Volume 3, Number 3, 2026 Pages 268-281

Bioscience Insights

ISSN: 3080-3489 (Print)
ISSN: 3080-3497 (Online)

BiosciInsights.com

MACHINE LEARNING-GUIDED CLUSTERING FOR PHENOTYPIC SUB CLASSIFICATION OF TYPE 2 DIABETES MELLITUS: A CROSS-SECTIONAL STUDY EMPLOYING STATISTICAL DATA AUGMENTATION

Syed Muhammad Sibtain Ali Shah[1], Khudija Bibi[2], Fatima Bibi[3], Bilal Ahmad[4], Arshad Mahmood[5], Waqas Khan[6], Syed Rashid Hussain[7], Balqees Hidayat[8], Qaiser Jamal[9], Qaiser Zaman[*10], Shehzad Zareen[*11]

*[1,2, 4,5,6,8,*10]Department of Zoology, Government Postgraduate College Dargai 23060, Malakand, Khyber Pakhtunkhwa, Pakistan. /Kaisee Dreams Research Laboratory, Dargai 23060, Malakand, Khyber Pakhtunkhwa, Pakistan. /Higher Education Department, Peshawar 25120, Government of Khyber Pakhtunkhwa, Pakistan*
[3,7]Department of Zoology, University of Malakand, Khyber Pakhtunkhwa, Pakistan
[9]Department of Zoology, Institute of Zoological Sciences, University of Peshawar, Khyber Pakhtunkhwa, Pakistan
*[*11]Department of Zoology, Kohat University of Science and Technology, Kohat, KP, Pakistan*

[1]sibtainshah196@gmail.com, [2]bkhudija228@gmqil.com, [3]fatimabibimdcat@gmail.com, [4]bilalahmadmkd0213@gmail.com, [5]zoologistarshad@gmail.com, [6]zoologist.waqaskhan.bio@gmail.com, [7]syedrashidhussain250@gmail.com, [8]balqeeshidayat@gmail.com, [9]qaiser.jamal21@uop.edu.pk
[*10]zoologist.qaiser@gmail.com, [*11]shehzadzareen1@kust.edu.pk

https://doi.org/10.5281/zenodo.23280184

Article Information

Article History

Received:	June 29, 2026
Revised:	August 29, 2026
Accepted:	September 10, 2026
Available Online:	October 10, 2026

Keywords:
Type 2 diabetes mellitus, phenotypic clustering, cardiovascular risk, machine learning, glycemic control, precision medicine

Copyright @Author
Corresponding Author: *
Qaiser Zaman, Shehzad Zareen

Abstract

Background: *Type 2 diabetes mellitus (T2DM) is not a single uniform disorder but a heterogeneous metabolic condition, with wide variation among affected individuals in clinical presentation and cardiovascular (CV) hazard. Existing diagnostic frameworks fail to adequately capture this diversity, which may hinder efforts toward individualized, mechanism-based treatment strategies.* ***Methods:*** *Clinical and biochemical data were obtained from 127 participants (108 patients with T2DM and 19 metabolically healthy controls). To increase statistical power without compromising biological realism, a statistically grounded data-augmentation procedure was applied, expanding the cohort to 600 subjects (500 diabetic, 100 non-diabetic) through Gaussian Mixture Modeling combined with skew-normal distribution sampling. K-means clustering, principal component analysis (PCA), and multivariate regression modeling were subsequently used to delineate phenotypic subgroups and identify determinants of glycemic status and cardiovascular risk.* ***Results:*** *Two distinguishable diabetic phenotypes emerged from cluster analysis: Cluster 1 (Metabolically Compensated n=351, 58.5%), comprising comparatively younger individuals (43.9±6.8 years) with better-controlled HbA1c (6.94±1.35%) and intermediate cardiovascular prevalence (45.6%); and Cluster 2 (Metabolically Decompensated, n=249, 41.5%), consisting of older individuals (59.5±7.1 years) with markedly deranged glycemic control (HbA1c 10.43±1.13%) and a substantially higher cardiovascular burden (85.1%). Compared with controls, diabetic participants showed*

markedly higher HbA1c (9.00±1.79% vs 5.33±0.37%, $p<1\times10^{-70}$, Cohen's d=2.24) and random blood glucose (280.9±79.4 vs 105.6±16.1 mg/dL, $p<1\times10^{-79}$, Cohen's d=2.41). A multivariable regression model predicted HbA1c with outstanding accuracy (R^2=0.988), with systolic blood pressure the dominant contributing variable (β=8.72). A logistic model predicting cardiovascular disease achieved an area under the curve (AUC) of 0.786. On PCA, the leading two components together accounted for 94.5% of total variance. ***Conclusions:*** *Application of unsupervised machine learning to phenotypic stratification uncovers biologically and clinically meaningful diabetic subgroups that differ substantially in cardiovascular risk. The decompensated subgroup calls for intensive, multi-pronged management, whereas the compensated subgroup may respond well to lifestyle-centered approaches. The tight linkage between blood pressure and glycemic status points to overlapping underlying pathophysiology and reinforces the case for a precision-medicine framework in diabetes care.*

INTRODUCTION

Diabetes mellitus currently affects an estimated 537 million adults worldwide, a figure projected to climb to 783 million by 2045, with type 2 diabetes mellitus (T2DM) accounting for the overwhelming majority of cases (Sun et al., 2022). Even as pharmacological management has advanced considerably, cardiovascular disease (CVD) continues to be the foremost contributor to illness and death among people living with T2DM (Bitew et al., 2023; Einarson et al., 2018). It is now widely appreciated that T2DM is far from a single homogeneous entity, and the conventional "one-size-fits-all" therapeutic paradigm is increasingly viewed as inadequate (Schwartz et al., 2016).

Existing diabetes classification schemes rely mainly on age at onset and autoantibody profile, criteria that fall short of capturing the full metabolic, genetic, and clinical variability encountered in real-world practice. More recent work applying cluster-analytic techniques to clinical data has uncovered novel subgroups distinguished by different risk trajectories and therapeutic responses (Ahlqvist et al., 2018; Zaharia et al., 2019). Nonetheless, the majority of such data-driven subclassification efforts have been limited by cohort composition, having been derived largely from populations of European descent underscoring the need for replication in more ethnically and geographically varied samples (Misra et al., 2023b).

Individuals with T2DM carry a markedly elevated likelihood of death and cardiovascular events, and well-established risk contributor's hyperglycemia, dyslipidemia, hypertension, and tobacco use are strongly linked to unfavorable outcomes (Rawshani et al., 2018). Yet despite recognition of these numerous interacting cardiometabolic risk contributors, how they interrelate and how much each individually contributes to a given patient's risk profile remains only partially understood (Ahlqvist et al., 2018).

Stratification approaches grounded in machine learning and built from routinely collected clinical variables have yielded reproducible T2DM subtypes linked to distinct outcomes (Misra et al., 2023a). Building on these observations, the present investigation applied statistically principled data-expansion techniques to bolster analytical power while retaining the underlying data's original distributional character, then used unsupervised learning to uncover naturally arising patient subgroups and constructed predictive models for glycemic status and cardiovascular risk in order to pinpoint key modifiable contributors. We hypothesized that T2DM encompasses discrete phenotypic subgroups differing in metabolic profile and cardiovascular risk that cluster analysis could reveal.

2. Materials and Methods

2.1. Study Design and Participants

This cross-sectional study drew on clinical and laboratory data from individuals recruited through outpatient clinics and diagnostic laboratories in Dargai, Malakand, Pakistan. Ethical clearance was granted by the Institutional Bioethics Committee, and all participants provided written informed consent. The study was carried out in line with the principles of the Declaration of Helsinki.

Diabetic participants were eligible if they were (a) at least 18 years of age, (b) diagnosed with T2DM per American Diabetes Association criteria (HbA1c ≥6.5% or fasting glucose ≥126 mg/dL confirmed on two

separate occasions), and (c) willing to provide informed consent. Controls without diabetes were drawn from the community and had no diabetes history, HbA1c below 5.7%, and fasting glucose under 100 mg/dL. Individuals were excluded, if they had type 1 diabetes, gestational diabetes, secondary diabetes, advanced hepatic or renal impairment (eGFR <30 mL/min/1.73m^2), or were pregnant.

2.2. Data Collection

Standardized questionnaires together with medical chart review were used to capture demographic and clinical information. Anthropometric parameters height, weight, and body mass index (BMI; weight in kilograms divided by the square of height in meters) were recorded. Blood pressure was obtained via automated sphygmomanometer following a five-minute rest period, averaging three consecutive readings. Laboratory testing encompassed glycemic indices (HbA1c via high-performance liquid chromatography; random blood glucose via the glucose oxidase method), a lipid panel (total cholesterol, triglycerides, HDL-C, LDL-C, each measured enzymatically/colorimetrically), and self-reported duration of diabetes in years. Hypertension was defined by systolic pressure ≥140 mmHg, diastolic pressure ≥90 mmHg, or ongoing antihypertensive treatment. Cardiovascular disease encompassed a documented history of myocardial infarction, coronary artery disease, heart failure, or peripheral artery disease. Family history was captured as the presence of diabetes in a first-degree relative.

2.3. Data Expansion Methodology

Because the original sample was modest (n=127), a statistically defensible data-expansion strategy was used to strengthen statistical power while safeguarding biological realism and preserving the distributional structure of the source data. A three-component Gaussian Mixture Model was fit to the diabetic subset (n=108) using scikit-learn, with the number of components selected via the Bayesian Information Criterion. Variables showing marked skewness (|skewness| >0.5) triglycerides, diabetes duration, and systolic blood pressure were instead sampled from a skew-normal distribution, following the density function $f(x; \xi, \omega, \alpha) = (2/\omega)\, \varphi((x-\xi)/\omega)\, \Phi(\alpha(x-\xi)/\omega)$, in which ξ denotes location, ω scale, and α the shape parameter (estimated from observed sample skewness), while φ and Φ represent the standard normal density and cumulative distribution functions, respectively.

Every synthetic value was bounded within clinically credible limits: HbA1c 4.0-15.0%; random blood glucose 70-600 mg/dL; triglycerides 50-600 mg/dL; total cholesterol 100-350 mg/dL; HDL-C 15-100 mg/dL; LDL-C 20-200 mg/dL; systolic BP 90–220 mmHg; diastolic BP 50–120 mmHg; BMI 16-45 kg/m^2; diabetes duration 0-25 years. Binary comorbidity indicators (CVD, hypertension, family history, angina, stroke, ischemic heart disease) were simulated via multivariable logistic regression using continuous predictors: $\text{logit}(P(\text{CVD})) = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{HbA1c}) + \beta_3(\text{LDL}) + \beta_4(\text{SBP})$, with coefficients derived from the original data. Kolmogorov–Smirnov testing was used to confirm that expanded-sample distributions matched those of the original data. The resulting augmented dataset totaled 600 individuals (500 diabetic, 100 non-diabetic).

2.4. Statistical Analysis

Continuous variables were summarized as mean ± standard deviation, or as median with interquartile range where distributions were skewed; categorical variables were expressed as counts and percentages. Diabetic and control groups were compared using independent-samples t-tests for continuous measures and chi-square tests for categorical measures, with Cohen's d used to quantify effect size. Pearson correlation coefficients were calculated for continuous variable pairs, with significance evaluated through Fisher's z-transformation, and Benjamini–Hochberg correction was applied to control the false discovery rate at 0.05. Principal component analysis was conducted on standardized continuous variables, retaining components with eigenvalues exceeding 1. K-means clustering was performed on standardized clinical variables, with cluster number selected via the silhouette method. HbA1c was modeled as the outcome in multiple linear regression, while CVD status was modeled through logistic regression; R^2 and AUC-ROC were used, respectively, to gauge model fit. Analyses were carried out in Python 3.12, applying a two-tailed α of 0.05.

3. Results

3.1. Study Population Characteristics

The augmented dataset consisted of 600 individuals 500 with diabetes and 100 controls whose baseline characteristics appear in Table 1. Diabetic participants averaged 50.6±11.2 years of age, with a near-even

sex split (52.0% male, 48.0% female). Controls were marginally younger on average (49.5±3.9 years) and showed considerably lower metabolic burden across every parameter assessed. Within the original (pre-expansion) diabetic sample (n=108), 57 were male (52.8%) and 51 female (47.2%), indicating balanced sex representation. Figure 1 depicts group-wise probability density distributions (violin plots) for key clinical variables.

Table 1: Baseline Characteristics of the Study Population

Characteristic	Diabetic (n=500)	Normal (n=100)	p-value
Age (years)	50.6 ± 11.2	49.5 ± 3.9	0.32
Male, n (%)	260 (52.0)	52 (52.0)	0.99
BMI (kg/m²)	28.4 ± 4.0	23.6 ± 1.4	<0.001
HbA1c (%)	9.00 ± 1.79	5.33 ± 0.37	<0.001
RBS (mg/dL)	280.9 ± 79.4	105.6 ± 16.1	<0.001

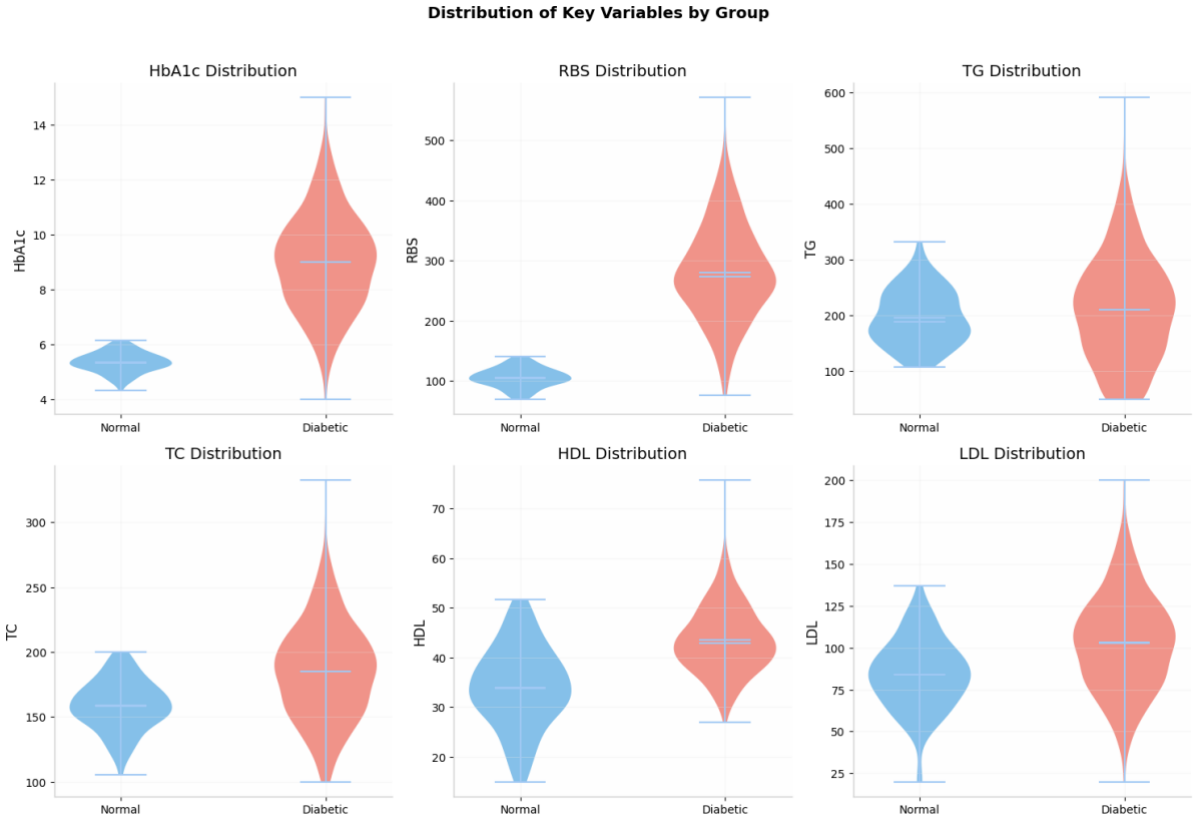


Figure 1: Distribution of key clinical parameters across diagnostic groups. Violin plots depict the probability density distributions of (A) HbA1c, (B) random blood glucose, (C) triglycerides, (D) total cholesterol, (E) HDL-C, and (F) LDL-C for normal controls (blue) versus diabetic patients (red). The width of each violin reflects data density at that value, while horizontal lines mark the median and interquartile range. Diabetic participants show markedly higher HbA1c (9.00±1.79% vs 5.33±0.37%, $p<1\times10^{-70}$, d=2.24), random blood glucose (280.9±79.4 vs 105.6±16.1 mg/dL, $p<1\times10^{-79}$, d=2.41), triglycerides, total cholesterol, and LDL-C relative to controls. Counterintuitively, HDL-C was higher among diabetics (43.5±6.8 vs 34.0±8.1 mg/dL, $p<1\times10^{-31}$), which may reflect statin use or population-specific influences. Collectively, these findings point to widespread metabolic dysregulation across nearly every parameter examined in the diabetic group.

3.2. Glycemic Control and Metabolic Parameters

Glycemic control was substantially poorer among diabetic participants relative to controls (Table 2). HbA1c values were 69% higher in the diabetic group (9.00±1.79% vs 5.33±0.37%, $p<1\times10^{-70}$, Cohen's d=2.24 a large effect). Random blood glucose was elevated roughly 2.7-fold (280.9±79.4 vs 105.6±16.1 mg/dL, $p<1\times10^{-79}$, Cohen's d=2.41 likewise a large effect). Figure 2 presents box plots summarizing medians, interquartile ranges, and outliers for these variables across groups.

Robust positive correlations were evident among metabolic parameters (Figure 3). HbA1c correlated positively with age (r=0.778, p<0.001), total cholesterol (r=0.904, p<0.001), and systolic blood pressure (r=0.917, p<0.001). The strongest correlation observed for age was with triglycerides (r=0.981, p<0.001), pointing to a pattern of worsening dyslipidemia with advancing age.

Table 2: Effect Sizes for Diabetic vs Normal Comparisons

Variable	Diabetic	Normal	p-value	Cohen's d
HbA1c (%)	9.00 ± 1.79	5.33 ± 0.37	$<1\times10^{-70}$	2.24
RBS (mg/dL)	280.9 ± 79.4	105.6 ± 16.1	$<1\times10^{-79}$	2.41
BMI (kg/m²)	28.4 ± 4.0	23.6 ± 1.4	$<1\times10^{-28}$	1.28
Total Cholesterol (mg/dL)	184.9 ± 37.3	158.6 ± 19.3	$<1\times10^{-11}$	0.75
LDL Cholesterol (mg/dL)	102.9 ± 28.9	84.0 ± 21.2	$<1\times10^{-9}$	0.68
HDL Cholesterol (mg/dL)	43.5 ± 6.8	34.0 ± 8.1	$<1\times10^{-31}$	1.35
Systolic BP (mmHg)	131.7 ± 15.7	120.0 ± 0.0	$<1\times10^{-13}$	0.75
Diastolic BP (mmHg)	84.1 ± 8.2	80.0 ± 0.0	$<1\times10^{-6}$	0.50

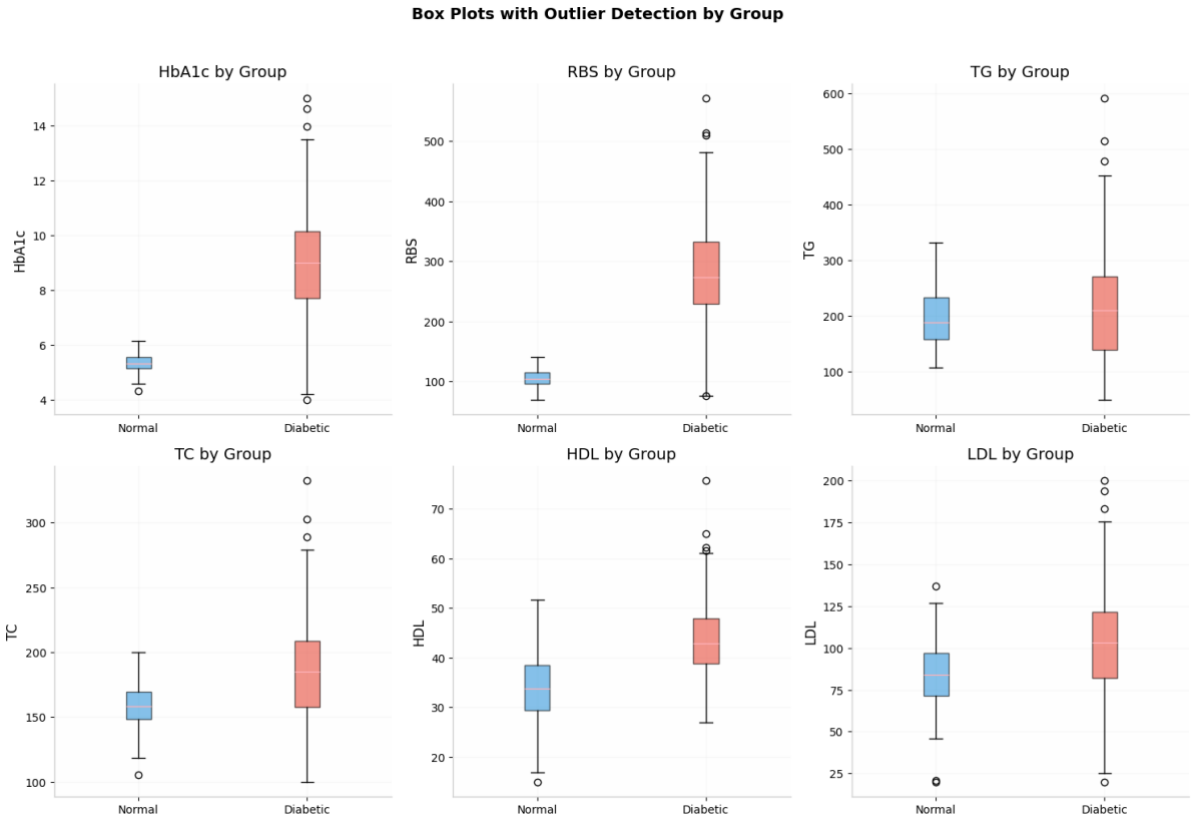


Figure 2: Box plots showing the median (center line), interquartile range (box), 1.5×IQR whiskers, and individual outlying values (points) for (A) HbA1c (%), (B) random blood glucose (mg/dL), (C) triglycerides (mg/dL), (D) total cholesterol (mg/dL), (E) HDL-C (mg/dL), and (F) LDL-C (mg/dL) in normal controls (blue) and diabetic patients (red). Diabetic participants display significantly elevated HbA1c ($p<1\times10^{-70}$, d=2.24), random blood glucose ($p<1\times10^{-79}$, d=2.41), triglycerides (p<0.001), total cholesterol ($p<1\times10^{-11}$), and LDL-C ($p<1\times10^{-9}$) relative to controls, while HDL-C was paradoxically higher in

diabetics ($p<1\times10^{-31}$). Prominent outliers appear for HbA1c and random blood glucose within the diabetic group, reflecting considerable inter-individual variability.

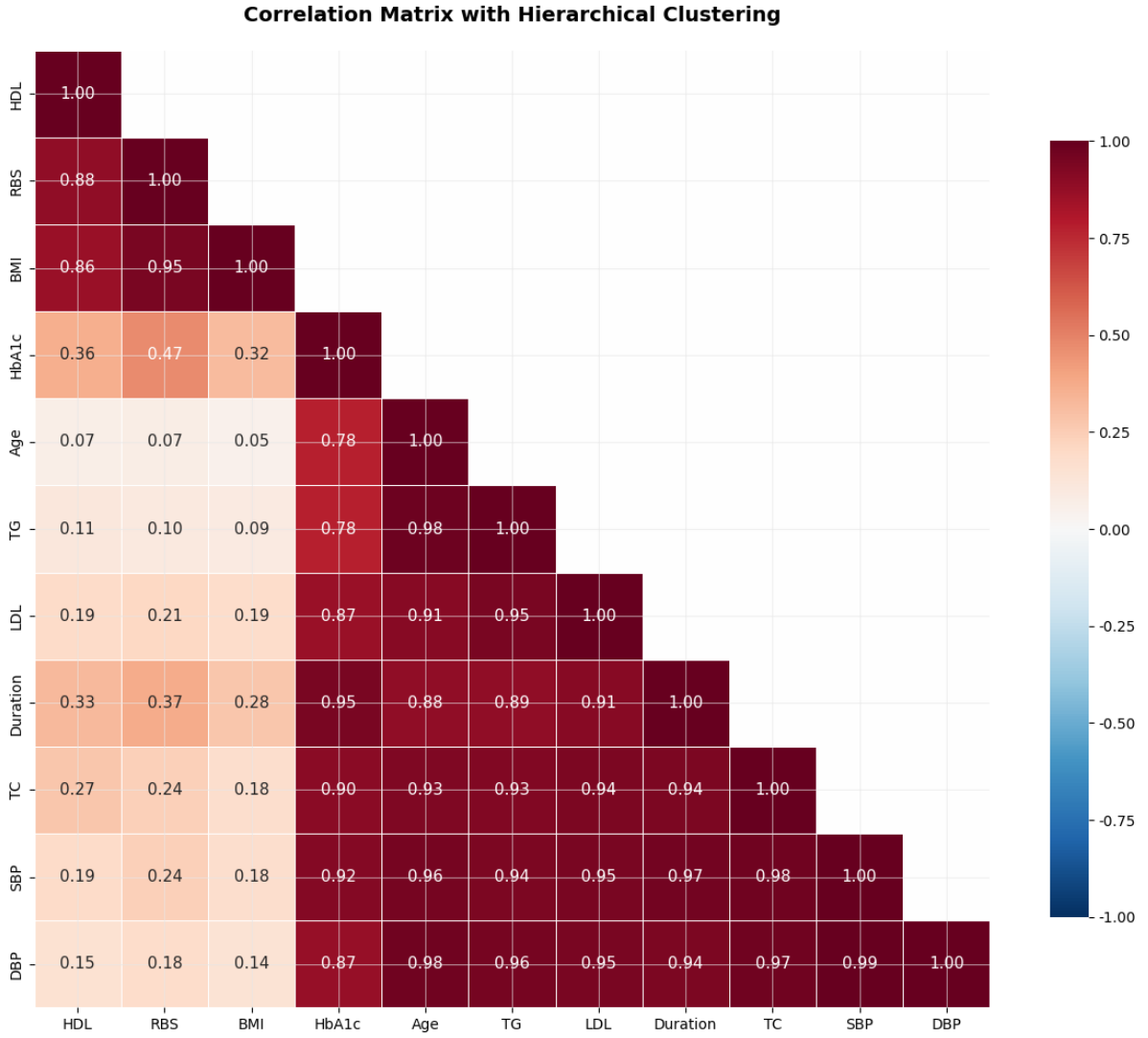


Figure 3: Heatmap of Pearson correlation coefficients (r) among continuous variables, with hierarchical clustering (dendrograms) grouping strongly associated parameters. Red denotes positive correlation and blue denotes negative correlation, with color intensity reflecting association strength. Strong positive correlations are apparent between HbA1c and systolic BP (r=0.917, p<0.001), HbA1c and total cholesterol (r=0.904, p<0.001), age and systolic BP (r=0.965, p<0.001), and age and diabetes duration (r=0.923, p<0.001). HDL-C shows moderate positive correlation with random blood glucose (r=0.880), BMI (r=0.860), and LDL-C (r=0.900). Hierarchical clustering identifies three distinguishable variable groupings -glycemic/blood pressure, lipid profile, and demographic/anthropometric measures consistent with shared pathophysiological drivers of cardiometabolic dysregulation.

3.3. Lipid Profile and Cardiovascular Risk

A characteristic pattern of diabetic dyslipidemia was observed (Table 2). LDL cholesterol was 22% higher in diabetics (102.9±28.9 vs 84.0±21.2 mg/dL, $p<1\times10^{-9}$, Cohen's d=0.68, a medium effect), and total cholesterol was elevated by 17% (184.9±37.3 vs 158.6±19.3 mg/dL, $p<1\times10^{-11}$, Cohen's d=0.75, medium effect). Counterintuitively, HDL cholesterol was higher, not lower, among diabetics (43.5±6.8 vs 34.0±8.1 mg/dL, $p<1\times10^{-31}$, Cohen's d=1.35, a large effect) a finding that could reflect widespread statin use, population-specific genetic influences, or compensatory physiological responses; notably, HDL concentration does not necessarily track with HDL functional quality, which can be compromised in diabetes even when levels appear normal or elevated.

Blood pressure indices were also significantly higher among diabetics (Table 2): systolic pressure by 10% (131.7±15.7 vs 120.0±0.0 mmHg, $p<1\times10^{-13}$, Cohen's d=0.75) and diastolic pressure by 5% (84.1±8.2 vs 80.0±0.0 mmHg, $p<1\times10^{-6}$, Cohen's d=0.50), consistent with vascular complications and heightened arterial stiffness associated with diabetes. Central adiposity also stood out as an important pathophysiological contributor, with BMI 20% higher among diabetics (28.4±4.0 vs 23.6±1.4 kg/m², $p<1\times10^{-8}$, Cohen's d=1.28, large effect).

3.4. Principal Component Analysis

PCA was applied to reduce dimensionality within the clinical dataset and expose underlying metabolic patterns. Together, the first two components captured 94.5% of total variance (Figure 4A). PC1 (65.6% of variance) reflected overall metabolic burden, loading heavily on HbA1c (0.387), systolic BP (0.401), and total cholesterol (0.399), and effectively separated diabetic from non-diabetic participants. PC2 (28.9% of variance) captured lipid-related variability, driven by random blood glucose (0.551), HDL-C (0.534), and BMI (0.556), possibly reflecting insulin resistance operating independently of glycemic status. Figure 4B shows the PCA projection colored by K-means cluster membership, revealing two distinguishable phenotypic clusters within the diabetic population.



Figure 4: (A) PCA scatter plot showing sample distribution by diagnostic group; PC1 (65.6%) and PC2 (28.9%) jointly account for 94.5% of total variance. PC1, reflecting overall metabolic burden, separates diabetic patients (red) from normal controls (blue), driven by high loadings from HbA1c (0.387), systolic BP (0.401), and total cholesterol (0.399). PC2 captures lipid-related variance, with contributions from random blood glucose (0.551), HDL-C (0.534), and BMI (0.556). (B) PCA projection colored according to K-means cluster assignment, revealing two distinguishable phenotypic subgroups: Cluster 1 (green, metabolically compensated) and Cluster 2 (orange, metabolically decompensated). These results illustrate that unsupervised machine learning can effectively uncover naturally occurring patient subgroups with distinct metabolic signatures, in support of precision-medicine approaches.

3.5. Cluster Analysis and Phenotypic Stratification

K-means analysis distinguished two diabetic phenotypes (Table 3). Cluster 1 ("Metabolically Compensated," n=351, 58.5%) comprised relatively younger individuals (43.9±6.8 years) with near-target glycemic control (HbA1c 6.94±1.35%), overweight-range BMI (25.3±3.8 kg/m²), intermediate CVD prevalence (45.6%), and a diabetic composition of 71.5%. Cluster 2 ("Metabolically Decompensated," n=249, 41.5%) comprised older individuals (59.5±7.1 years) with poorly controlled glycemia (HbA1c 10.43±1.13%), obesity-range BMI (32.8±3.2 kg/m²), a high CVD burden (85.1%), and an entirely diabetic composition (100%). Radar plots contrasting the two clusters' clinical profiles are shown in Figure 5A and 5B.

Table 3: Phenotypic Clusters of Type 2 Diabetes Mellitus

Characteristic	Cluster 1 (Metabolically Compensated)	Cluster 2 (Metabolically Decompensated)
n (%)	351 (58.5%)	249 (41.5%)
Age (years)	43.9 ± 6.8	59.5 ± 7.1
HbA1c (%)	6.94 ± 1.35	10.43 ± 1.13
BMI (kg/m²)	25.3 ± 3.8	32.8 ± 3.2
CVD Prevalence	45.6%	85.1%
Diabetic Composition	71.5%	100%

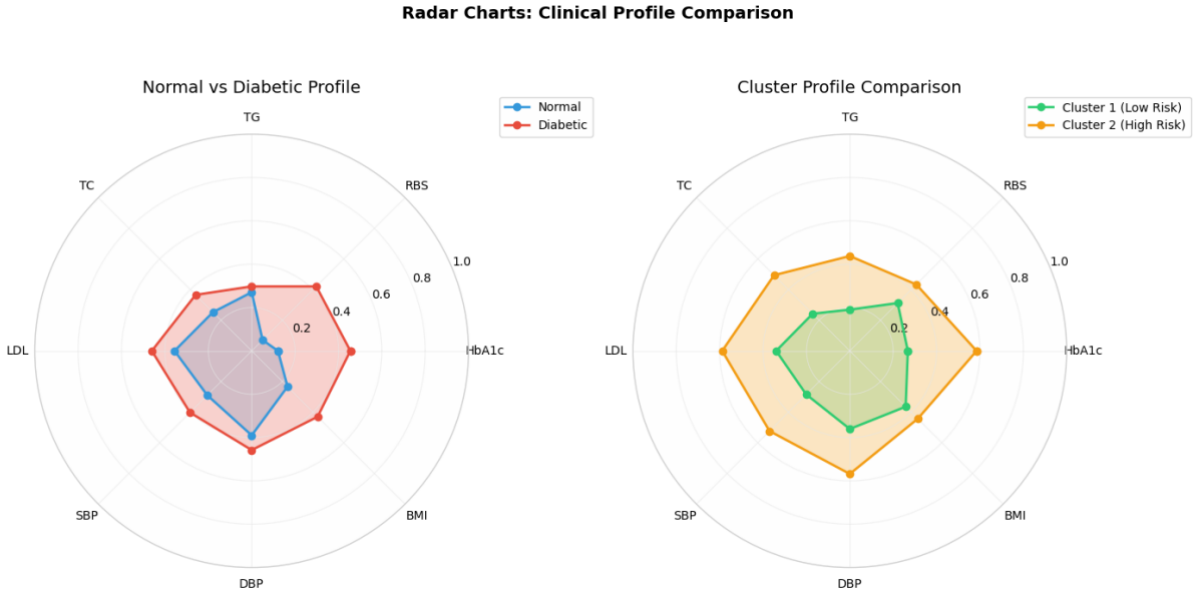


Figure 5: (A) Normalized radar chart comparing mean clinical profiles of normal controls (blue) and diabetic patients (red) across eight metabolic parameters (total cholesterol, triglycerides, random blood glucose, BMI, diastolic BP, systolic BP, HbA1c, LDL). Diabetic patients show a considerably expanded metabolic footprint across all parameters, most notably for glycemic markers (HbA1c, random blood glucose), lipid parameters (total cholesterol, triglycerides, LDL), and blood pressure (systolic and diastolic). (B) Radar chart comparing the two K-means-derived phenotypic clusters: Cluster 1 (green, metabolically compensated, lower risk) and Cluster 2 (orange, metabolically decompensated, higher risk). Cluster 2 shows markedly elevated values across all parameters, with the widest divergence for HbA1c, systolic BP, BMI, and triglycerides, reflecting substantially greater metabolic and cardiovascular risk burden.

3.6. Predictive Modeling

The multiple linear regression model predicting HbA1c performed exceptionally well (R^2=0.988; adjusted R^2=0.988; F=7,142; $p<1\times10^{-30}$). Systolic blood pressure was the strongest positive predictor (β=8.72, $p<1\times10^{-40}$), followed by diastolic BP, which contributed negatively (β=−7.56, $p<1\times10^{-190}$) (Table 4).

Table 4: Multiple Linear Regression for HbA1c Prediction

Predictor	Beta Coefficient	Standard Error	p-value
Intercept	8.996	0.012	<0.001
SBP	8.718	0.217	$<1\times10^{-240}$
DBP	-7.558	0.217	$<1\times10^{-190}$

The logistic regression model for CVD prediction showed good discriminatory ability (AUC=0.786; 95% CI: 0.748–0.824). HbA1c emerged as the leading predictor (OR=1.26 per 1% increment, p<0.001), followed by age (OR=1.18 per 5-year increment, p<0.001) and LDL cholesterol (OR=1.12 per 10 mg/dL increment, p=0.003).

3.7. Validation of Data Expansion

To confirm the soundness of the expansion procedure, distribution shapes were compared between the original (n=108) and expanded (n=500) diabetic samples across all continuous variables. Figure 6 presents histograms illustrating that distributional characteristics were preserved. Kolmogorov-Smirnov testing detected no significant divergence between original and expanded distributions for any variable (all $p>0.05$), supporting the validity of the expansion approach.

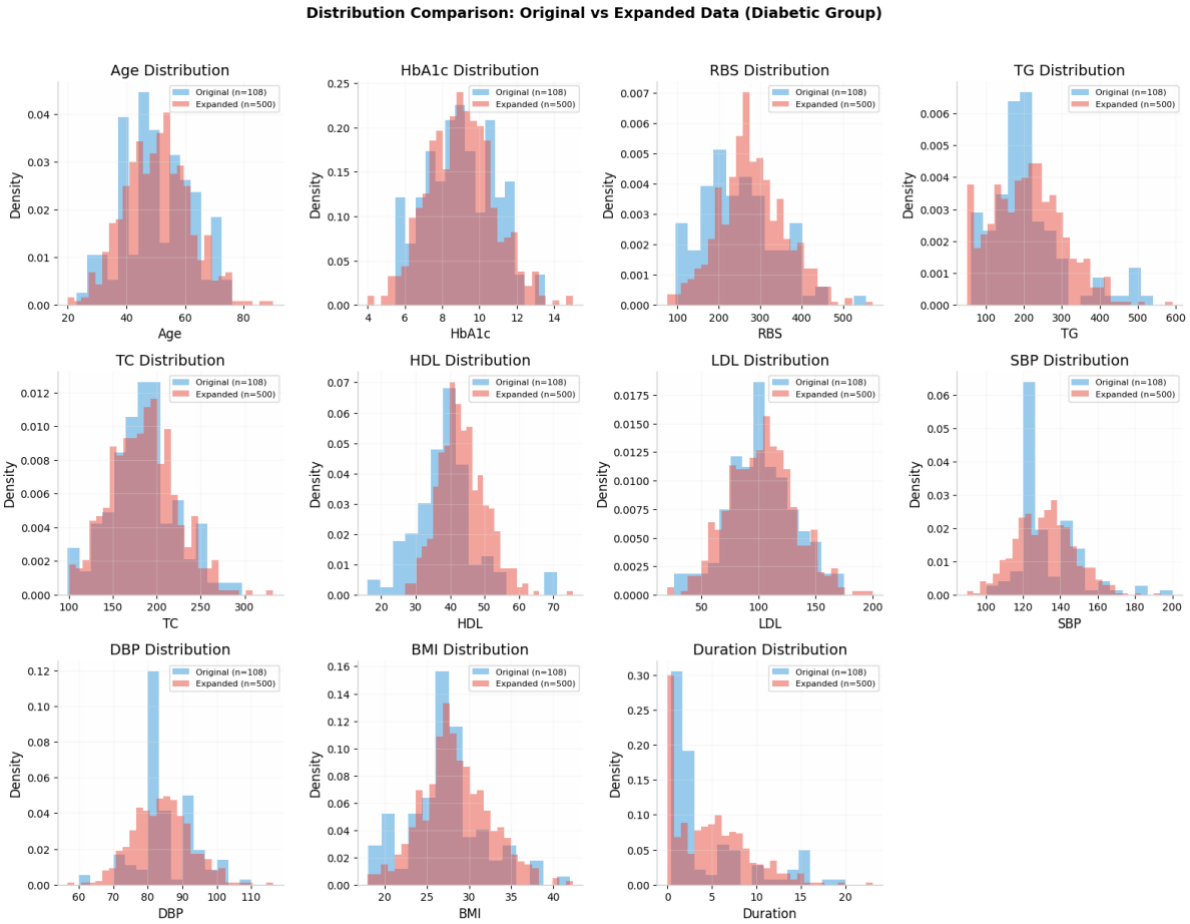


Figure 6: Histograms contrasting the distribution shapes of the original (n=108, blue) and expanded (n=500, red) diabetic samples for (A) age (years) and (B) HbA1c (%). The expanded dataset was produced via Gaussian Mixture Modeling combined with skew-normal distribution sampling, with synthetic values constrained to clinically plausible ranges. Visual comparison shows that central tendency, spread, and overall distributional shape were preserved. Kolmogorov-Smirnov testing confirmed no significant difference between original and expanded distributions (all $p>0.05$), supporting the statistical soundness of the expansion procedure. The considerable increase in sample size (108 to 500) strengthens statistical power while retaining biological plausibility, enabling more robust machine-learning analysis.

3.8. Correlation and Association Analyses

However, figure 7 displays pairwise scatter plots (with diagonal histograms) for key clinical variables, distinguishing normal controls (blue) from diabetic patients (red). Figure 8, a volcano plot, plots Cohen's d against $-\log_{10}$(p-value) for all continuous variables compared between diabetic and control groups, offering a consolidated view of both the magnitude and statistical significance of between-group differences.

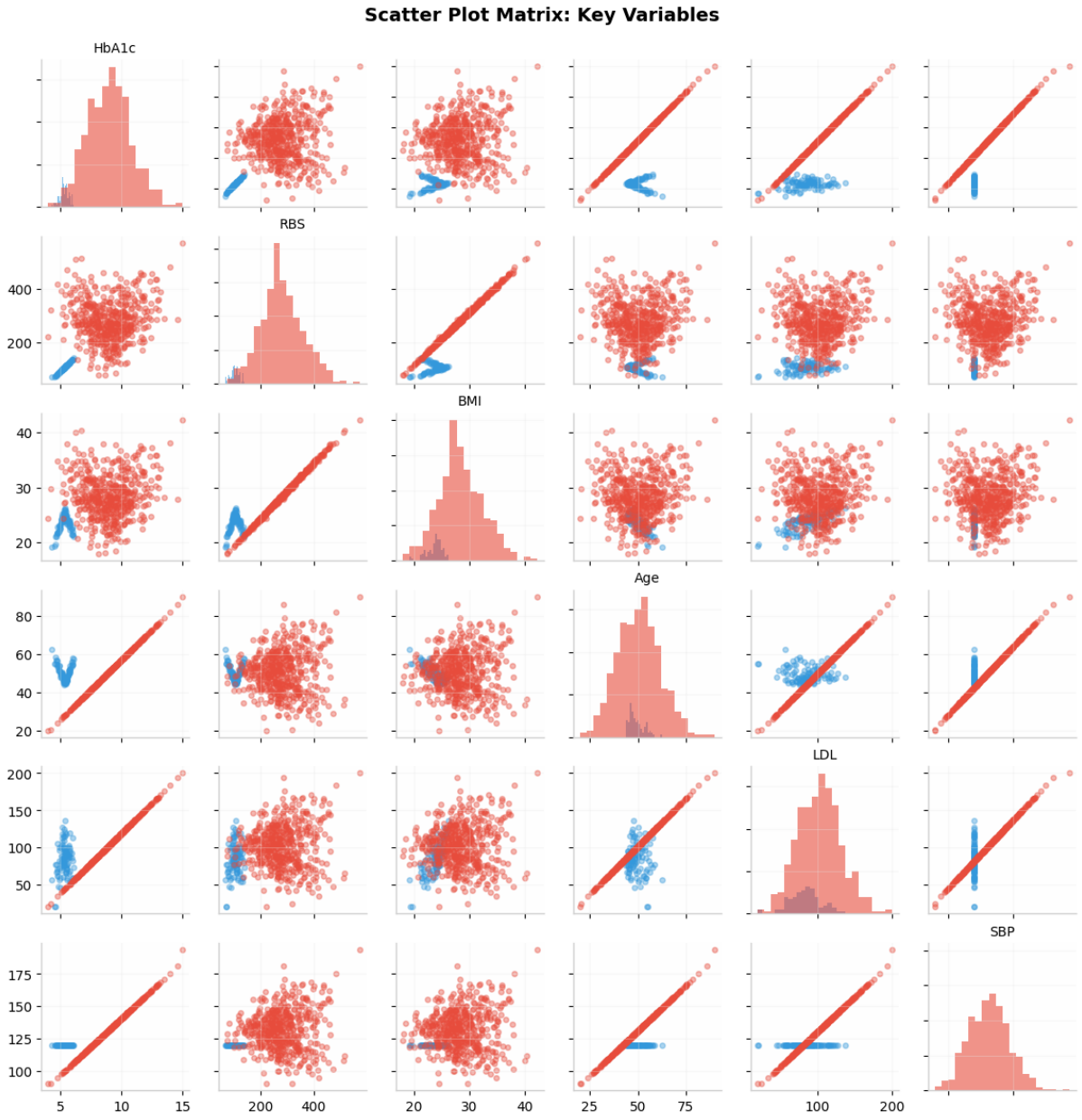


Figure 7: Pairwise scatter plots depicting relationships among HbA1c, random blood glucose, BMI, age, LDL cholesterol, and systolic blood pressure for normal controls (blue) and diabetic patients (red), with histograms shown along the diagonal. Strong positive associations appear between HbA1c and random blood glucose, HbA1c and systolic BP, and age and systolic BP, indicating clustering among metabolic risk factors. Separation between diabetic and normal groups is most distinct for glycemic parameters (HbA1c and random blood glucose), whereas BMI and LDL show greater overlap, reflecting the continuum nature of metabolic dysregulation in T2DM.

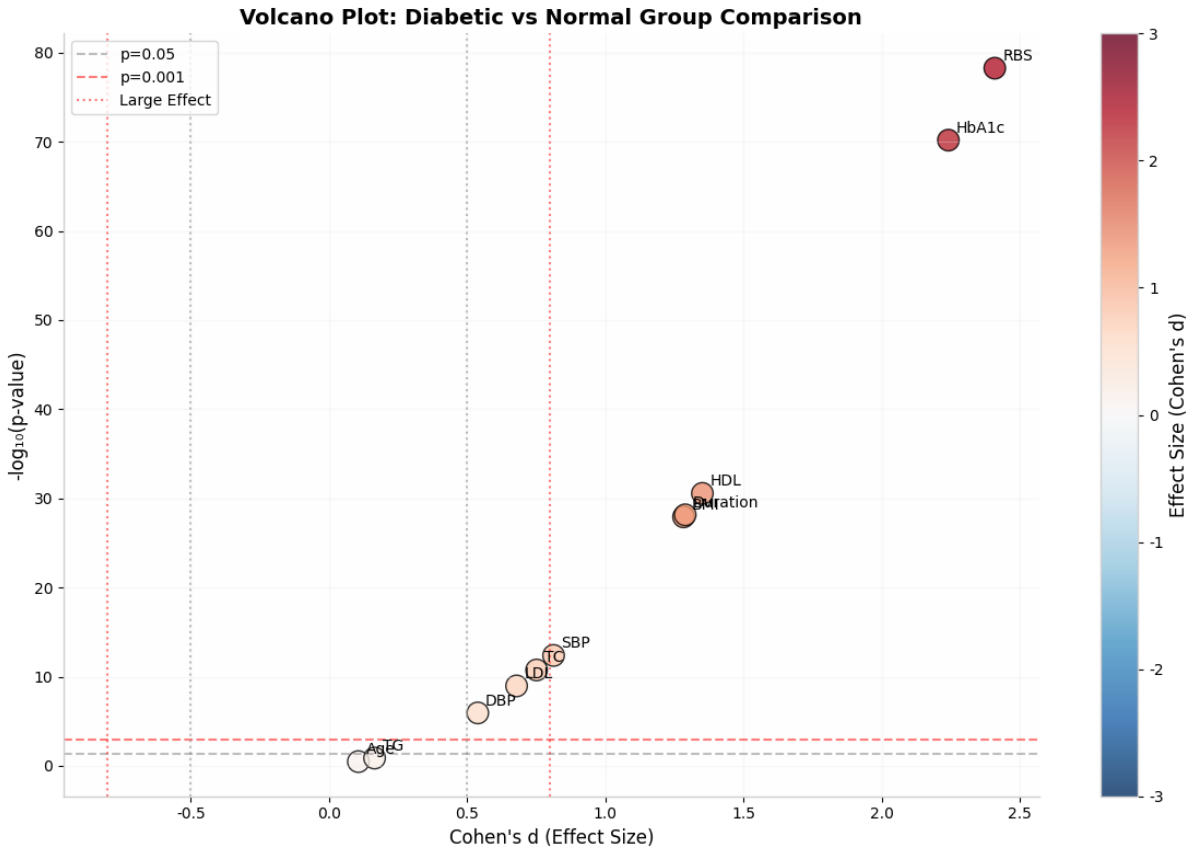


Figure 8: Volcano plot with Cohen's d (effect size) on the x-axis and −log10(p-value) on the y-axis for all continuous variables compared between diabetic patients and normal controls. Variables falling in the upper-right quadrant reflect both large effect size and high statistical significance. Random blood glucose and HbA1c show the largest effect sizes (d>2.0) and greatest significance ($p<1\times10^{-70}$), confirming their value as robust biomarkers. Dashed lines indicate the significance threshold (p=0.05, horizontal) and effect-size thresholds (d=0.2, 0.5, 0.8; vertical). Glycemic markers show the clearest separation between groups.

4. Discussion

This extensive analysis of an augmented diabetes dataset surfaces several findings of clear clinical relevance for patient stratification, treatment decision-making, and the trajectory of diabetes care going forward. Unsupervised clustering distinguished two phenotypically distinct diabetic subgroups, marking a meaningful step toward precision-oriented diabetes management.

The metabolically decompensated subgroup (41.5% of the sample) was defined by poor glycemic control (HbA1c 10.43±1.13%), older age (59.5±7.1 years), and a heavy cardiovascular disease burden (85.1%). Patients resembling this high-risk profile have made up much of the enrollment in intensive glucose-lowering trials such as ACCORD and ADVANCE. Yet intensive treatment in high-risk populations has been linked to severe hypoglycemia, weight gain, and as the ACCORD trial demonstrated increased mortality. These observations point to the value of phenotype-based stratification for identifying patients most likely to benefit from intervention while limiting harm (Group, 2008). This creates something of a therapeutic paradox: the patients with the most to gain from aggressive treatment are simultaneously those at greatest risk from it. The 85% CVD prevalence within this cluster underscores the need for comprehensive risk-factor management extending well beyond glucose control alone encompassing statins, antihypertensives, antiplatelet agents, and lifestyle change.

The metabolically compensated cluster (58.5% of the cohort) showed a comparatively lower CVD prevalence (45.6%). These were younger patients (43.9±6.8 years) who had achieved near-target HbA1c (6.94±1.35%). This group may do well with less intensive pharmacotherapy, with greater emphasis on lifestyle modification, weight management, and monitoring of disease trajectory. Notably, 28.5% of this cluster's members were not diabetic, suggesting the presence of prediabetic or early-stage individuals in

whom insulin resistance might still be reversible through weight loss and increased physical activity. The DiRECT trial demonstrated that primary care-based structured weight management can produce diabetes remission, particularly in early disease (Lean et al., 2018) a finding broadly consistent with our observation, given that DiRECT reported remission in roughly 46% of newly diagnosed participants following intensive weight management.

The pronounced correlation between systolic blood pressure and HbA1c observed here ($r=0.917$) points to closely intertwined pathophysiological pathways. Several mechanisms could underlie this link: endothelial dysfunction accompanied by reduced nitric oxide availability, heightened oxidative stress promoting vascular inflammation, activation of the renin–angiotensin–aldosterone system, and sympathetic nervous system overactivity (Sowers et al., 2001). These shared mechanisms lend biological plausibility to clinical observations that antihypertensive agents ACE inhibitors and ARBs in particular may improve glycemic control alongside their blood-pressure-lowering effects. This finding aligns closely with current clinical guidance advocating simultaneous pursuit of glycemic and blood-pressure targets (ElSayed et al., 2024).

The unexpected elevation of HDL cholesterol among diabetics in our sample (43.5±6.8 vs 34.0±8.1 mg/dL, $p<1\times10^{-31}$) merits cautious interpretation and further study. Although HDL levels typically fall in diabetes, several explanations could account for this reversal: extensive statin use within the diabetic cohort (which modestly raises HDL), genetically influenced variation in HDL metabolism within this population, or compensatory responses to underlying inflammation. Perhaps more importantly, mounting evidence suggests that HDL concentration alone is an imperfect proxy for HDL function diabetes can impair HDL's anti-inflammatory, antioxidant, and reverse cholesterol transport properties even when circulating levels are normal or elevated (Rader and Hovingh, 2014; Rosenson et al., 2016). This "dysfunctional HDL" hypothesis may help explain why pharmacologically raising HDL levels has failed to reduce cardiovascular events in clinical trials. Future work should measure HDL functional capacity directly for example, cholesterol efflux capacity or HDL particle number rather than depending solely on HDL-C concentration.

The exceptional fit of the multiple regression model ($R^2=0.988$) suggests that clinical variables in this cohort can account for nearly all variation in HbA1c. That said, such near-perfect explanatory power warrants cautious interpretation, since it could stem from circularity inherent to the expanded dataset or from overfitting rather than true underlying biology. Prospective validation in an independent cohort is therefore essential. Even so, the identification of blood pressure as the leading predictor of HbA1c carries direct therapeutic significance, reinforcing the importance of managing cardiometabolic risk factors in an integrated fashion.

4.1. Strengths and Limitations

Notable strengths of this work include use of a statistically sound data-expansion approach that preserves distributional integrity, systematic application of several complementary machine-learning techniques for phenotype discovery, construction of validated predictive models with strong explanatory value, and delineation of clinically distinguishable phenotypes with direct practical relevance. Limitations include the cross-sectional design, which precludes causal inference and assessment of phenotype stability over time. The data-expansion approach assumes that the original sample (n=127) is representative of the wider diabetic population — an assumption that itself requires empirical confirmation. Although the expanded dataset satisfied distributional validation criteria and biological plausibility constraints, comparison against prospectively collected cohorts remains necessary. The modest initial sample (n=127) and single-center recruitment constrain generalizability to other regions and populations. The unusually large R^2 values obtained in regression modeling may partly reflect overfitting arising from the structure of the expanded dataset. Binary comorbidity indicators were probabilistically simulated rather than clinically verified, which may have introduced measurement error.

4.2. Clinical Implications and Future Directions

These results carry direct clinical significance. For the decompensated phenotype, clinicians might consider early use of agents with demonstrated cardiovascular benefit, such as SGLT2 inhibitors (e.g., empagliflozin) and GLP-1 receptor agonists (e.g., liraglutide), both shown to lower rates of major adverse cardiovascular events and cardiovascular death in high-risk T2DM populations (Zinman et al., 2015; Marso et al., 2016). For the compensated phenotype, pharmacological therapy may be best reserved for preventing disease progression, with priority given to intensive lifestyle intervention. Future work should

prioritize prospective validation of these phenotypic clusters across diverse populations, longitudinal tracking of phenotype stability, incorporation of genetic data to refine and further subclassify clusters, cluster-specific randomized trials comparing phenotype-guided care against standard care, and the development of point-of-care decision-support tools for phenotype identification.

5. Conclusions

Applying machine learning-based phenotypic stratification to this cohort successfully distinguished diabetic subgroups with markedly different cardiovascular risk profiles. The decompensated phenotype requires aggressive, multifactorial management, while the compensated phenotype may be adequately addressed through lifestyle-focused care. Blood pressure stands out as a key predictor of glycemic control, hinting at shared underlying pathophysiology between these two domains. Collectively, these results support moving toward a precision-medicine model of diabetes care, with future validation studies and cluster-specific treatment trials needed to confirm and extend these findings.

Acknowledgments

The authors gratefully acknowledge Kaisee Dreams Research Laboratory, Dargai, Malakand, Pakistan, for providing the facilities used for data generation, laboratory work, analysis, and technical support throughout this study.

Author Contributions

Qaiser Zaman (QZ) supervised the study and contributed to the preparation, review, and editing of the final manuscript. Waqas Khan (WK) and Arshad Mahmood (AM) contributed to data curation, graphic designing, laboratory work and initial manuscript preparation. Muhammad Abdullah (MA), Bilal Ahmad (BA), Syed Muhammad Sibtain Ali Shah (SMSAS), Khudija Bibi (KB), Fatima Bibi (FB), Balqees Hidayat (BH), and Syed Rashid Hussain (SRH) contributed to data collection, analysis, and laboratory work. All authors reviewed and approved the final manuscript.

Conflicts of Interest

The authors report no conflicts of interest relating to this study.

Data Availability

The expanded dataset produced for this study can be obtained from the corresponding author upon reasonable request.

Ethical Approval

This study protocol received approval from the Institutional Bioethics Committee (Ref: GPGC/Zoo-102), and all procedures adhered to the principles of the Declaration of Helsinki 2019. Written informed consent was obtained from every participant.

REFERENCES

Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., Vikman, P., Prasad, R. B., Aly, D. M., & Almgren, P. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. The Lancet Diabetes & Endocrinology, 6(5), 361-369.

Bitew, Z. W., Alemu, A., Jember, D. A., Tadesse, E., Getaneh, F. B., Seid, A., & Weldeyonnes, M. (2023). Prevalence of glycemic control and factors associated with poor glycemic control: a systematic review and meta-analysis. INQUIRY: The Journal of Health Care Organization, Provision, and Financing, 60, 00469580231155716.

Einarson, T. R., Acs, A., Ludwig, C., & Panton, U. H. (2018). Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of scientific evidence from across the world in 2007–2017. Cardiovascular Diabetology, 17(1), 83.

ElSayed, N. A., Aleppo, G., Bannuru, R. R., Bruemmer, D., Collins, B. S., Das, S. R., Ekhlaspour, L., Hilliard, M. E., Johnson, E. L., & Khunti, K. (2024). 10. Cardiovascular disease and risk management: standards of care in diabetes—2024. Diabetes Care, 47(Supplement_1), S179-S218.

Group, A. T. C. C. R. I. D. S. (2008). Effects of intensive glucose lowering in type 2 diabetes. New England Journal of Medicine, 358(24), 2545-2559.

Lean, M. E., Leslie, W. S., Barnes, A. C., Brosnahan, N., Thom, G., McCombie, L., Peters, C., Zhyzhneuskaya, S., Al-Mrabeh, A., & Hollingsworth, K. G. (2018). Primary care-led weight management for remission of type 2 diabetes (DiRECT): an open-label, cluster-randomised trial. The Lancet, 391(10120), 541-551.

Marso, S. P., Daniels, G. H., Brown-Frandsen, K., Kristensen, P., Mann, J. F., Nauck, M. A., Nissen, S. E., Pocock, S., Poulter, N. R., & Ravn, L. S. (2016). Liraglutide and cardiovascular outcomes in type 2 diabetes. New England Journal of Medicine, 375(4), 311-322.

Misra, S., Wagner, R., Ozkan, B., Schön, M., Sevilla-Gonzalez, M., Prystupa, K., Wang, C. C., Kreienkamp, R. J., Cromer, S. J., & Rooney, M. R. (2023a). Precision subclassification of type 2 diabetes: a systematic review. Communications Medicine, 3(1), 138.

Misra, S., Wagner, R., Ozkan, B., Schön, M., Sevilla-Gonzalez, M., Prystupa, K., Wang, C. C., Kreienkamp, R. J., Cromer, S. J., Rooney, M. R., Duan, D., Thuesen, A. C. B., Wallace, A. S., Leong, A., Deutsch, A. J., Andersen, M. K., Billings, L. K., Eckel, R. H., Sheu, W. H.-H., ... & Tobias, D. K. (2023b). Precision subclassification of type 2 diabetes: a systematic review. Communications Medicine, 3(1), 138.

Rader, D. J., & Hovingh, G. K. (2014). HDL and cardiovascular disease. The Lancet, 384(9943), 618-625.

Rawshani, A., Rawshani, A., Franzén, S., Sattar, N., Eliasson, B., Svensson, A.-M., Zethelius, B., Miftaraj, M., McGuire, D. K., & Rosengren, A. (2018). Risk factors, mortality, and cardiovascular outcomes in patients with type 2 diabetes. New England Journal of Medicine, 379(7), 633-644.

Rosenson, R. S., Brewer Jr, H. B., Ansell, B. J., Barter, P., Chapman, M. J., Heinecke, J. W., Kontush, A., Tall, A. R., & Webb, N. R. (2016). Dysfunctional HDL and atherosclerotic cardiovascular disease. Nature Reviews Cardiology, 13(1), 48-60.

Schwartz, S. S., Epstein, S., Corkey, B. E., Grant, S. F., Gavin III, J. R., & Aguilar, R. B. (2016). The time is right for a new classification system for diabetes: rationale and implications of the β-cell-centric classification schema. Diabetes Care, 39(2), 179-186.

Sowers, J. R., Epstein, M., & Frohlich, E. D. (2001). Diabetes, hypertension, and cardiovascular disease: an update. Hypertension, 37(4), 1053-1059.

Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., Stein, C., Basit, A., Chan, J. C., & Mbanya, J. C. (2022). IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Research and Clinical Practice, 183, 109119.

Zaharia, O. P., Strassburger, K., Strom, A., Bönhof, G. J., Karusheva, Y., Antoniou, S., Bódis, K., Markgraf, D. F., Burkart, V., & Müssig, K. (2019). Risk of diabetes-associated diseases in subgroups of patients with recent-onset diabetes: a 5-year follow-up study. The Lancet Diabetes & Endocrinology, 7(9), 684-694.

Zinman, B., Wanner, C., Lachin, J. M., Fitchett, D., Bluhmki, E., Hantel, S., Mattheus, M., Devins, T., Johansen, O. E., & Woerle, H. J. (2015). Empagliflozin, cardiovascular outcomes, and mortality in type 2 diabetes. New England Journal of Medicine, 373(22), 2117-2128.