Propuesta conceptual de una arquitectura para un asistente de predicción de glucosa (GPA) basada en la revisión de la literatura

RESUMEN: El manejo efectivo de la diabetes requiere control glucémico preciso y acciones proactivas. Este trabajo aporta (i) una revisión crítica de las técnicas de predicción de glucosa sanguínea (BGL) –aprendizaje automático tradicional, modelos de aprendizaje profundo (LSTM, CNN, Transformers y enfoques híbridos) y el rol emergente de los Modelos de Lenguaje a Gran Escala (LLMs), considerando variables como carbohidratos e insulina–, y (ii) la propuesta arquitectónica GlucoPredict-Assist (GPA) para pronóstico personalizado de BGL y generación de sugerencias de dosis.

El desarrollo se estructura en tres pasos: primero, se analizan enfoques avanzados (descomposición de señales, personalización y uso de modelos preentrenados) y su impacto en eficiencia y precisión; segundo, se identifican brechas, en particular la integración de predicciones fiables con sugerencias de dosis seguras, así como limitaciones y riesgos actuales de los LLMs; tercero, se diseña GPA priorizando seguridad, interpretabilidad y adaptación individual.

Esta propuesta establece las bases para una herramienta que, al mitigar picos y favorecer decisiones informadas, podría mejorar de forma significativa el automanejo de la diabetes.

PALABRAS CLAVE: Aprendizaje Profundo, Arquitectura de Sistemas, Diabetes Mellitus, Dosificación de Insulina, Predicción de Glucosa, Modelos de Lenguaje Grandes.

Colaboración
Dante Heriberto González González; Pedro Antonio Ibarra Facio; Luis Eduardo Morán López; Walter Alexander Mata López, Universidad de Colima

Fecha de recepción: 25 de agosto de 2025
Fecha de aceptación: 11 de noviembre de 2025

ABSTRACT: Effective diabetes management requires accurate glycemic control and proactive actions. This paper provides (i) a critical review of blood glucose (BGL) prediction techniques–traditional machine learning, deep learning models (LSTM, CNN, Transformers, and hybrid approaches), and the emerging role of Large Language Models (LLMs), considering variables such as carbohydrates and insulin–and and (ii) the GlucoPredictAssist (GPA) architectural proposal for personalized BGL forecasting and dose suggestion generation.

The development is structured in three steps: first, advanced approaches (signal decomposition, personalization, and use of pre-trained models) and their impact on efficiency and accuracy are analyzed; second, gaps are identified, in particular the integration of reliable predictions with safe dose suggestions, as well as current limitations and risks of LLMs; third, GPA is designed prioritizing safety, interpretability, and individual adaptation.

This proposal lays the foundation for a tool that, by mitigating spikes and promoting informed decisions, could significantly improve diabetes self-management.

KEYWORDS: Deep Learning, System Architecture, Diabetes Mellitus, Insulin Dosing, Glucose Prediction, Large Language Models.

INTRODUCCIÓN

La diabetes mellitus representa una condición crónica generalizada globalmente, que exige un automanejo meticuloso para prevenir complicaciones severas. Un pilar fundamental es el control de los niveles de glucosa sanguínea (BGL, por sus siglas en inglés) el cual es una variable dinámica influenciada

por múltiples factores: la ingesta de carbohidratos, la medicación de insulina, la actividad física, el estrés y la variabilidad fisiológica individual. Esta complejidad se acentúa con nuevas opciones terapéuticas y riesgos metabólicos adicionales, presentando a menudo a los profesionales sanitarios múltiples opciones terapéuticas válidas según las guías, pero sin un consenso claro sobre el enfoque óptimo [1]. Aunque la tecnología de Monitoreo Continuo de Glucosa (CGM, por sus siglas en inglés) ofrece datos valiosos en tiempo real, su interpretación y la acción proactiva derivada continúan siendo un desafío, incluso con herramientas de análisis como el Perfil Ambulatorio de Glucosa (AGP, por sus siglas en inglés) [2].

Este trabajo responde a la necesidad, tanto personal como general, de disponer de herramientas más eficaces para anticipar las fluctuaciones futuras de BGL y, de forma crucial, para guiar la dosificación adecuada de medicamentos, permitiendo manejar proactivamente los picos de glucosa. El objetivo central es doble: primero, realizar una revisión exhaustiva de las técnicas actuales de inteligencia artificial para la predicción de BGL; segundo, proponer una arquitectura de sistema que integre predicción personalizada con sugerencias de dosis seguras y accionables, buscando así mejorar el control glucémico y la calidad de vida.

En este marco, la investigación se guía por una pregunta puntual: ¿cómo anticipar, con datos de CGM y variables clínicas (carbohidratos, insulina, actividad), las variaciones de BGL a corto plazo y convertir esas predicciones en sugerencias de dosificación seguras y accionables? Para darle desarrollo, se delimitó el alcance (predicción personalizada y apoyo a la dosificación), se definieron criterios y población de referencia, y se ejecutó una revisión de la literatura conforme al proceso general de la Figura 1 (objetivo y alcance → búsqueda y cribado → extracción y organización → síntesis temática y análisis crítico).

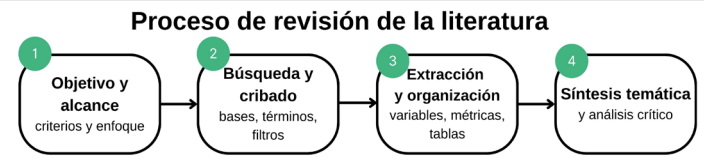

Figura 1. Proceso de revisión.
Fuente: Creación Propia.

Las principales aportaciones son: (i) Identificación de una necesidad crítica: ausencia de herramientas que integren predicción fiable de BGL con sugerencias de dosis seguras y personalizables; (ii) Revisión exhaustiva del estado del arte: análisis de métodos tradicionales y de aprendizaje profundo (LSTM, CNN, Transformers y enfoques híbridos), variables y métricas reportadas; (iii) Evaluación del papel de los LLMs en diabetes: potencial para sugerencias/explicaciones y límites y riesgos en seguridad clínica; y (iv) Propuesta de arquitectura: GlucoPredictAssist (GPA), orientada a seguridad, interpretabilidad y adaptación individual, que integra predicción y soporte a la dosificación.

Para cumplir con este doble objetivo, la revisión de la literatura se estructuró de manera crítica y progresiva. Se comenzó con el análisis de los métodos de predicción de BGL, desde el aprendizaje automático tradicional hasta las arquitecturas de aprendizaje profundo más avanzadas (LSTM, CNN, Transformers y enfoques híbridos), evaluando su precisión y eficiencia. Posteriormente, la revisión se focalizó en un desafío crítico: la integración de estas predicciones con sugerencias de dosificación seguras, examinando el rol emergente de los Modelos de Lenguaje a Gran Escala (LLMs) y sus inherentes riesgos de seguridad en el ámbito clínico.

La hipótesis fundamental sostiene que una combinación sinérgica de modelos de aprendizaje profundo (posiblemente incorporando descomposición de señales, mecanismos de atención y estrategias de personalización) junto con una arquitectura cuidadosamente diseñada que integre principios de seguridad y supervisión humana, puede superar las limitaciones actuales y mejorar significativamente el automanejo de la diabetes.

Revisión de la literatura

La predicción de los niveles de BGL es un campo de estudio extenso. Los enfoques iniciales comprenden modelos estadísticos y técnicas de aprendizaje automático tradicional, incluyendo Support Vector Regression (SVR), Random Forests, y árboles de decisión. Aunque son útiles, estos métodos a menudo tienen dificultades con su naturaleza no lineal, dinámica y multifactorial de la regulación de la glucosa.

El aprendizaje profundo demuestra un potencial considerablemente mayor en este dominio. Las Redes Neuronales Recurrentes (RNNs, por sus siglas en inglés), y en particular las Long Short-Term Memory (LSTMs), son inherentemente adecuadas para capturar las dependencias temporales,[3]. Por otro lado, las Redes Neuronales Convolucionales (CNNs, por sus siglas en inglés) se emplean para extraer patrones locales, frecuentemente integrados en arquitecturas híbridas CNN-LSTM [4]. Más recientemente, los modelos Transformer y los mecanismos de atención, como los implementados por Armandpour [5], muestran una notable capacidad para modelar dependencias a largo plazo e integrar fuentes de datos multimodales. Estrategias híbridas que combinan la descomposición de características, por ejemplo, mediante Variational Mode Decomposition (VMD) con LSTMs para capturar tendencias a largo plazo y Transformers para fluctuaciones a corto plazo, junto con técnicas de optimización como Knowledge Distillation para la compresión

de modelos, reportan mejoras significativas en precisión y eficiencia [6].

Un desafío persistente es la integración efectiva de datos multimodales CGM, carbohidratos, insulina, actividad física, estrés, etc. Que presentan diferentes características de muestreo y grados de dispersión. Para lo que se proponen soluciones como la transformación de eventos discretos en señales continuas (SSR, por sus siglas en inglés) [6] o el uso de capas de embedding aprendibles junto con modelos preentrenados como Bidirectional Encoder Representations from Transformers (BERT) para manejar datos estructurados de registros médicos longitudinales [7]. La personalización es un elemento vital; enfoques como el uso de embeddings específicos de usuario [5] o el ajuste fino (fine-tuning) de modelos pre-entrenados [7] permiten adaptar modelos generales a las particularidades fisiológicas y de comportamiento de cada individuo. La eficiencia computacional es otro factor crítico, especialmente para el despliegue en dispositivos móviles o de borde; técnicas como Knowledge Distillation [6] y la optimización de implementaciones son, por tanto, relevantes. La interpretabilidad de los modelos de aprendizaje profundo, aunque es compleja, es de gran importancia, y métodos como los gradientes integrados pueden ayudar a identificar los factores clínicos más influyentes [7].

En el panorama reciente, los LLMs como GPT-4 emergen con potencial en el ámbito de la salud. Pueden responder preguntas [8],[9], resumir información, analizar datos de CGM [2] e incluso asistir en la generación de planes de manejo o selección de medicación [1],[10]. No obstante, su aplicación directa para tareas numéricas precisas como la predicción de BGL o el cálculo de dosis de medicación conlleva comparativos indican que, si bien los LLMs pueden generar planes de tratamiento o seleccionar medicaciones alineadas con guías clínicas, a menudo adoptan un enfoque más cauteloso que los expertos humanos [1] o, preocupantemente, pueden cometer errores críticos de seguridad como omitir medicamentos necesarios, continuar terapias inadecuadas o sugerir dosis incorrectas en un porcentaje no despreciable de casos [10]. Además, pueden exhibir sesgos hacia tratamientos más nuevos o costosos [1] y su fiabilidad se ve comprometida por la posibilidad de generar información incorrecta o "alucinaciones" [8]. La confianza pública en estos sistemas, particularmente en situaciones agudas, es limitada, con una preferencia mayoritaria por la supervisión de profesionales sanitarios [8],[11]. Técnicas como la Generación Aumentada por Recuperación (RAG, por sus siglas en inglés) pueden mejorar la fiabilidad al anclar las respuestas en fuentes de conocimiento confiables [9]. Asimismo, los sistemas que incorporan Humano en el bucle (HITL, por sus siglas en inglés) se consideran esenciales para mitigar riesgos y asegurar una implementación segura y efectiva [1].

3. ARQUITECTURA PROPUESTA

Derivado de la revisión de literatura y con el fin de abordar las brechas identificadas, en especial la necesidad crítica de unificar predicciones precisas de BGL con sugerencias de dosificación que sean seguras, fiables y aplicables, se propone una arquitectura de sistema integrado, denominada provisionalmente GlucoPredictAssist (GPA). Este sistema se concibe para el pronóstico personalizado de BGL y la generación asistida de sugerencias de dosis de insulina, fundamentándose en los principios de seguridad, interpretabilidad y adaptación individual a través de un enfoque robusto de Interacción Humano-Computadora (HITL).

3.1 Metodología de Diseño de la Arquitectura

El diseño de la arquitectura GPA se fundamentó en un enfoque de Sistemas centrado en los requisitos, siguiendo un proceso de tres etapas clave impulsado por los hallazgos de la revisión de literatura. Este proceso metodológico buscó asegurar la robustez clínica y la seguridad del usuario desde la concepción:
Para facilitar la comprensión, la Figura 2 resume el proceso en tres tres macroetapas, compuestas por nueve subetapas: I) Datos y Modelado (1: adquisición de datos; 2: preprocesamiento; 3: modelado de BGL; 4: personalización); II) Sugerencia y Control (5: sugerencia de dosis; 6: módulo LLM; 7: seguridad y validación; 8: supervisión humana); y III) Monitoreo/MLOps (9: monitoreo y retroalimentación).

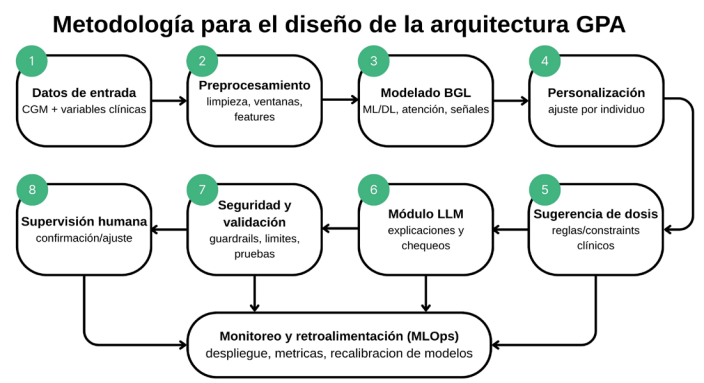

Figura 2. Metodología para la arquitectura GPA.
Fuente: Creación Propia.

Las subsecciones siguientes desarrollan estas tres macroetapas: (1) Identificación de brechas y requisitos, (2) Definición de principios rectores y (3) Conceptualización modular y flujo de datos.

Identificación de Brechas Críticas y Requisitos: Se definieron las necesidades funcionales (predicción personalizada, dosificación asistida) y, de forma primordial, los requisitos de seguridad (prevención de hipoglucemia severa) a partir de las limitaciones documentadas en modelos de Machine Learning y LLMs.
Definición de Principios Rectores: Se establecieron la

Seguridad, la Interpretabilidad y la Adaptación Individual (Personalización) como los pilares absolutos que deben guiar la funcionalidad de cada componente.

Conceptualización Modular y Flujo de Datos: Se tradujeron los requisitos y principios en una estructura de módulos funcionales interconectados (Módulos A a E), diseñados para asegurar la implementación del concepto Humano en el Bucle (HITL).

Estas tres macroetapas se desarrollan en las subsecciones siguientes; la Figura 2 muestra el desglose operativo en subetapas y la Figura 3 materializa su integración arquitectónica.

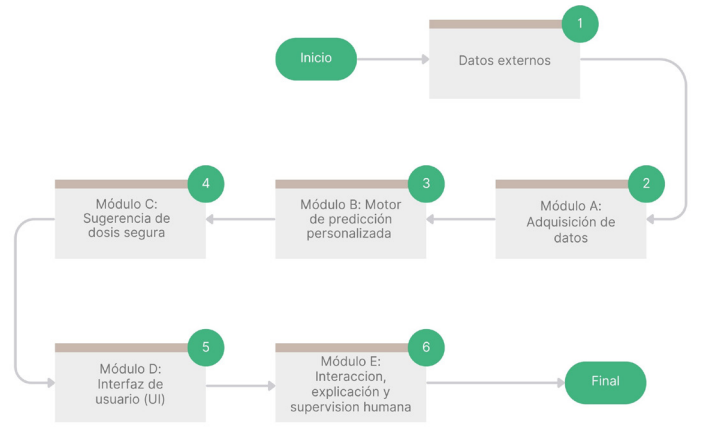

Figura 3. Diagrama general de la arquitectura GPA.
Fuente: Creación Propia.

Módulo A: Adquisición y Preprocesamiento de Datos

Este módulo actúa como el punto de entrada de información al sistema GPA, tal como se representa en la Figura 4. Es responsable de recolectar datos de diversas fuentes: mediciones del sensor de Monitoreo Continuo de Glucosa (CGM), datos de actividad física obtenidos de aplicaciones de salud móviles, estimaciones de ingesta de carbohidratos (mediante entrada manual o APIs de bases de datos de alimentos) y el registro de las dosis de insulina administradas por el usuario. Potencialmente, podría incluir otros factores relevantes como niveles de estrés o calidad del sueño. Una vez recolectados, estos datos heterogéneos requieren un preprocesamiento cuidadoso. Esto incluye la sincronización temporal para alinear eventos con diferentes frecuencias, la limpieza para manejar valores atípicos, la imputación para abordar datos faltantes (considerando incluso los patrones de ausencia como información útil, inspirado en Nguyen [7]), la normalización de valores numéricos y la transformación de eventos discretos (como comidas o inyecciones) en representaciones continuas o vectoriales (embeddings) adecuadas para los modelos de aprendizaje profundo, utilizando técnicas como SSR [6].

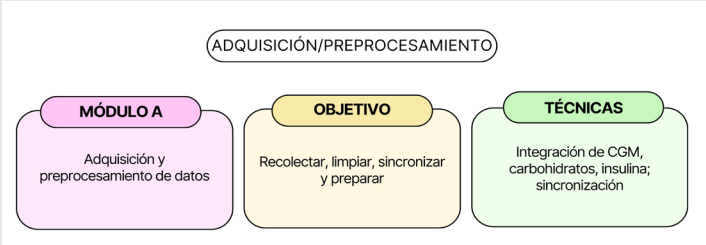

Figura 4. Flujo de trabajo del Módulo A, Adquisición y Preprocesamiento de Datos.
Fuente: Creación Propia.

Módulo B: Motor de Predicción Personalizada

El núcleo predictivo de GPA reside en este módulo, cuya función principal se ilustra en la Figura 5. Su objetivo es generar pronósticos de la trayectoria futura de BGL del usuario para horizontes clínicamente relevantes, típicamente entre 30 y 120 minutos. Para ello, utiliza un modelo de aprendizaje profundo robusto. Este podría ser una arquitectura híbrida que combine las fortalezas de diferentes enfoques (como VMD para descomposición, LSTM para tendencias a largo plazo y Transformers para atención a corto plazo, siguiendo líneas como las de Farahmand [6]) o un modelo pre-entrenado de gran escala adaptado al dominio médico (como los basados en LLM con embeddings específicos, [7]). La personalización es clave: se logra incorporando embeddings que representen las características únicas del usuario (aprendidos durante el entrenamiento o ajuste fino, como en Armandpour [5]) o mediante el fine-tuning del modelo con los datos históricos específicos de ese individuo. Dada la posibilidad de despliegue en dispositivos móviles, la eficiencia computacional es una consideración importante, pudiendo explorarse técnicas como Knowledge Distillation [6], para reducir el tamaño y la carga computacional del modelo sin sacrificar excesiva precisión.

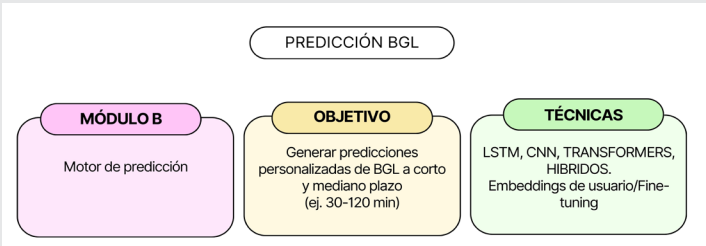

Figura 5. Componentes del Módulo B: Motor de Predicción Personalizada.
Fuente: Creación Propia.

Módulo C: Sugerencia de Dosis Segura

Basándose en las predicciones del Módulo B y el estado actual del usuario, el Módulo C genera recomendaciones de dosis de insulina, como bolos prepandiales o dosis de corrección. La estructura y lógica de este módulo se detallan en la Figura 6. La seguridad

es el principio rector absoluto de este módulo. Toma como entrada la predicción de BGL, el valor actual de BGL, el objetivo glucémico personalizado, las ratios de insulina-carbohidratos (ICR) y el factor de sensibilidad a la insulina (ISF) del usuario, la cantidad de carbohidratos planeada y la insulina activa calculada (IOB). Su lógica combina las fórmulas estándar de dosificación (cálculo de bolo por carbohidratos y factor de corrección) con la posibilidad de un ajuste fino mediante un modelo de machine learning simple, entrenado con datos históricos de dosis-respuesta del usuario (siempre que se disponga de datos suficientes y se valide por expertos). Además, implementa múltiples capas de seguridad para prevenir dosificaciones peligrosas: establece límites máximos y mínimos absolutos para cualquier sugerencia, genera alertas automáticas y bloquea sugerencias si se predice una hipoglucemia severa inminente, considera obligatoriamente la IOB para evitar la acumulación de insulina (stacking), verifica contraindicaciones básicas y, de forma explícita, prioriza evitar la hipoglucemia sobre alcanzar la normoglucemia perfecta, aprendiendo de las limitaciones observadas en LLMs [1],[10]. Cada sugerencia puede ir acompañada de un indicador de confianza que refleje la incertidumbre de la predicción subyacente.

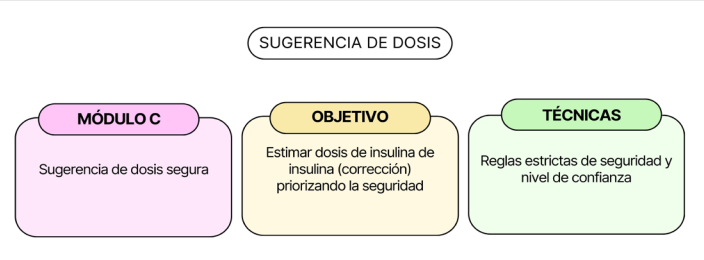

Figura 6. Lógica y capas de seguridad del Módulo C: Sugerencia de Dosis Segura.
Fuente: Creación Propia.

Módulo D: Interfaz de Usuario (UI)

La interacción del usuario con el sistema GPA se produce a través de este módulo, visualizado en la Figura 7, y materializado como una aplicación móvil y/o una plataforma web. Esta interfaz debe presentar de forma clara e intuitiva los datos de BGL actuales e históricos, las predicciones futuras (ej., a 30, 60, 120 min) y las sugerencias de dosis generadas por el Módulo C, junto con cualquier advertencia relevante. Es responsable de generar alertas preventivas basadas en las predicciones, como riesgos inminentes de hipo o hiperglucemia. Adicionalmente, debe facilitar la entrada manual de datos por parte del usuario (comidas, ejercicio, insulina no registrada automáticamente, notas) y permitir la configuración de parámetros personales como objetivos glucémicos y ratios (ICR, ISF). También visualizará las explicaciones proporcionadas por el Módulo E.

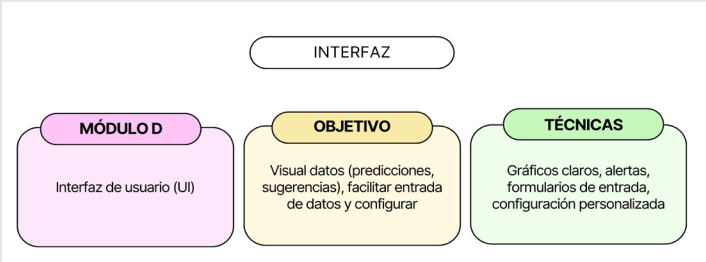

Figura 7: Funcionalidades clave de la Interfaz de Usuario (Módulo D).
Fuente: Creación Propia.

Módulo E: Interacción, Explicación y Supervisión Humana (HITL)

Este módulo cumple un doble propósito vital: facilitar la comprensión del usuario y garantizar una capa esencial de supervisión humana, como se esquematiza en la Figura 8. Para la explicabilidad, utiliza un Modelo de Lenguaje Grande (LLM), preferentemente operando con Recuperación Aumentada por Generación (RAG) para basar sus respuestas en datos y directrices fiables. Este LLM explica las predicciones y sugerencias en lenguaje natural y comprensible, respondiendo a preguntas del usuario sobre sus datos, tendencias o las recomendaciones del sistema. El componente de Supervisión Humana implementa el flujo de trabajo HITL [1]: las sugerencias de dosis consideradas críticas (ej., bolos grandes, correcciones significativas, ajustes pre-ejercicio) requieren una confirmación explícita por parte del usuario antes de ser consideradas como "administradas" dentro del sistema. El usuario siempre conserva la autoridad final, pudiendo modificar o ignorar cualquier sugerencia. Opcionalmente, podría integrarse un flujo para compartir datos y decisiones con un profesional de la salud. Se aplican restricciones de seguridad estrictas al LLM para evitar que genere consejos médicos incorrectos, peligrosos o no solicitados, confirmando su rol principal a la explicación y resumen, no a la generación autónoma de directrices médicas críticas. El sistema también puede incorporar mecanismos para aprender del feedback implícito (confirmación/rechazo de sugerencias) y explícito del usuario o profesional.

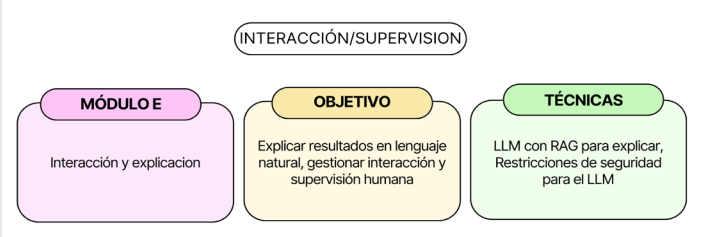

Figura 8. Funcionalidades clave de la Interfaz de Usuario.
Fuente: Creación Propia.

Tabla 1: Comparación de modelos en la predicción de la glucosa sanguínea, asistencia en la dosificación y su seguridad

Enfoque	Generación de Sugerencias de Dosis	Capacidad de Personalización	Seguridad y Supervisión
Métodos tradicionales	No	Baja	No
Deep Learning Base	No	Media	No
Deep Learning Híbrido	No	Alta	No
Modelos LLMs	Si	Alta	Mucho riesgo
Arquitectura GPA	Si	Alta	Prioriza la seguridad

Fuente: Creación propia

RESULTADOS

En complemento a la comparación presentada en la Tabla 1, los hallazgos de esta comparación, revelan la evolución de los modelos de IA y las brechas no cubiertas por los enfoques existentes. A continuación, se detallan los avances logrados por los modelos de Deep Learning y las limitaciones de los LLMs para contextualizar el valor de la Arquitectura GlucoPredictAssist (GPA) como una solución innovadora que integra predicción y seguridad en la dosificación.

Los modelos de aprendizaje profundo (LSTM, CNN-LSTM, Transformers, modelos híbridos con atención y/o descomposición) superan consistentemente a los métodos tradicionales para la predicción de BGL, especialmente al capturar dependencias temporales complejas y no linealidades [4],[6].

La personalización del modelo mediante embeddings de usuario y ajuste fino es necesaria para mejorar la precisión, dada la alta variabilidad interindividual [5],[7].

Asimismo, la integración de datos multimodales (CGM, carbohidratos, insulina, actividad, etc.) mejora la predicción, pero requiere técnicas robustas para manejar la irregularidad, dispersión y diferentes frecuencias de muestreo [6],[7]. En este contexto, las técnicas de atención y la descomposición de señales como VMD son útiles para analizar cómo los niveles de glucosa dependen de eventos pasados y para modelar dependencias a largo plazo [5],[6]; sin embargo, la eficiencia computacional es un factor importante para la implementación en tiempo real, y técnicas como Knowledge Distillation o arquitecturas optimizadas son relevantes [6].

Por su parte, los LLMs ofrecen nuevas vías para la interacción, explicación, educación y potencial asistencia en la toma de decisiones [10],[1]. Sin embargo, su aplicación directa en tareas numéricas críticas como la predicción BGL o la dosificación de insulina presenta riesgos significativos de errores y seguridad [1],[8],[10] y exige extrema precaución, validación rigurosa y, preferiblemente, supervisión humana (HITL)[1].

Aunque la arquitectura GlucoPredictAssist (GPA) permanece en fase conceptual y aún no se ha evaluado con usuarios reales, se ha trazado una ruta ético-regulatoria acorde con la normativa mexicana vigente. Al no involucrar sujetos humanos ni datos personales identificables, esta etapa preliminar no requiere la revisión formal de un Comité de Ética en Investigación; sin embargo, se aconseja solicitar una "carta de exención" o "no objeción" al CEI institucional para respaldar la trazabilidad del proyecto y preparar su transición a fases posteriores. De manera complementaria, se han considerado el marco de protección de datos personales (LFPDPPP) y las disposiciones de la NOM-004-SSA3-2012 para expedientes clínicos, así como las regulaciones aplicables a software como dispositivo médico (SaMD) reconocidas por COFEPRIS —en particular la NOM-241-SSA1-2021 y el Suplemento 5.0 de la Farmacopea de México. En caso de implementarse clínicamente, GPA deberá contar con protocolo aprobado por CEI, notificación o autorización sanitaria correspondiente y validaciones técnicas y clínicas conforme a estándares nacionales e internacionales (IMDRF, OMS, FDA).

Discusión

La revisión de la literatura confirma la superioridad de los modelos de aprendizaje profundo (LSTM, Transformers, arquitecturas híbridas) sobre métodos tradicionales para la predicción de BGL, gracias a su capacidad para modelar las complejas dinámicas temporales y no lineales de la glucosa. La personalización (embeddings, fine-tuning) y la integración robusta de datos multimodales (CGM, carbohidratos, insulina, actividad) surgen como factores cruciales para mejorar la precisión predictiva, aunque presentan desafíos técnicos relacionados con la heterogeneidad y dispersión de los datos.

Si bien los LLMs muestran un potencial considerable para mejorar la interacción, explicación y educación del paciente, su aplicación directa en tareas numéricas críticas como la predicción BGL o la sugerencia de dosis de insulina conlleva riesgos significativos de seguridad y precisión. El principal problema identificado es la falta de sistemas que integren de forma fiable y segura predicciones precisas con recomendaciones de dosificación personalizadas y accionables.

La propuesta arquitectónica, GlucoPredict-Assist, busca abordar esta brecha combinando modelos predictivos avanzados, potencialmente Transformer, con un módulo de sugerencia de dosis diseñado con estrictas reglas de seguridad como consideración de IOB, priorización de evitación de hipoglucemia y un enfoque HITL. Se propone el uso de LLMs, específica-

mente con RAG, para tareas de explicación y soporte a la interacción, pero no para la generación directa de recomendaciones médicas críticas, mitigando así los riesgos identificados.

CONCLUSIONES

Este trabajo propone la arquitectura GlucoPredictAssist (GPA), que vincula la predicción personalizada de BGL con la generación de sugerencias de dosificación bajo salvaguardas clínicas (guardrails) y supervisión humana (HITL). La aportación principal es conectar de forma explícita la predicción con un mecanismo de dosificación seguro, trazable y personalizable, operable en entornos reales mediante monitoreo y recalibración. Además, el papel de los LLMs se limita a funciones de explicación y verificación, evitando su uso directo en tareas numéricas críticas. Hasta donde alcanza nuestra revisión, no identificamos propuestas que integren simultáneamente predicción personalizada, dosificación con salvaguardas clínicas y operación con monitoreo y recalibración bajo supervisión humana.

La propuesta permanece en fase conceptual. Como trabajo futuro se requiere validación clínica prospectiva, evaluación de seguridad (p. ej., reducción de hipoglucemias), usabilidad y factores humanos, y pruebas de generalización entre perfiles de pacientes y contextos de uso. Estos pasos definirán su madurez y posible adopción en práctica clínica.

BIBLIOGRAFÍA

[1] Pavon, J. M., Schlientz, D., Maciejewski, M. L., Economou-Zavlanos, N., & Lee, R. H. (2025). Large language models in diabetes management: The need for human and artificial intelligence collaboration. Diabetes Care, 48(2), 182–184.

[2] Healey, E., Tan, A. L. M., Flint, K. L., Ruiz, J. L., & Kohane, I. (2024). A case study on using a large language model to analyze continuous glucose monitoring data. Scientific Reports, 14(1), 1143.

[3] Martinsson, J., Schliep, A., Eliasson, B., & Mogren, O. (2020). Blood glucose prediction with variance estimation using recurrent neural networks. Journal of Healthcare Informatics Research, 4, 1-18.

[4] Rabby, M. F., Tu, Y., Hossen, M. I., Lee, I., Maida, A. S., & Hei, X. (2021). Stacked LSTM based deep recurrent neural network with Kalman smoothing for blood glucose prediction. BMC Medical Informatics and Decision Making, 21, 1-15.

[5] Armandpour, M., Kidd, B., Du, Y., & Huang, J. Z. (2021). Deep personalized glucose level forecasting using attention-based recurrent neural networks. arXiv preprint arXiv:2106.00884v2.

[6] Farahmand, E., Soumma, S. B., Chatrudi, N. T., & Ghasemzadeh, H. (2024). Hybrid attention model using feature decomposition and knowledge distillation for blood glucose forecasting. arXiv preprint arXiv:2403.10703.

[7] Nguyen, P. B. H., Hungele, A., Holl, R. W., & Menden, M. P. (2025). Leveraging pretrained large language model for prognosis of type 2 diabetes using longitudinal medical records [Preprint]. medRxiv.

[8] Hulman, A., Dollerup, O. L., Mortensen, J. F., Fenech, M. E., Norman, K., Støvring, H., & Hansen, T. K. (2023). ChatGPT- versus human-generated answers to frequently asked questions about diabetes: A Turing test-inspired survey among employees of a Danish diabetes center. PLoS ONE, 18(8), e0290773.

[9] Wang, D., Liang, J., Ye, J., Li, J., Li, J., Zhang, Q., ... Zheng, Y. (2024). Enhancement of large language models' performance in diabetes education: Retrieval-augmented generation approach. JMIR Preprints, 58041. doi: 10.2196/preprints.58041.

[10] Mondal, A., & Naskar, A. (2024). Evaluating the effectiveness and safety of large language model in generating type 2 diabetes mellitus management plans: A comparative study with medical experts based on real patient records [Preprint]. medRxiv.

[11] Schaaruup, J. R., Isaksen, A. A., Norman, K., Bjerg, L., & Hulman, A. (2025). Trust in large language model-based solutions in healthcare among people with and without diabetes: A cross-sectional survey from the Health in Central Denmark cohort [Preprint]. medRxiv.