From Rules to Reports: Enhancing Diabetes Prediction Interpretability with Anchor and LLMs

Francesco Festa[1,†], Antonio Della Porta[1,†], Viviana Pentangelo[1,*,†] and Fabio Palomba[1,†]

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

Abstract

Diabetes is a widespread chronic disease that requires timely and accurate diagnosis to prevent severe complications. Machine Learning (ML) models have shown great potential in predicting diabetes risk, but their lack of interpretability remains a major barrier to clinical adoption. Explainable AI (XAI) techniques, such as Anchor, offer rule-based insights that are closer to natural language but still fall short of full transparency. With the emergence of Large Language Models (LLMs), it is now possible to enhance these explanations and make them more accessible. In this study, we present a pipeline that combines ML prediction, Anchor-based explanation, and LLM-augmented natural language reporting. In this study, we trained four ML models and selected the best-performing one—SVM with an accuracy of 82%—which we then paired with Anchor to generate rule-based explanations. These were refined through five iterations of prompt tuning with ChatGPT 3.5, evaluated qualitatively for clarity and precision. The resulting natural-language reports were integrated into DIA, a web-based tool designed to deliver interpretable, human-centered diabetes predictions.

Keywords

Diabetes Prediction, Explainable Machine Learning, Prompt Refinement.

1. Introduction

Diabetes is a chronic condition that, according to the World Health Organization, ranked among the top ten causes of death worldwide in 2020 [1]. It is characterized by elevated blood glucose levels resulting from impaired insulin production or action [2]. If not diagnosed and managed promptly, diabetes can lead to severe complications such as cardiovascular disease, kidney failure, vision loss, and neuropathy [3, 4, 5]. Recently, Artificial Intelligence (AI) has been even more integrated in the medical context [6, 7]. Its rapidly evolving capability to analyze patient data and assist clinical decision-making is being exploited to support clinicians and patients in early diagnosis and continuous monitoring.

A key factor to consider when adopting AI solutions is that most machine learning (ML) models employed for this purpose function as *black boxes*, offering accurate predictions without providing insights into the underlying reasoning. In healthcare, a domain inherently sensitive and high-stakes, this lack of transparency limits the adoption of AI-based systems, as practitioners are reluctant to trust outputs they cannot interpret [8]. To address this challenge, the field of *eXplainable Artificial Intelligence* (XAI) has emerged, aiming to make model predictions more transparent and, consequently, interpretable by human users [9]. Simultaneously, the rapid spread of *Large Language Models* (LLMs) can further support the translation of complex model outputs into natural language explanations. Combining XAI techniques with the linguistic power of an LLM can considerably enhance accessibility for non-expert users and their trust in the results of ML predictions [8, 10].

Despite the promising potential of AI for diabetes prediction, real-world adoption remains limited, not only due to the black-box nature of predictive models but also due to the lack of empirical studies on how to effectively integrate explainability into this domain. Existing research has primarily focused on identifying accurate classifiers [11, 12] or on evaluating XAI methods in isolation [13, 14], without fully exploring how their outputs can be made usable and accessible to non-technical stakeholders. 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.

f.festa19@studenti.unisa.it (F. Festa); adellaporta@unisa.it (A. Della Porta); vpentangelo@unisa.it (V. Pentangelo); fpalomba@unisa.it (F. Palomba)

particular, little empirical knowledge exists on how Large Language Models (LLMs) can enhance the interpretability of XAI techniques in the context of clinical prediction tasks.

This work aims to contribute to the field by empirically evaluating a pipeline that integrates machine learning models, rule-based explanations via Anchor [15], and natural language generation through an LLM. In particular, we put the focus of our study on the iterative refinement of the LLM prompt, designed to transform technical rule-based outputs into human-readable explanations. Through five rounds of prompt tuning and qualitative evaluation, we explored how language-driven interventions can enhance the clarity and usefulness of explanations for non-technical users. The resulting pipeline is implemented in DIA, a web-based tool that enables users to input clinical data and receive both a model prediction and a natural-language explanation. Our findings provide practical insights into how LLM-augmented XAI, when guided by careful prompt engineering, can bridge the interpretability gap and support more transparent, trustworthy decision-making in healthcare.

2. Background & Related Work

Artificial Intelligence (AI) is increasingly adopted in healthcare to support disease diagnosis and management, with diabetes being one of its most widely studied applications [16]. Machine Learning (ML) models have shown strong predictive capabilities using structured clinical and biometric data, but their "black-box" nature limits adoption in clinical practice [9]. To address this, Explainable AI (XAI) techniques have been proposed to make model behavior more transparent and interpretable to both clinicians and patients [8]. Commonly used XAI methods include SHAP [17], which assigns feature importance scores using game theory; LIME [18], which creates local surrogate models to highlight influential features; and Anchor [15], which extracts if–then rules that "anchor" a prediction and are easier for humans to understand. However, these approaches often still require technical interpretation. Recently, Large Language Models (LLMs) have been explored to enhance XAI by translating such outputs into natural language, making predictions more accessible and meaningful to end users [10].

Following the growing trend of applying XAI to diabetes prediction, several empirical studies have begun to investigate the performance of various machine learning models and explanation techniques on patient data. A common approach is to apply model-agnostic explanation methods, such as SHAP and LIME, to classical ML models. Lee et al.[19] used SHAP with gradient boosting to identify key features such as glucose and BMI, enhancing alignment with clinical expectations. Hasan et al.[20] integrated AutoML and multiple XAI methods—including SHAP, LIME, and Counterfactual Analysis—into an interactive tool for clinicians. Curia [21] applied LIME to interpret various ML classifiers for Type 1 diabetes prediction, emphasizing transparency in clinical decision support. Annuzzi et al.[22] focused on predicting postprandial glucose levels in Type 1 patients, using SHAP to explain neural network predictions. Ahmed et al.[14] compared SHAP and LIME explanations for logistic regression and random forest models, analyzing their consistency and interpretability. Hossain et al.[23] proposed an ensemble approach for diabetes management and visualized feature importance across models. Vivek Khanna et al.[13] predicted gestational diabetes using multiple XAI techniques—including Anchor and "Explain Like I'm 5"—to highlight influential clinical markers. Finally, Tasin et al. [24] developed a real-time prediction system using LIME and SHAP for web and mobile deployment in low-resource settings.

3. Experiment Setup

The goal of our experiment was to evaluate different combinations of ML models and associated explanations, produced via XAI techniques and augmented with a LLM, to assess whether non-technical users can meaningfully understand diabetes risk predictions. To this end, we experimented through a multi-phase pipeline involving model training, explanation generation, LLM-based natural language summarization, and interactive delivery via a web application.

▸ **Data Preprocessing.** We used the Pima Indians Diabetes Dataset (PIDD), consisting of 768 samples

with eight clinical features—e.g., glucose concentration, blood pressure, BMI—and a binary outcome [25] to train our models. Missing values were handled based on the distribution of each variable, applying the mean for normally distributed features and the median otherwise. To address class imbalance, we applied SMOTE, generating a balanced dataset of 1000 instances (500 per class). Feature scaling was performed using standard normalization, which significantly improved performance after initial hyperparameter tuning proved ineffective.

▸ **Models' Training.** After preprocessing the data, we selected and trained four widely used classifiers in the relaed literature [13, 14]—Random Forest, Support Vector Machine, XGBoost and a feed-forward Neural Network. Such models were trained and subsequently evaluated using stratified 80/20 train-test splits and standard metrics including accuracy, precision, recall, and F1-score.

▸ **Anchor-Based Model Explanation** We explained the model's predictions using Anchor [26], a model-agnostic technique that explains the behavior of complex models with high-precision rules called anchors, representing local, sufficient conditions for predictions. The anchors rules are simple *if–then* rules–ideal for an LLM to translate directly into clear, complete explanations. We choose not to use other more commonly used XAI techniques like SHAP and LIME, since they produce numeric feature scores to explain the outcome of a model, and are hard to process for LLMs. To select the most broadly applicable anchors—and by extension the best model for our LLM pipeline—we measured the **coverage of anchors**, the fraction of perturbed samples around each instance for which an anchor's condition holds. A high coverage score shows that a rule applies to many similar cases, not just a single outlier, balancing against precision to ensure explanations are both accurate and general. By choosing anchors with strong coverage, we guaranteed that our plain-language summaries reflect real, repeatable model behavior rather than one-off quirks.

▸ **LLM-Driven Explanation Generation** In the final stage, an LLM translates each anchor rule into a clear, plain-language summary. We began by fixing a random seed to select a representative sample of data points; for each, the pipeline generated a prediction and its corresponding Anchor rule, then the LLM produced an explanation. The authors independently reviewed these natural-language summaries and wrote feedback on the overall clarity of the answer. Based on that feedback, we iteratively refined the LLM prompt—adjusting instructions, examples and formatting—to enhance the quality and consistency of the outputs. The stopping condition of the process was that all the authors agreed that the result was satisfactory under two main metrics: clarity and usefulness. All the code used to perform the experiment and all the prompts used throughout the experiment are available in the Appendix A.

4. Analysis of the Results

Preliminarily to our work, we performed a data-cleaning phase to improve the quality of the data since there were some missing values in the dataset. The dataset consists of 768 observations and nine features: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and Outcome. An initial data audit uncovered implausible zero values in five predictors—Glucose (5 cases), BloodPressure (35), SkinThickness (227), Insulin (374), and BMI (11)—which were treated as missing. To guide imputation, feature distributions were examined via histograms: Glucose and BloodPressure showed approximately Gaussian behavior and were imputed using the mean, while the right-skewed distributions of SkinThickness, Insulin, and BMI warranted median imputation. Post-imputation histograms confirmed that the underlying distributions remained consistent. Subsequently, class imbalance (65.1 % non-diabetic vs. 34.9 % diabetic) was addressed using SMOTE to synthesize minority-class examples, yielding a perfectly balanced set of 1000 instances (500 per class).

4.1. Training Results

We trained our black-box classifiers—Random Forest, Support Vector Machine (SVM), XGBoost, and a feed-forward neural network with four dense layers—on the fully preprocessed feature set. Initial

assessment using accuracy, precision, recall, and F_1-score yielded performances clustered around 80% accuracy, with the neural network outperforming marginally. A comprehensive grid search over hyperparameters failed to produce substantive gains; consequently, all features were standardized, and the models were retrained. The final metrics of the models are described in Table 1.

Table 1
Performance comparison of classification models

Model	Accuracy	Precision	Recall	F1 Score	Anchors Coverage
Random Forest	0.8250	0.7982	0.8700	0.8325	0.068
SVM	0.8200	0.7909	0.8700	0.8286	0.106
XGBoost	0.8100	0.7981	0.8300	0.8137	0.058
Neural Network	0.8350	0.8131	0.8700	0.8406	0.097

4.2. Explanations Augmentation through an LLM

For the core part of our study, we selected OpenAI's ChatGPT 3.5-Turbo as the LLM responsible for transforming rule-based explanations into natural-language summaries. This choice was driven by a trade-off between performance and cost-efficiency, as ChatGPT 3.5-Turbo offers fast response times and stable outputs at a significantly lower cost than more advanced models such as GPT-4. To evaluate whether the LLM's linguistic capabilities could improve the clarity and usability of the explanations, we conducted an iterative refinement process on the prompt. Using a representative Anchor rule (Glucose > 147.034 AND BMI > 32.772) applied to a true-positive prediction, we tested five successive versions of the prompt. Each iteration was qualitatively assessed by the authors to ensure that the generated explanations were both faithful to the original rule and accessible to non-technical users.

The process took five iterations until all the authors were satisfied with the results. Throughout the iterations, we observed several recurring issues. Initially, the LLM struggled to correctly interpret and convey the meaning of the anchor's coverage, often misrepresenting it with probabilistic or statistical expressions, or directly quoting technical terms such as "coverage" and "confidence" that were not meaningful to lay users. In other cases, the model tended to restate the rule literally rather than abstracting it into a user-friendly form, which limited the interpretive value of the explanation. By refining the prompt wording and progressively constraining undesired behaviors—e.g., instructing the LLM to avoid numbers, rules, or terms like "model"—we were able to produce a final prompt that reliably generated empathetic, informative, and linguistically clear explanations. These outputs avoided technical jargon, generalized the condition in relatable terms—e.g., "high blood sugar and weight"—and emphasized the rationale behind the model's prediction.

This iterative process underscores the critical role of prompt engineering in leveraging LLMs for explainability tasks, and supports the hypothesis that **LLMs can effectively bridge the interpretability gap** by transforming technical model outputs into clear, meaningful narratives—particularly in the context of diabetes prediction.

4.3. The DIA Web App

To demonstrate the practical applicability of our approach, we developed DIA–Diabetes Diagnoser, a web-based tool that allows users to interact with the predictive model and receive LLM-enhanced explanations in real time. The application was implemented using the Streamlit framework and provides a user-friendly interface, and it is entirely available in the repository linked in Appendix A.

Upon accessing the interface, users are presented with a series of numerical input boxes corresponding to clinical features from the diabetes dataset. Each input is accompanied by a tooltip that provides a short description of its medical meaning, ensuring clarity for non-expert users. After submitting the data, the system returns the model's binary prediction (diabetic or not) along with a natural language explanation generated via ChatGPT 3.5 based on the corresponding Anchor rule. The explanation

is formatted into readable paragraphs and numbered bullet points for each condition. The web app also includes a reliability score derived from the anchor's coverage, helping users identify potentially unreliable predictions. Notably, the system highlights cases where the explanation becomes overly specific or convoluted—often a signal of low model confidence. These features support both usability and transparency, making DIA suitable for both patients and healthcare professionals.

5. Conclusions

In this work, we combine Anchor-extracted "if–then" rules from an SVM model that predicts diabetes presence with an LLM to generate clear, on-demand explanations of individual predictions without manual effort. To make results actionable, we created a web-app called DIA that turns complex decision logic into concise, user-friendly narratives that improve transparency and trust.

Future work will focus on formal user studies to assess trust and comprehension in real-world settings, refining prompt strategies for clearer narratives, and extending the framework to other high-stakes classification tasks.

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

During the preparation of this work, the authors used GPT-o4-mini-high for grammar and spelling check and text improvement. After using this service, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] World Health Organization, The top 10 causes of death of 2020, ???? URL: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

[2] P. Z. Zimmet, D. J. Magliano, W. H. Herman, J. E. Shaw, Diabetes: a 21st century challenge, The lancet Diabetes & endocrinology 2 (2014) 56–64.

[3] A. Khan, I. Petropoulos, G. Ponirakis, R. Malik, Visual complications in diabetes mellitus: beyond retinopathy, Diabetic medicine 34 (2017) 478–484.

[4] D. Glovaci, W. Fan, N. D. Wong, Epidemiology of diabetes mellitus and cardiovascular disease, Current cardiology reports 21 (2019) 1–8.

[5] L. Mayeda, R. Katz, I. Ahmad, N. Bansal, Z. Batacchi, I. B. Hirsch, N. Robinson, D. L. Trence, L. Zelnick, I. H. De Boer, Glucose time in range and peripheral neuropathy in type 2 diabetes mellitus and chronic kidney disease, BMJ Open Diabetes Research and Care 8 (2020) e000991.

[6] P.-r. Liu, L. Lu, J.-y. Zhang, T.-t. Huo, S.-x. Liu, Z.-w. Ye, Application of artificial intelligence in medicine: an overview, Current Medical Science 41 (2021) 1105–1115.

[7] Z. Guan, H. Li, R. Liu, C. Cai, Y. Liu, J. Li, X. Wang, S. Huang, L. Wu, D. Liu, et al., Artificial intelligence in diabetes management: advancements, opportunities, and challenges, Cell Reports Medicine 4 (2023).

[8] R. Rosenbacke, Å. Melhus, M. McKee, D. Stuckler, How explainable artificial intelligence can increase or decrease clinicians' trust in ai applications in health care: Systematic review, JMIR AI 3 (2024) e53207.

[9] S. Reddy, Explainability and artificial intelligence in medicine, The Lancet Digital Health 4 (2022) e214–e215.
[10] T. Mirzaei, L. Amini, P. Esmaeilzadeh, Clinician voices on ethics of llm integration in healthcare: A thematic analysis of ethical concerns and implications, BMC Medical Informatics and Decision Making 24 (2024) 250.
[11] N. Poria, A. Jaiswal, Empirical analysis of diabetes prediction using machine learning techniques, in: Applications of Artificial Intelligence and Machine Learning: Select Proceedings of ICAAAIML 2021, Springer, 2022, pp. 391–401.
[12] U. E. Laila, K. Mahboob, A. W. Khan, F. Khan, W. Taekeun, An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study, Sensors 22 (2022) 5247.
[13] V. Vivek Khanna, K. Chadaga, N. Sampathila, S. Prabhu, R. Chadaga P, D. Bhat, S. KS, Explainable artificial intelligence-driven gestational diabetes mellitus prediction using clinical and laboratory markers, Cogent Engineering 11 (2024) 2330266.
[14] S. Ahmed, M. S. Kaiser, M. S. Hossain, K. Andersson, A comparative analysis of lime and shap interpreters with explainable ml-based diabetes predictions, IEEE Access (2024).
[15] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.
[16] 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.
[17] S. Lundberg, A unified approach to interpreting model predictions, arXiv preprint arXiv:1705.07874 (2017).
[18] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[19] S.-Y. Lee, W. C.-C. Chu, Y.-H. Tseng, Y.-G. Zhang, H.-L. Tsai, Explainable ai applied in healthcare: A case study of diabetes prediction, in: 2024 IEEE 24th International Conference on Software Quality, Reliability, and Security Companion (QRS-C), IEEE, 2024, pp. 336–344.
[20] R. Hasan, V. Dattana, S. Mahmood, S. Hussain, Towards transparent diabetes prediction: Combining automl and explainable ai for improved clinical insights, Information 16 (2024) 7.
[21] F. Curia, Explainable and transparency machine learning approach to predict diabetes develop, Health and Technology 13 (2023) 769–780.
[22] G. Annuzzi, P. Arpaia, L. Bozzetto, S. Criscuolo, S. Giugliano, M. Pesola, Assessing the features on blood glucose level prediction in type 1 diabetes patients through explainable artificial intelligence, in: 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), IEEE, 2023, pp. 278–283.
[23] R. Ganguly, D. Singh, Explainable artificial intelligence (xai) for the prediction of diabetes management: An ensemble approach, International Journal of Advanced Computer Science and Applications 14 (2023).
[24] I. Tasin, T. U. Nabil, S. Islam, R. Khan, Diabetes prediction using machine learning and explainable ai techniques, Healthcare technology letters 10 (2023) 1–10.
[25] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, R. S. Johannes, Using the adap learning algorithm to forecast the onset of diabetes mellitus, in: Proceedings of the annual symposium on computer application in medical care, 1988, p. 261.
[26] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

A. Online Resources

- GitHub repository containing the DIA web app.