Transparent Machine Learning for Type 1 Diabetes Diagnosis from Gene Expression Data

Rosa Carotenuto[1,†], Viviana Pentangelo[1,*,†], Antonio Della Porta[1,†] and Fabio Palomba[1,†]

[1]Software Engineering (SeSa) Lab, Department of Computer Science, University of Salerno, Salerno, Italy

Abstract

Early diagnosis of Type 1 Diabetes (T1D) is essential for effective intervention but remains challenging with conventional clinical methods. In this study, we investigate the potential of machine learning (ML) models to classify pediatric subjects as T1D or healthy based on microarray gene expression data. We train and evaluate three models—Support Vector Machine, Random Forest, and XGBoost—and assess their predictive performance and interpretability. The best-performing model (SVM) achieves an accuracy of 80.8% and an AUC-ROC of 87.6%. To understand model behavior, we apply SHAP and Anchor explanations, which identify key genes such as *PTPRN2* and *HLA-DQB1* as major contributors to classification outcomes. These results demonstrate the feasibility of combining predictive accuracy with model transparency, laying the groundwork for clinically meaningful and explainable decision support systems for early T1D detection.

Keywords

Type 1 Diabetes, Explainable Machine Learning, Gene Expression Profiling.

1. Introduction

Type 1 Diabetes (T1D) is a chronic autoimmune disease caused by the destruction of insulin-producing β-cells in the pancreas, typically emerging in childhood or adolescence and resulting in lifelong insulin dependence and elevated risk of cardiovascular complications [1, 2]. Early diagnosis is critical to prevent acute complications and preserve residual β-cell function. However, traditional diagnostic approaches often rely on clinical symptoms and glycemic measurements, which appear only after significant immunological damage has occurred [3]. There is thus a pressing need for molecular tools that support earlier and more precise diagnosis.

Recent advances in transcriptomic technologies, such as microarrays and single-cell RNA sequencing, have opened new opportunities to characterize gene expression changes associated with T1D pathogenesis [4, 5, 6]. In particular, gene expression profiling from peripheral blood mononuclear cells (PBMCs) has shown promise as a non-invasive source of disease-related biomarkers. However, the high dimensionality and heterogeneity of such data present challenges for traditional statistical techniques.

Machine learning (ML) methods are well-suited to handle complex, high-dimensional biological datasets and have shown promising results in disease classification [7, 8, 9]. Yet **their integration into clinical settings is still limited, largely due to concerns about model interpretability and trust** [10]. Moreover, most prior work on diabetes prediction has focused on type 2 diabetes or adult populations, with very limited attention to pediatric T1D or to explainability [11, 12, 13].

To address these gaps, this study investigates whether ML models can accurately and transparently classify pediatric subjects as T1D or healthy based on microarray gene expression data. We experiment with three state-of-the-art models, i.e., Support Vector Machine (SVM), Random Forest (RF), and XGBoost, and interpret their predictions using SHAP and Anchor, two complementary explainable AI techniques. Unlike previous work, we focus specifically on pediatric cohorts and integrate biological knowledge in

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

*Corresponding author.

†These authors contributed equally.

r.carotenuto16@studenti.unisa.it (R. Carotenuto); vpentangelo@unisa.it (V. Pentangelo); adellaporta@unisa.it (A. Della Porta); fpalomba@unisa.it (F. Palomba)

0009-0003-1425-9398 (V. Pentangelo); 0000-0003-1860-8404 (A. Della Porta); 0000-0001-9337-5116 (F. Palomba)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2025-12-07

CEUR-WS.org/Vol-4121/Ital-IA_2025_paper_21.pdf

both data preprocessing and post-hoc explanation. Our goal is to contribute a clinically meaningful and interpretable pipeline for early T1D detection, offering new insights at the intersection of computational modeling and biomedical application.

2. Related Work

This study investigates the accuracy of machine learning (ML) models in predicting Type 1 Diabetes (T1D) from genomic data. To contextualize this goal, we briefly outline the genetic underpinnings of T1D and review relevant literature. T1D is an autoimmune disease marked by the immune-mediated destruction of pancreatic β-cells, leading to insulin deficiency [1]. Its onset is driven by both environmental and genetic factors, with strong associations found in the HLA class II region (e.g., HLA-DQA1, HLA-DQB1, and HLA-DRB1) and non-HLA loci such as INS, PTPN22, CTLA4, and IFIH1 [3]. These genes are thought to influence disease progression via immune modulation and antigen presentation [6]. Recent advances in transcriptomic technologies, including microarrays and single-cell RNA sequencing [4], have enabled genome-wide profiling of such genetic signatures [5], making gene expression data a promising target for predictive modeling. ML offers powerful tools to exploit these high-dimensional datasets, but a persistent challenge is the lack of interpretability, which hinders clinical adoption [10]. Despite the importance of transparent decision-making, few studies have addressed explainability in the context of T1D. AlRefaai et al.[11] proposed an ML pipeline for classifying T1D using gene expression data, achieving high accuracy but relying only on feature ranking for interpretability. Patil et al.[13] applied XGBoost to scRNA-seq data from T1D patients and identified key gene signatures, yet their approach lacked explicit explainability techniques such as SHAP or rule-based methods. Broader work in diabetes prediction, mostly targeting type 2 diabetes (T2D), shows strong model performance [7, 8, 9], though these studies typically rely on structured clinical features and do not explore genomic data or interpretability in depth. In contrast, our study provides the first in-depth application of explainable ML techniques to microarray-based gene expression data in pediatric T1D cohorts. By integrating SHAP and Anchor with high-performing classifiers, we aim to make predictive outcomes transparent and biologically meaningful, supporting future use in trustworthy clinical decision support tools.

3. Research Method

The *goal* of this study is to assess whether ML models can accurately and transparently classify individuals with T1D using gene expression data. The *purpose* is twofold: to evaluate the predictive performance of ML models trained on pediatric genomic profiles, and to interpret their decisions using explainable AI (XAI) techniques. The *perspective* is that of researchers investigating how reliably and transparently T1D can be classified through ML. To structure this investigation, we define two research questions (**RQs**). First, we ask whether ML models can effectively distinguish between healthy and T1D subjects based solely on gene expression. This step is preliminary but crucial: if classification performance is poor, explainability would offer little value, as it would be grounded in non-informative input data. Therefore, we first asked:

> **RQ$_1$.** *How accurately can ML models classify subjects as T1D or healthy based on microarray gene expression data?*

After evaluating the predictive capabilities of the models, we turn to the second, core aspect of the study: the interpretability of the learned decision processes. Here, the goal is to assess whether the model predictions can be explained in a meaningful and biologically plausible way, and which genes contribute most significantly to the classification. This led to the second question:

> **RQ$_2$.** *How interpretable are the predictions of these models, and which genes most influence the classification of T1D?*

3.1. Dataset Construction and Preprocessing

To develop and evaluate machine learning models for T1D classification, we integrated two pediatric microarray studies from the Gene Expression Omnibus (GEO) [14]: GSE9006 and GSE43488. Both profile gene expression in peripheral blood mononuclear cells (PBMCs), a non-invasive source of immune-related biomarkers. Pediatric and adolescent subjects are the focus of both datasets, an underrepresented yet clinically important group in T1D research [2]. GSE9006 includes 43 T1D patients (some with follow-ups), 24 healthy controls, and 12 T2D patients. To reduce technical variability, we retained only samples measured with the Affymetrix U133A platform (GPL96). GSE43488 contains 18 prediabetic children and 18 matched controls, profiled using the distinct Affymetrix U219 platform (GPL13667). Because mixing platforms can introduce batch effects, we applied a harmonized preprocessing pipeline—including probe-to-gene mapping, annotation, and batch effect correction—to align both datasets to a shared set of gene-level expression values. This ensured compatibility for downstream ML analysis.

Several preprocessing steps were applied to ensure data quality, comparability, and biological relevance. As part of the harmonization process, we used $\log_2$ transformation and quantile normalization to standardize expression levels, applying the Robust Multi-array Average (RMA) method where appropriate. Probes were mapped to gene symbols using platform-specific Bioconductor annotations (`hgu133a.db`, `hgu219.db`), and batch effect correction was performed to reduce platform-induced variability, with Principal Component Analysis (PCA) confirming improved biological clustering. To reduce noise, genes with low expression ($\log_2 < 5$) were removed. Finally, we selected biologically meaningful features by intersecting genes from the KEGG Pathway (`hsa04940`) and Disease (`H00408`) entries related to T1D. This preprocessing pipeline ensured a technically robust and biologically informed feature set for downstream analysis.

3.2. Selection and Training of Machine Learning Models

To evaluate the classification of T1D based on gene expression data, we selected three machine learning models that are commonly used in biomedical research for their ability to balance accuracy with a degree of interpretability [7]: Support Vector Machine (SVM) with a radial basis function (RBF) kernel, Random Forest (RF), and XGBoost (Extreme Gradient Boosting). These algorithms were chosen for complementary reasons. SVMs are well-suited for small-sample, high-dimensional problems like gene expression analysis due to their margin-based optimization and ability to handle nonlinear patterns. Random Forest and XGBoost, instead, are ensemble tree-based models known for their robustness, strong predictive performance, and built-in mechanisms to estimate feature importance. All models were implemented using the Python libraries `scikit-learn` and `xgboost`. To identify optimal hyperparameters, we conducted a grid search with stratified 10-fold cross-validation on 80% of the available data. This ensured a fair balance of class representation in each fold and helped avoid overfitting. Once tuned, the final models were trained on the full training set and evaluated on a separate 20% hold-out test set to estimate their generalization performance.

3.3. Data Collection and Analysis

To address $\mathbf{RQ}_1$, which investigates whether machine learning models can accurately classify subjects as T1D or healthy based on gene expression data, we trained each model using an 80%-20% split between training and test sets. Model evaluation was performed using a variety of established performance metrics. Specifically, we computed accuracy to assess the overall correctness of the model, while precision and recall were used to capture the model's ability to avoid false positives and false negatives, respectively. To complement these metrics, we also computed the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which reflects the model's capacity to discriminate between T1D and healthy individuals across different thresholds. These metrics collectively provided a comprehensive assessment of each model's predictive ability, thereby directly addressing $\mathbf{RQ}_1$.

To address $\mathbf{RQ}_2$, which concerns the interpretability of the model predictions and the identification of influential genes, we applied two state-of-the-art explainable AI (XAI) techniques [10]: SHAP (SHapley

Additive exPlanations) and Anchor. SHAP is a unified framework based on cooperative game theory that assigns each feature a contribution score for individual predictions. In our context, this allowed us to estimate the importance of each gene both globally (across all samples) and locally (for individual predictions), thereby identifying which genes consistently influenced the classification outcome. SHAP explanations were particularly useful for understanding how different features interacted and contributed to the decision boundary learned by the model. Complementing this, we used Anchor explanations to derive local, rule-based interpretations in the form of "if-then" conditions that anchor a prediction with high precision. These explanations highlight minimal subsets of features whose presence suffices to support a model's decision, providing interpretable patterns that are easily understood by clinicians. Unlike SHAP, which offers continuous-valued importance scores, Anchor provides discrete and actionable rules that help verify whether a model's reasoning aligns with domain knowledge.

Together, the combination of quantitative performance evaluation and qualitative explanation techniques enabled us to thoroughly address both $\mathbf{RQ}_1$ and $\mathbf{RQ}_2$. This dual focus not only confirmed that the models could learn meaningful classification boundaries from genomic data but also ensured that their predictions could be interpreted and potentially trusted in clinical contexts.

4. Analysis of the Results

Table 1
Classification performance of ML models on the hold-out test set.

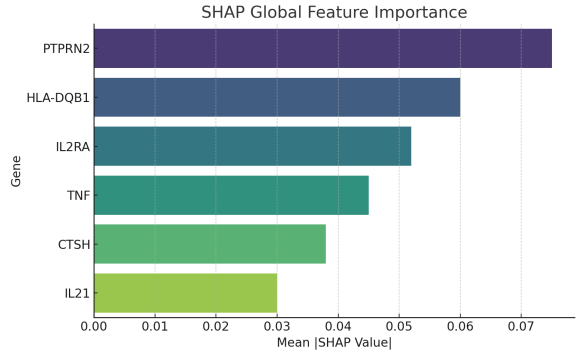
Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
SVM	0.808	0.813	0.808	0.807	0.876
Random Forest	0.795	0.803	0.795	0.794	0.874
XGBoost	0.795	0.799	0.795	0.793	0.872

4.1. $\mathbf{RQ}_1$: Predictive Performance

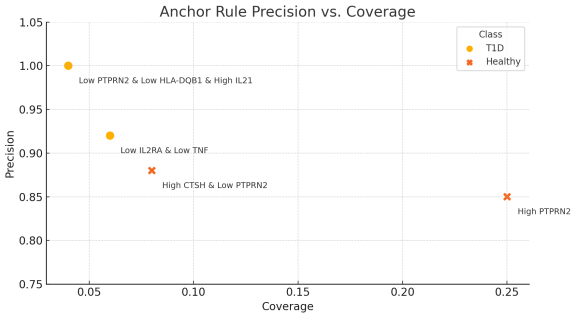
Table 1 shows the results achieved when assessing the predictive capabilities of the considered models. As shown, SVM achieved the best overall performance, with an accuracy of 80.8%, precision of 81.3%, recall of 80.8%, F1-score of 80.7%, and an AUC-ROC of 87.6%. These metrics indicate a strong balance between sensitivity and specificity, which is critical in clinical settings where both false positives and false negatives can have serious consequences. The other models, i.e., Random Forest and XGBoost, performed slightly below SVM but still demonstrated high and comparable accuracy (79.5%) and AUC-ROC values (87.4% and 87.2%, respectively). This confirms that all models were able to extract meaningful patterns from the microarray data. The strong predictive results justified moving forward with the explainability analysis, as the foundation of a reliable model is a prerequisite for extracting trustworthy explanations. The performance also demonstrates that pediatric gene expression data, when properly curated and preprocessed, can effectively support ML-based T1D prediction, reinforcing the potential of genomic diagnostics in early-stage autoimmune disorders. This observation will be the basis for our future investigations into the matter.

4.2. $\mathbf{RQ}_2$: Explainability of Predictions

SHAP summary plots (Figure 1(a)) revealed consistent patterns of gene importance across all three models, with PTPRN2 standing out as the most influential gene. The plot shows that lower expression levels of PTPRN2 (depicted in blue) are strongly associated with predictions of T1D, while higher expression levels (in red) contribute to healthy classifications. This clear separation reinforces its discriminative power and its role as a key biomarker in the dataset. Other genes, such as IL2RA, TNF, and CTSH, also contributed significantly to the model's predictions, each displaying a distinct pattern



(a) SHAP global feature importance.



(b) Anchor rule precision vs. coverage.

Figure 1: Explainability results addressing RQ2. SHAP reveals globally important genes, while Anchor provides interpretable rules with quantified precision and coverage.

of influence on classification outcomes. Notably, HLA-DQB1, a gene with a well-established link to T1D susceptibility, emerged as the top global predictor in the SVM model's SHAP output, although it showed less prominence in other models or methods.

Complementing these insights, Anchor explanations (Figure 1(b)) offered localized, rule-based interpretations that align with clinical reasoning. The plot summarizes a set of interpretable decision rules and compares their precision and coverage. One high-confidence rule for predicting T1D combines low expression of PTPRN2 and HLA-DQB1 with high expression of IL21, achieving perfect precision (100%) but with limited coverage (4%). This suggests the rule is highly reliable for a small subset of clear-cut cases. Conversely, a more general rule for healthy classification based solely on high PTPRN2 expression achieves 85% precision and applies to 25% of the dataset, offering broader coverage with slightly reduced certainty.

Taken together, Figures 1(a) and 1(b) show that both SHAP and Anchor consistently highlight a core subset of genes—including PTPRN2, HLA-DQB1, and IL2RA—as critical to the model's decision-making process. The convergence of these methods adds credibility to the interpretability findings, and the alignment with known T1D-related genes supports the biological plausibility of the results. Importantly, the rules derived by Anchor offer an interpretable format for explaining individual predictions, which could aid clinical understanding and decision support. The integration of global (SHAP) and local (Anchor) perspectives thus reinforces the potential of explainable machine learning for translational use in genomic diagnostics.

4.3. Joint Interpretation and Implications

When analyzed together, the results from $\mathbf{RQ}_1$ and $\mathbf{RQ}_2$ offer some valuable insights into the prediction and explainability of T1D. The models, particularly SVM, demonstrated robust predictive ability, indicating that transcriptomic data from pediatric cohorts contain sufficient discriminatory information to support T1D diagnosis. At the same time, the application of SHAP and Anchor enabled transparent interpretation of model behavior, facilitating the identification of key gene markers that align with existing biomedical literature. The mutual reinforcement of global (SHAP) and local (Anchor) explanations enhances trust in the model outputs and suggests that explainable machine learning is not only feasible but also valuable in high-stakes medical contexts. Moreover, our approach highlights specific gene combinations that could serve as diagnostic indicators or therapeutic targets in the future. In summary, **this study validates the use of explainable ML models for early T1D detection based on genomic features and represents one of the first efforts to combine pediatric gene expression data with both performance- and explanation-focused evaluation**. The findings contribute to ongoing efforts to build interpretable, data-driven tools that can be responsibly deployed in clinical research and practice.

5. Conclusion

This study showed that machine learning models can effectively classify pediatric subjects as T1D or healthy using microarray gene expression data, with SVM achieving the best performance. By applying SHAP and Anchor, we obtained interpretable predictions, demonstrating the potential of combining accuracy and transparency in genomic ML. As future work, we plan to refine our pipeline by incorporating longitudinal data, and expanding to other omics modalities.

Acknowledgments

This work has been partially supported by the European Union through the Italian Ministry of University and Research, Project PNRR "D3-4Health: Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care". PNC 0000001. CUP B53C22006090001.

Declaration on Generative AI

The authors used GPT-4 for grammar and spelling check and text improvement. After using it, the authors reviewed and edited the content as needed and take full responsibility for the paper's content.

References

[1] D. Glovaci, W. Fan, N. D. Wong, Epidemiology of diabetes mellitus and cardiovascular disease, Current cardiology reports 21 (2019) 1–8.
[2] M. A. Atkinson, G. S. Eisenbarth, A. W. Michels, Type 1 diabetes, The lancet 383 (2014) 69–82.
[3] E. Kawasaki, Anti-islet autoantibodies in type 1 diabetes, International Journal of Molecular Sciences 24 (2023) 10012.
[4] R. Moncada, D. Barkley, F. Wagner, M. Chiodin, J. C. Devlin, M. Baron, C. H. Hajdu, D. M. Simeone, I. Yanai, Integrating microarray-based spatial transcriptomics and single-cell rna-seq reveals tissue architecture in pancreatic ductal adenocarcinomas, Nature biotechnology 38 (2020) 333–342.
[5] F. Pociot, Type 1 diabetes genome-wide association studies: not to be lost in translation, Clinical & translational immunology 6 (2017) e162.
[6] A. Zajec, K. Trebušak Podkrajšek, T. Tesovnik, R. Šket, B. Čugalj Kern, B. Jenko Bizjan, D. Šmigoc Schweiger, T. Battelino, J. Kovač, Pathogenesis of type 1 diabetes: established facts and new insights, Genes 13 (2022) 706.
[7] P. B. Khokhar, C. Gravino, F. Palomba, Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review, Artificial Intelligence in Medicine (2025) 103132.
[8] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, Predicting diabetes mellitus with machine learning techniques, Frontiers in genetics 9 (2018) 515.
[9] L. Muhammad, E. A. Algehyne, S. S. Usman, Predictive supervised machine learning models for diabetes mellitus, SN Computer Science 1 (2020) 240.
[10] N. Burkart, M. F. Huber, A survey on the explainability of supervised machine learning, Journal of Artificial Intelligence Research 70 (2021) 245–317.
[11] N. AlRefaai, S. Z. AlRashid, Classification of gene expression dataset for type 1 diabetes using machine learning methods, Bulletin of Electrical Engineering and Informatics 12 (2023) 2986–2992.
[12] V. Chang, J. Bailey, Q. A. Xu, Z. Sun, Pima indians diabetes mellitus classification based on machine learning (ml) algorithms, Neural Computing and Applications 35 (2023) 16157–16173.
[13] A. R. Patil, J. Schug, C. Liu, D. Lahori, H. C. Descamps, A. Naji, K. H. Kaestner, R. B. Faryabi, G. Vahedi, Modeling type 1 diabetes progression using machine learning and single-cell transcriptomic measurements in human islets, Cell Reports Medicine 5 (2024).
[14] T. Barrett, R. Edgar, [19] gene expression omnibus: microarray data storage, submission, retrieval, and analysis, Methods in enzymology 411 (2006) 352–369.