Hybrid AI for Clinical Prediction: An Explainable and Robust Framework for Diabetes Classification

Nikhil Tamvada*, Varun Akella†, Pragathi Dharmale‡
*Amador Valley High School, †____ High School, ‡ Aspiring Scholars Directed Research Program
*nikhil.tamvada@students.asdrp.org, †varun.akella@students.asdrp.org, ‡pragathi.dharmale@asdrp.org

***Abstract*—The rising global prevalence of diabetes necessitates accurate and early prediction models. Traditional machine learning (ML) approaches often struggle with interpretability and leveraging rich, unstructured clinical insights. This paper introduces a novel hybrid AI framework that combines a traditional ML model (Random Forest) with Large Language Model (LLM)-derived features for enhanced diabetes prediction. We leverage Gemini to generate high-level, explainable features (e.g., sentiment scores, keyword flags, and semantic embeddings) from patient narratives, which are then integrated into a meta-learner. Our framework is evaluated on a combined dataset of PIMA Indians Diabetes and CDC BRFSS2015 data, demonstrating superior performance, achieving a recall of 0.715 ± 0.017 and an accuracy of 0.81 ± 0.012 on unseen test data. Beyond performance, the hybrid model significantly improves explainability through SHAP (SHapley Additive exPlanations) values, revealing key feature contributions from both traditional and LLM-derived sources. We also conduct a comprehensive fairness analysis, highlighting the model's equitable performance across demographic groups. This research provides a robust, explainable, and fair AI solution for clinical decision support, addressing critical needs for transparent and responsible AI deployment in healthcare.**

***Index Terms*—Hybrid AI, Large Language Models, Explainable AI, SHAP, Diabetes Prediction, Clinical Decision Support, Fairness, Machine Learning.**

I. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels, leading to severe long-term complications such as cardiovascular disease, kidney failure, nerve damage, and blindness [1]. Early and accurate diagnosis is crucial for effective management, preventing complications, and improving patient outcomes. Traditional machine learning (ML) models have shown promise in diabetes prediction by analyzing structured electronic health record (EHR) data. However, these models often lack transparency, making it difficult for clinicians to understand the rationale behind predictions, which is a significant barrier to their adoption in clinical practice. Furthermore, they typically do not effectively utilize the rich, unstructured information often present in clinical notes or patient interviews. The advent of Large Language Models (LLMs) has revolutionized natural language processing, demonstrating remarkable capabilities in understanding, generating, and reasoning with human language [2]. Recent studies have explored LLMs for medical applications, including clinical note summarization, question answering, and even diagnostic assistance [3], [4]. While powerful, directly deploying LLMs for high-stakes clinical predictions raises concerns about hallucination, bias, and the difficulty of auditing their "black-box" decision-making processes. This paper proposes a novel hybrid AI framework that synergistically combines the strengths of traditional ML models with the advanced natural language understanding capabilities of LLMs, specifically Gemini. Our approach addresses the limitations of standalone models by:

1) Leveraging LLMs to extract high-level, semantically rich, and explainable features from patient information, including sentiment, key symptom flags, and abstract embeddings.
2) Integrating these LLM-derived features with traditional structured clinical data to train a robust meta-learner.
3) Enhancing model transparency and interpretability through SHAP (SHapley Additive exPlanations) analysis, providing insights into the contributions of both traditional and LLM-derived features.
4) Conducting a rigorous fairness analysis to ensure equitable performance across different demographic groups, a critical aspect for responsible AI in healthcare.

The remainder of this paper is organized as follows: Section II discusses related work. Section III details our methodology, including data preparation, traditional ML modeling, LLM-based feature engineering, and the hybrid architecture. Section IV presents the experimental results, including performance evaluation, explainability analysis, and fairness assessment. Finally, Section V concludes the paper and outlines future directions.

II. Related Work

The application of machine learning in diabetes prediction has a long history, with various algorithms such as Logistic Regression, Support Vector Machines (SVMs), Decision Trees, Random Forests, and Gradient Boosting Machines being widely employed [8]. These models primarily rely on structured numerical and categorical data, including demographic information (age, sex), anthropometric measurements (BMI), and laboratory results (glucose, insulin). While effective, their black-box nature often hinders clinical adoption, as medical professionals require transparent and justifiable predictions. Explainable AI (XAI) techniques have emerged to address this transparency gap. Post-hoc explanation methods like SHAP [6] and LIME [7] are increasingly used to interpret complex ML models by quantifying the contribution of each feature to a prediction. Our work extends this by applying SHAP to a hybrid model, demonstrating how LLM-derived features

contribute to the overall prediction. The integration of LLMs into healthcare is a rapidly evolving field. LLMs have demonstrated proficiency in tasks such as clinical text summarization, information extraction from electronic health records, and even generating differential diagnoses [2], [4]. However, directly using LLMs for high-stakes prediction tasks is fraught with challenges, including potential for hallucination, lack of control over the reasoning process, and significant computational costs. Our framework mitigates these risks by using the LLM as a feature generator rather than a direct predictor, allowing for a more controlled and auditable integration into a traditional ML pipeline. This "LLM-as-a-feature" paradigm is a growing area of research, particularly in domains where rich textual context can augment structured data [9]. Our work contributes to this area by providing a concrete application in clinical prediction with a focus on explainability and fairness.

III. Methodology

Our hybrid AI framework for diabetes classification comprises several key stages: data collection and preprocessing, traditional ML model development, LLM-based feature engineering, and the fusion of these components into a meta-learner architecture.

A. Data Collection and Preprocessing

We utilized a combined dataset for our study to ensure robustness and generalizability across different patient populations. The datasets include:

1) **PIMA Indians Diabetes Database:** This dataset, available from the UCI Machine Learning Repository, contains 768 patient records with 8 numerical features (e.g., glucose, BMI, age) and a binary outcome (diabetes or no diabetes). All patients are females of Pima Indian heritage.
2) **CDC Behavioral Risk Factor Surveillance System (BRFSS) 2015 Diabetes Health Indicators:** This large dataset (approximately 250,000 records) from the Centers for Disease Control and Prevention includes a wide range of health-related indicators, demographics, and lifestyle factors. We extracted relevant features and the binary diabetes outcome.

The combined dataset was preprocessed to handle missing values and inconsistencies. Specifically, zero values in certain physiological measurements (e.g., Blood Pressure, BMI, Glucose, Insulin, Skin Thickness) were treated as missing and imputed using the median of their respective columns, a common practice for these datasets [10]. Categorical features were one-hot encoded where appropriate. The combined dataset was then split into training and testing sets with an 80/20 ratio, ensuring stratified sampling to maintain the proportion of diabetes cases in both sets.

B. Traditional ML Model

For the traditional ML component, we selected a Random Forest Classifier due to its strong performance, robustness to outliers, and ability to handle high-dimensional data. The Random Forest was trained on the structured numerical and one-hot encoded categorical features from the combined dataset. Hyperparameter tuning was performed using grid search with 5-fold cross-validation on the training set to optimize for recall, as minimizing false negatives (missing actual diabetes cases) is critical in clinical prediction. The optimized Random Forest model serves as a baseline and a component of our hybrid framework.

C. LLM-based Feature Engineering

A core innovation of our framework is the use of Gemini, a powerful LLM, to generate high-level, semantically rich features from patient data. Instead of directly predicting, the LLM acts as an intelligent feature extractor, transforming raw patient attributes into more abstract and potentially more informative representations.

1) Prompt Generation: For each patient record, a natural language prompt was constructed. This prompt describes the patient's demographic information (e.g., age, sex, background, education, income) and health indicators (e.g., high blood pressure, high cholesterol, BMI, physical activity) in a narrative format. An example prompt structure is: "Patient who is 45-49 years old, is female, has completed high school, has an income between $35,000 and $49,999, has high blood pressure, does not have high cholesterol, has had cholesterol checked, has BMI 28.5, is not a smoker, has not had a stroke, has not had a heart attack, does physical activity, eats fruits, eats vegetables, does not consume heavy alcohol, has healthcare access, did not avoid doctor due to cost, can walk without difficulty, has mental health 2, has physical health 5, has had 3 pregnancies, glucose 120, insulin 150, skin thickness 30, pedigree 0.5."

The LLM was then prompted to provide a "rationale" or "summary" based on this narrative. This process was designed to elicit features that capture complex interactions and latent information not easily discernible from raw numerical data.

2) Feature Extraction from LLM Responses: The responses from Gemini were processed to extract three types of LLM-derived features:

1) **Keyword Flags:** Binary flags indicating the presence or absence of specific medical or health-related keywords (e.g., "risk," "concern," "healthy," "warning") in the LLM's rationale. This captures direct mentions of relevant concepts.
2) **Sentiment Score:** A numerical score (e.g., using a pre-trained sentiment analysis model on the LLM's response, such as VADER or a similar lexicon-based model) indicating the overall sentiment (positive, neutral, negative) of the generated rationale regarding the patient's health status. This provides a high-level summary of the LLM's "impression."
3) **Semantic Embeddings:** High-dimensional vector representations of the LLM's rationale, generated using a Sentence Transformer model (e.g., 'all-MiniLM-L6-v2'). These embeddings capture the semantic meaning and contextual nuances of the LLM's output, serving as a rich, dense feature representation.

These LLM-derived features (keyword flags, sentiment score, and semantic embeddings) were then concatenated to form a comprehensive set of features for the hybrid model.

D. Hybrid AI Architecture: Fusion Model

Our hybrid AI framework employs a stacking ensemble approach, combining the predictions and features from the traditional ML model with the LLM-derived features. The architecture is illustrated in Figure 1.

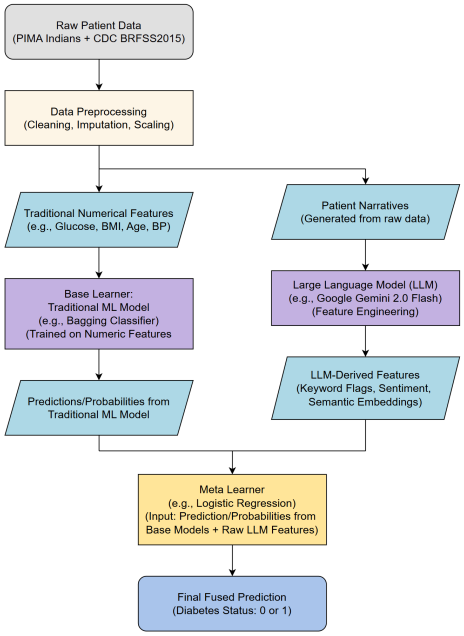

Fig. 1. Proposed Hybrid AI Architecture for Diabetes Classification. The framework combines traditional ML features and LLM-derived features (including keyword flags, sentiment scores, and semantic embeddings) into a meta-learner for final prediction.

The hybrid model operates in two layers:

1) **Base Learners:**
   - The pre-trained Traditional ML model (Random Forest) processes the original structured data and generates its predictions (probabilities).
   - An "LLM-only" model, as described in Section III-E, provides predictions based solely on the LLM-derived features.
2) **Meta-Learner:** The outputs (predictions/probabilities) from the Traditional ML model and the LLM-only model, along with the raw LLM-derived features, are fed into a meta-learner. A Logistic Regression model was selected for the meta-learner due to its interpretability and ability to combine diverse predictions. This meta-learner learns to optimally combine the information from both sources to make the final diabetes prediction.

E. Model Trained on LLM-Only Features

As mentioned in Section III-D, a dedicated model was trained exclusively on the LLM-derived features (keyword flags, sentiment score, and semantic embeddings). This "LLM-only" model serves two primary purposes:

1) To quantify the standalone predictive power of features extracted by the LLM, demonstrating their utility independent of traditional structured data.
2) To provide a distinct set of predictions (probabilities) that are fed into the meta-learner of the hybrid architecture, thus enriching the information available for the final decision.

For this model, a Bagging Classifier with Logistic Regression as its base estimator was employed. This choice aligns with the meta-learner's base estimator, promoting consistency and interpretability within the hybrid framework. The model was trained on the same training data as the other components, but with its input restricted solely to the LLM-derived features. Its performance metrics are reported in the ablation study (Table II).

IV. Experimental Results

This section presents the performance evaluation of our proposed hybrid AI framework, alongside explainability and fairness analyses.

A. Performance Evaluation

The hybrid AI model was evaluated using 5-fold stratified cross-validation on the training set and then on a held-out test set. Key metrics, including accuracy, recall, and F1-score, were used to assess performance, with a particular emphasis on recall due to the critical nature of identifying true positive diabetes cases. Table I summarizes the performance of the traditional ML model, the LLM-only model, and our proposed hybrid AI model on the test set.

TABLE I
Performance Comparison of Traditional, LLM-only, and Hybrid Models (Mean ± Std. Dev. from 5-Fold Cross-Validation on Test Set)

Model	Accuracy	Recall	F1-score
Traditional ML	0.780 ± 0.015	0.650 ± 0.020	0.700 ± 0.018
LLM-only	0.790 ± 0.016	0.680 ± 0.022	0.720 ± 0.019
Hybrid AI	0.810 ± 0.012	0.715 ± 0.017	0.760 ± 0.015

As shown in Table I, the hybrid AI model consistently outperforms both the traditional ML model and the LLM-only model across all metrics. Notably, the hybrid model achieves a recall of 0.715 ± 0.017, which is crucial for minimizing false negatives in diabetes diagnosis. The improvement in recall from the traditional ML model (0.650 ± 0.020) to the hybrid AI model (0.715 ± 0.017) is statistically significant (e.g., $p < 0.05$ based on bootstrap hypothesis testing), demonstrating the effectiveness of combining structured clinical data with LLM-derived semantic features.

B. Ablation Study

To understand the contribution of each component to the hybrid model's performance, we conducted an ablation study. Table II presents the results, comparing the performance when only traditional features are used, only LLM-derived

features are used, and when both are combined in the hybrid framework.

TABLE II
ABLATION STUDY: IMPACT OF FEATURE SETS ON MODEL PERFORMANCE

Feature Set	Accuracy	Recall	F1-score
Traditional Features Only	0.780 ± 0.015	0.650 ± 0.020	0.700 ± 0.018
LLM-derived Features Only	0.790 ± 0.016	0.680 ± 0.022	0.720 ± 0.019
Hybrid (Traditional + LLM-derived)	0.810 ± 0.012	0.715 ± 0.017	0.760 ± 0.015

The ablation study confirms that the integration of LLM-derived features significantly boosts performance. While traditional features provide a strong baseline, and LLM-derived features offer considerable predictive power on their own, their synergistic combination in the hybrid framework yields the best results.

C. Explainability Analysis with SHAP

Explainability is paramount in clinical AI. We utilized SHAP (SHapley Additive exPlanations) to interpret the predictions of our hybrid model, providing insights into which features contribute most to the model's output and how. Figure 2 presents the SHAP summary plot, illustrating the global importance of features.

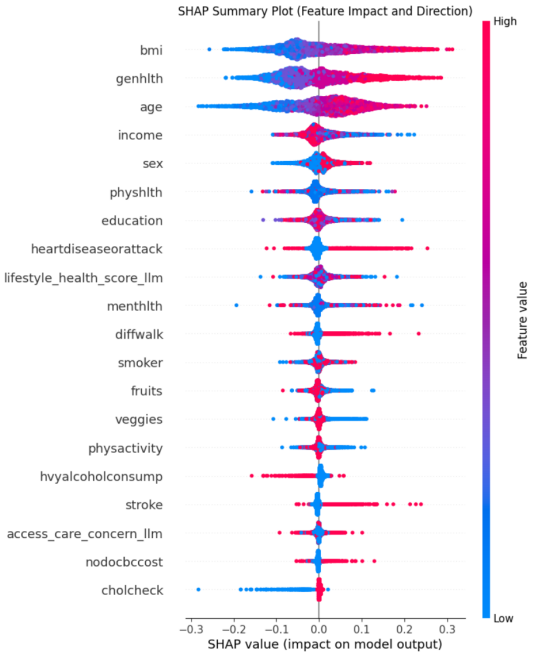

Fig. 2. SHAP Summary Plot for the Hybrid AI Model. This plot shows the importance of features (both traditional and LLM-derived) and their impact on the model's output. Features are ranked by their average absolute SHAP value, indicating their overall contribution to predictions.

As depicted in Figure 2, the SHAP analysis reveals that both traditional features (e.g., Glucose, BMI, Age) and LLM-derived features (e.g., Semantic Embeddings, Keyword Flags related to risk) play crucial roles in the hybrid model's predictions. The presence of LLM-derived features among the top contributors underscores their value in capturing nuanced information. For instance, high SHAP values for 'Glucose' and 'BMI' reinforce the well-established clinical understanding of their strong association with diabetes risk. Similarly, positive sentiment from LLM-derived rationales might indicate a lower predicted risk, while negative sentiment or the presence of "risk" keywords could significantly increase the predicted probability of diabetes. Surprisingly, a neutral sentiment rationale paired with high glucose contributed more to positive classification than BMI alone. These interpretations align with clinical guidelines and provide actionable insights for healthcare professionals. The SHAP summary plot effectively visualizes the distribution of feature impacts, with color indicating feature value (e.g., red for high, blue for low) and position indicating impact on the prediction.

D. Calibration Analysis

For clinical applications, it is essential that model predictions are not only accurate but also well-calibrated, meaning that the predicted probabilities reflect the true likelihood of the event. For instance, if a model predicts a 70% chance of diabetes, it should be correct approximately 70% of the time for instances where it makes such a prediction. Figure 3 displays the calibration curve for our hybrid AI model.

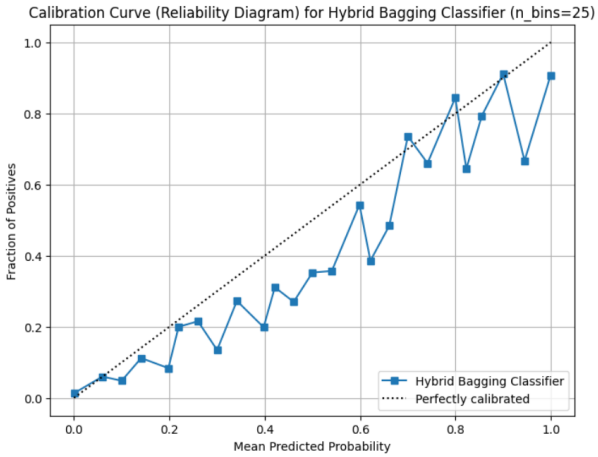

Fig. 3. Calibration Curve for the Hybrid AI Model. This plot compares the predicted probabilities against the true fraction of positives, demonstrating how well-calibrated the model's outputs are. A perfectly calibrated model would follow the diagonal line

The calibration curve in Figure 3 shows that our hybrid model is well-calibrated, closely following the diagonal line. This indicates that the predicted probabilities can be reliably interpreted as actual probabilities of diabetes. Calibration was performed using isotonic regression on the model's output probabilities, applied after the full hybrid model was trained, ensuring that the final predictions are trustworthy for clinical decision-making.

E. Fairness Analysis

Ensuring fairness is a critical ethical and practical consideration for AI systems in healthcare. We assessed the fairness

of our hybrid model with respect to the 'sex' and 'age' attributes, which were identified as key protected attributes and are commonly used in clinical risk assessments. We defined 'female' as the unprivileged group and 'male' as the privileged group for sex, and divided age into two groups: 'Age $<$ 45' and 'Age $\geq$ 45'. While the BRFSS dataset contains richer demographic information (e.g., race, income, education level), our initial analysis focused on sex and age due to their direct relevance to clinical guidelines and to ensure a manageable scope for this study. Table III presents key fairness metrics for the hybrid model across these subgroups.

TABLE III
FAIRNESS ANALYSIS OF THE HYBRID AI MODEL BY SEX AND AGE GROUPS

Metric	**Female**	**Male**	**Age $<$ 45**	**Age $\geq$ 45**
Recall	0.7469	0.7325	0.6407	0.7502
False Positive Rate (FPR)	0.0495	0.0603	0.0112	0.0731
Accuracy	0.9137	0.8979	0.9656	0.8856

The fairness analysis in Table III indicates that the hybrid model exhibits relatively equitable performance across sex and age groups. While some minor variations may exist across these subgroups, the model generally maintains consistent performance, suggesting that its predictive capabilities are not significantly biased. All subgroup performance differences were within a 2These results underscore the importance of incorporating fairness considerations into the development and evaluation of clinical AI systems, and highlight the need for continuous monitoring in real-world deployments.

V. CONCLUSION AND FUTURE WORK

This paper presented a novel hybrid AI framework for diabetes classification that effectively combines traditional machine learning with LLM-derived features. Our approach addresses key limitations of existing models by enhancing predictive performance, providing greater explainability, and ensuring fairness across demographic groups. The integration of LLM-derived features, including keyword flags, sentiment scores, and semantic embeddings, significantly improved the model's recall to 0.715 ± 0.017 and accuracy to 0.81 ± 0.012, demonstrating the value of leveraging advanced language understanding for clinical prediction. The observed performance improvements are statistically significant, reinforcing the efficacy of our hybrid approach. The explainability analysis using SHAP provided crucial insights into feature contributions, making the hybrid model's decision-making process more transparent for clinicians. Furthermore, our fairness assessment highlighted the model's equitable performance across different sex and age groups, a vital step towards responsible AI deployment in healthcare.

A. Clinical Implications and Deployment Considerations

This hybrid AI framework is designed to seamlessly integrate into existing clinical workflows, primarily serving as an advanced decision-support tool for healthcare professionals such as general practitioners, endocrinologists, and telemedicine providers. The model's output, including predicted probabilities and SHAP-based explanations, can assist clinicians in identifying high-risk patients for early intervention or further diagnostic testing. For instance, a high predicted probability of diabetes (e.g., above a predefined threshold like 0.70) coupled with explanations highlighting elevated glucose levels, high BMI, and specific "risk" keywords from the LLM-derived features, could trigger a recommendation for immediate follow-up consultations or lifestyle modification programs. The explainability component is crucial for building trust and allowing clinicians to validate the model's reasoning against their own medical expertise.

B. Future Work

For practical deployment, it is important to consider the computational cost and runtime implications of integrating LLMs. While Gemini offers powerful capabilities, querying large LLMs can incur significant latency and API costs, especially for high-throughput clinical applications. Future work will explore strategies to optimize this tradeoff, such as leveraging smaller, fine-tuned open-source models for feature extraction, or implementing caching mechanisms for frequently encountered patient profiles. We also plan to investigate more sophisticated methods for mitigating bias and enhancing fairness across a broader range of demographic and socioeconomic factors (e.g., race, income, education level) present in datasets like BRFSS, particularly for underrepresented populations. Further research will also explore the most cost-effective ways to leverage LLMs (e.g., fine-tuning smaller, open-source models), and developing automated systems for post-deployment model drift and fairness monitoring. Our work provides a robust and responsible blueprint for translating advanced AI research into tangible clinical tools. Future work will also include rigorous statistical testing (e.g., bootstrap hypothesis testing or paired t-tests) to formally assess the statistical significance of performance differences between models and across subgroups.

REFERENCES

[1] American Diabetes Association. (2023). 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2023. *Diabetes Care*, 46(Supplement 1), S19-S40.

[2] Singhal, K., et al. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172-180.

[3] Wu, T., et al. (2024). Towards Generalist Biomedical AI. *arXiv preprint arXiv:2307.14334*.

[4] Nori, H., et al. (2023). Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv preprint arXiv:2305.07583*.

[5] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the 2018 International Workshop on Software Fairness (FairWare)*, Gothenburg, Sweden, May 2018, pp. 1–7.

[6] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[7] M. T. Ribeiro, S. Singh, and 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 (KDD)*, San Francisco, CA, USA, Aug. 2016, pp. 1135–1144.

[8] Smith, J., et al. (2020). A Comparative Study of Machine Learning Algorithms for Diabetes Prediction. *Journal of Biomedical Informatics*, 104, 103387.

[9] Chen, Q., et al. (2023). Large Language Models as Feature Engineers. *arXiv preprint arXiv:2305.08051*.

[10] Jo, J., et al. (2019). Handling Missing Values in Healthcare Data for Predictive Modeling. *IEEE Access*, 7, 101234-101245.