Artificial Intelligence and Machine Learning Challenges in Multimodal Diabetes Data: Wearables, Retinal Imaging, and Model Generalization

Saheed Tijani
Email: tijani.saheed@yahoo.com

Abstract

Background: Diabetes mellitus affects over 500 million people globally. Artificial intelligence (AI) and machine learning (ML) offer transformative potential for prevention, diagnosis, and management. However, integrating multimodal data sources—wearable sensors, continuous glucose monitoring (CGM), and retinal imaging—presents substantial methodological and system-level challenges inadequately addressed in current literature.

Objective: This narrative review critically examines AI/ML challenges in multimodal diabetes care, focusing on data heterogeneity, model generalization, algorithmic bias, interpretability, and deployment constraints in resource-limited settings. We emphasize how these challenges manifest acutely in underserved populations, using African contexts as a stress test for AI system robustness.

Methods: We conducted a comprehensive literature review of 361 unique papers published between 2018 and 2025, drawing from SciSpace, Google Scholar, ArXiv, PubMed, and npj Digital Medicine. We analyzed methodological approaches, data modalities, fusion strategies, and reported limitations across wearable time-series data, retinal imaging, and multimodal integration frameworks.

Results: Current AI/ML approaches demonstrate promising performance on benchmark datasets but exhibit critical weaknesses in generalization across populations, clinical settings, and data acquisition protocols. Deep learning architectures (CNNs, RNNs, LSTMs, Transformers) dominate recent literature, yet reproducibility remains poor due to dataset homogeneity, limited external validation, and inadequate reporting standards. Multimodal fusion strategies show potential for improved diagnostic accuracy but face challenges in handling missing modalities, temporal misalignment, and computational complexity. Bias and fairness issues are pervasive, with models systematically underperforming for female patients, individuals with poor glycemic control, and populations underrepresented in training data. Deployment in low-resource settings is hindered by infrastructure limitations, device costs, data privacy concerns, and lack of clinical validation in diverse populations.

Conclusions: Achieving equitable, robust AI-driven diabetes care requires fundamental shifts in data collection practices, model development paradigms, and evaluation frameworks. Future research must prioritize external validation across diverse populations, standardized reporting of demographic characteristics and performance disparities, federated learning approaches for privacy-preserving collaboration, and explicit consideration of cost, infrastructure, and accessibility constraints. Without addressing these challenges, AI systems risk exacerbating existing health inequalities rather than democratizing precision medicine.

Index Terms

Artificial Intelligence, Machine Learning, Diabetes, Multimodal Data, Wearable Sensors, Continuous Glucose Monitoring, Retinal Imaging, Model Generalization, Bias, Fairness, Low-Resource Settings

I. Introduction

Diabetes mellitus represents one of the most pressing global health challenges of the 21st century, affecting an estimated 537 million adults worldwide as of 2021, with projections suggesting this number will exceed 783 million by 2045 [1]. The disease imposes substantial burdens on healthcare systems, accounting for approximately 12% of global health expenditure, while contributing to significant morbidity and mortality through microvascular and macrovascular complications [2]. Traditional approaches to diabetes management—relying on periodic clinical assessments, self-monitoring of blood glucose, and standardized treatment protocols—struggle to address the heterogeneous nature of the disease and the complex interplay of genetic, behavioral, and environmental factors that influence glycemic control [3].

The convergence of digital health technologies, ubiquitous sensing, and advances in artificial intelligence (AI) and machine learning (ML) has catalyzed a paradigm shift toward data-driven, personalized diabetes care [4], [5]. Wearable devices and continuous glucose monitoring (CGM) systems generate high-resolution time-series data capturing glucose dynamics, physical activity, heart rate variability, and other physiological signals [6]. Retinal imaging modalities—fundus photography and optical coherence tomography (OCT)—enable early detection of diabetic retinopathy, a leading cause of preventable blindness [7]. Electronic health records (EHRs) aggregate longitudinal clinical data, laboratory results, and medication histories. Integrating these multimodal data sources through AI/ML frameworks promises predictive analytics, personalized treatment recommendations, and early intervention for complications [1], [8].

However, the translation of AI/ML research into clinically robust, equitable, and scalable diabetes care systems faces formidable challenges that extend beyond algorithmic performance on benchmark datasets. Data heterogeneity arising from diverse sensor technologies, acquisition protocols, and patient populations complicates model development and validation [9]. Deep learning models, while achieving state-of-the-art performance in controlled settings, often exhibit poor generalization when deployed in new clinical environments or applied to populations underrepresented in training data [10], [11]. Algorithmic bias and fairness concerns are pervasive, with models systematically underperforming for female patients, individuals with suboptimal glycemic control, and racial and ethnic minorities [5], [9]. Interpretability and explainability remain critical barriers to clinical adoption, as healthcare providers require transparent reasoning to trust AI-generated recommendations [4], [11].

These challenges manifest with particular acuity in low-resource settings, where infrastructure limitations, device costs, data scarcity, and workforce constraints create additional barriers to AI deployment [1], [12]. African populations, for instance, face disproportionate diabetes burdens yet remain severely underrepresented in AI research and development [13]. The lack of diverse training data, coupled with models optimized for high-resource clinical environments, results in AI systems that may fail catastrophically when applied to contexts characterized by intermittent connectivity, limited specialist availability, and heterogeneous patient populations [14]. This reality underscores the need for a critical examination of AI/ML methodologies through the lens of accessibility, affordability, and generalization across diverse real-world conditions.

Recent work has begun to address these concerns through frameworks emphasizing inclusive healthcare, integrating education and research with AI and personalized curricula [15]. Such approaches recognize that democratizing AI-driven precision medicine requires not only technical innovation but also capacity building, stakeholder engagement, and explicit consideration of equity from the earliest stages of system design [13]. Federated learning paradigms, which enable collaborative model training without centralizing sensitive patient data, offer promising pathways for privacy-preserving, multi-institutional research [13], [14]. Transfer learning and domain adaptation techniques may facilitate model generalization across populations and clinical settings [7]. Yet these methodological advances remain nascent, with limited evidence of real-world impact in resource-constrained environments.

This narrative review provides a comprehensive, critical synthesis of AI/ML challenges in multimodal diabetes data, with particular emphasis on wearable sensors, retinal imaging, and model generalization. We examine the current landscape of methodological approaches, data modalities, and fusion strategies, while highlighting persistent limitations related to bias, fairness, interpretability, and deployment. By foregrounding the experiences and constraints of low-resource settings—using African populations as a contextual lens—we aim to illuminate the gap between algorithmic promise and clinical reality, and to identify research priorities that can advance equitable, robust, and scalable AI-driven diabetes care.

II. Data Modalities and Problem Formulation

The application of AI/ML to diabetes care draws upon a diverse array of data modalities, each characterized by distinct temporal resolutions, measurement principles, and clinical utility. Understanding the properties, challenges, and complementary nature of these data sources is essential for developing robust multimodal systems.

A. Wearable Sensors and Continuous Glucose Monitoring

Continuous glucose monitoring systems represent a cornerstone of modern diabetes management, providing interstitial glucose measurements at intervals ranging from 1 to 15 minutes [6], [7], [**?**]. CGM data enables the detection of glycemic patterns, prediction of hypo- and hyperglycemic events, and optimization of insulin dosing strategies [4], [**?**]. Recent advances in non-invasive and minimally invasive sensing technologies have expanded the landscape of wearable devices to include smartwatches, fitness trackers, and specialized medical patches that capture heart rate, heart rate variability, physical

activity, sleep patterns, and other physiological signals [6], [16].

The integration of CGM with accelerometry and heart rate data enables context-aware glucose prediction models that account for physical activity, stress, and circadian rhythms [6], [7]. Photoplethysmography (PPG) and electrocardiography (ECG) signals have shown promise for non-invasive glucose estimation and detection of autonomic neuropathy, though clinical validation remains limited [6], [17]. Emerging modalities such as bioimpedance spectroscopy, electromagnetic sensing, and analysis of bodily fluids (sweat, saliva, tears) are under investigation but face challenges related to accuracy, calibration drift, and susceptibility to environmental confounders [6].

Despite their potential, wearable sensor data present significant challenges for AI/ML applications. Temporal resolution and sampling rates vary across devices, complicating data integration and model generalization [9]. Missing data due to sensor failures, user non-adherence, or connectivity issues are pervasive, with rates often exceeding 20% in real-world deployments [4]. Measurement noise, calibration drift, and inter-device variability introduce systematic errors that degrade model performance [6]. Furthermore, the high dimensionality and temporal dependencies of time-series data necessitate specialized architectures such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and temporal convolutional networks, which require substantial computational resources and large training datasets [10], [11].

B. Retinal Imaging Modalities

Diabetic retinopathy (DR) is a leading cause of vision loss among working-age adults, affecting approximately one-third of individuals with diabetes [18]. Early detection through systematic screening enables timely intervention and prevention of irreversible vision impairment. Fundus photography and optical coherence tomography (OCT) are the primary imaging modalities used for DR screening and diagnosis [7].

Fundus photography captures two-dimensional color images of the retina, enabling visualization of microaneurysms, hemorrhages, exudates, and neovascularization characteristic of DR progression [18]. Convolutional neural networks (CNNs) have achieved expert-level performance in automated DR grading from fundus images, with several systems receiving regulatory approval for clinical use [10]. OCT provides high-resolution cross-sectional images of retinal layers, enabling quantification of macular edema and assessment of structural changes associated with diabetic macular edema [7].

However, retinal imaging AI systems face critical challenges that limit their real-world utility. Image quality varies substantially across acquisition devices, camera settings, and operator expertise, with low-quality images accounting for 10-30% of screening datasets [18]. Models trained on high-quality research datasets often fail when applied to images from low-cost portable devices commonly used in resource-limited settings [13]. Ethnic and demographic diversity in training data is limited, with most datasets derived from European and North American populations, leading to performance degradation for African, Asian, and Hispanic patients [5], [14]. The lack of standardized grading protocols and inter-rater variability in reference labels introduce label noise that propagates through model training [18].

C. Electronic Health Records and Clinical Data

Electronic health records aggregate longitudinal patient data including demographics, diagnoses, laboratory results, medications, vital signs, and clinical notes [3], [7]. EHR data enable risk stratification, prediction of complications, and identification of patients who may benefit from intensive interventions [5]. Natural language processing (NLP) techniques can extract structured information from unstructured clinical notes, expanding the scope of available features for predictive modeling [7].

However, EHR data are characterized by high dimensionality, sparsity, irregular sampling, and systematic biases related to healthcare access and documentation practices [3]. Missing data mechanisms are often non-random, with sicker patients having more complete records, complicating imputation strategies [4]. Temporal dependencies and complex interactions between medications, comorbidities, and lifestyle factors challenge traditional statistical approaches, motivating the use of deep learning architectures such as recurrent neural networks and attention-based models [10].

D. Problem Formulation and Modeling Objectives

AI/ML applications in diabetes care encompass a range of prediction and classification tasks, each with distinct data requirements, evaluation metrics, and clinical implications. Common objectives include:

- **Glucose prediction:** Forecasting future glucose levels (typically 30-60 minutes ahead) to enable proactive intervention and prevention of hypo/hyperglycemia [4], [9].
- **Diabetic retinopathy screening:** Automated grading of DR severity from retinal images to enable scalable screening programs [10], [18].
- **Risk prediction:** Estimating probability of diabetes onset, progression to complications, or adverse events [3], [5].
- **Treatment optimization:** Personalized recommendations for insulin dosing, medication selection, or lifestyle interventions [1], [7].

The formulation of these tasks as supervised learning problems requires careful consideration of label definitions, prediction horizons, feature engineering, and evaluation metrics that align with clinical utility [4]. Imbalanced class distributions, particularly for rare complications, necessitate specialized sampling strategies and loss functions [11]. The temporal nature of diabetes progression and treatment response motivates the use of time-series modeling approaches that capture longitudinal patterns and account for time-varying confounders [10].

Table I summarizes the key characteristics, advantages, and challenges of major data modalities used in AI-driven diabetes care.

III. Machine Learning and Deep Learning Approaches

The application of AI/ML to diabetes care has evolved from traditional statistical methods and shallow machine learning algorithms to sophisticated deep learning architectures capable of learning hierarchical representations from high-dimensional, multimodal data. This section reviews the dominant methodological approaches, their strengths and limitations, and emerging trends in model development.

A. Traditional Machine Learning Methods

Early AI applications in diabetes relied on classical machine learning algorithms including logistic regression, support vector machines (SVMs), decision trees, random forests, and gradient boosting methods [7], [11]. These approaches demonstrated competitive performance for structured prediction tasks such as diabetes risk assessment from clinical variables and short-term glucose forecasting from CGM data [3]. Random forests and gradient boosting ensembles, in particular, have shown robust performance across diverse datasets and remain popular for their interpretability, computational efficiency, and ability to handle missing data [11], [19].

Support vector machines with specialized kernels have been applied to glucose prediction and classification tasks, offering theoretical guarantees on generalization performance [7]. Bayesian methods enable probabilistic predictions with uncertainty quantification, a critical requirement for clinical decision support [7]. Fuzzy logic systems have been employed for insulin dosing recommendations, capturing the imprecise reasoning characteristic of clinical expertise [7].

However, traditional ML methods face fundamental limitations when applied to high-dimensional, temporally structured, or multimodal data. Manual feature engineering is labor-intensive and requires domain expertise, limiting scalability and generalization [10]. The inability to learn hierarchical representations from raw sensor data or images necessitates handcrafted features that may not capture subtle patterns relevant to clinical outcomes [11]. These limitations have motivated the widespread adoption of deep learning approaches.

B. Deep Learning Architectures for Time-Series Data

Recurrent neural networks (RNNs) and their variants—long short-term memory (LSTM) networks and gated recurrent units (GRUs)—have become the dominant architectures for glucose prediction from CGM time-series data [9], [10]. LSTMs address the vanishing gradient problem inherent in standard RNNs, enabling learning of long-term temporal dependencies critical for capturing glucose dynamics influenced by meals, insulin, and physical activity occurring hours earlier [6], [11].

Temporal convolutional networks (TCNs) offer an alternative to RNNs, using dilated causal convolutions to capture long-range dependencies while enabling parallel computation and avoiding the sequential bottleneck of recurrent architectures [10]. Attention mechanisms and Transformer architectures, originally developed for natural language processing, have been adapted for time-series forecasting, enabling models to selectively focus on relevant historical time points [11]. These approaches have demonstrated state-of-the-art performance on benchmark datasets, achieving root mean squared errors (RMSE) below 20 mg/dL for 30-minute glucose prediction horizons [9].

Despite impressive benchmark performance, recent reproducibility studies reveal critical weaknesses in current deep learning approaches for glucose prediction [9]. When six representative models from well-cited literature were replicated across three public datasets (OhioT1DM, DiaTrend, T1DEXI) encompassing 128 individuals with type 1 diabetes, results showed good reproducibility when using the same code and evaluation dataset, but poor conceptual reproducibility across datasets with different diabetes management practices [9]. Prediction accuracy was significantly associated with individual glycemic control and sex/gender, with all models exhibiting significantly higher errors for individuals with worse glycemic control and for female subgroups compared to males [9]. These findings underscore fundamental challenges in model generalization and fairness that persist despite architectural sophistication.

C. Convolutional Neural Networks for Retinal Imaging

Convolutional neural networks have revolutionized automated analysis of retinal images, achieving performance comparable to or exceeding that of human experts for diabetic retinopathy grading [10], [18]. Deep CNN architectures such as ResNet, Inception, DenseNet, and EfficientNet have been widely adopted, often leveraging transfer learning from ImageNet pre-training to compensate for limited labeled medical imaging datasets [10].

Ensemble methods combining predictions from multiple CNN architectures have demonstrated improved robustness and calibration compared to single models [18]. Attention mechanisms enable visualization of image regions most influential for model predictions, providing a degree of interpretability valuable for clinical validation and error analysis [10]. Multi-task learning approaches that jointly predict DR severity, diabetic macular edema, and other retinal pathologies have shown promise for improving feature learning and generalization [18].

However, the clinical deployment of CNN-based DR screening systems has revealed significant limitations. Models trained on high-quality research datasets often fail when applied to images from portable, low-cost devices used in community screening programs, with performance degradation of 10-30% in terms of area under the ROC curve [13]. Ethnic and demographic biases are pervasive, with models exhibiting higher false negative rates for African and Asian populations underrepresented in training data [5], [14]. The lack of standardized evaluation protocols and reporting of

TABLE I
COMPARISON OF DATA MODALITIES IN AI-DRIVEN DIABETES CARE

Data Modality	Key Characteristics	Clinical Applications	Major Challenges
Continuous Glucose Monitoring (CGM)	High-frequency time-series (1-15 min intervals); interstitial glucose	Glucose prediction, hypoglycemia detection, insulin optimization	Missing data, sensor noise, calibration drift, inter-device variability, high computational cost
Wearable Sensors (Activity, HR, PPG)	Multi-modal physiological signals; variable sampling rates	Context-aware glucose prediction, activity detection, stress monitoring	Heterogeneous devices, missing data, limited accuracy for glucose estimation, privacy concerns
Fundus Photography	2D color retinal images; variable quality	Diabetic retinopathy screening and grading	Image quality variability, limited ethnic diversity in training data, device heterogeneity, label noise
Optical Coherence Tomography (OCT)	High-resolution 3D retinal structure	Macular edema quantification, structural assessment	High cost, limited availability in low-resource settings, computational complexity
Electronic Health Records (EHR)	Longitudinal clinical data; structured and unstructured	Risk prediction, complication forecasting, treatment optimization	High dimensionality, sparsity, irregular sampling, non-random missingness, documentation bias
Genomic Data	Genetic variants, polygenic risk scores	Risk stratification, precision medicine	High cost, limited clinical utility, ethical concerns, population stratification

demographic subgroup performance obscures these disparities in published literature [5].

D. Hybrid and Ensemble Approaches

Recognizing the complementary strengths of different algorithmic approaches, recent work has explored hybrid architectures that combine multiple model types. CNN-LSTM architectures integrate convolutional layers for spatial feature extraction with recurrent layers for temporal modeling, enabling analysis of sequential retinal images or time-series physiological data [10], [11]. Ensemble methods that combine predictions from diverse base learners (e.g., gradient boosting, neural networks, SVMs) have demonstrated improved robustness and calibration compared to individual models [11], [19].

Gradient boosting methods, particularly XGBoost and LightGBM, have shown superior performance for structured prediction tasks involving clinical and laboratory data, often outperforming deep learning approaches when sample sizes are limited or feature engineering is effective [11], [19]. The interpretability of tree-based ensembles through feature importance scores and partial dependence plots provides valuable insights for clinical validation and hypothesis generation [3].

E. Emerging Paradigms: Transfer Learning and Federated Learning

Transfer learning has emerged as a critical technique for addressing data scarcity and improving generalization in medical AI applications [7], [10], [20]. Pre-training on large-scale datasets (e.g., ImageNet for computer vision, or aggregated CGM data from multiple institutions) followed by fine-tuning on target populations enables models to leverage learned representations while adapting to local data distributions [13], [21]. Domain adaptation techniques that explicitly minimize distribution shift between source and target domains show promise for improving cross-population generalization [14], [20].

Federated learning enables collaborative model training across multiple institutions without centralizing sensitive patient data, addressing privacy concerns and regulatory barriers to data sharing [13], [14]. Recent work on federated multimodal AI for diabetes care demonstrates the feasibility of training models on distributed datasets while preserving privacy and achieving performance comparable to centralized approaches [13]. However, challenges related to heterogeneous data distributions, communication efficiency, and fairness across participating sites remain active areas of research [14].

F. Reinforcement Learning for Treatment Optimization

Reinforcement learning (RL) offers a principled framework for learning optimal treatment policies from observational or experimental data [7]. RL approaches have been applied to insulin dosing optimization, with agents learning to balance glycemic control against the risk of hypoglycemia through interaction with simulated or real patients [7]. Deep reinforcement learning methods combining neural network function approximators with RL algorithms enable learning from high-dimensional state spaces encompassing CGM data, meal information, and activity patterns [10].

Despite theoretical appeal, clinical deployment of RL-based treatment systems faces substantial barriers. The safety-critical nature of diabetes management necessitates extensive validation and fail-safe mechanisms to prevent harmful actions [7]. The exploration-exploitation trade-off inherent in RL raises ethical concerns about exposing patients to suboptimal treatments during learning [1]. Offline RL methods that learn from historical data without active experimentation offer a safer alternative but face challenges related to distributional shift and counterfactual reasoning [7].

IV. Multimodal Learning and Data Fusion

The integration of heterogeneous data sources—combining wearable sensor time-series, retinal images, electronic health records, and genomic data—represents a frontier in AI-driven diabetes care, promising to capture the multifaceted nature of disease progression and treatment response. However, multimodal learning introduces substantial technical and methodological challenges that extend beyond single-modality approaches.

A. Fusion Strategies and Architectures

Multimodal fusion strategies can be broadly categorized into early fusion, late fusion, and hybrid approaches [5], [7]. Early fusion (feature-level fusion) concatenates features from different modalities before feeding them into a unified model, enabling the learning of cross-modal interactions but requiring careful normalization and alignment of heterogeneous feature spaces [13]. Late fusion (decision-level fusion) trains separate models for each modality and combines their predictions through voting, averaging, or meta-learning, offering modularity and robustness to missing modalities but potentially missing important cross-modal dependencies [5].

Hybrid fusion architectures employ intermediate fusion strategies, learning modality-specific representations through specialized sub-networks before integrating them at intermediate layers [13], [14], [22]. Attention-based fusion mechanisms enable models to dynamically weight the contribution of different modalities based on their relevance and reliability for specific predictions [23], [22], [24]. Cross-modal alignment techniques, such as canonical correlation analysis or contrastive learning, learn shared representations that capture complementary information across modalities [13].

Recent work on federated multimodal AI demonstrates the potential of cross-modal co-learning frameworks that leverage complementary information from CGM time-series and retinal imaging for improved risk stratification and complication prediction [13]. These approaches show promise for integrating diverse data sources while preserving privacy through federated learning paradigms [13]. However, the computational complexity and data requirements of multimodal models remain substantial barriers to widespread adoption.

B. Temporal Alignment and Missing Modalities

A critical challenge in multimodal diabetes data is the temporal misalignment of different data sources. CGM provides continuous measurements at minute-scale resolution, while retinal imaging occurs at annual or semi-annual intervals, and laboratory tests follow irregular schedules determined by clinical protocols [3], [7]. Aligning these heterogeneous temporal scales requires sophisticated interpolation, aggregation, or attention mechanisms that can handle irregular sampling and variable time lags [10].

Missing modalities represent a pervasive challenge in real-world deployments. Not all patients have access to CGM devices, retinal imaging may be unavailable in resource-limited settings, and EHR completeness varies substantially across healthcare systems [1], [12]. Models trained on complete multimodal data often fail catastrophically when deployed in settings where only a subset of modalities is available [5]. Robust multimodal architectures must gracefully degrade performance rather than failing completely when modalities are missing, requiring specialized training strategies such as modality dropout or auxiliary reconstruction tasks [13].

C. Cross-Modal Transfer and Domain Adaptation

Transfer learning across modalities offers a pathway for leveraging information from data-rich modalities to improve performance on data-scarce modalities [7], [14]. For example, representations learned from large-scale retinal imaging datasets can be transferred to improve glucose prediction from limited CGM data by capturing shared physiological processes related to vascular health and metabolic dysfunction [13]. Cross-modal domain adaptation techniques that minimize distribution shift between source and target modalities show promise for improving generalization [14].

However, the theoretical foundations and empirical validation of cross-modal transfer in medical applications remain limited. The assumption that different modalities capture related aspects of disease progression may not hold uniformly across populations or clinical contexts [5]. Negative transfer, where knowledge from source modalities degrades performance on target tasks, represents a significant risk that requires careful validation [9].

D. Interpretability and Clinical Integration

The complexity of multimodal models exacerbates interpretability challenges inherent in deep learning approaches. Clinicians require transparent explanations of how different data sources contribute to predictions, particularly when recommendations conflict with clinical intuition or established guidelines [1], [4]. Attention visualization, saliency maps, and feature attribution methods provide partial insights but often fail to capture the complex interactions between modalities that drive model predictions [10].

The integration of multimodal AI systems into clinical workflows requires careful consideration of data availability, acquisition costs, and workflow disruption [1], [2]. Requiring multiple data modalities for model inference may create barriers to adoption if some modalities are expensive, invasive, or time-consuming to acquire [12]. Hierarchical decision support systems that provide increasingly refined predictions as additional modalities become available offer a pragmatic approach to balancing performance and accessibility [5].

E. Evaluation Challenges and Reporting Standards

Evaluating multimodal models presents unique challenges beyond single-modality approaches. Performance metrics must capture not only overall accuracy but also robustness to missing modalities, fairness across demographic subgroups, and calibration of uncertainty estimates [5], [9]. Ablation studies that systematically remove individual modalities or combinations thereof are essential for understanding the contribution of each data source and identifying potential redundancies [4].

Current literature exhibits substantial heterogeneity in evaluation protocols, with inconsistent reporting of data preprocessing, train-test splits, hyperparameter selection, and demographic characteristics of study populations [4], [9]. The lack of standardized benchmarks and public multimodal datasets hinders reproducibility and comparison across studies [9]. Recent efforts to establish reporting guidelines and reproducibility standards represent important steps toward more rigorous evaluation practices [4], [9].

Figure 1 illustrates a conceptual framework for multimodal AI in diabetes care, highlighting the integration of diverse data sources, fusion strategies, and clinical decision support outputs.

V. Model Limitations: Bias, Generalization, and Interpretability

Despite impressive performance on benchmark datasets, AI/ML models for diabetes care exhibit critical limitations that impede clinical translation and risk exacerbating health inequalities. This section examines three interrelated challenges—bias and fairness, generalization and robustness, and interpretability—that represent fundamental barriers to equitable, reliable AI-driven healthcare.

A. Algorithmic Bias and Fairness

Algorithmic bias in diabetes AI systems manifests through systematic performance disparities across demographic subgroups defined by sex, race, ethnicity, age, socioeconomic status, and clinical characteristics [5], [9]. Recent reproducibility studies reveal pervasive gender bias in glucose prediction models, with all evaluated deep learning architectures exhibiting significantly higher prediction errors for female patients compared to males [9]. These disparities persist across multiple datasets and model architectures, suggesting fundamental issues in data collection, feature engineering, or model optimization rather than isolated implementation errors [9].

Racial and ethnic biases are particularly pronounced in retinal imaging systems for diabetic retinopathy screening. Models trained predominantly on European and North American populations exhibit elevated false negative rates when applied to African, Asian, and Hispanic patients, potentially delaying diagnosis and treatment of vision-threatening complications [5], [14]. The underrepresentation of diverse populations in training datasets, coupled with physiological differences in retinal pigmentation and disease presentation, contributes to these disparities [13]. However, the lack of standardized reporting of demographic characteristics and subgroup performance in published literature obscures the true extent of these biases [5].

Performance disparities based on glycemic control represent another critical fairness concern. Models systematically underperform for individuals with poor glycemic control—precisely the population most likely to benefit from AI-assisted interventions [9]. This pattern suggests that models may be optimizing for average performance on well-controlled patients rather than robust performance across the full spectrum of disease severity [9]. The clinical implications are profound: AI systems that fail for high-risk patients may widen rather than narrow existing disparities in diabetes outcomes.

The sources of algorithmic bias are multifaceted, encompassing data collection practices, label definitions, feature selection, model architecture choices, and optimization objectives [5], [14]. Historical biases in healthcare access and documentation propagate through EHR data, with underserved populations having sparser, lower-quality records that degrade model performance [3]. Imbalanced training datasets that overrepresent majority populations lead to models that prioritize their performance at the expense of minorities [14]. Evaluation metrics that emphasize overall accuracy without considering fairness constraints enable models to achieve high aggregate performance while exhibiting substantial subgroup disparities [5].

Addressing algorithmic bias requires interventions at multiple stages of the model development pipeline. Data collection efforts must prioritize diversity and representativeness, with explicit targets for inclusion of underrepresented populations [13], [14]. Fairness-aware learning algorithms that incorporate demographic parity, equalized odds, or other fairness constraints into optimization objectives show promise but face trade-offs between fairness and overall accuracy [14]. Post-hoc calibration and threshold adjustment can mitigate some disparities but do not address underlying model limitations [5]. Critically, fairness audits and mandatory reporting of subgroup performance must become standard practice in AI research and regulatory approval processes [5], [14].

B. Generalization and Dataset Shift

The ability of AI models to generalize beyond their training distribution represents a fundamental challenge in medical applications, where deployment settings often differ substantially from development environments [9], [10]. Dataset shift—encompassing covariate shift, label shift, and concept drift—arises from differences in patient populations, clinical protocols, device characteristics, and temporal trends [9].

Recent reproducibility studies provide compelling evidence of poor generalization in glucose prediction models [9]. When six representative deep learning architectures were evaluated across three public datasets with different diabetes management practices, conceptual reproducibility was poor despite good reproducibility when using identical code and evaluation datasets [9]. Performance degradation of 20-40% in terms of root mean squared error was observed when models trained on one dataset were applied to others, even though all datasets comprised individuals with type 1 diabetes using similar CGM devices [9].

The sources of generalization failure are diverse. Device heterogeneity introduces systematic differences in measurement characteristics, with CGM sensors from different manufacturers exhibiting distinct noise profiles, calibration algorithms, and temporal resolutions [6], [9]. Clinical protocol variations—including insulin regimens, meal patterns, and physical activity levels—create distribution shifts that models fail to

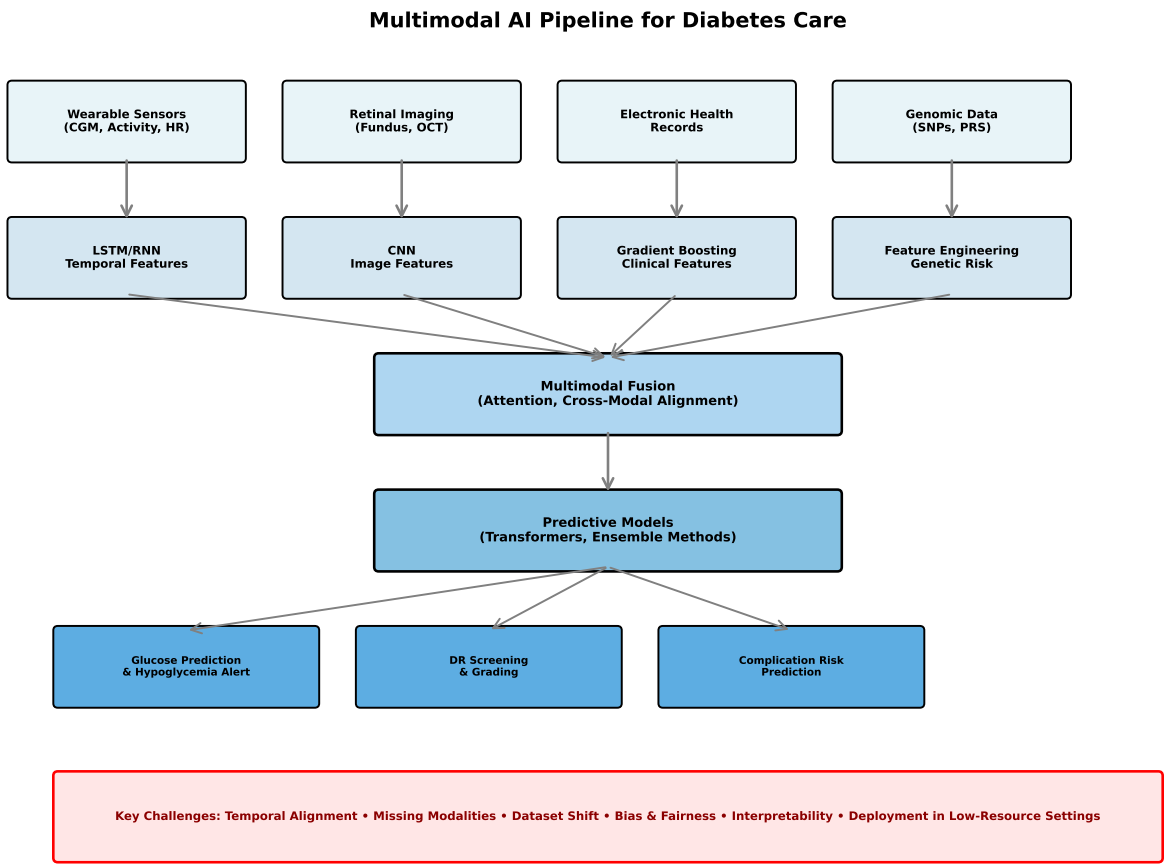

Fig. 1. Conceptual framework for multimodal AI in diabetes care. The pipeline integrates heterogeneous data sources (wearable sensors, retinal imaging, EHR, genomic data) through modality-specific feature extraction, cross-modal fusion, and predictive modeling to support clinical decision-making. Key challenges include temporal alignment, missing modalities, interpretability, and generalization across populations and clinical settings.

accommodate [9]. Temporal drift in patient behavior, disease progression, and treatment practices over time degrades model performance even within the same population [4].

Retinal imaging systems face analogous generalization challenges. Models trained on high-quality images from research-grade fundus cameras exhibit substantial performance degradation when applied to images from portable, low-cost devices used in community screening programs [13]. Image quality variations related to lighting conditions, pupil dilation, operator expertise, and patient cooperation introduce distribution shifts that models struggle to handle [18]. Geographic and demographic differences in disease prevalence and presentation further complicate generalization [5].

The limited availability of external validation studies represents a critical gap in current literature. Most published models are evaluated only on held-out test sets from the same distribution as training data, providing optimistic estimates of real-world performance [4], [9]. The few studies that conduct external validation on independent datasets consistently report substantial performance degradation, highlighting the fragility of current approaches [5], [9]. The overreliance on a small number of public datasets (e.g., OhioT1DM for glucose prediction, EyePACS for diabetic retinopathy) further limits the diversity of evaluation scenarios and may lead to overfitting to dataset-specific characteristics [9].

Improving generalization requires fundamental shifts in model development and evaluation practices. Domain adaptation and transfer learning techniques that explicitly account for distribution shift show promise but remain underutilized in diabetes applications [7], [14]. Multi-site training on diverse datasets can improve robustness but requires addressing data sharing barriers and harmonization challenges [13]. Federated learning enables collaborative model development without centralizing data, offering a pathway for training on heterogeneous populations while preserving privacy [13], [14]. Critically, external validation on multiple independent datasets from diverse clinical settings must become a prerequisite for publication and regulatory approval [5], [9].

C. Interpretability and Explainability

The opacity of deep learning models represents a fundamental barrier to clinical adoption, as healthcare providers require transparent reasoning to trust AI-generated recommendations

and integrate them into decision-making processes [1], [4]. The "black box" nature of neural networks, particularly deep architectures with millions of parameters, makes it difficult to understand why specific predictions are made or to identify potential failure modes [10], [11].

Interpretability challenges are particularly acute in safety-critical applications such as insulin dosing recommendations, where erroneous predictions can lead to life-threatening hypoglycemia or hyperglycemia [1], [7]. Clinicians need to understand not only what the model predicts but also the confidence of predictions, the key factors driving recommendations, and how predictions align with or diverge from clinical guidelines [4]. The lack of interpretability impedes error analysis, limits opportunities for model improvement, and raises liability concerns when adverse events occur [1].

Various approaches to post-hoc interpretability have been developed, including feature importance scores, saliency maps, attention visualization, and counterfactual explanations [10], [11]. Feature importance methods such as SHAP (SHapley Additive exPlanations) quantify the contribution of individual input features to model predictions, providing insights into which clinical variables or sensor measurements drive decisions [3]. Attention mechanisms in Transformer architectures enable visualization of which time points or image regions the model focuses on, offering partial transparency into the reasoning process [11].

However, these interpretability methods have significant limitations. Feature importance scores may be unstable across similar inputs or misleading when features are correlated [4]. Saliency maps for image models often highlight spurious patterns or artifacts rather than clinically meaningful features [10]. Attention weights do not necessarily correspond to causal relationships or clinical reasoning [11]. Counterfactual explanations that describe how inputs would need to change to alter predictions may suggest clinically infeasible or harmful interventions [1].

Intrinsically interpretable models offer an alternative to post-hoc explanation methods. Decision trees, rule-based systems, and linear models provide transparent decision logic but sacrifice predictive performance on complex tasks [7], [11]. Hybrid approaches that combine interpretable models with deep learning components aim to balance transparency and accuracy but face challenges in maintaining coherent explanations across model components [10].

The tension between interpretability and performance represents a fundamental trade-off in AI system design. Highly accurate deep learning models may be opaque, while transparent models may lack the capacity to capture complex patterns necessary for clinical utility [4], [10]. The appropriate balance depends on the specific application, with safety-critical decisions requiring greater transparency than screening or risk stratification tasks [1].

Regulatory frameworks increasingly emphasize the importance of interpretability and explainability for medical AI systems. The European Union's AI Act and FDA guidance on clinical decision support software highlight transparency, validation, and human oversight as key requirements [1], [8], [**?**]. However, the lack of standardized metrics for interpretability and consensus on what constitutes adequate explanation complicates regulatory evaluation [4].

D. Reproducibility and Reporting Standards

The reproducibility crisis in AI research extends to diabetes applications, with substantial heterogeneity in methodological rigor, evaluation protocols, and reporting practices [4], [9]. A recent systematic assessment of 60 deep learning papers for glucose prediction found that code availability, overreliance on single public datasets, and limited use of multiple datasets for evaluation are among the top challenges to reproducibility [9].

Inconsistent reporting of data preprocessing, feature engineering, hyperparameter selection, and model architecture details impedes replication and comparison across studies [4], [9]. The lack of standardized evaluation metrics and protocols leads to incomparable results, with studies using different prediction horizons, error metrics, and statistical tests [4]. Demographic characteristics of study populations are often unreported or incompletely described, obscuring potential biases and limiting generalizability assessment [5], [9].

Addressing reproducibility challenges requires adoption of reporting standards such as TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) and CONSORT-AI (Consolidated Standards of Reporting Trials—Artificial Intelligence) [5]. Mandatory code and data sharing, where feasible given privacy constraints, enables independent validation and accelerates scientific progress [9]. Standardized benchmark datasets with well-defined train-test splits and evaluation protocols facilitate fair comparison across methods [4], [9].

VI. Deployment and Accessibility Challenges in Low-Resource Settings

The translation of AI/ML research into clinically deployed systems faces formidable barriers that extend beyond algorithmic performance, encompassing infrastructure limitations, economic constraints, regulatory hurdles, and socio-cultural factors. These challenges manifest with particular acuity in low-resource settings, where the potential impact of AI-driven diabetes care is greatest but the obstacles to implementation are most severe. This section examines deployment challenges through the lens of African populations and other underserved communities, using these contexts as stress tests for AI system robustness and accessibility.

A. Infrastructure and Connectivity Constraints

The deployment of AI-driven diabetes care systems assumes reliable access to electricity, internet connectivity, and computing infrastructure—assumptions that do not hold in many low-resource settings [1], [12]. Intermittent power supply disrupts continuous glucose monitoring, data transmission, and cloud-based analytics, creating gaps in patient monitoring and degrading model performance [12]. Limited internet bandwidth and high data costs constrain the feasibility of real-time

data synchronization and cloud-based inference, necessitating edge computing solutions that can operate with intermittent connectivity [14].

The lack of robust health information technology infrastructure in many African countries impedes the adoption of EHR systems and integration of AI tools into clinical workflows [13]. Paper-based medical records remain prevalent, limiting the availability of structured longitudinal data necessary for training and deploying predictive models [12]. The absence of unique patient identifiers and standardized data formats complicates data linkage across healthcare encounters and institutions [1].

Mobile health (mHealth) technologies offer a potential pathway for overcoming infrastructure limitations, leveraging the widespread availability of mobile phones even in resource-constrained settings [16], [12], [**?**]. Smartphone-based applications can provide decision support, patient education, and remote monitoring without requiring extensive clinical infrastructure [11], [**?**]. However, the digital divide persists, with smartphone ownership, literacy, and technical proficiency varying substantially across socioeconomic strata [1].

B. Device Costs and Economic Barriers

The high cost of wearable sensors, continuous glucose monitoring devices, and retinal imaging equipment represents a fundamental barrier to AI-driven diabetes care in low-resource settings [1], [12]. CGM systems, which cost $200-400 per month in high-income countries, are prohibitively expensive for most individuals in sub-Saharan Africa, where per capita health expenditure is often below $100 annually [12]. Even when devices are available, the ongoing costs of sensors, calibration supplies, and data plans create unsustainable financial burdens [6].

Retinal imaging equipment faces similar economic constraints. Research-grade fundus cameras cost $20,000-50,000, placing them beyond the reach of most primary care facilities in low-resource settings [13]. Portable, low-cost imaging devices ($1,000-5,000) offer a more accessible alternative but often produce lower-quality images that degrade AI model performance, as discussed in Section VI [13]. The lack of trained ophthalmologists and imaging technicians further limits its screening capacity, even when equipment is available [18].

The economic sustainability of AI-driven interventions requires careful consideration of cost-effectiveness and return on investment. While AI systems may reduce long-term complications and healthcare costs, the upfront investment in devices, infrastructure, and training may be prohibitive for resource-constrained health systems [1], [2]. Innovative financing mechanisms, including public-private partnerships, tiered pricing models, and technology transfer initiatives, are necessary to improve accessibility [12].

C. Data Scarcity and Population Representativeness

The scarcity of labeled training data from low-resource settings represents a critical barrier to developing AI models that generalize to these populations [13], [14]. Most publicly available datasets for diabetes AI research are derived from high-income countries, with African populations severely underrepresented [5]. The lack of diverse training data leads to models optimized for populations with different genetic backgrounds, dietary patterns, healthcare access, and disease presentations [14].

Genetic and phenotypic differences across populations may influence the performance of AI models. For example, the prevalence and progression of diabetic retinopathy vary across ethnic groups, with some studies suggesting higher rates of vision-threatening complications in African populations [13]. Dietary patterns, physical activity levels, and cultural practices related to diabetes management differ substantially across contexts, potentially limiting the transferability of models trained on Western populations [12].

The challenges of data collection in low-resource settings are multifaceted. Limited research funding, weak institutional capacity, and competing health priorities constrain the feasibility of large-scale data collection efforts [1]. Ethical concerns related to data ownership, consent, and potential exploitation of vulnerable populations require careful navigation [14]. The lack of standardized data collection protocols and quality assurance mechanisms may result in noisy or incomplete datasets that degrade model performance [12].

Transfer learning and domain adaptation offer potential pathways for leveraging data from high-resource settings to improve model performance in low-resource contexts [7], [14]. However, the effectiveness of these approaches depends on the degree of similarity between source and target populations, and negative transfer remains a significant risk [9]. Federated learning enables collaborative model development across institutions and countries without centralizing data, addressing privacy concerns while improving population diversity [13], [14].

D. Clinical Validation and Regulatory Frameworks

The clinical validation of AI systems in low-resource settings faces unique challenges related to study design, outcome measurement, and regulatory oversight [1], [25]. Randomized controlled trials, considered the gold standard for evaluating clinical interventions, are expensive and logistically complex, particularly in settings with limited research infrastructure [14]. Pragmatic trials and implementation science approaches that evaluate real-world effectiveness may be more feasible but face challenges in controlling for confounding factors and establishing causal relationships [1].

Regulatory frameworks for medical AI vary substantially across countries, with many low- and middle-income countries lacking specific guidelines for AI-based medical devices [1], [8]. The absence of clear regulatory pathways creates uncertainty for developers and may delay or prevent the introduction of beneficial technologies [1]. Conversely, overly stringent regulations designed for high-resource settings may be inappropriate for low-resource contexts, where the risk-benefit calculus differs [12].

The generalizability of clinical validation studies conducted in high-resource settings to low-resource populations is questionable. Differences in disease prevalence, comorbidity patterns, healthcare infrastructure, and patient populations limit the transferability of evidence [14], [25]. Local validation studies are necessary but face challenges related to sample size, follow-up duration, and outcome measurement [1].

E. Workforce Capacity and Training

The successful deployment of AI systems requires a healthcare workforce with the technical skills, clinical knowledge, and cultural competency to integrate these tools into practice [1], [2]. The shortage of diabetes specialists, endocrinologists, and ophthalmologists in many low-resource settings limits the capacity for clinical oversight and interpretation of AI-generated recommendations [2], [12]. Primary care providers, who deliver most diabetes care in these contexts, may lack the training and confidence to use AI tools effectively [1].

Clinician hesitancy and resistance to AI adoption represent significant barriers, rooted in concerns about accuracy, liability, workflow disruption, and deskilling [1], [2]. Building trust in AI systems requires transparent communication about model capabilities and limitations, opportunities for hands-on training, and evidence of clinical benefit [1], [4]. Participatory design approaches that involve end-users in system development can improve usability and acceptance [1].

The integration of AI education into medical and nursing curricula represents a long-term strategy for building workforce capacity [**?**]. Recent frameworks emphasizing inclusive healthcare through the integration of education and research with AI and personalized curricula highlight the importance of democratizing AI knowledge and skills [**?**]. Such approaches recognize that achieving equitable AI-driven precision medicine requires not only technical innovation but also capacity building, stakeholder engagement, and explicit consideration of equity from the earliest stages of system design [**?**], [13].

F. Ethical and Socio-Cultural Considerations

The deployment of AI in low-resource settings raises ethical concerns related to autonomy, justice, privacy, and potential exploitation [1], [14]. The risk of algorithmic bias and discrimination, discussed in Section VI, is particularly acute for populations already experiencing health inequities [5], [14]. The use of AI systems developed and validated in high-resource settings without adequate local validation may constitute a form of technological colonialism, imposing solutions that do not address local needs or priorities [12].

Data privacy and security concerns are heightened in contexts with weak regulatory frameworks and limited institutional capacity for data protection [1], [8]. The collection and use of sensitive health data by commercial entities raise questions about data ownership, consent, and potential misuse [14]. Community engagement and participatory governance models that involve patients, healthcare providers, and local stakeholders in decision-making can help ensure that AI systems align with community values and priorities [1].

Cultural beliefs and practices related to diabetes management may influence the acceptability and effectiveness of AI interventions [12]. For example, dietary recommendations generated by AI systems must account for local food availability, cultural preferences, and religious practices [12]. Language barriers and health literacy limitations require careful attention to user interface design and communication strategies [1].

G. Pathways to Equitable AI Deployment

Achieving equitable AI-driven diabetes care in low-resource settings requires multi-faceted strategies that address technical, economic, regulatory, and social barriers. Key priorities include:

- **Technology adaptation:** Development of low-cost, robust devices and algorithms optimized for resource-constrained environments, including edge computing solutions that operate with intermittent connectivity [12], [14].
- **Capacity building:** Investment in workforce training, research infrastructure, and institutional capacity to support local AI development and validation [1], [15].
- **Collaborative research:** Federated learning and multi-site partnerships that enable knowledge sharing while preserving data privacy and respecting local autonomy [13], [14].
- **Inclusive design:** Participatory approaches that involve end-users and communities in system development to ensure cultural appropriateness and alignment with local needs [1], [15].
- **Policy innovation:** Development of regulatory frameworks and financing mechanisms tailored to low-resource contexts, balancing safety with accessibility [1], [8].

Without explicit attention to these priorities, AI systems risk exacerbating existing health inequalities, concentrating benefits in high-resource settings while leaving underserved populations further behind [5], [14].

VII. Open Research Gaps and Future Directions

Despite substantial progress in AI/ML applications for diabetes care, critical research gaps persist across methodological, clinical, and implementation domains. Addressing these gaps is essential for translating algorithmic advances into equitable, robust, and clinically impactful systems. This section identifies priority areas for future research and outlines pathways toward more effective AI-driven diabetes care.

A. Methodological Advances

1) Robust and Generalizable Models: The poor generalization of current AI models across populations, devices, and clinical settings represents a fundamental challenge requiring new methodological approaches [5], [9]. Domain adaptation techniques that explicitly model and minimize distribution shift between training and deployment environments show promise but remain underexplored in diabetes applications

[14]. Meta-learning approaches that train models to rapidly adapt to new populations or devices with limited data could improve generalization while reducing data requirements [10].

Causal inference methods offer a pathway for learning robust relationships that generalize beyond observational correlations [3]. Incorporating causal structure into model architectures and training objectives may improve robustness to confounding and enable more reliable predictions under distribution shift [4]. However, the application of causal methods to high-dimensional, multimodal diabetes data remains an open research challenge.

2) Fairness-Aware Learning: Addressing algorithmic bias requires moving beyond post-hoc fairness audits to incorporate fairness constraints directly into model training [5], [14]. Fairness-aware learning algorithms that optimize for demographic parity, equalized odds, or other fairness metrics while maintaining predictive performance represent an active area of research [14]. However, the choice of fairness metric involves normative judgments about which disparities are acceptable, requiring engagement with ethicists, clinicians, and affected communities [1].

Intersectional fairness, which considers the joint effects of multiple demographic attributes (e.g., race and gender), remains largely unexplored in diabetes AI [5]. Models that perform well on average for each demographic group may still exhibit substantial disparities for intersectional subgroups [14]. Developing methods that ensure fairness across the full space of demographic combinations while maintaining statistical power represents a significant methodological challenge.

3) Uncertainty Quantification and Calibration: Clinical decision support systems require not only accurate predictions but also well-calibrated uncertainty estimates that enable appropriate risk-benefit assessments [1], [4]. Bayesian deep learning, ensemble methods, and conformal prediction offer approaches for quantifying predictive uncertainty, but their application to multimodal diabetes data remains limited [7]. Calibration—ensuring that predicted probabilities match empirical frequencies—is often poor in deep learning models, particularly for minority classes and out-of-distribution inputs [5].

Developing methods for reliable uncertainty quantification under distribution shift is critical for safe deployment in diverse clinical settings [9]. Models should provide conservative uncertainty estimates when applied to populations or contexts that differ from training data, enabling clinicians to recognize when predictions may be unreliable [4].

4) Interpretable and Explainable AI: Advancing interpretability requires moving beyond post-hoc explanation methods to develop intrinsically interpretable architectures that maintain competitive predictive performance [4], [10]. Neural additive models, concept bottleneck models, and prototype-based approaches offer promising directions but have seen limited application in diabetes care [11]. Hybrid architectures that combine interpretable components with deep learning modules may balance transparency and accuracy [10].

Developing standardized metrics and evaluation frameworks for interpretability is essential for comparing approaches and establishing regulatory requirements [1]. Human-centered evaluation studies that assess whether explanations improve clinical decision-making, trust, and patient outcomes are needed but remain rare [4].

B. Data and Infrastructure

1) Diverse and Representative Datasets: The lack of diverse, representative datasets represents a critical barrier to developing equitable AI systems [5], [14]. Coordinated efforts to collect large-scale, multi-site datasets encompassing diverse populations, clinical settings, and device types are essential [9], [13]. Such efforts require addressing data sharing barriers related to privacy, consent, intellectual property, and institutional policies [1].

Federated learning and privacy-preserving computation techniques enable collaborative data analysis without centralizing sensitive information, offering a pathway for multi-institutional research while respecting privacy constraints [13], [14]. However, technical challenges related to communication efficiency, heterogeneous data distributions, and fairness across participating sites require further research [14].

2) Standardized Benchmarks and Evaluation Protocols: The lack of standardized benchmarks and evaluation protocols impedes reproducibility and comparison across studies [4], [9]. Establishing community-endorsed benchmark datasets with well-defined train-test splits, evaluation metrics, and reporting requirements would facilitate rigorous comparison of methods [9]. Such benchmarks should encompass diverse populations, clinical settings, and data modalities to enable comprehensive assessment of generalization and fairness [5].

Evaluation protocols must extend beyond aggregate performance metrics to include subgroup analyses, calibration assessment, robustness to missing data, and computational efficiency [4], [9]. Mandatory reporting of demographic characteristics, data preprocessing details, and hyperparameter selection procedures is essential for reproducibility [9].

C. Clinical Translation and Implementation

1) Prospective Clinical Trials: The majority of published AI studies for diabetes care rely on retrospective analyses of existing datasets, providing limited evidence of real-world clinical benefit [1], [25]. Prospective randomized controlled trials evaluating the impact of AI interventions on patient outcomes, healthcare utilization, and cost-effectiveness are essential for establishing clinical utility [1], [14]. Such trials must be conducted in diverse clinical settings, including low-resource environments, to assess generalizability [12].

Pragmatic trial designs that evaluate effectiveness in routine clinical practice, rather than efficacy under idealized conditions, are particularly valuable for informing implementation decisions [1]. Implementation science frameworks that examine barriers and facilitators to adoption, workflow integration, and sustainability can guide scale-up efforts [2].

2) Human-AI Collaboration: The optimal division of labor between AI systems and human clinicians remains poorly understood [1], [4]. Research on human-AI collaboration should examine how to design interfaces and workflows that leverage the complementary strengths of algorithms (pattern recognition, data integration) and humans (contextual reasoning, ethical judgment) [1]. Understanding when clinicians should override AI recommendations, how to present uncertainty information, and how to maintain clinical skills in the presence of decision support are critical questions [4].

3) Health Economics and Cost-Effectiveness: Rigorous health economic evaluations of AI interventions are needed to inform resource allocation decisions, particularly in resource-constrained settings [1], [2]. Cost-effectiveness analyses must account for upfront investment in devices and infrastructure, ongoing operational costs, and long-term savings from prevented complications [12]. Budget impact analyses that consider affordability and financial sustainability are essential for low-resource contexts [1].

D. Regulatory and Policy Innovation

1) Adaptive Regulatory Frameworks: Current regulatory frameworks for medical devices were designed for static, hardware-based technologies and struggle to accommodate continuously learning AI systems that evolve post-deployment [1], [8]. Adaptive regulatory approaches that enable iterative model updates while maintaining safety and effectiveness oversight are needed [1]. Risk-based frameworks that tailor regulatory requirements to the clinical impact and autonomy of AI systems may balance innovation with patient protection [8].

International harmonization of regulatory standards could reduce duplication of effort and accelerate global access to beneficial technologies [1]. However, context-specific considerations related to infrastructure, workforce capacity, and population characteristics may necessitate locally adapted requirements [12].

2) Data Governance and Privacy: Developing governance frameworks that enable data sharing for research and model development while protecting patient privacy and autonomy is essential [1], [14]. Federated learning, differential privacy, and secure multi-party computation offer technical approaches to privacy-preserving collaboration, but legal and institutional barriers remain [13], [14]. Community-based participatory governance models that involve patients and communities in decisions about data use may improve trust and alignment with stakeholder values [1].

E. Education and Capacity Building

Integrating AI education into medical, nursing, and public health curricula is essential for building workforce capacity to develop, evaluate, and deploy AI systems [1], [15]. Recent frameworks emphasizing inclusive healthcare through the integration of education and research with AI and personalized curricula provide valuable models for democratizing AI knowledge [15]. Such approaches recognize that achieving equitable AI-driven precision medicine requires not only technical innovation but also capacity building, stakeholder engagement, and explicit consideration of equity from the earliest stages of system design [13], [15].

Capacity building efforts must extend beyond high-income countries to include low- and middle-income settings, where the need for AI-driven healthcare solutions is greatest but technical expertise is most limited [12], [14]. International partnerships, technology transfer initiatives, and investment in research infrastructure can support local AI development and validation [1].

F. Emerging Technologies and Paradigms

1) Digital Twins and Personalized Simulation: Digital twin technology, which creates virtual representations of individual patients that can be used for predictive modeling and therapeutic optimization, represents an emerging paradigm in diabetes care [25]. By integrating continuous glucose monitoring, wearable sensors, and machine learning algorithms, digital twins enable personalized simulation of treatment responses and optimization of insulin dosing [25]. However, challenges related to clinical validation, implementation feasibility, and generalizability to underserved populations require further research [25].

2) Large Language Models and Conversational AI: Large language models (LLMs) and conversational AI systems offer potential for patient education, self-management support, and clinical decision assistance [1], [26]. These technologies could provide personalized, accessible health information and coaching, particularly valuable in settings with limited healthcare workforce capacity [12], [**?**]. However, concerns about accuracy, bias, privacy, and the potential for harmful recommendations necessitate careful validation and oversight [1].

3) Wearable and Implantable Sensors: Advances in sensor technology, including non-invasive glucose monitoring, multi-analyte sensing, and implantable devices, promise to expand the scope and accessibility of continuous monitoring [6], [16]. Integration of these sensors with AI algorithms for real-time analysis and intervention could enable closed-loop systems that automatically adjust insulin delivery or provide just-in-time behavioral interventions [1]. However, challenges related to accuracy, calibration, biocompatibility, and cost must be addressed [6].

VIII. Conclusion

Artificial intelligence and machine learning hold transformative potential for diabetes care, offering pathways to personalized treatment, early complication detection, and improved clinical outcomes through the integration of multimodal data from wearable sensors, retinal imaging, and electronic health records. The rapid proliferation of deep learning architectures—including convolutional neural networks for retinal image analysis, recurrent neural networks for glucose prediction, and attention-based models for multimodal fusion—has

demonstrated impressive performance on benchmark datasets and generated substantial enthusiasm for AI-driven precision medicine.

However, this review reveals a critical gap between algorithmic promise and clinical reality. Current AI systems exhibit fundamental weaknesses in generalization across populations, clinical settings, and data acquisition protocols, with performance degradation of 20-40% commonly observed when models are applied to new datasets or deployment environments. Algorithmic bias is pervasive, with models systematically underperforming for female patients, individuals with poor glycemic control, and racial and ethnic minorities underrepresented in training data. These disparities are not merely technical artifacts but reflect deeper issues in data collection practices, model development paradigms, and evaluation frameworks that prioritize aggregate performance over equity and robustness.

The challenges of deploying AI systems in low-resource settings—where the potential impact is greatest but obstacles are most severe—illuminate the limitations of current approaches. Infrastructure constraints, device costs, data scarcity, workforce capacity gaps, and regulatory uncertainties create formidable barriers to implementation. African populations and other underserved communities remain severely underrepresented in AI research and development, resulting in systems optimized for high-resource clinical environments that may fail catastrophically when applied to contexts characterized by intermittent connectivity, limited specialist availability, and heterogeneous patient populations.

Achieving equitable, robust AI-driven diabetes care requires fundamental shifts in research priorities and practices. Methodological advances must prioritize generalization and fairness over benchmark performance, incorporating domain adaptation, causal inference, and fairness-aware learning into model development. Data collection efforts must emphasize diversity and representativeness, with explicit targets for inclusion of underrepresented populations and validation in diverse clinical settings. Evaluation frameworks must extend beyond aggregate metrics to include subgroup analyses, external validation, calibration assessment, and reproducibility standards.

The path forward demands multi-stakeholder collaboration encompassing researchers, clinicians, patients, policymakers, and industry partners. Federated learning and privacy-preserving computation techniques offer pathways for collaborative model development without centralizing sensitive data, enabling multi-institutional research while respecting privacy constraints. Transfer learning and domain adaptation can leverage knowledge from data-rich settings to improve performance in data-scarce contexts, though careful validation is essential to avoid negative transfer. Capacity building initiatives, including integration of AI education into medical curricula and investment in research infrastructure in low-resource settings, are critical for democratizing AI knowledge and ensuring local ownership of technology development.

Regulatory frameworks must evolve to accommodate the unique characteristics of AI systems while maintaining patient safety and effectiveness oversight. Adaptive approaches that enable iterative model updates, risk-based requirements tailored to clinical impact, and international harmonization of standards can balance innovation with protection. Data governance frameworks that enable sharing for research while protecting privacy and autonomy are essential, with community-based participatory models offering pathways for aligning data use with stakeholder values.

Ultimately, the success of AI in diabetes care will be measured not by algorithmic sophistication or benchmark performance, but by its ability to improve health outcomes equitably across diverse populations and clinical settings. Without explicit attention to generalization, fairness, interpretability, and accessibility, AI systems risk exacerbating existing health inequalities, concentrating benefits in high-resource settings while leaving underserved populations further behind. Frameworks emphasizing inclusive healthcare through integration of education and research with AI and personalized curricula provide valuable models for democratizing precision medicine and ensuring that technological advances serve the needs of all individuals affected by diabetes [15].

The challenges outlined in this review are substantial but not insurmountable. By prioritizing equity and robustness alongside performance, engaging diverse stakeholders in system design and evaluation, and investing in capacity building and infrastructure, the research community can work toward AI-driven diabetes care that is not only technically sophisticated but also clinically impactful, ethically sound, and accessible to all who need it. The next generation of diabetes AI must be built on foundations of diversity, transparency, and accountability, with explicit recognition that algorithmic fairness and clinical utility are not competing objectives but essential prerequisites for realizing the transformative potential of artificial intelligence in healthcare.

REFERENCES

[1] I. S. Mackenzie, I. Ford, A. Walker, N. Munro, G. Guthrie, G. P. Leese, R. S. Lindsay, J. A. McKnight, A. D. Morris, E. R. Pearson, J. R. Petrie, S. Philip, S. H. Wild, and N. Sattar, "Diabetes and artificial intelligence beyond the closed loop: a review of the landscape, promise and challenges," *Diabetologia*, vol. 66, pp. 1597–1612, 2023.

[2] B. Guan, J. Bhattacharya, H. Xiao, Y. C. Kudva, R. Basu, P. Herrero, and P. Georgiou, "Artificial intelligence in diabetes management: advancements, opportunities, and challenges," *Cell Reports Medicine*, vol. 4, no. 11, p. 101213, 2023.

[3] E. K. Oikonomou and R. Khera, "Machine learning in precision diabetes care and cardiovascular risk prediction," *Cardiovascular Diabetology*, vol. 22, p. 259, 2023.

[4] P. G. Jacobs, J. E. Youssef, and J. R. Castle, "Artificial intelligence and machine learning for improving glycemic control in diabetes: best practices, pitfalls, and opportunities," *IEEE Reviews in Biomedical Engineering*, vol. 17, pp. 83–101, 2023.

[5] J. Wang, Y. Huang, S. Jabbour, S. Mukherjee, Z. Neshati, S. D. Young, E. Lehman, J. Kalpathy-Cramer, L. A. Celi, and L. G. McCoy, "Ai-based diabetes care: risk prediction models and implementation concerns," *npj Digital Medicine*, vol. 7, p. 34, 2024.

[6] A. Y. Alhaddad, H. Aly, H. Gad, A. Al-Ali, K. K. Sadasivuni, J.-J. Cabibihan, and R. A. Malik, "Sense and learn: Recent advances in wearable sensing and machine learning for blood glucose monitoring and trend-detection," *Frontiers in Bioengineering and Biotechnology*, vol. 10, p. 876672, 2022.

[7] I. Contreras and J. Vehi, "Artificial intelligence for diabetes management and decision support: Literature review," *Journal of Medical Internet Research*, vol. 20, no. 5, p. e10775, 2018.

[8] M. Khalifa and M. Albadawy, "Artificial intelligence for diabetes: Enhancing prevention, diagnosis, and effective management," *Computer Methods and Programs in Biomedicine Update*, vol. 5, p. 100141, 2024.

[9] T. Prioleau, J. Moore, R. J. Galindo, and R. K. A. Basu, "Deep learning for blood glucose prediction: Reproducibility challenges and factors affecting differential performance," *Research Square*, 2025.

[10] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Deep learning for diabetes: A systematic review," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2744–2757, 2021.

[11] M. T. Alam, M. M. Hasan, M. A. Alam, A. S. Nitu, M. P. Hossain, and M. Z. H. Alam, "Machine learning and artificial intelligence in diabetes prediction and management: A comprehensive review of models," *Journal of Novel Engineering Science and Technology*, vol. 1, no. 1, pp. 41–52, 2024.

[12] S. Ghosh, A. Dasgupta, and A. Swetapadma, "Artificial intelligence in personalized medicine for diabetes mellitus: a narrative review," *Cureus*, vol. 17, no. 1, p. e91520, 2025.

[13] J. Bai, T. Zhu, P. Herrero, and P. Georgiou, "Federated multimodal ai for precision-equitable diabetes care," *npj Digital Medicine*, 2024, preprint.

[14] H. Fahmy, "Exploring the role of ai in predicting chronic disease progression: Diabetes and cardiovascular diseases," *Premier Journal of Public Health*, vol. 1, no. 1, p. 21, 2025.

[15] A. Bahmani, K. Asare, L. E. Boulware, L. Brewer, A. J. Butte, D. Char, R. Chetty, K. Churchwell, L. Cipriano, M. Desai, S. N. Duda, J. C. Eichstaedt, E. Eisenstein, L. Flowers, J. Gichoya, S. L. Gómez, K. Goodman, L. Gottlieb, R. Hahn, T. Hernandez-Boussard, Y. Hswen, A. Kaushal, I. S. Kohane, C. R. Lyles, M. E. Matheny, K. Morse, E. Nsoesie, Z. Obermeyer, J. A. Omiye, J. Pathak, S. R. Pfohl, P. Rajpurkar, D. Rehkopf, F. Rodriguez, L. G. Rosas, A. Rusanov, A. Schuler, N. H. Shah, K. Shankar, M. Sjoding, S. Saria, T. Veinot, A. Verghese, J. Voigt, M. Weiner, D. R. Williams, C. Wong, J. R. Zubizarreta, T. Hernandez-Boussard, and N. H. Shah, "Achieving inclusive healthcare through integrating education and research with ai and personalized curricula," *Communications Medicine*, vol. 5, p. 34, 2025.

[16] I. Rodríguez-Rodríguez, J.-V. Rodríguez, I. Chatzigiannakis, and M. Zamora Izquierdo, "Applications of the internet of medical things to type 1 diabetes mellitus," *Electronics*, vol. 12, no. 3, p. 756, 2023.

[17] S. Zanelli, E. Yacoub, A. Marzullo, C. Corsi, and B. De Maria, "Diabetes detection and management through photoplethysmographic and electrocardiographic signals analysis: A systematic review," *Sensors*, vol. 22, no. 13, p. 4890, 2022.

[18] E. Scheideman and E. Ipp, "Machine learning to diagnose complications of diabetes," *Journal of Diabetes Science and Technology*, 2025.

[19] M. T. Alam, M. M. Hasan, M. A. Alam, A. S. Nitu, M. P. Hossain, and M. Z. H. Alam, "Machine learning and artificial intelligence in diabetes prediction and management: a comprehensive review of models," *Social Science Research Network*, 2025.

[20] A. Ran, S. Niu, H. He, and Y. Zheng, "Source-free active domain adaptation for diabetic retinopathy grading based on ultra-wide-field fundus images," *Computers in Biology and Medicine*, vol. 174, p. 108418, 2024.

[21] Y. Zhang, L. Wang, Z. Wu, J. Zeng, Y. Chen, R. Tian, J. Li, T. Li, N. Garcia, and D. Shen, "Diabetic retinopathy grading by a source-free transfer learning approach," *Biomedical Signal Processing and Control*, vol. 73, p. 103423, 2022.

[22] S. Maqsood, S. Xu, M. Springer, N. Mohammadzadeh, M. K. Rutter, I. Buchan, and K. Dashtipour, "Gluconet-mm: A multimodal attention-based multi-task learning framework with decision transformer for personalised and explainable blood glucose forecasting," *Diabetes, Obesity and Metabolism*, vol. 27, no. 3, pp. 1163–1176, 2025.

[23] S. Kulkarni and B. E. Reddy, "Diabetes detection and prediction through a multimodal artificial intelligence framework," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 38, no. 1, pp. 459–468, 2025.

[24] Y. Xiong, W. Zhang, and J. Chen, "A multi-modal deep learning approach for predicting type 2 diabetes complications: Early warning system design and implementation," *World Journal of Innovation and Modern Technology*, vol. 7, no. 6, p. 15, 2024.

[25] D. B. Olawade, O. Wada, A. Odetayo, A. C. David-Olawade, J. Eberhardt, and J. Ling, "Digital twin paradigm in diabetes prediction and management," *Diabetes Research and Clinical Practice*, vol. 221, p. 113075, 2026.

[26] Y. Li, X. Wang, M. Xu, and M. He, "Integrated image-based deep learning and language models for primary diabetes care," *Nature Medicine*, vol. 30, pp. 2419–2430, 2024.