

BACKGROUND

Type 1 Diabetes (T1D) is a growing global health concern, with an estimated 8.4 million cases worldwide and projections reaching 13.5-17.4 million by 2040. Early and accurate prediction is critical for clinical intervention, particularly in pediatric populations. Machine learning has emerged as a promising tool for T1D risk stratification, with ensemble methods such as Random Forest and Gradient Boosting achieving reported accuracies exceeding 93%.

A 2025 IEEE conference paper introduced an Ensemble-Based Hybrid Voting Strategy for feature selection, combining Mutual Information, Random Forest Importance, RF-RFE, and LASSO across stratified subsets of a 322,895-record pediatric dataset sourced from hospital records in West Java, Indonesia.

The study reported a Final Hybrid Score ranking 10 clinical features, with Pancreatic and BMI ranked highest, and achieved classification accuracies of 93.58-94.13%. However, no distributional validation was performed to confirm that the selected features carry meaningful signal relative to the diagnostic outcome. This omission reflects a broader pattern in clinical ML research, where feature selection and model accuracy are prioritized over upstream data validation.

PURPOSE

This poster applies the EDA Toolkit, an open-source Python library for exploratory data analysis, to independently evaluate the dataset and feature rankings reported in a 2025 IEEE paper on pediatric T1D risk prediction.

High classifier accuracy is frequently cited as evidence of model validity, yet accuracy alone cannot confirm that a dataset carries genuine clinical signal. Without distributional validation performed prior to modeling, inflated performance metrics may reflect artifacts of synthetic data generation rather than learned clinical relationships.

Systematic EDA, applied before any modeling step, is essential for validating dataset integrity. The EDA Toolkit provides standardized functions for distributional profiling, group-level comparisons, and summary table generation that can be applied across clinical datasets without custom scripting. It further illustrates how reproducible, open-source tooling can serve as a critical checkpoint in clinical machine learning pipelines, surfacing data quality concerns that downstream model evaluation cannot detect.

METHODS

All analyses were conducted using the EDA Toolkit (v0.0.28), an open-source Python library for reproducible exploratory data analysis. The source dataset comprised 322,895 pediatric patient records with 27 features across genetic, clinical, lifestyle, and demographic categories, sourced from a 2025 IEEE conference paper on ensemble feature selection for T1D risk prediction.

Data Profile

- Generated a structural summary of null counts, unique value distributions, and dominant value frequencies across all 27 features
- Disease name and medication each contain ~25% missing values, flagging data completeness concerns in a purportedly clinical dataset
- Autoantibody markers IAA, IA-2A, and ZnT8A each contain ~319,000 unique values across 322,895 rows; this is consistent with randomly generated continuous values rather than real lab measurements, which cluster around assay thresholds

Distributional Separability Analysis

- Compared probability densities of key features across diagnostic classes

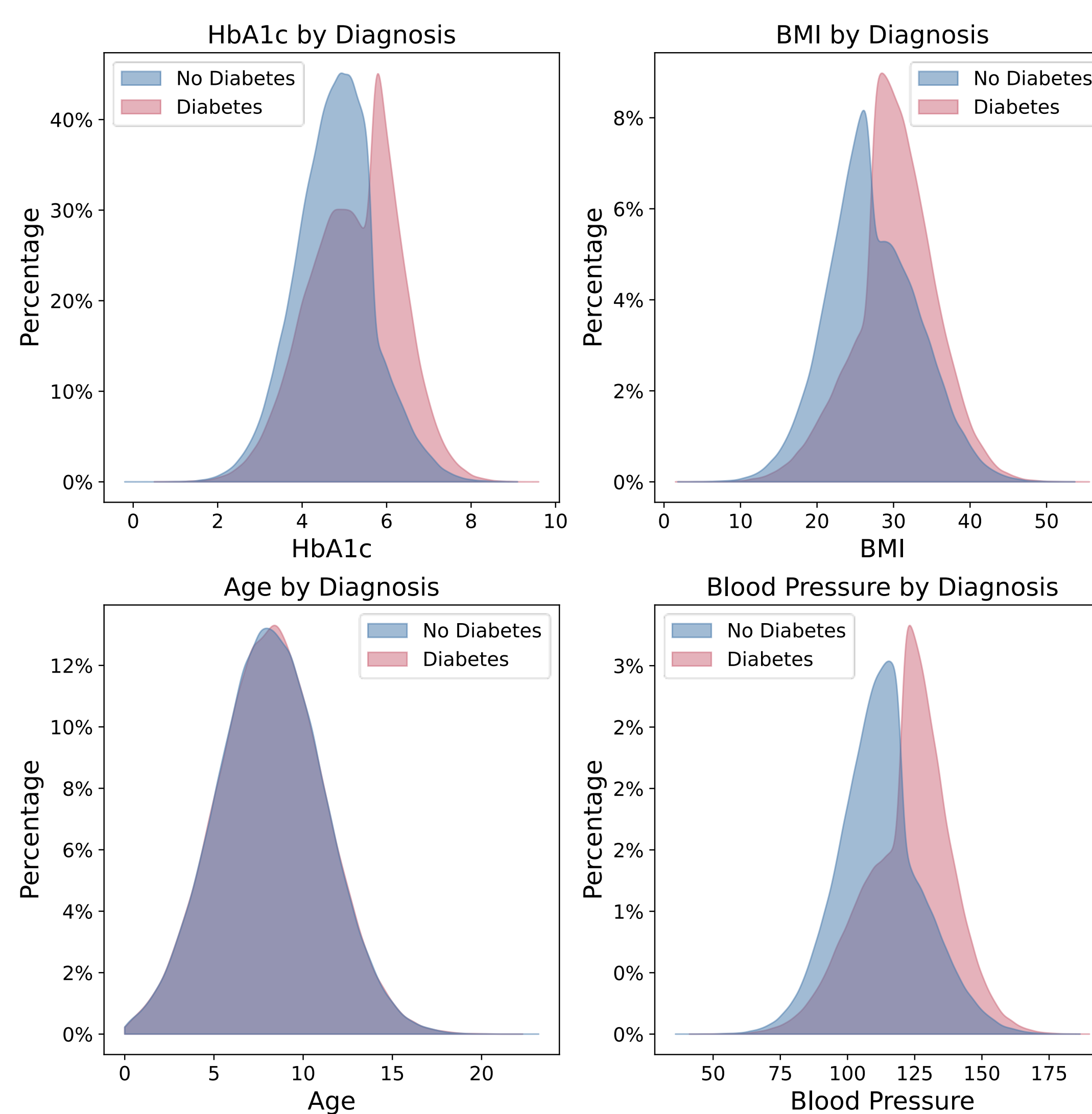


Figure 1. Probability density distributions of HbA1c, BMI, Age, and blood pressure by diagnostic class. Near-perfect overlap across all four features suggests diagnosis labels were generated independently of clinical values.

RESULTS

Table 1. Participant characteristics by diagnostic class

Variable	Overall	Diabetes (n = 133,768)	No Diabetes (n = 189,127)	P-value
BMI	28.50 (5.87)	30.24 (5.37)	27.27 (5.90)	< 0.001
HbA1c	4.97 (1.02)	5.26 (1.07)	4.77 (0.93)	< 0.001
Blood pressure	116.41 (16.67)	122.45 (16.21)	112.14 (15.64)	< 0.001
Pancreatic	0.97 (0.73)	0.97 (0.73)	0.97 (0.74)	0.36
Gender	322,895 (100%)	133,768 (41.43%)	189,127 (58.57%)	0.07
Male	161,577 (50.04%)	67,189 (50.23%)	94,388 (49.91%)	
Female	161,318 (49.96%)	66,579 (49.77%)	94,739 (50.09%)	
Age group	322,895 (100%)	133,768 (41.43%)	189,127 (58.57%)	0.43
Infant (0-1)	5,421 (1.68%)	2,233 (1.67%)	3,188 (1.69%)	
Toddler (2-4)	40,173 (12.44%)	16,733 (12.51%)	23,440 (12.39%)	
Child (5-11)	243,485 (75.41%)	100,697 (75.28%)	142,788 (75.50%)	
Adolescent (12-17)	33,622 (10.41%)	14,014 (10.48%)	19,608 (10.37%)	
Young adult (18-20)	189 (0.06%)	89 (0.07%)	100 (0.05%)	
Adult (21+)	5 (0.00%)	2 (0.00%)	3 (0.00%)	

Continuous variables presented as mean (SD) by diagnostic class. Continuous variables: Welch's t-test. Categorical variables: chi-squared test. SD = standard deviation. Red P-values indicate statistical significance (< 0.001); however, in this context significance reflects synthetic data artifacts rather than genuine clinical signal. Identical group proportions across both parent and child rows indicate distributional noise, not clinical differentiation.

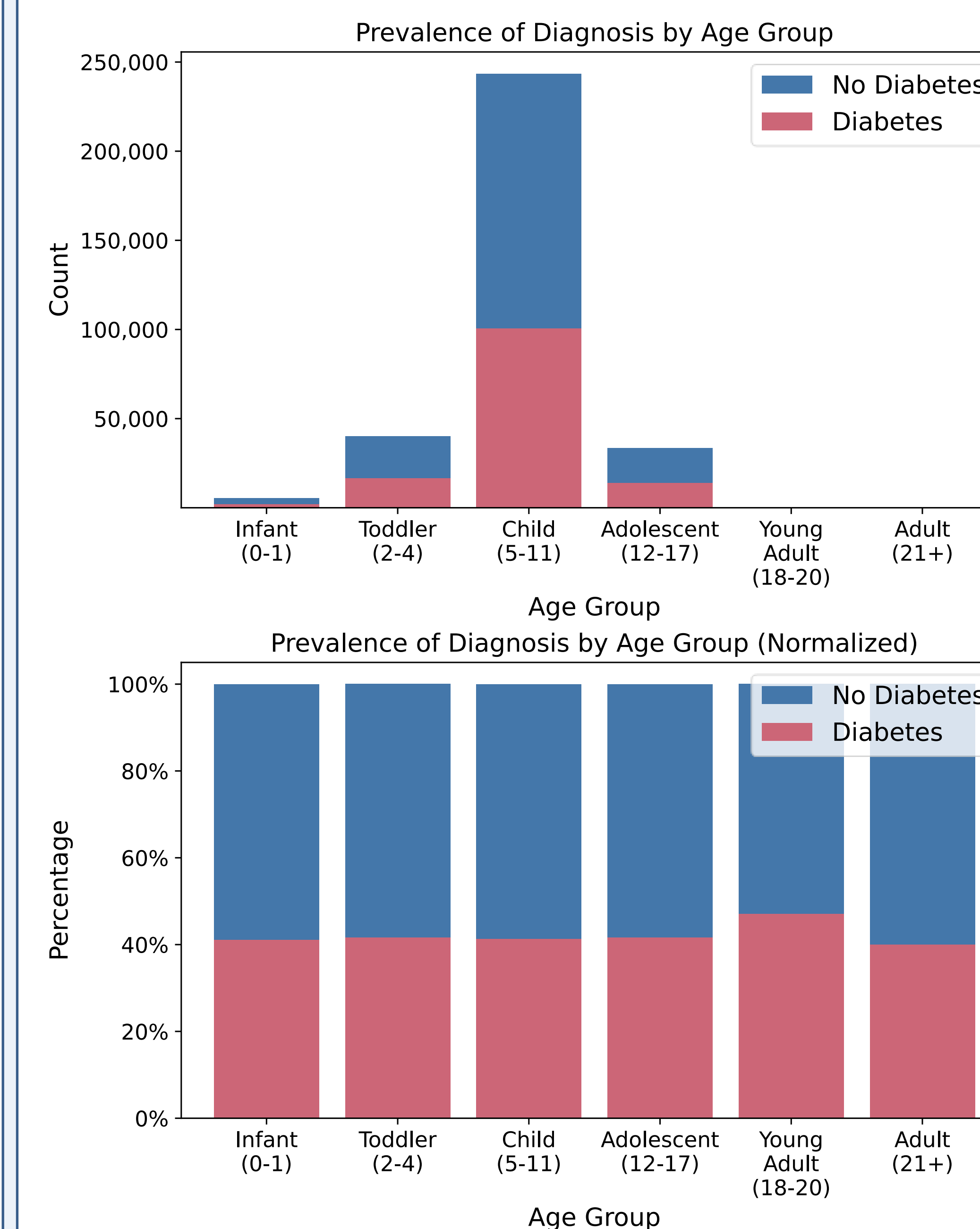


Figure 2. T1D diagnosis prevalence by age group. Diabetes rate is uniformly ~41% across all age groups.

CONCLUSIONS

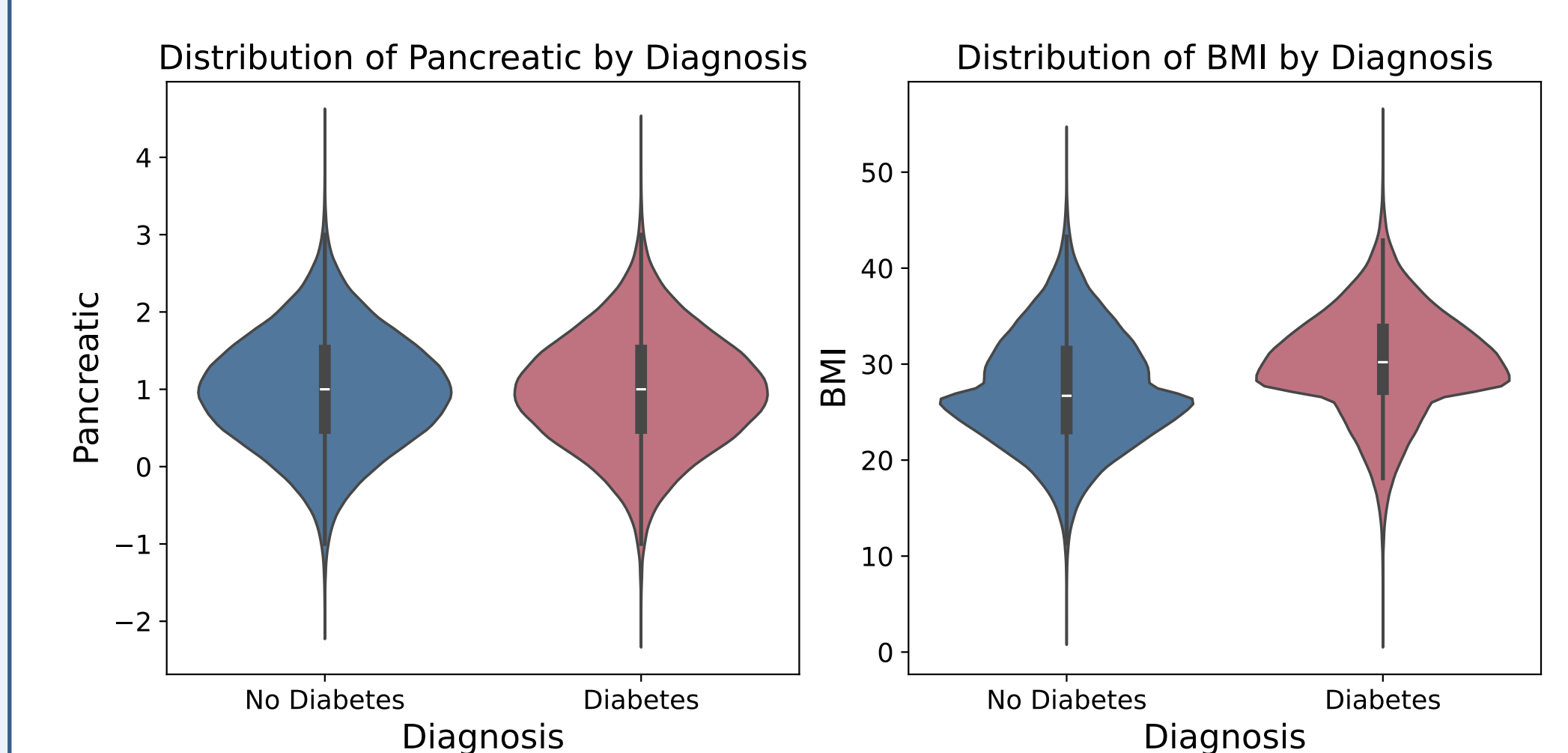


Figure 3. Distribution of the paper's top-ranked features (Pancreatic, BMI) by diagnostic class. Near-identical shape and IQR across both groups indicates no class separability.

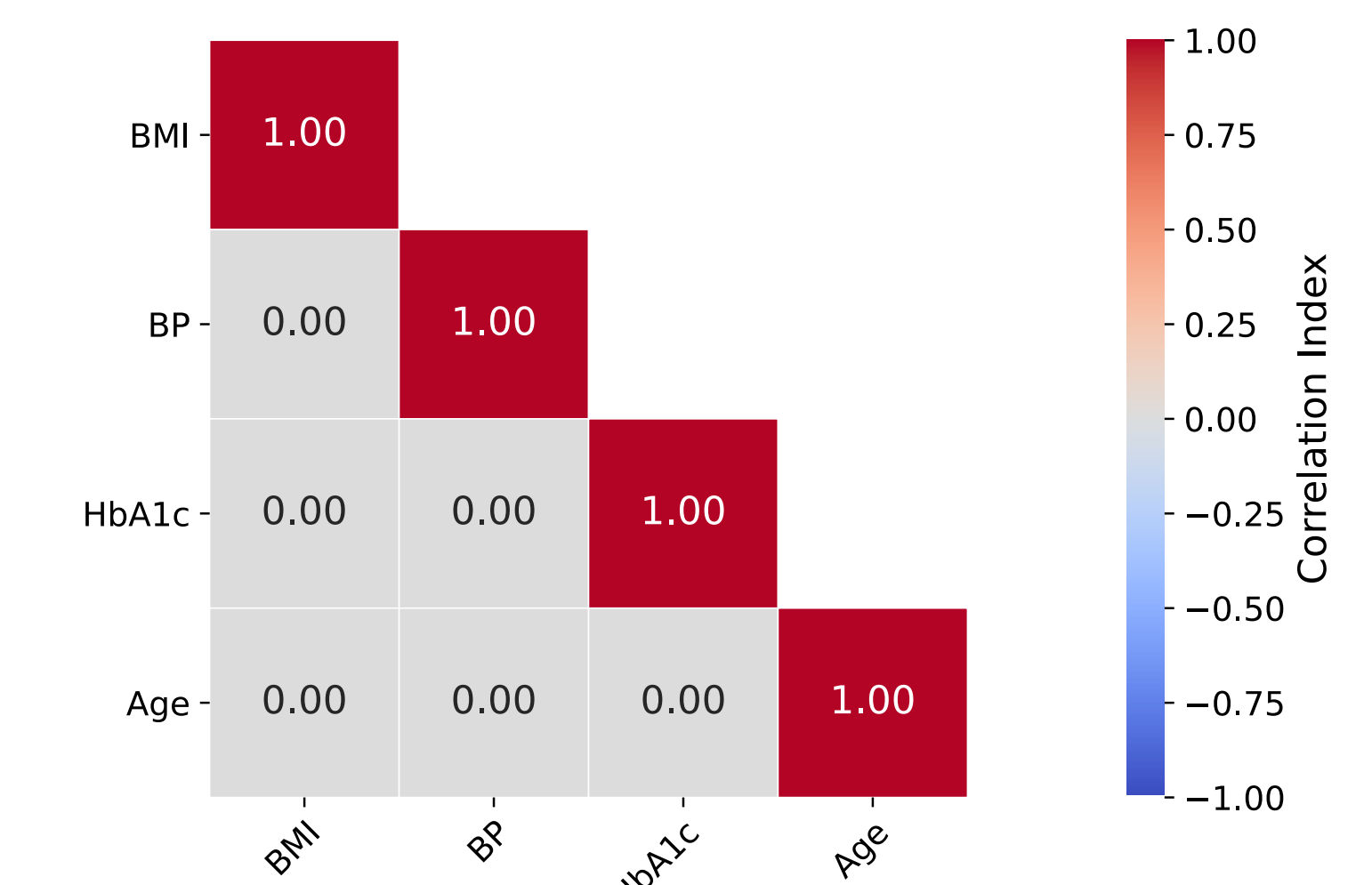


Figure 4. Pairwise correlations among clinically related features. Values approximate zero, indicating features were generated independently with no biological covariance structure.

- HbA1c, the primary T1D biomarker, showed no class separability; this is inconsistent with any genuine clinical dataset
- Pancreatic and BMI, the paper's top-ranked features, show identical distributions across diagnostic classes
- Diagnosis labels appear generated independently of clinical feature values
- Reported 93-94% model accuracies reflect synthetic data artifacts, not learned clinical relationships
- Findings highlight the critical role of EDA in validating dataset integrity prior to model development

REFERENCES

- Syahidin Y, Maulidevi NU, Surendro K. Ensemble feature selection: a hybrid approach for pediatric type 1 diabetes risk factors. In: 2025 International Conference on Smart-Green Technology in Electrical and Information Systems (ICSGTEIS); 2025 Oct 9-10; Bali, Indonesia. IEEE; 2025. p. 30-5. DOI: 10.1109/ICSGTEIS68532.2025.11284421
 - Shpaner L, Gil O. eda_toolkit [software]. Version 0.0.28. 2024. Available from: https://pypi.org/project/eda_toolkit
- Corresponding author: shpaner@ucla.edu

28

DATA SCIENCE DYNAMICS

Reproducible EDA Frameworks for Clinical Research A Pediatric Type 1 Diabetes Case Study

Leonid Shpaner, M.S.¹, Oscar Gil, M.S.²

CHOC Research
GO BEYOND

EDA-Toolkit

BACKGROUND

Type 1 Diabetes (T1D) is a growing global health concern, with an estimated 8.4 million cases worldwide and projections reaching 13.5-17.4 million by 2040. Early and accurate prediction is critical for clinical intervention, particularly in pediatric populations. Machine learning has emerged as a promising tool for T1D risk stratification, with ensemble methods such as Random Forest and Gradient Boosting achieving reported accuracies exceeding 93%.

A 2025 IEEE conference paper introduced an Ensemble-Based Hybrid Voting Strategy for feature selection, combining Mutual Information, Random Forest Importance, RF-RFE, and LASSO across stratified subsets of a 322,895-record pediatric dataset sourced from hospital records in West Java, Indonesia. The study reported a Final Hybrid Score ranking 10 clinical features, with Pancreatic and BMI ranked highest, and achieved classification accuracies of 93.58-94.13%. However, no distributional validation was performed to confirm that the selected features carry meaningful signal relative to the diagnostic outcome. This omission reflects a broader pattern in clinical ML research, where feature selection and model accuracy are prioritized over upstream data validation.

PURPOSE

This poster applies the EDA Toolkit, an open-source Python library for exploratory data analysis, to independently evaluate the dataset and feature rankings reported in a 2025 IEEE paper on pediatric T1D risk prediction.

High classifier accuracy is frequently cited as evidence of model validity, yet accuracy alone cannot confirm that a dataset carries genuine clinical signal. Without distributional validation performed prior to modeling, inflated performance metrics may reflect artifacts of synthetic data generation rather than learned clinical relationships.

Systematic EDA, applied before any modeling step, is essential for validating dataset integrity. The EDA Toolkit provides standardized functions for distributional profiling, group-level comparisons, and summary table generation that can be applied across clinical datasets without custom scripting. It further illustrates how reproducible, open-source tooling can serve as a critical checkpoint in clinical machine learning pipelines, surfacing data quality concerns that downstream model evaluation cannot detect.

1. University of California, Los Angeles; Data Science Dynamics
2. University of California, Riverside; Data Science Dynamics

METHODS

All analyses were conducted using the EDA Toolkit (v0.0.28), an open-source Python library for reproducible exploratory data analysis. The source dataset comprised 322,895 pediatric patient records with 27 features across genetic, clinical, lifestyle, and demographic categories, sourced from a 2025 IEEE conference paper on ensemble feature selection for T1D risk prediction.

Data Profile

- Generated a structural summary of null counts, unique value distributions, and dominant value frequencies across all 27 features
- Disease name and medication each contain ~25% missing values, flagging data completeness concerns in a purportedly clinical dataset
- Autoantibody markers (IAA, IA-2A, and ZntBA each contain ~319,000 unique values across 322,895 rows, this is consistent with randomly generated continuous values rather than real lab measurements, which cluster around assay thresholds

Distributional Separability Analysis

- Compared probability densities of key features across diagnostic classes

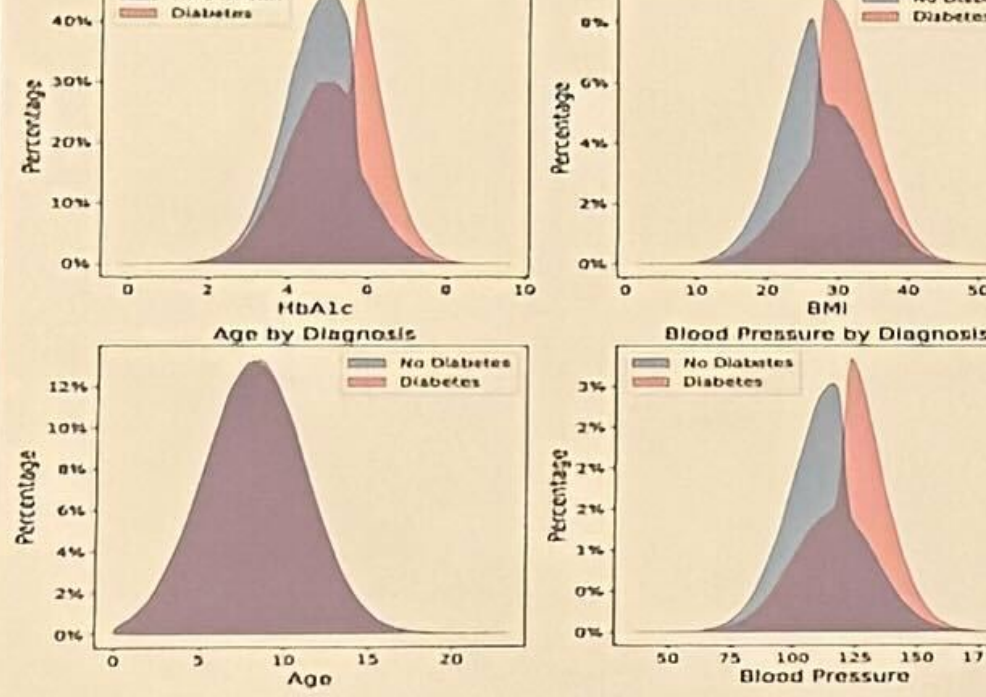


Figure 1. Probability density distributions of HbA1c, BMI, Age, and blood pressure by diagnostic class. Near-perfect overlap across all four features suggests diagnosis labels were generated independently of clinical values.

RESULTS

Variable	Overall	Diabetes (N=153,768)	No Diabetes (N=169,127)	P-value
T1D	28.56 (0.87)	30.24 (0.37)	27.27 (0.30)	< 0.001
HbA1c	4.87 (1.02)	5.26 (1.07)	4.77 (0.93)	< 0.001
Blood pressure	118.41 (16.67)	122.45 (16.21)	112.14 (15.64)	< 0.001
Pancreatic	0.97 (0.73)	0.97 (0.73)	0.97 (0.74)	0.30
Gender	322,895 (100%)	153,768 (47.6%)	169,127 (52.4%)	0.07
Male	101,577 (50.04%)	47,180 (30.73%)	54,380 (40.91%)	
Female	161,318 (49.96%)	46,579 (49.77%)	54,739 (50.09%)	
Age group	322,895 (100%)	153,768 (47.6%)	169,127 (52.4%)	0.43
Infant (0-1)	5,421 (1.68%)	2,233 (1.47%)	3,188 (1.89%)	
Toddler (2-4)	40,173 (12.44%)	16,733 (12.51%)	23,440 (12.39%)	
Child (5-11)	243,488 (76.41%)	100,897 (76.26%)	142,591 (75.90%)	
Adolescent (12-17)	33,822 (10.41%)	14,914 (10.49%)	18,908 (10.37%)	
Young adult (18-20)	189 (0.06%)	89 (0.07%)	100 (0.05%)	
Adult (21+)	9 (0.00%)	2 (0.00%)	3 (0.00%)	

Continuous variables presented as mean (SD) by diagnostic class. Continuous variables: Welch's t-test. Categorical variables: chi-squared test. SD = standard deviation. Red P-values indicate statistical significance ($\alpha = 0.001$); however, in this context significance reflects synthetic data artifacts rather than genuine clinical signal. Identical group proportions across both parent and child rows indicate distributional noise, not clinical differentiation.

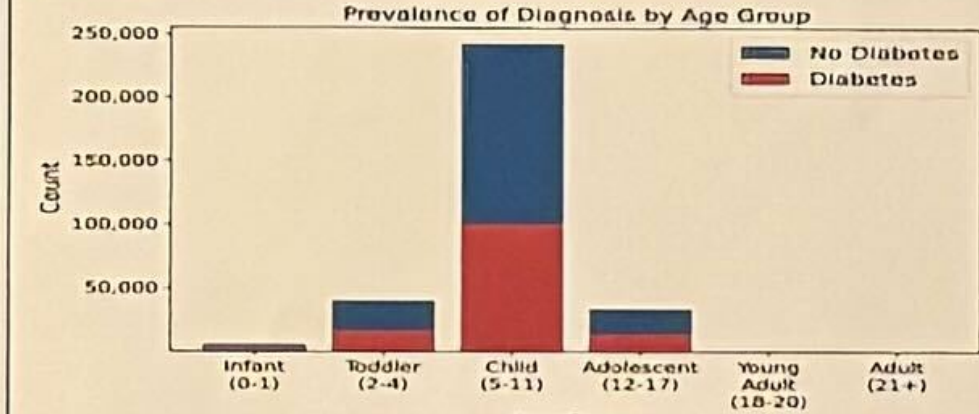


Figure 2. T1D prevalence by age group. Diabetes rate is uniformly ~41% across all age groups.

CONCLUSIONS

Distribution of Pancreatic by Diagnosis. Distribution of BMI by Diagnosis.

Figure 3. Distribution of the paper's top-ranked features (Pancreatic, BMI) by diagnostic class. Near-identical shape and IQR across both groups indicates no class separability.

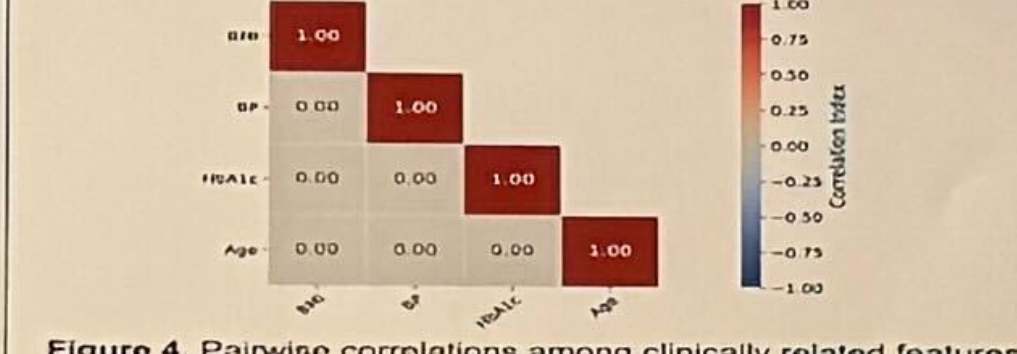


Figure 4. Pairwise correlations among clinically related features. Values approximate zero, indicating features were generated independently with no biological covariance structure.

- HbA1c, the primary T1D biomarker, showed no class separability; this is inconsistent with any genuine clinical dataset
- Pancreatic and BMI, the paper's top-ranked features, show identical distributions across diagnostic classes
- Diagnosis labels appear generated independently of clinical feature values
- Reported 93-94% model accuracies reflect synthetic data artifacts, not learned clinical relationships
- Findings highlight the critical role of EDA in validating dataset integrity prior to model development

REFERENCES

1. Syahidin Y, Maulideli NU, Surendro K. Ensemble feature selection: a hybrid approach for pediatric type 1 diabetes risk factors. In: 2025 International Conference on Smart-Green Technology in Electrical and Information Systems (ICSGTEIS); 2025 Oct 9-10; Bali, Indonesia. IEEE; 2025. p. 30-5. DOI: 10.1109/ICSGTEIS568532.2025.1284421
2. Shpaner L, Gil O. eda_toolkit [software]. Version 0.0.28. 2024. Available from: https://pypi.org/project/eda_toolkit/

Children's Hospital of Orange County, CA
April 29, 2026



Certificate Poster Award

This award is proudly given to:

LEON SHPANER



Terence Sanger, MD, PhD
Chief Scientific Officer
CHOC Research Institute



Louis Ehwerhemuepha, PhD
Chair
CHOC Research Institute