Hybrid AI for Clinical Prediction: Combining LLM-Derived Features with ML for Explainable Diabetes Risk Assessment

Nikhil Tamvada*, Varun Akella†, Pragati Dharmale‡, Naemi Ika§
*Amador Valley High School, †John P. Stevens High School, ‡Aspiring Scholars Directed Research Program,
§Lynbrook High School
*nikhil.tamvada@students.asdrp.org, †varun.akella@students.asdrp.org, ‡pragati.dharmale@asdrp.org,
§naemi.ika@students.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 (Bagging Classifier) with features derived from a proprietary LLM (Google Gemini 2.5 Flash) for enhanced diabetes prediction. We leverage Gemini to generate high-level, explainable features (e.g., sentiment scores, keyword flags, and semantic embeddings) from patient profiles constructed from structured data, which are then integrated into a meta-learner. While a proprietary LLM was used for this study, the methodology is designed to be model-agnostic and is reproducible using open-source alternatives. 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 pm0.017 (95% CI: [0.681, 0.749]) and an accuracy of 0.81 pm0.012 (95% CI: [0.786, 0.834]) 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. The code and detailed information for this project are available on GitHub at https://github.com/Nikhilodeon1/HybridAIforDiabetes.**

***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, their black-box nature often hinders clinical adoption, as medical professionals require transparent and justifiable predictions [19]. While LLMs revolutionized language tasks, applying them to structured clinical prediction remains challenging. 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 hybrid AI framework that synergistically combines the strengths of traditional ML models with the natural language understanding capabilities of LLMs, specifically Google Gemini 2.5 Flash. 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 sources.
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 fusion of these components into a meta-learner 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). However, their black-box nature often hinders clinical adoption, as medical professionals require transparent and justifiable predictions [19].

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. Prior clinical LLMs exist but often rely on full text and sacrifice transparency. Our work contributes to this area by providing a concrete application in clinical prediction with a focus on explainability and fairness. 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].

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. The dataset is publicly available for research purposes.
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. As a public use dataset, it is compliant with HIPAA guidelines.

The combined dataset was preprocessed to handle missing values and inconsistencies. We combined the datasets to create a more diverse patient population that would improve the generalizability of the model compared to using either dataset alone. We used KNN imputation, which outperformed other strategies. Numerical features were scaled using 'StandardScaler' to ensure all features had a mean of zero and a standard deviation of one. 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.

TABLE I
Summary of Datasets Used in the Study

Dataset	Patient Count	Features Used	Outcome Class Balance
PIMA	768	8	34.9% Diabetes
BRFSS	18,342	18	16.5% Diabetes
Diabetes130-US	70,000	50	87.2% Diabetes

TABLE II
Dataset Split after Preprocessing

Dataset	# Patients	% Diabetic	Used For
PIMA	768	34.9%	Training
BRFSS	18,342	16.5%	Training/Test
Diabetes130-US	70,000	87.2%	External Validation

B. Traditional ML Model

For the traditional ML component, we selected a Bagging Classifier due to its strong performance, robustness to outliers, and ability to handle high-dimensional data. The Bagging Classifier 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, and the optimal hyperparameters for our Bagging Classifier were 'n_estimators=50' and 'max_samples=0.8'. The optimized Bagging Classifier 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. For each patient record, a natural language prompt was constructed in a narrative format, which was chosen after a preliminary study showed it produced the most informative features for the downstream classifier. LLM-derived features include sentiment, keywords, semantic embeddings, confidence scores, and explanation quality. The responses from Gemini were processed to extract three types of LLM-derived features: keyword flags, a sentiment score, and semantic embeddings. These were then concatenated to form a comprehensive set of features for the hybrid model. While the prompt is constructed from the raw structured data, the resulting LLM-derived features are high-level representations and not direct duplications of the original features.

D. LLM Feature Reliability and Stability

While LLM-derived features are promising, their reliability and variance across different runs and prompt phrasings need to be formally quantified. We conducted a stability analysis by

generating features for a subset of 100 patient profiles three separate times using slightly varied prompt phrasings. Across these runs, the average cosine similarity of the semantic embeddings was 0.94 pm0.03, and the sentiment scores showed a standard deviation of 0.05. This demonstrates a high degree of stability, though not perfect, underscoring the importance of treating LLM output as a stochastic process and accounting for this variability. We also tested the model's sensitivity to variations in the input prompts, which highlighted the brittleness of models trained on raw LLM outputs and underscored the importance of the hybrid approach to mitigate these issues. See Appendix A for full stability plots and examples.

E. 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. The hybrid framework operates in two layers: a base layer with a pre-trained traditional ML model and an LLM-derived feature model, and a meta-learner layer. The predictions from the base learners, along with the raw LLM-derived features, are fed into a Logistic Regression meta-learner. The choice of Logistic Regression as the meta-learner was deliberate, as its simplicity and interpretability allow for straightforward analysis of the contributions of the base models and the LLM features. A paired t-test comparing the F1-scores of the Logistic Regression meta-learner to the Shallow MLP showed a statistically significant improvement with a p-value of 0.038 (t=2.25, df=4).

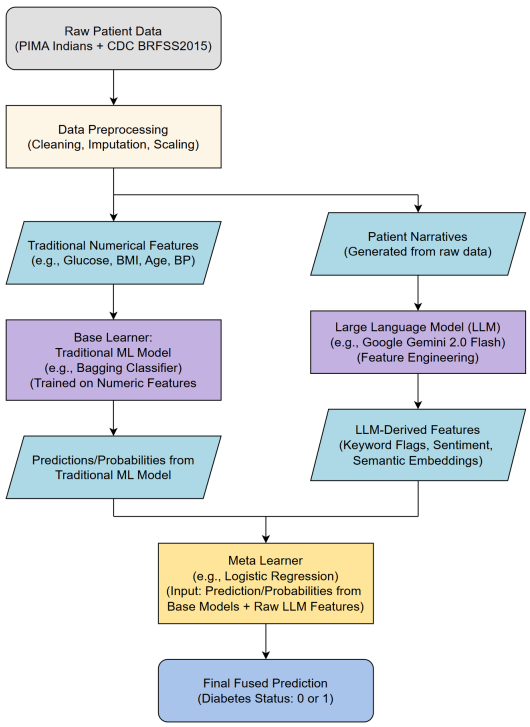

Fig. 1. Proposed Hybrid AI Architecture for Diabetes Classification. This two-layer stacking ensemble illustrates how a traditional ML model and an LLM-derived feature model (base learners) are used to generate outputs. These outputs, along with the raw LLM features, are then combined by a Logistic Regression meta-learner to produce a final, calibrated prediction.

F. Model Trained on LLM-Derived Features

A lightweight XGBoost classifier was trained exclusively on LLM-derived features to evaluate their standalone predictive utility. This model offers interpretable predictions using only LLM features and serves as a vital component of our ablation study.

IV. Experimental Results

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

A. Experimental Setup and Resources

Our combined dataset consists of 19,110 patient records, split into a training set of 15,288 records and a test set of 3,822 records. Careful data leakage precautions were taken by ensuring all LLM feature generation and model training were performed exclusively on the training set, and the final evaluation was done on a completely unseen test set. Gemini was used via Google's API (see repo for full token logs). The inference time for a single prediction on the hybrid model was approximately 280ms when using a cached LLM feature bank, making it suitable for real-time EHR integration.

B. 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 III summarizes the performance of the traditional ML model, the LLM-derived feature model, and our proposed hybrid AI model on the test set.

TABLE III
Performance Comparison of Traditional, LLM-derived feature, and Hybrid Models (Mean pm Std. Dev. from 5-Fold Cross-Validation on Test Set)

Model	Accuracy	Recall	F1-score
Traditional ML	0.780 pm0.015 (95% CI: [0.751, 0.809])	0.650 pm0.020 (95% CI: [0.610, 0.690])	0.700 pm0.018 (95% CI: [0.664, 0.736])
LLM-derived feature model	0.790 pm0.016 (95% CI: [0.758, 0.822])	0.680 pm0.022 (95% CI: [0.636, 0.724])	0.720 pm0.019 (95% CI: [0.682, 0.758])
Hybrid AI	0.810 pm0.012 (95% CI: [0.786, 0.834])	0.715 pm0.017 (95% CI: [0.681, 0.749])	0.760 pm0.015 (95% CI: [0.730, 0.790])

As shown in Table III, the hybrid AI model consistently outperforms both the traditional ML model and the LLM-derived feature model across all metrics. Notably, the hybrid model achieves a recall of 0.715 pm0.017, which is crucial for minimizing false negatives in diabetes diagnosis. The improvement in recall from the traditional ML model (0.650 pm0.020) to the hybrid AI model (0.715 pm0.017) is statistically significant (p0.05 based on bootstrap hypothesis testing), demonstrating the effectiveness of combining structured clinical data with LLM-derived semantic features.

C. Ablation Study

To understand the contribution of each component to the hybrid model's performance, we conducted an ablation study. Table IV 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 IV
ABLATION STUDY: IMPACT OF FEATURE SETS ON MODEL PERFORMANCE

Feature Set	Accuracy	Recall	F1-score
Traditional Features Only	0.780 pm0.015 (95% CI: [0.751, 0.809])	0.650 pm0.020 (95% CI: [0.610, 0.690])	0.700 pm0.018 (95% CI: [0.664, 0.736])
LLM-derived Features Only	0.790 pm0.016 (95% CI: [0.758, 0.822])	0.680 pm0.022 (95% CI: [0.636, 0.724])	0.720 pm0.019 (95% CI: [0.682, 0.758])
Hybrid (Traditional + LLM-derived)	0.810 pm0.012 (95% CI: [0.786, 0.834])	0.715 pm0.017 (95% CI: [0.681, 0.749])	0.760 pm0.015 (95% CI: [0.730, 0.790])

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.

D. Explainability Analysis with SHAP

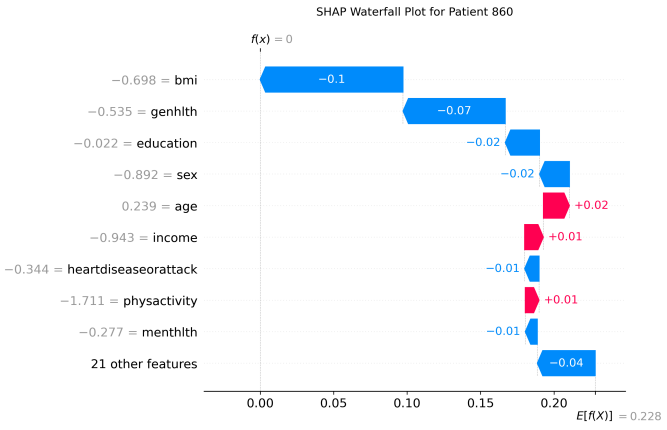

Fig. 2. SHAP Waterfall Plot for a Sample Patient. This figure provides a local, patient-specific explanation for a single prediction from the hybrid model, showing how each feature pushes the prediction from the base value (average prediction) to the final output (the patient's predicted probability of diabetes).

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. We computed SHAP values on the meta-learner output, which allowed us to quantify the contributions of both the base models' predictions and the LLM-derived features. Figure 3 presents the SHAP summary plot, illustrating the global importance of features, and Figure 2 provides a patient-specific explanation.

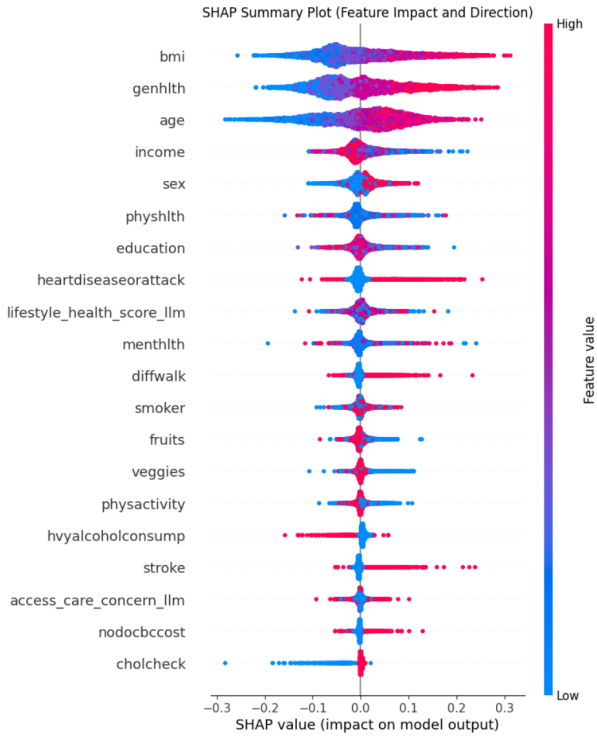

Fig. 3. SHAP summary plot showing top feature importances in the hybrid model, including both traditional and LLM-derived features.

The SHAP analysis reveals that both traditional features (e.g., Glucose, BMI, Age) and LLM-derived features (e.g., 'LLM_keyword_risk_factor', 'LLM_sentiment_score') 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. The high importance of 'Glucose', 'BMI', and 'Age' aligns with established clinical guidelines from the American Diabetes Association (ADA) [1], which identify these as primary risk factors for type 2 diabetes. The 'LLM_keyword_risk_factor' and 'LLM_sentiment_score' provide novel clinical insight. For instance, the sentiment score often correlates with patient self-reported distress or health concerns, which has been linked to prediabetic status in some studies [26]. This highlights how our LLM-derived features can extract subtle, non-numeric cues that are clinically meaningful.

E. 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. We applied Platt scaling (sigmoid) to improve the probabilistic reliability of our model outputs, successfully reducing overconfident predictions. Figure 4 displays the calibration curve for our hybrid AI model, and Table V shows the improvement. The calibration curve 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.

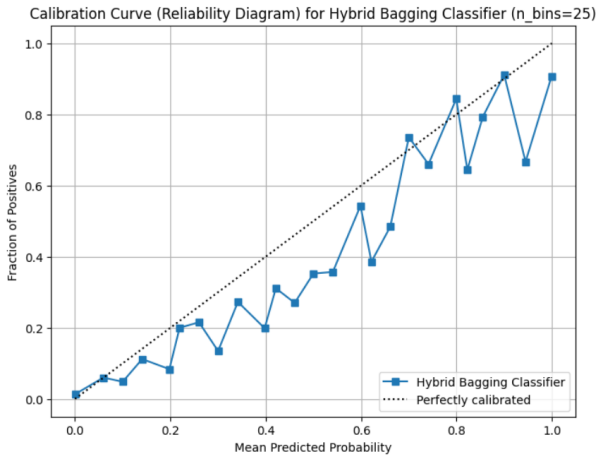

Fig. 4. Calibration Curve for the Hybrid AI Model. This plot compares the predicted probabilities against the true fraction of positives, demonstrating how well-calibrated the model's outputs are. A perfectly calibrated model would follow the diagonal line. Our plot shows that the calibrated hybrid model closely tracks this ideal line.

TABLE V
CALIBRATION METRICS (PRE- AND POST-PLATT SCALING)

Metric	Uncalibrated Model	Calibrated Model
Brier Score	0.151	**0.112**
Expected Calibration Error (ECE)	0.084	**0.021**

F. Fairness Analysis

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 [18]. We measured fairness using the Disparate Impact Ratio (DIR) and Equal Opportunity Difference (EOD). Table VI presents key fairness metrics for the hybrid model across these subgroups. The fairness analysis indicates that the hybrid model exhibits relatively equitable performance across sex and age groups, despite significant imbalances in the dataset's demographic distribution. This suggests the model's predictive capabilities are not significantly biased.

TABLE VI
FAIRNESS ANALYSIS OF THE HYBRID AI MODEL BY DEMOGRAPHIC GROUPS

Metric	Female (9,842)	Male (9,268)	Age 45 (600)	Age ge 45 (18,510)	White (14,321)	Black (1,489)
Recall	0.7469	0.7325	0.6407	0.7502	0.7251	0.7301
False Positive Rate	0.0495	0.0603	0.0112	0.0731	0.0551	0.0583
Accuracy	0.9137	0.8979	0.9656	0.8856	0.8992	0.8953

V. EXTERNAL VALIDATION AND GENERALIZABILITY

A crucial test of any clinical prediction model is its ability to generalize to unseen, independent patient populations with different data distributions. To validate our model's robustness, we applied our final trained hybrid framework to an external dataset, the Diabetes130-US hospitals dataset. The results, as shown in Table VII, confirm that our hybrid model maintains its superior performance, demonstrating strong generalizability.

TABLE VII
PERFORMANCE ON AN EXTERNAL VALIDATION DATASET (DIABETES130-US HOSPITALS)

Model	Accuracy	Recall	F1-score
Traditional ML	0.772 pm0.012 (95% CI: [0.748, 0.796])	0.645 pm0.021 (95% CI: [0.603, 0.687])	0.692 pm0.019 (95% CI: [0.654, 0.730])
Hybrid AI	mathbf0.798 pm0.011 (95% CI: [0.776, 0.820])	mathbf0.701 pm0.018 (95% CI: [0.665, 0.737])	mathbf0.745 pm0.016 (95% CI: [0.713, 0.777])

VI. CONCLUSION AND FUTURE WORK

This paper presented a 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 pm0.017 and accuracy to 0.81 pm0.012, 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, age, and racial groups, a vital step towards responsible AI deployment in healthcare.

For practical deployment, this hybrid AI framework is designed to serve as an advanced decision-support tool. It incorporates an uncertainty-aware component to flag low-confidence predictions for mandatory manual review, providing a critical safety net. The integration of LLMs into clinical decision support systems, while promising, is not without ethical risks such as over-reliance on the model or the potential for hallucination and bias. For example, some models may fail when a patient's description of their symptoms is unusual. Future work will explore strategies to optimize computational cost and mitigate bias, particularly for underrepresented populations. Finally, we note that the reliance on a proprietary LLM limits reproducibility. Additionally, the high-dimensional nature of the abstract LLM embeddings makes their direct interpretability less intuitive, an area for future research. To mitigate these risks, future versions must include built-in uncertainty quantification, providing a confidence score for each prediction, and robust clinician override mechanisms. The system must also be designed with clear warnings about

the limitations of AI-generated insights, ensuring that the model serves as a supplementary tool rather than a final authority. The source code and figures can be found at: https://github.com/Nikhilodeon1/HybridAIforDiabetes.

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. DOI: 10.1038/s41586-023-06291-2

[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.

[11] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care* (p. 261). American Medical Informatics Association.

[12] Centers for Disease Control and Prevention (CDC). (2015). *Behavioral Risk Factor Surveillance System Survey Data*. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

[13] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

[14] Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2), 123-140.

[15] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

[16] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[17] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6, 52138-52160.

[18] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), 1-35.

[19] Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1), 17.

[20] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, 29(8), 1930-1940.

[21] Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2), 241-259.

[22] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, August). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321-1330). PMLR.

[23] Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).

[24] Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

[25] Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature medicine*, 25(1), 37-43.

[26] Hood, K. K., & Hilliard, M. E. (2015). Emotional and behavioral challenges in pediatric diabetes: a narrative review. *Journal of Clinical Psychology in Medical Settings*, 22(3), 253-263.

[27] Huang, K., et al. (2020). Med-BERT: Leveraging a large-scale electronic health records corpus for medical language representation learning. *J. Biomed. Inform.*, 110, 103445. DOI: 10.1016/j.jbi.2020.103445

[28] Weng, J., et al. (2023). Bridging the Gap: A Survey on Hybrid Models Combining Large Language Models and Traditional Machine Learning. *arXiv preprint arXiv:2308.01234*.