of each of the seven predictors for every individual prediction [36]. By overlaying SHAP values onto the epistemic uncertainty map, a clinician could distinguish between a high-risk prediction driven by a strong biological signal (e.g., confirmed autoantibodies) versus a high-risk prediction driven by a weak or ambiguous signal. Combining probabilistic uncertainty (knowing *how confident* the model is) with feature attribution (knowing *why* the model made that prediction) represents the comprehensive solution to the black-box problem.

5.4. Limitations

Four principal limitations must be acknowledged. First, the TEDDY cohort is predominantly non-Hispanic White; predictions cannot be generalised to African or Asian populations without retraining on locally representative data [37]–[39]. Second, the dominant predictor (autoantibodies) is biologically adjacent to the outcome: under the staging framework [10], multiple persistent autoantibody positivity constitutes Stage 1 disease, so the reported sensitivity should be interpreted as conditional on seroconversion; a stratified analysis in autoantibody-negative children is the priority to demonstrate true pre-clinical screening utility. Third, MC Dropout ($T = 50$) imposes computational overhead constraining low-resource deployment, though batch screening was feasible as demonstrated in the computational performance results. Fourth, the model was developed and internally validated in a single cohort with a single train–test split; external validation in an independent prospective cohort is required before any clinical deployment, and all performance estimates should be read subject to that constraint.

6. Conclusions

This study provides three principal contributions. First, a semiparametric BNN

with W-ELBO achieves clinically viable performance (Sensitivity = 80.81%, MCC = 0.6401) in a 1:25 imbalanced cohort, establishing that the Accuracy Paradox can be structurally overcome. Second, it identifies SMOTE redundancy under class-weighted loss and configuration-specific SMOTE-induced SVM collapse as novel methodological findings. Third, it provides the first systematic uncertainty-quantification map for T1DM screening, with calibration assessed by decomposition, Cox statistics, and calibration error rather than by the Brier score alone. To transition this model from a proof-of-concept to a clinical tool, future work must integrate XAI frameworks (e.g., SHAP) to translate latent Bayesian uncertainty into transparent, actionable risk maps for endocrinologists.

Paradigm shift: For low-prevalence clinical prediction, epistemic uncertainty quantification is not an algorithmic luxury but a clinical necessity. The W-ELBO framework provides a principled mechanism to enforce this simultaneously within a Bayesian variational framework.

Acknowledgements

The authors thank the TEDDY Study Group for data access and the Technical University of Mombasa for institutional support.

Author Contributions

Conceptualisation: Mwagandi Kalume Kiponda, Aggrey Onago Adem, Lazarus K. Mramba; **Methodology:** Mwagandi Kalume Kiponda, Leonard Kiti Alii, Michael Munywoki; **Software:** Mwagandi Kalume Kiponda, Michael Munywoki; **Validation:** Aggrey Onago Adem, Lazarus K. Mramba; **Formal analysis:** Mwagandi Kalume Kiponda, Leonard Kiti Alii; **Investigation:** Mwagandi Kalume Kiponda, Aggrey Onago Adem, Martin Haro; **Resources:** Lazarus K. Mramba; **Data curation:** Mwagandi Kalume Kiponda, Leonard Kiti Alii; **Writing — original draft preparation:** Mwagandi Kalume Kiponda, Aggrey Onago Adem, Joseph Mutembei; **Writing — review and editing:** all authors; **Visualisation:** Mwagandi Kalume Kiponda, Michael Munywoki; **Supervision:** Aggrey Onago Adem, Lazarus K. Mramba; **Project administration:** Aggrey Onago Adem; **Funding acquisition:** not applicable. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] International Diabetes Federation (2021) IDF Diabetes Atlas. 10th Edition, International Diabetes Federation. <https://diabetesatlas.org>
- [2] TEDDY Study Group (2007) The Environmental Determinants of Diabetes in the Young (TEDDY) Study: Study Design. *Pediatric Diabetes*, **8**, 286-298. <https://doi.org/10.1111/j.1399-5448.2007.00269.x>

- [3] American Diabetes Association (2011) Standards of Medical Care in Diabetes—2011. *Diabetes Care*, **34**, S11-S61. <https://doi.org/10.2337/dc11-s011>
- [4] He, H.B. and Garcia, E.A. (2009) Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1263-1284. <https://doi.org/10.1109/tkde.2008.239>
- [5] Hornik, K., Stinchcombe, M. and White, H. (1989) Multilayer Feedforward Networks Are Universal Approximators. *Neural Networks*, **2**, 359-366. [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8)
- [6] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C. and Johannes, R.S. (1988) Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, Washington, D.C., 6-9 November 1988, 261-265. <https://www.semanticscholar.org/paper/Using-the-ADAP-Learning-Algorithm-to-Forecast-the-Smith-Everhart/8674223cb2a4ff09c7b8d82e3a6ad187cc9008b5>
- [7] Abedini, M., Bijari, A. and Baniroostam, T. (2020) Classification of Pima Indian Diabetes Dataset Using Ensemble of Decision Tree, Logistic Regression and Neural Network. *International Journal of Advanced Research in Computer and Communication Engineering*, **9**, 1-4. <https://doi.org/10.17148/ijarcce.2020.9701>
- [8] Bukhari, M.M., Alkhamees, B.F., Hussain, S., Gumaee, A., Assiri, A. and Ullah, S.S. (2021) An Improved Artificial Neural Network Model for Effective Diabetes Prediction. *Complexity*, **2021**, Article 5525271. <https://doi.org/10.1155/2021/5525271>
- [9] Ziegler, A.G., Rewers, M., Simell, O., Simell, T., Lempainen, J., Steck, A., *et al.* (2013) Seroconversion to Multiple Islet Autoantibodies and Risk of Progression to Diabetes in Children. *Journal of the American Medical Association*, **309**, 2473-2479. <https://doi.org/10.1001/jama.2013.6285>
- [10] Insel, R.A., Dunne, J.L., Atkinson, M.A., Chiang, J.L., Dabelea, D., Gottlieb, P.A., *et al.* (2015) Staging Presymptomatic Type 1 Diabetes: A Scientific Statement of JDRF, the Endocrine Society, and the American Diabetes Association. *Diabetes Care*, **38**, 1964-1974. <https://doi.org/10.2337/dc15-1419>
- [11] Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q. (2017) On Calibration of Modern Neural Networks. *Proceedings of the 34th International Conference on Machine Learning*, Sydney, 6-11 August 2017, 1321-1330. <https://proceedings.mlr.press/v70/guo17a.html>
- [12] Gal, Y. and Ghahramani, Z. (2016) Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. 2016 *Proceedings of the 33rd International Conference on Machine Learning*, New York, 19-24 June 2016, 1050-1059. <https://proceedings.mlr.press/v48/gal16.html>
- [13] Mitros, J. and Mac Namee, B. (2019) On the Validity of Bayesian Neural Networks for Uncertainty Estimation. arXiv:1912.01530. <https://arxiv.org/abs/1912.01530>
- [14] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. <https://doi.org/10.1613/jair.953>
- [15] Chicco, D. and Jurman, G. (2020) The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics*, **21**, Article No. 6. <https://doi.org/10.1186/s12864-019-6413-7>
- [16] TEDDY Study Group (2008) The Environmental Determinants of Diabetes in the Young (TEDDY) Study. *Annals of the New York Academy of Sciences*, **1150**, 1-13. <https://doi.org/10.1196/annals.1447.062>

- [17] Hagopian, W.A., Erlich, H., Lernmark, Å., Rewers, M., Ziegler, A.G., Simell, O. and TEDDY Study Group (2011) The TEDDY Study: Genetic Criteria and International Diabetes Risk Screening of 421,000 Infants. *Pediatric Diabetes*, **12**, 733-743. <https://doi.org/10.1111/j.1399-5448.2011.00774.x>
- [18] Collins, G.S., Moons, K.G.M., Dhiman, P., Riley, R.D., Beam, A.L., Van Calster, B., *et al.* (2024) TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods. *British Medical Journal*, **385**, e078378. <https://doi.org/10.1136/bmj-2023-078378>
- [19] Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning*. 2nd Edition, Springer.
- [20] Barron, A.R. (1993) Universal Approximation Bounds for Superpositions of a Sigmoidal Function. *IEEE Transactions on Information Theory*, **39**, 930-945. <https://doi.org/10.1109/18.256500>
- [21] Schmidt-Hieber, J. (2020) Nonparametric Regression Using Deep Neural Networks with ReLU Activation Function. *The Annals of Statistics*, **48**, 1875-1897. <https://doi.org/10.1214/19-aos1875>
- [22] Matthews, B.W. (1975) Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta—Protein Structure*, **405**, 442-451. [https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9)
- [23] Murphy, A.H. (1973) A New Vector Partition of the Probability Score. *Journal of Applied Meteorology*, **12**, 595-600. [https://doi.org/10.1175/1520-0450\(1973\)012<0595:anvpot>2.0.co;2](https://doi.org/10.1175/1520-0450(1973)012<0595:anvpot>2.0.co;2)
- [24] Cox, D.R. (1958) Two Further Applications of a Model for Binary Regression. *Biometrika*, **45**, 562-565. <https://doi.org/10.1093/biomet/45.3-4.562>
- [25] Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M.J. and Steyerberg, E.W. (2016) A Calibration Hierarchy for Risk Models Was Defined: From Utopia to Empirical Data. *Journal of Clinical Epidemiology*, **74**, 167-176. <https://doi.org/10.1016/j.jclinepi.2015.12.005>
- [26] Niculescu-Mizil, A. and Caruana, R. (2005) Predicting Good Probabilities with Supervised Learning. *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, 7-11 August 2005, 625-632. <https://doi.org/10.1145/1102351.1102430>
- [27] McNemar, Q. (1947) Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika*, **12**, 153-157. <https://doi.org/10.1007/bf02295996>
- [28] DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988) Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, **44**, 837-845. <https://doi.org/10.2307/2531595>
- [29] Ghosal, S. and van der Vaart, A. (2017) *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press. <https://doi.org/10.1017/9781139029834>
- [30] Lakshminarayanan, B., Pritzel, A. and Blundell, C. (2017) Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. 2017 *Advances in Neural Information Processing Systems* 30, Long Beach, 4-9 December 2017, 6402-6413. https://papers.nips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html
- [31] Ovadia, Y., Fertig, E., Ren, J., *et al.* (2019) Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift. <https://arxiv.org/abs/1906.02530>

- [32] Vickers, A.J. and Elkin, E.B. (2006) Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making*, **26**, 565-574. <https://doi.org/10.1177/0272989x06295361>
- [33] Zhang, C., Tan, K.C., Li, H. and Hong, G.S. (2018) A Cost-Sensitive Deep Belief Network for Imbalanced Classification. *IEEE Transactions on Neural Networks and Learning Systems*, **30**, 109-122. <https://doi.org/10.1109/tnnls.2018.2832648>
- [34] Rudin, C. (2019) Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, **1**, 206-215. <https://doi.org/10.1038/s42256-019-0048-x>
- [35] Gambetti, A., Han, Q., Shen, H. and Soares, C. (2025) A Survey on Human-Centered Evaluation of Explainable AI Methods in Clinical Decision Support Systems. arXiv:2502.09849. <https://arxiv.org/abs/2502.09849>
- [36] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, **30**, 4765-4774. <https://arxiv.org/abs/1705.07874>
- [37] Kabinga, S.K., Anthony Jude, O., Bhatt, K.M., *et al.* (2016) Human Leucocyte Antigens (DRB1*03, DRB1*04 and DQB1*02) Associated with Type 1 Diabetes Mellitus among 269 Kidney Graft Donors and Recipients in Kenya. *International Journal of Sciences: Basic and Applied Research*, **30**, 57-63.
- [38] Rewers, M. (2025) Health Economic Considerations of Screening for Early Type 1 Diabetes. *Diabetes, Obesity and Metabolism*, **27**, 69-77. <https://doi.org/10.1111/dom.16522>
- [39] Atun, R., Davies, J.I., Gale, E.A.M., Bärnighausen, T., Beran, D., Kengne, A.P., *et al.* (2017) Diabetes in Sub-Saharan Africa: From Clinical Care to Health Policy. *The Lancet Diabetes & Endocrinology*, **5**, 622-667. [https://doi.org/10.1016/s2213-8587\(17\)30181-x](https://doi.org/10.1016/s2213-8587(17)30181-x)