The Indonesian Journal of Computer Science
www.ijcs.net
Volume 15, Issue 2, April 2026
https://doi.org/10.33022/ijcs.v15i2.5110

A Digital Twin–Driven Machine Learning Framework for Diabetes Risk Prediction and Short-Term Health Trajectory Simulation

Belinda Ndlovu[1]
belinda.ndlovu@nust.ac.zw[1]
[1]Department of Informatics and Analytics, National University of Science and Technology, Bulawayo, Zimbabwe

Article Information

Received : 29 Mar 2026
Revised : 7 Apr 2026
Accepted : 13 Apr 2026

Keywords

Diabetes risk prediction, Digital twin, Machine learning, Predictive analytics, Health trajectory simulation, Personalized healthcare, Clinical decision support, Data-driven healthcare

Abstract

Diabetes remains a major global health challenge, requiring early risk detection and proactive management to reduce long-term complications. However, existing approaches are largely reactive and rely on static clinical indicators, limiting their ability to support personalized and forward-looking care. While machine learning (ML) has improved predictive accuracy, most models provide point-in-time risk estimates without accounting for how risk evolves. Similarly, digital twin (DT) technologies enable simulation of patient states but are often implemented independently of robust predictive modelling frameworks. This study proposes a unified framework that integrates ML and DT technologies to enable diabetes risk prediction and short-term health trajectory simulation. Using the Centres for Disease Control and Prevention (CDC) Diabetes Health Indicators dataset, a structured CRISP-DM methodology was applied to preprocess the data, select features, develop and evaluate models. Class imbalance (13.9% minority class) was addressed using the Synthetic Minority Over-sampling Technique (SMOTE). Five machine learning models were evaluated, with Gradient Boosting achieving the best performance (ROC-AUC = 0.797; F1-score = 0.415), indicating moderate discriminative capability under imbalanced conditions. Building on this predictive layer, a digital twin framework was developed to translate risk predictions into short-term (90-day) trajectory simulations. The system was implemented using a modular, web-based architecture that integrates prediction, simulation, and visualization. The results demonstrate that linking predictive modelling with simulation enables trajectory-aware interpretation of diabetes risk rather than static assessment. While the simulation is based on model-driven assumptions rather than real-time physiological data, it provides an additional analytical layer to support anticipatory decision-making. This study contributes a scalable framework that bridges predictive analytics and simulation, positioning the framework as a baseline for future integration of predictive analytics and simulation in data-constrained healthcare environments.

A. Introduction

Diabetes is a chronic and progressively debilitating metabolic disorder that continues to impose a significant burden on global healthcare systems, affecting over 500 million individuals worldwide and contributing substantially to morbidity, mortality, and healthcare expenditure [1][2]. The complexity of diabetes stems from its multifactorial nature, with genetic, behavioural, and environmental determinants that interact dynamically over time. Consequently, effective management requires continuous monitoring, early risk detection, and personalized intervention strategies [3][4]. However, conventional clinical approaches remain largely reactive, relying on periodic assessments and static measurements, which are insufficient for capturing the temporal evolution of the disease and enabling proactive care [5][6][7].

The emergence of machine learning (ML) has transformed predictive healthcare by enabling the analysis of large-scale, heterogeneous datasets to uncover complex, non-linear relationships between risk factors and disease outcomes [8][9]. In diabetes research, ML models have been widely applied for predicting disease onset, glycaemic control, and complications using algorithms such as random forests, gradient boosting, support vector machines, and deep neural networks [10][11]. These models have demonstrated improved predictive performance compared to traditional statistical methods, particularly in handling high-dimensional data and capturing complex interactions [9][6][12]. Despite these advances, most ML-based systems focus on static prediction, providing risk estimates at a single point in time without accounting for how risk evolves.

Parallel to these developments, digital twin (DT) technology has emerged as a transformative paradigm for personalized healthcare [13]. A digital twin is a dynamic virtual representation of a physical entity that evolves through real-time data integration, enabling simulation, prediction, and optimization of system behaviour [13][14]. In healthcare, DTs have been applied to model patient-specific conditions, simulate treatment outcomes, and support clinical decision-making [16][13][16][17]. In the context of diabetes, DTs enable the creation of virtual patient profiles that integrate physiological, behavioural, and environmental data to simulate disease progression and personalize treatment strategies [19][20]. However, many existing DT implementations rely on high-resolution physiological data or specialized clinical infrastructures, limiting their accessibility and scalability in broader healthcare settings.

Despite the individual advances in ML and DT technologies, current research remains fragmented. Many studies focus exclusively on predictive modelling without incorporating simulation capabilities, while others explore digital twin concepts without robust integration of advanced machine learning techniques [21][22]. Furthermore, existing approaches often rely on limited variables and lack standardized evaluation frameworks, which constrains their generalisability and practical applicability [23][24]. Critically, there is a lack of integrated frameworks that combine predictive analytics with short-term simulation to support proactive and trajectory-aware healthcare decision-making.

This study addresses these gaps by proposing a unified framework that integrates machine learning and digital twin technologies to enable diabetes risk prediction and short-term health trajectory simulation. The framework extends conventional predictive modelling by linking classification outputs to a simplified digital twin environment, enabling trajectory-aware interpretation of patient risk rather than static point-in-time estimates. Specifically, the study (1) develops a machine learning model for population-level diabetes risk prediction using accessible health indicators, (2) introduces a digital twin component that translates these predictions into short-term patient-specific risk trajectories, and (3) integrates both components within a scalable architecture. Unlike prior approaches that treat prediction and simulation independently, the proposed framework establishes a continuous interaction between risk estimation and dynamic state updates under a unified data representation. The framework is evaluated using structured health indicators to demonstrate its feasibility as a decision-support prototype.

B. Materials and Methods

This section outlines the methodological approach adopted to develop and evaluate the proposed machine learning and digital twin framework for diabetes risk prediction and simulation. The study follows a structured, systematic process encompassing data acquisition, preprocessing, model development and evaluation, and the design and integration of the system architecture. The methodological choices were guided by the need to ensure reproducibility, robustness, and alignment with established data-driven modelling practices. The research design is presented first to provide an overview of the analytical workflow underpinning the study.

Research Design and Methodological Framework

This study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework to guide the end-to-end development of the machine learning and digital twin system[25]. The CRISP-DM framework was used to structure the analytical workflow, guiding the sequential processes of data understanding, preparation, modelling, evaluation, and deployment. Each phase was operationalized within the study to ensure a consistent progression from data preprocessing to model evaluation and system integration.

Dataset Description and Source

The dataset utilized in this study is the Centers for Disease Control and Prevention (CDC) Diabetes Health Indicators dataset, which contains a comprehensive set of demographic, behavioural, and health-related variables associated with diabetes risk. The dataset includes attributes such as age, body mass index (BMI), income, lifestyle behaviours, and pre-existing health conditions.

The dataset was selected due to its scale, diversity, and relevance to population-level diabetes risk modelling. Initial exploration revealed class imbalance, with the minority class (individuals diagnosed with diabetes) representing approximately 13.9% of the dataset. This imbalance necessitated corrective measures during preprocessing to ensure the model learned in an unbiased manner.

System Architecture and Overall Workflow

The overall system architecture integrates machine learning and digital twin components within a unified analytical pipeline designed to support diabetes risk prediction and short-term simulation. The system's end-to-end workflow is illustrated in Figure 1.

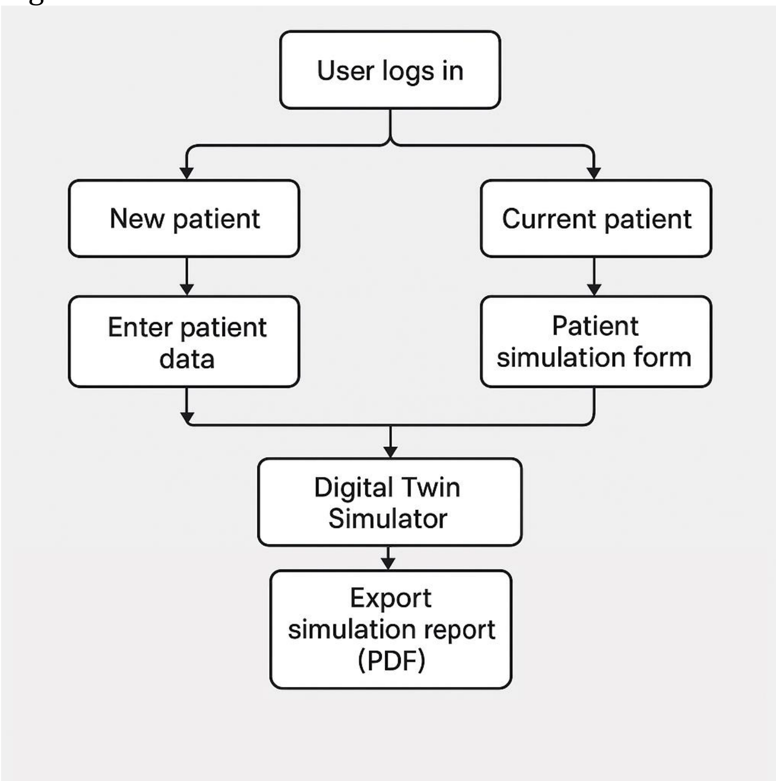

Figure 1. Model/System Workflow Diagram

Figure 1 presents a structured representation of the system pipeline, beginning with data input and progressing through preprocessing, model inference, and simulation. The workflow begins with the ingestion of structured patient data, which is then preprocessed to ensure consistency, normalization, and compatibility with machine learning algorithms.

Following preprocessing, the data is passed to the machine learning module, where trained models generate a probabilistic diabetes risk score. This prediction is subsequently transmitted to the digital twin simulation module, which uses it to initialize a virtual patient model.

The workflow concludes with the generation of structured outputs, including prediction scores and simulation-ready data, which are passed to the system interface for visualization. The architecture distinguishes between offline model training and real-time inference, ensuring efficient deployment and scalability.

Overall, the system architecture operationalizes the proposed framework by linking the predictive model outputs to a dynamic simulation layer. The predicted risk scores serve as inputs to the digital twin module, which updates the patient's risk trajectory over discrete time steps.

Use Case Characteristics and System Interaction Model

The system interaction model defines how users engage with the machine learning and digital twin framework within an operational environment. This interaction is illustrated in Figure 2, which represents the functional relationships between the user and the system.

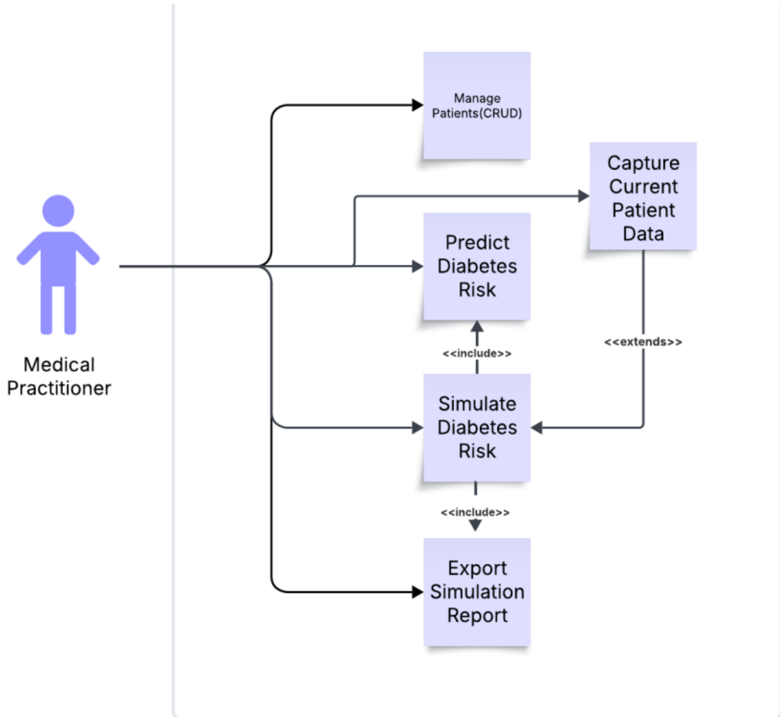

Figure 2. System Use Case Diagram

Figure 2 presents a structured view of the system's functional behaviour by identifying the primary actor, system boundaries, and core interaction processes. The primary actor in this system is the healthcare practitioner or analyst, who interacts with it through a user interface to perform data entry, prediction, and simulation tasks.

The primary actor in the system is the healthcare practitioner or analyst, who interacts with the system through a user interface. The system boundary encompasses internal processes such as data preprocessing, machine-learning inference, digital-twin simulation, and output generation.

The interaction begins with the input of patient data, which triggers the internal processing pipeline. The system then generates a diabetes risk prediction, followed by a simulation within the digital twin module. The outputs are subsequently made available for visualization and reporting.

This use case model complements the workflow diagram by focusing on user-system interaction and providing a functional perspective on the system's operation.

Data Preprocessing and Cleaning

Data preprocessing was conducted to ensure the dataset was of high quality and suitable for machine learning modelling. The preprocessing pipeline included handling missing values, normalizing numerical features, and encoding categorical variables.

Missing values were addressed using appropriate imputation techniques to preserve dataset integrity. Numerical features were scaled to ensure uniformity across variables, improving model convergence and stability. Categorical variables were transformed into numerical representations using encoding techniques suitable for machine learning algorithms.

This process reduced noise, improved data consistency, and enhanced the reliability of subsequent modelling stages.

Handling Class Imbalance

The dataset exhibited significant class imbalance, with the minority class representing approximately 13.9% of observations. To address class imbalance (13.9% minority class), the Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic minority samples while preserving all original observations. This approach improves class representation without discarding majority-class data, which is particularly important in healthcare datasets where data loss may affect clinical relevance.

Feature Selection Strategy

Feature selection was performed to identify the most relevant predictors and reduce dimensionality. A Random Forest-based feature-importance approach was employed to evaluate each variable's contribution to model performance.

Feature importance scores were computed using the training dataset after preprocessing and class balancing. A threshold-based selection criterion was applied, retaining only features with an importance score above 0.02 for model development, ensuring the final feature set captured the most informative predictors while reducing noise and dimensionality.

The resulting feature rankings were used to guide model optimization and are presented in the Results section.

Model Development and Training

Multiple machine learning models were developed to predict diabetes risk, including Logistic Regression, Random Forest, Support Vector Machine, k-Nearest Neighbours, and Gradient Boosting. These models were selected to capture a range of learning paradigms, including linear, distance-based, and ensemble approaches.

The dataset was split into training and test sets at 80:20. Model training was conducted using the Scikit-learn library in Python (version 3.10). The following hyperparameters were applied for each model:

- Logistic Regression: L2 regularisation, solver = *lbfgs*, maximum iterations = 1000.
- Random Forest: number of estimators = 100, maximum depth = 10, minimum samples split = 2, random state = 42.
- Support Vector Machine (SVM): kernel = radial basis function (RBF), regularisation parameter (C) = 1.0, gamma = *scale*.
- k-Nearest Neighbours (k-NN): number of neighbours (k) = 5, distance metric = Euclidean.

https://doi.org/10.33022/ijcs.v15i2.5110

- Gradient Boosting: number of estimators = 100, learning rate = 0.1, maximum depth = 3, random state = 42.

Default Scikit-learn parameters were retained for all other settings unless explicitly stated. Hyperparameter values were selected based on standard practices in the literature to balance model performance and computational efficiency.

All experiments were conducted on a system with an Intel Core i7 processor and 16 GB of RAM, running a 64-bit operating system.

Evaluation was performed on the held-out test dataset to ensure unbiased performance estimation.

To enhance robustness and reduce the risk of overfitting, k-fold cross-validation (k = 5) was applied during model training. The training dataset was partitioned into five equal subsets, where in each iteration, four folds were used for training and one fold for validation. This process was repeated five times, ensuring that each subset served as a validation set once.

The average performance across the folds was used to assess model stability and guide model selection. Cross-validation was performed prior to final evaluation on the held-out test dataset to ensure an unbiased estimate of performance.

Limited hyperparameter tuning was applied to maintain model comparability and reduce the risk of overfitting, with parameter values selected based on established practices in the literature.

Model Evaluation and Performance Metrics

Model performance was evaluated using accuracy, F1-score, and Receiver Operating Characteristic Area Under the Curve (ROC-AUC). These metrics provide a comprehensive assessment of classification performance, particularly in imbalanced datasets.

The F1-score was emphasized due to its ability to balance precision and recall, while ROC-AUC was used to assess the model's discriminative capability across different classification thresholds. Evaluation was conducted on the test dataset to ensure generalisability.

Comparative performance results and detailed analysis are presented in the Results section.

Digital Twin Design and Simulation Framework

The digital twin component extends machine learning predictions into a simulation environment capable of modelling patient-specific diabetes risk trajectories. The simulation process is governed by a deterministic update function defined as: $R_{t+1}=R_t+\alpha\cdot\Delta X$, where R_t represents the risk at time t, α is a scaling coefficient derived from model output, and ΔX represents changes in input features over time. This update function represents a simplified linear approximation intended to demonstrate short-term trajectory dynamics rather than to model physiological processes with clinical precision. The conceptual design of the digital twin is illustrated in Figure 3.

Core > digital_twin.py > PatientDigitalTwin > run_simulation

```python
import numpy as np
import pandas as pd
from Core.custom_funcs import preprocess_user_input, predict_multiple_instances

class PatientDigitalTwin:
    """
    A class to simulate the trajectory of patient parameters and predict diabetes probabilities over time.
    """
    def __init__(self): ···

    def preprocess_initial_data(self, initial_data): ···

    def simulate_patient_trajectory(self, initial_data, days=365): ···

    def analyze_historical_data(self, historical_data): ···

    def simulate_patient_trajectory_with_history(self, historical_data, days=365): ···

    def predict_diabetes_trajectory(self, trajectory):
        """
        Predicts diabetes probabilities for each day in the simulated trajectory using reusable prediction functions.

        Parameters:
            trajectory (list of dict): Simulated patient trajectory in model-compatible format.

        Returns:
            list of float: List of predicted diabetes probabilities for each day.
        """
        # Use the reusable predict_multiple_instances function for batch prediction
        predictions = predict_multiple_instances(trajectory)

        # Extract only the probabilities as floats (e.g., 0.55 instead of "55.34%")
        probabilities = [
            float(result['Probability'].strip('%')) / 100 for result in predictions
        ]

        return probabilities
```

Figure 3. Digital Twin Construct

Figure 3 presents the conceptual architecture of the digital twin, comprising three core components: the input state representation, the predictive model interface, and the simulation engine. The input state representation captures patient-specific attributes, including demographic, behavioural, and health-related variables derived from the dataset. These attributes constitute the digital twin's initial state.

The predictive model interface connects the trained machine learning model to the digital twin, enabling it to transform input features into a diabetes risk probability score. This probabilistic output serves as the baseline parameter for simulation.

The simulation engine updates the patient's risk state iteratively over discrete time intervals. At each step, the system recalculates the risk score based on the initial prediction and predefined progression assumptions derived from the model output. This process generates a temporal sequence of risk values representing the evolution of diabetes risk over 90 days.

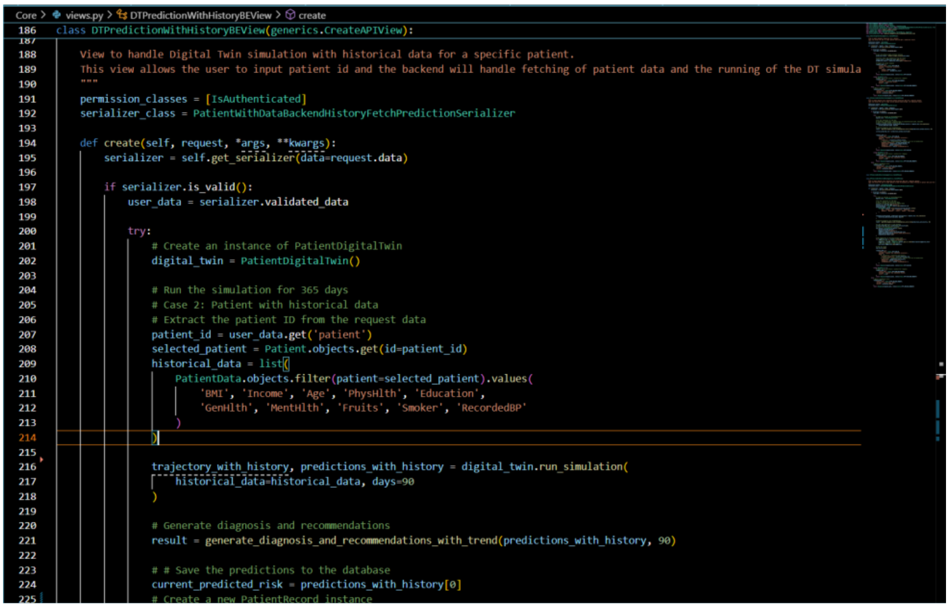

Figure 4. Digital Twin Construct Integration into the REST API Backend-1

Figure4 illustrates the system-level integration of the digital twin within the backend architecture. The integration is implemented using a Representational State Transfer (REST) API that facilitates communication among the machine learning module, the digital twin simulation engine, and the user interface.

The workflow begins with a client request from the frontend interface, where patient data is submitted to the backend system. The backend processes the input data and passes it to the trained machine learning model, which generates a risk prediction. This prediction is then forwarded to the digital twin module via API endpoints.

Within the digital twin module, the simulation engine processes the prediction and generates a sequence of risk values over the defined time horizon. The simulation output is then returned to the backend and transmitted to the frontend for visualization.

The separation of components through API-based communication ensures the system's modularity, scalability, and flexibility. It also enables independent updates to the machine learning model and digital twin components without affecting overall system functionality.

Importantly, this framework distinguishes between prediction (static inference) and simulation (dynamic modelling), enabling the system to support both immediate risk assessment and forward-looking analysis within a unified architecture.

The simulation engine produces a sequence of time-dependent risk values based on the initial prediction and model-driven progression assumptions. These outputs are structured as temporal data points representing the evolution of diabetes risk over the defined simulation horizon. The generated simulation results are subsequently visualized in the system interface and presented and analyzed in the Results section.

System Implementation, Tools, and Technologies

The system was implemented using a modular technology stack to support machine learning, simulation, and system integration. The backend was developed using the Django REST Framework, which enables the creation of scalable API endpoints for data processing and communication between system components.

Machine learning models were implemented in Python using Scikit-learn, supported by additional libraries for data preprocessing, feature engineering, and class balancing. The digital twin simulation module was integrated within the backend as a service layer responsible for generating time-dependent risk trajectories.

The frontend interface was developed using Angular, providing a platform for data input, interaction, and visualization. Communication between the frontend and backend components is facilitated through RESTful API calls, ensuring real-time responsiveness.

The combination of these technologies ensures system scalability, modularity, and efficient integration between machine learning and digital twin components within a unified architecture.

C. Results

The results of the study are presented to evaluate the performance of the proposed machine learning and digital twin framework and to examine the relative contribution of input features to diabetes risk prediction. The analysis is structured as follows: first, assessing the importance of predictor variables; second, comparing model performance; and third, examining system outputs, including simulation results and interface functionality. This progression enables a comprehensive understanding of both the models' predictive capability and the practical implications of integrating prediction with simulation in a unified system.

Feature Selection and Predictor Importance Analysis

The feature importance analysis identified body mass index (BMI) and age as the dominant predictors of diabetes risk (Figure 5). These variables contribute substantially more to model decisions than other features, indicating a strong concentration of predictive signal within a limited subset of attributes. Socio-economic variables, including income, also exhibit a measurable influence, though to a lesser extent. This concentration of predictive importance suggests that a reduced feature set may be sufficient for effective model performance, which has implications for model simplification and deployment in resource-constrained environments.

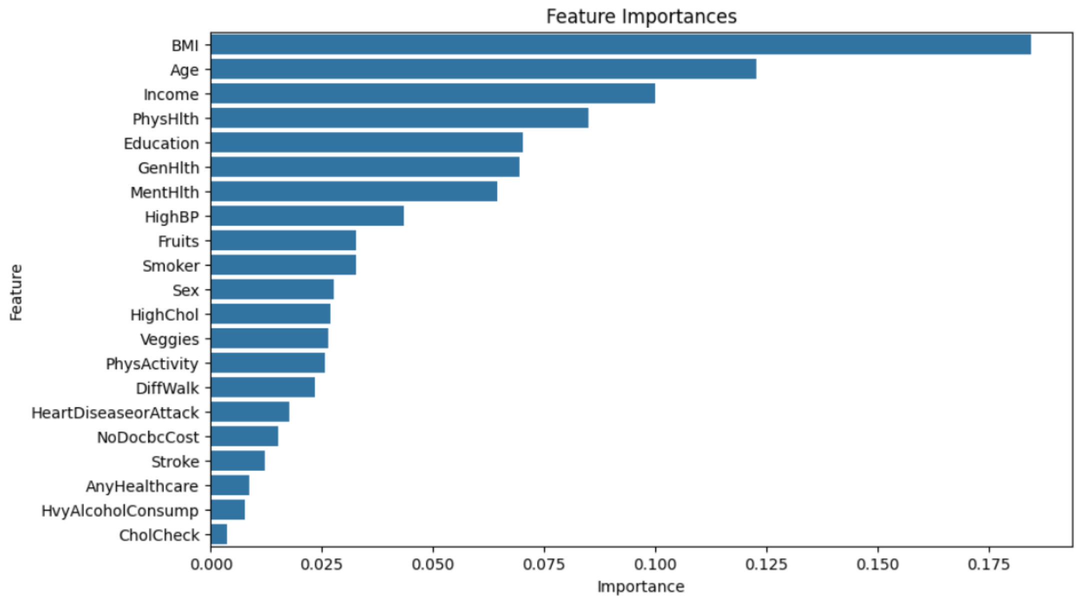

Figure 5. Feature importance ranking using Random Forest

This distribution suggests that the model captures both physiological and contextual determinants of diabetes risk. The prominence of BMI and age aligns with established epidemiological evidence, providing an additional layer of interpretability to the model outputs. Importantly, reducing the number of low-importance features improved computational efficiency without noticeable degradation in this study

Machine Learning Model Performance and Comparative Analysis

The comparative performance of the machine learning models is presented in Figure 6 and Table 1. Ensemble-based methods demonstrate superior performance relative to traditional models, with Gradient Boosting achieving the highest overall scores.

Summary Table:

	Model	F1 Score	ROC AUC Score	Accuracy
0	Logistic Regression	0.412348	0.790472	0.706579
1	Random Forest	0.355320	0.726541	0.755538
2	Gradient Boosting	0.415000	0.796885	0.704214
3	Decision Tree	0.309136	0.605893	0.733069
4	ANN	0.422483	0.795904	0.792219

Figure 6. Model Evaluation and Selection-2

Table 1. Model Evaluation

Model	F1 Score	ROC AUC Score	Accuracy
Logistic Regression	0.412348	0.790472	0.706579
Random Forest	0.355320	0.726541	0.755538
Gradient Boosting	0.415000	0.796885	0.704214
Decision Tree	0.309136	0.605893	0.733069

Specifically, Gradient Boosting achieved an ROC-AUC of 0.797 and an F1-score of 0.415, indicating a balanced trade-off between discrimination and classification performance. While Logistic Regression produced a comparable ROC-AUC (0.790), its slightly lower F1-score indicates reduced effectiveness in handling class imbalance.

Random Forest performed moderately but did not outperform Gradient Boosting, suggesting that boosting-based approaches may better capture feature interactions in this dataset. The Decision Tree model performed the worst, reinforcing the limitations of single-tree models in complex healthcare prediction tasks.

The marginal difference between Gradient Boosting and Logistic Regression suggests that while ensemble methods provide performance gains, simpler models may still offer competitive performance with reduced computational complexity.
The cross-validation results further indicated stable performance across folds, supporting the reliability of the selected model and suggesting limited sensitivity to data partitioning.

Model Evaluation Using ROC Curve Analysis

As summarised in Table 1, Gradient Boosting achieved the highest overall performance across evaluation metrics. To further assess discriminative capability, the ROC curve for the Gradient Boosting model is presented in Figure 7. The curve illustrates the trade-off between sensitivity and specificity across different classification thresholds. The relatively smooth curve and area under the curve confirm that the model maintains consistent performance across varying thresholds.

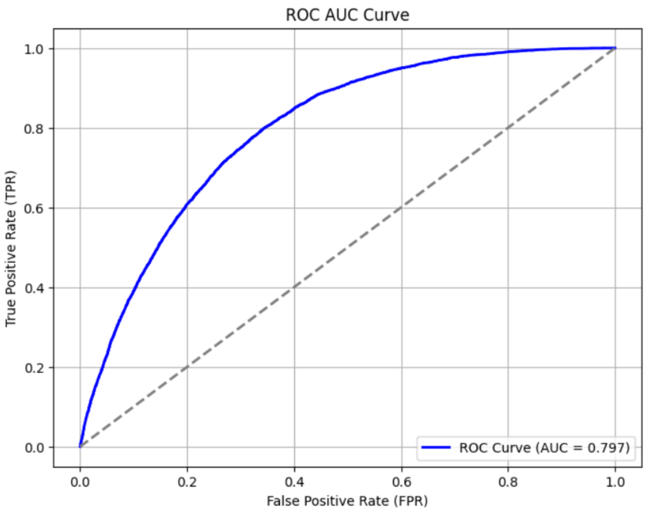

Figure 7. ROC curve for the Gradient Boosting model

While the ROC curve provides a global view of model performance, it does not reflect classification outcomes at a specific decision threshold. To address this, the confusion matrix for the Gradient Boosting model is presented in Figure 8.

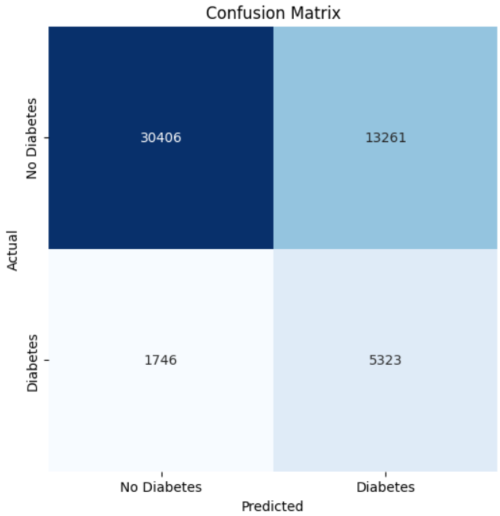

Figure 8. Confusion matrix for the Gradient Boosting model

The confusion matrix provides a detailed breakdown of true positives, true negatives, false positives, and false negatives. The results indicate that the model can correctly identify a substantial proportion of non-diabetic cases while maintaining reasonable sensitivity for detecting diabetic cases. However, the presence of false positives and false negatives reflects the inherent trade-offs in classification performance, particularly in the context of imbalanced datasets.

In practical deployment contexts, false negatives may delay early identification of high-risk individuals. In contrast, false positives may lead to unnecessary follow-up interventions, highlighting the importance of threshold calibration.

Collectively, Figures 7 and 8 provide a comprehensive evaluation of model performance. The ROC-AUC confirms overall discriminative strength, while the confusion matrix contextualizes classification errors, offering insight into the practical implications of model deployment in healthcare settings.

Digital Twin Simulation Results

The digital twin component extends the predictive model by generating time-dependent risk trajectories, as illustrated in Figure 9. The simulation outputs represent the evolution of predicted diabetes risk over 90 days for an individual patient. Figure 9 illustrates the temporal variability in risk trajectories, highlighting the sensitivity of predictions to input parameters.

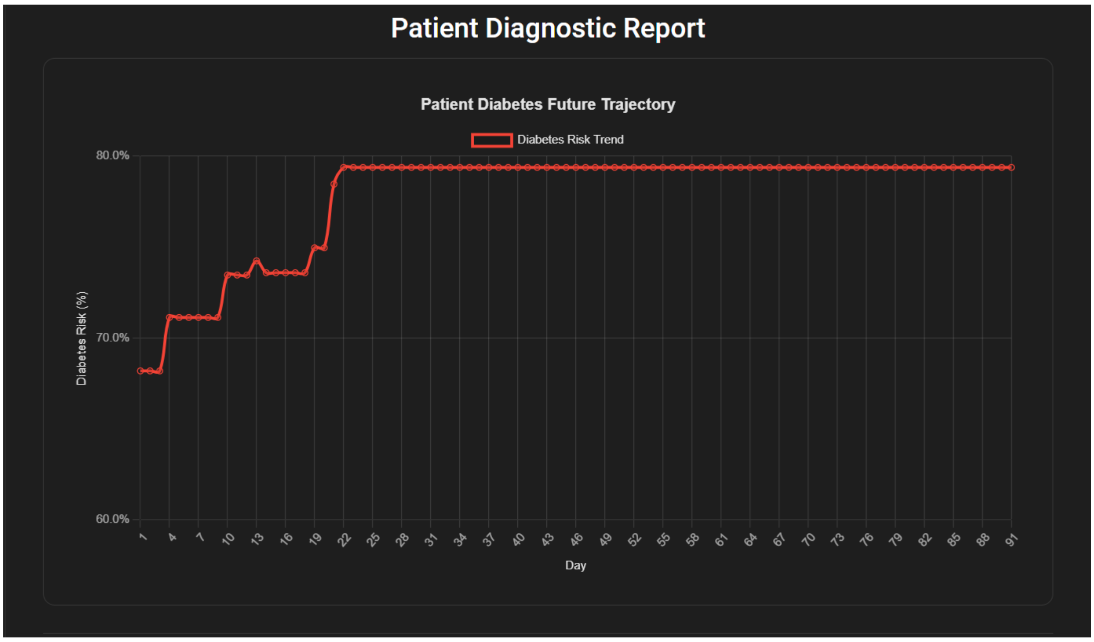

Figure 9. Digital twin–based diabetes risk trajectory simulation

The results demonstrate that risk is not static but varies over time, driven by model assumptions and input parameters. This temporal representation provides an additional layer of insight beyond single-point predictions, enabling a more nuanced understanding of potential risk progression.

It is important to note that these trajectories are indicative rather than predictive of exact clinical outcomes, as the simulation does not incorporate real-time physiological updates. Nevertheless, the ability to visualize short-term trends enhances the interpretability and practical relevance of the model outputs.

System Architecture Validation

The system architecture, illustrated in Figure 10, confirms the successful integration of machine-learning and digital-twin components into a REST API-based backend. The architecture enables the transfer of prediction outputs to the simulation engine in real time, supporting continuous interaction between system components.

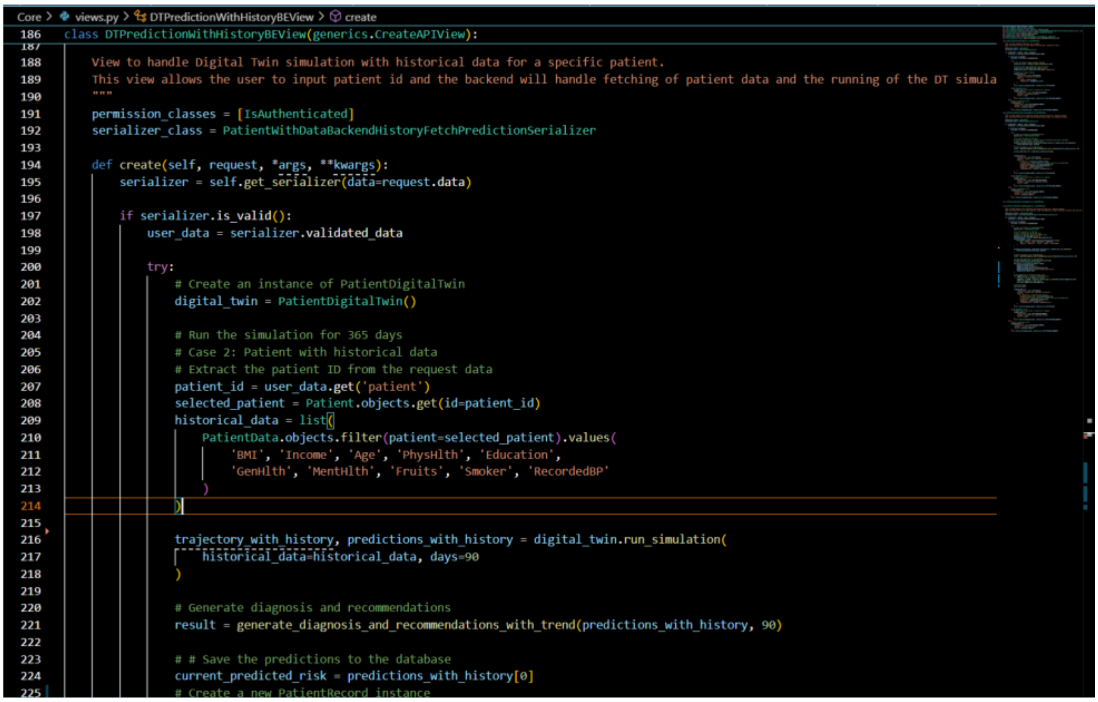

Figure 10. Example API implementation for integrating prediction and simulation components

The architecture follows a modular design, with the prediction engine and simulation module operating as independent yet interconnected services.

User inputs are processed through the REST API, which invokes the trained machine learning model to generate risk predictions. These predictions are subsequently passed to the digital twin module for simulation of short-term risk trajectories. The outputs are then returned to the user interface for visualization and interaction.

This modular structure supports scalability and flexibility, allowing individual components to be updated or replaced without affecting the overall system functionality.

Interface Demonstration

The system interface (Figure 11) provides a unified environment for data input, prediction, and simulation visualization. The dashboard integrates multiple system components, allowing users to interact with the model and observe outputs in real time.

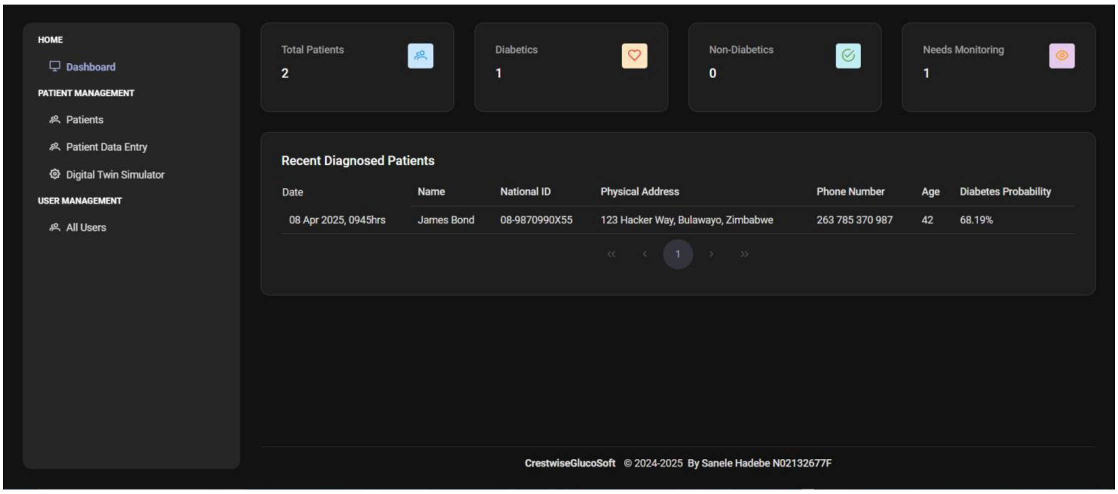

Figure 11. User Dashboard

The interface supports dynamic interaction, enabling users to modify input variables and immediately observe corresponding changes in predicted risk and simulated trajectories. This functionality enhances usability and supports exploratory analysis of risk factors.

System Functionality and Output Generation

The system's reporting functionality is illustrated in Figure 12, which shows the generation of structured diagnostic outputs. The ability to export results enhances the system's practical utility, allowing outputs to be stored, shared, and integrated into broader analytical workflows.

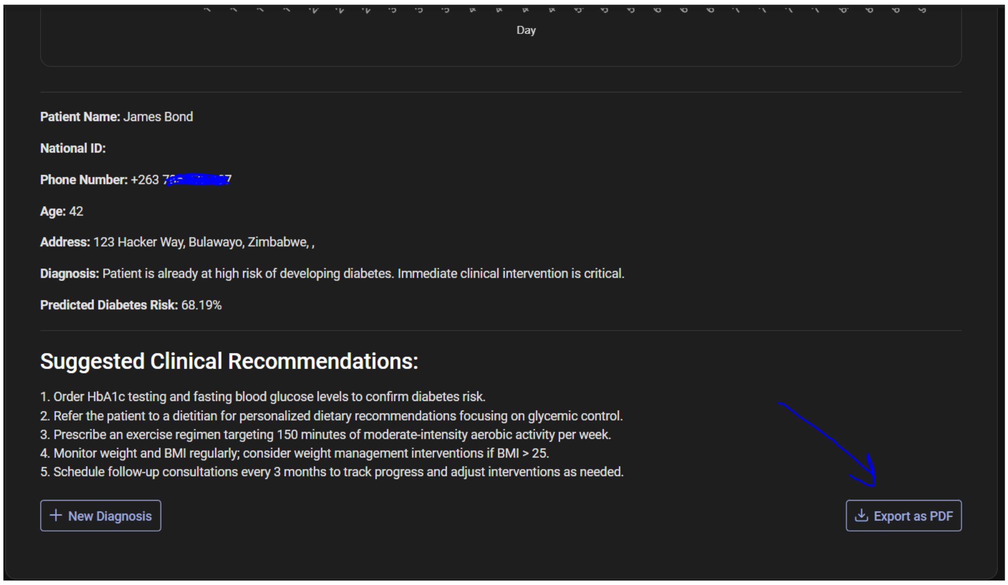

Figure 12. Diagnosis Report export Option

D. Discussion

The findings of this study demonstrate that integrating machine learning with digital twin technology provides a structured approach for diabetes risk prediction and short-term health trajectory modelling. Rather than treating prediction and simulation as separate processes, the proposed framework links them within a unified system, enabling both immediate risk assessment and forward-looking analysis. This integration aligns with the growing emphasis on predictive and preventive healthcare, in which early identification of risk is essential to reducing long-term complications [26][21].

Interpretation of Machine Learning Performance

The comparative evaluation of machine learning models indicates that ensemble-based methods, particularly Gradient Boosting, offer improved predictive capability relative to traditional approaches. The ROC-AUC score of 0.797 indicates acceptable discriminative performance in distinguishing high-risk from low-risk individuals. This finding is consistent with prior studies showing that ensemble techniques are better suited to modelling non-linear relationships in healthcare datasets [27][28]

However, the relatively moderate F1-score (0.415) highlights the persistent challenge of class imbalance, even after applying SMOTE. This level of performance is consistent with prior studies [23][24] on imbalanced healthcare datasets, where F1-scores in the range of 0.35–0.50 are commonly reported despite the use of resampling techniques, reflecting the inherent difficulty of predicting the minority class in clinical contexts. As such, while the model demonstrates reasonable predictive capability, its outputs should be interpreted as supportive rather than definitive in clinical decision-making contexts.

The feature importance results further reinforce the model's validity. The prominence of body mass index (BMI) and age aligns with established epidemiological evidence linking these variables to diabetes risk [29][30]. This consistency enhances model interpretability, addressing a common concern in healthcare machine learning applications regarding the transparency of predictive outputs [31][32].

The application of k-fold cross-validation further supports the robustness of the developed models by demonstrating consistent performance across different subsets of the training data. This indicates that the models are not overly sensitive to specific data partitions and are less likely to overfit. The use of a separate held-out test dataset for final evaluation strengthens this position by providing an unbiased estimate of model performance.

However, while the results suggest reasonable generalizability within the dataset, caution is warranted when extending these findings to other populations or clinical settings. The dataset is derived from population-level health indicators, which may not fully capture variations in clinical practices, demographic distributions, or underlying health conditions across different regions. As noted in prior studies, model performance in healthcare is often context-dependent, and external validation remains essential for confirming generalisability [23][24]

This reinforces the importance of combining multiple evaluation metrics when assessing model performance in imbalanced healthcare datasets, where reliance on a single metric may lead to misleading conclusions.

The digital twin developed in this study represents a model-driven abstraction rather than a fully synchronized clinical replica. As such, it should be interpreted as a scenario simulation tool that explores plausible short-term risk trajectories based on model outputs, rather than a real-time physiological digital twin.

Digital Twin Simulation and Added Value

The findings demonstrate how integrating predictive modelling and simulation can be used to interpret risk progression over time, rather than relying solely on static risk estimates.

This capability represents a practical extension of existing work in digital health, where many machine learning applications focus on classification without incorporating temporal dynamics[6][33]. Similarly, while digital twin applications have been explored in healthcare, they often rely on high-resolution physiological data or specialized clinical settings [16]. In contrast, this study demonstrates that meaningful simulation can be achieved using population-level health indicators, thereby lowering the barrier to implementation.

The digital twin component in this study is designed as a model-driven simulation rather than a clinically validated physiological replica. As such, the generated trajectories should be interpreted as scenario-based projections rather than precise forecasts. This distinction is important, as it positions the digital twin as a decision-support tool rather than a predictive clinical instrument.

Importantly, the simulation results should be interpreted as indicative rather than predictive of exact future outcomes. The absence of real-time physiological data means that the digital twin operates on model-driven assumptions rather than continuously updated patient states. Nevertheless, the ability to visualize potential risk trajectories provides an additional layer of insight that supports anticipatory decision-making.

This positions the framework as a baseline for future work exploring the integration of predictive analytics and lightweight simulation in data-constrained healthcare settings.

System Integration and Practical Considerations

The successful integration of machine learning and digital twin components within a unified architecture highlights the feasibility of deploying such systems in practical settings. The use of a REST API-based backend ensures modularity and supports real-time interaction between system components, consistent with contemporary digital health infrastructures.

The dashboard interface further demonstrates how predictive and simulation outputs can be presented in an accessible format. While the interface itself is not the primary contribution of this study, its inclusion illustrates how technical outputs can be translated into usable insights for practitioners. This is particularly relevant, as usability and interpretability are key factors influencing the adoption of AI-driven healthcare systems.

Positioning Within the Literature and Contribution

When positioned within the broader literature, this study contributes by bridging a gap between predictive modelling and simulation-based approaches. Previous research has largely treated these as distinct areas, focusing either on classification models [27] or on digital twin conceptualizations and applications [16] [19].

This study contributes by demonstrating a practical integration of these approaches within a single framework. In doing so, it supports a shift from static prediction towards more dynamic representations of patient risk. However, the contribution should be understood as an incremental advance rather than a complete solution, given the limitations of data availability and simulation assumptions. This positions the work as a methodological bridge rather than a standalone predictive breakthrough.

These findings should therefore be interpreted as dataset-specific, with external validation required to confirm applicability across different populations and healthcare contexts.

Theoretical Implications

This study contributes to the evolving discourse on artificial intelligence in healthcare by demonstrating how predictive modelling can be extended into a simulation context through the integration of digital twin technology. While prior research has predominantly focused on either machine learning-based prediction or conceptual applications of digital twins [15][18], this study provides an empirical illustration of how these approaches can be operationalized within a unified framework.

The findings suggest a conceptual shift from static risk prediction towards temporally aware modelling, in which risk is not only estimated at a single point but also explored as a short-term trajectory. This positions digital twins as a complementary analytical layer rather than a replacement for predictive models. In doing so, the study advances the understanding of how simulation can augment traditional machine learning outputs in healthcare contexts.

Furthermore, by demonstrating that meaningful simulation can be achieved using population-level health indicators, the study challenges the prevailing assumption that digital twins require high-frequency physiological or sensor-based data. This contributes to a more accessible conceptualization of digital twins within resource-constrained environments.

E. Implications

This section discusses the implications of the study's findings, with a focus on how the proposed integration of machine learning and digital twin technologies informs both practice and broader healthcare contexts. The implications are presented by considering how the framework can support decision-making, system design, and the adoption of data-driven approaches in healthcare environments. The discussion begins with practical implications, highlighting how the findings may be applied in real-world settings.

Practical Implications

From a practical perspective, the proposed framework offers a structured approach for supporting early-stage risk assessment and short-term monitoring of diabetes using readily available data. The results indicate that ensemble machine learning models can achieve acceptable predictive performance, while the digital twin component enables visualization of potential risk progression over time.

This combination has implications for healthcare practitioners and analysts, particularly in contexts where continuous monitoring technologies are not available. The ability to generate both a risk estimate and a corresponding trajectory can support more informed decision-making by highlighting potential future risk patterns rather than relying solely on static predictions.

The system architecture, implemented with a modular, API-driven design, further demonstrates the feasibility of integrating predictive and simulation components into existing digital health infrastructures. While the interface component is illustrative, it underscores the importance of presenting analytical outputs in an interpretable, user-friendly manner to support adoption.

Policy and Healthcare System Implications

At a broader level, the findings have implications for digital health strategies aimed at improving chronic disease management. The study demonstrates that predictive and simulation-based tools can be developed using readily available datasets and standard computational resources, which is particularly relevant for healthcare systems operating under resource constraints.

The integration of machine learning and digital twin technologies aligns with policy directions that emphasize preventive, data-driven healthcare.By enabling earlier risk identification and providing a mechanism for exploring short-term trajectories, such systems can support more proactive intervention strategies.

However, the findings also highlight the importance of cautious implementation. Given the limitations of the dataset and the absence of real-time physiological data, the system's outputs should be used to complement, rather than replace, clinical judgement. Policymakers and practitioners should therefore consider such frameworks as decision-support tools within a broader clinical ecosystem.

More broadly, the framework contributes to the ongoing transition from reactive to anticipatory healthcare systems, where decision-making is informed not only by current patient states but also by plausible short-term trajectories. While further validation and refinement are required, this work provides a practical foundation for future research and implementation efforts to bridge the gap between predictive analytics and actionable clinical insight.

Collectively, the implications of this study suggest that integrating machine learning and digital twin technologies offers a pragmatic pathway towards more anticipatory, data-driven healthcare practices. Rather than replacing existing clinical processes, the proposed framework complements them by enabling the exploration of short-term risk trajectories alongside conventional prediction. This combined perspective has the potential to support more informed decision-making, particularly in settings where continuous monitoring infrastructure is limited. At the same time, the findings highlight the importance of cautious implementation, as the effectiveness of such systems depends on data quality, contextual relevance, and appropriate interpretation of outputs. Overall, the study provides a foundation for advancing the use of integrated predictive and simulation-based approaches in healthcare, while recognizing the need for continued validation and refinement.

F. Limitations

Despite the contributions of this study, several limitations must be acknowledged to contextualize the findings and guide future research. First, reliance on the CDC Diabetes Health Indicators dataset imposes constraints on the scope and representativeness of the data. While the dataset provides a broad range of demographic and behavioural variables, it does not include real-time physiological measurements such as continuous glucose monitoring or wearable sensor data. As a result, the developed model captures static and self-reported health indicators rather than dynamically evolving physiological states, which may limit its sensitivity to short-term fluctuations in patient health.

Second, although class imbalance was addressed using SMOTE, the relatively moderate F1-score suggests that the challenge of imbalanced classification was not entirely eliminated. This limitation is consistent with prior studies indicating that oversampling techniques, while effective, may introduce synthetic noise and fail to fully replicate the complexity of minority-class patterns in healthcare datasets. In addition, the use of SMOTE may introduce synthetic patterns that do not fully reflect real-world data distributions, potentially influencing model behaviour on unseen datasets.

Third, the digital twin simulation component is constrained to a 90-day prediction horizon. While this timeframe is sufficient to demonstrate short-term risk trajectory modelling, it does not capture long-term disease progression, which is critical for chronic conditions such as diabetes. Extending the simulation horizon would require more granular temporal data and longitudinal patient records. Moreover, the simulation component was not validated against longitudinal patient data, which limits the ability to assess its predictive fidelity over time.

Finally, the study did not incorporate external validation with independent datasets or in clinical settings. As a result, the generalisability of the model across different populations and healthcare contexts remains to be empirically established. These limitations highlight the need for caution in interpreting the results and underscore opportunities for further refinement.

G. Future Work

Building on the identified limitations, several avenues for future research emerge to enhance the robustness, applicability, and impact of the proposed framework. A primary direction is to integrate real-time data sources, such as wearable devices, Internet of Things (IoT) sensors, and electronic health records. Incorporating continuous physiological data would enable the digital twin to evolve dynamically, thereby improving the accuracy and responsiveness of both prediction and simulation components.

Another important area for future work is extending the digital twin simulation beyond the current 90-day horizon. Longitudinal modelling of disease progression would provide deeper insights into the long-term dynamics of diabetes and support more strategic planning for interventions. This would require incorporating time-series modelling techniques and access to longitudinal datasets.

Further research should also explore the application of advanced machine learning techniques, including deep learning architectures and hybrid models, to enhance predictive performance. In particular, integrating explainable artificial intelligence (XAI) methods would improve model transparency and build clinical trust, addressing a key barrier to the adoption of AI-driven healthcare systems.

In addition, future studies should prioritize external validation of the proposed framework across diverse populations and healthcare settings. Conducting pilot implementations in clinical environments would provide valuable insights into system usability, effectiveness, and integration with existing healthcare workflows.

While the selected models provide a representative comparison of classical and ensemble-based approaches, the study does not include recent deep learning architectures or transformer-based models that have shown promise in healthcare prediction tasks. Future work should incorporate such models to provide a more comprehensive benchmarking framework.

Future work may explore hybrid data-driven and physiological simulation models to enhance the fidelity of digital twin representations.

Finally, the conceptual framework developed in this study can be extended beyond diabetes to cardiovascular conditions and hypertension, where continuous monitoring and personalized intervention are equally critical. Such extensions would further demonstrate the scalability and versatility of the integrated ML–DT approach in advancing digital health and precision medicine.

H. Conclusion

The study demonstrates the feasibility of linking predictive modelling with simplified simulation to support trajectory-based interpretation of diabetes risk. The results highlight the potential of such integrated approaches for enhancing decision support in data-constrained healthcare environments. By leveraging a structured CRISP-DM methodology, the study systematically developed and evaluated multiple machine learning models, identifying Gradient Boosting as the most effective approach for predicting diabetes risk within the given dataset. The integration of the digital twin component extended the system's predictive capability by enabling dynamic simulation of patient-specific risk trajectories over a 90-day period.

The findings demonstrate that combining predictive analytics with simulation provides a more comprehensive and actionable approach to diabetes management. While machine learning models provide accurate risk predictions, the digital twin framework transforms these predictions into forward-looking insights, thereby enabling proactive, personalized healthcare decision-making. This integration addresses a critical gap in existing research, where predictive models are often limited to static outputs and lack temporal context.

From a theoretical perspective, the study contributes to the growing body of knowledge on digital health by proposing a hybrid ML–DT framework that bridges the gap between prediction and simulation. From a practical standpoint, the system offers a scalable architectural framework that can be implemented in resource-constrained environments.

Overall, this work demonstrates that integrating machine learning and digital twin technologies offers a practical pathway to more anticipatory, data-driven healthcare systems. While further validation is required, the framework provides a foundation for bridging predictive analytics and simulation, supporting a gradual transition from reactive to forward-looking approaches in chronic disease management.

I. References

[1] M. J. Hossain, M. Al-Mamun, and M. R. Islam, "Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused," *Heal. Sci. Reports*, vol. 7, no. 3, pp. 5–9, 2024, doi: 10.1002/hsr2.2004.

[2] H. Sun *et al.*, "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 183, p. 109119, 2022, doi: 10.1016/j.diabres.2021.109119.

[3] B. Mutunhu, B. Chipangura, and S. Singh, "Towards a quantified-self technology conceptual framework for monitoring diabetes," *South African J. Sci. Technol.*, vol. 43, no. 1, pp. 69–84, 2024, doi: 10.36303/SATNT.2024.43.1.970.

[4] B. Mutunhu, B. Chipangura, and H. Twinomurinzi, "Internet of Things in the Monitoring of Diabetes," *Int. J. Heal. Syst. Transl. Med.*, vol. 2, no. 1, pp. 1–20, 2022, doi: 10.4018/ijhstm.300336.

[5] Q. Zhang and L. Ma, "From Perceptions to Actions: Unveiling Factors Influencing Diabetic Patients' Attitudes and Compliance Intention with Smart Blood Glucose Management Technology," *Int. J. Hum. Comput. Interact.*, pp. 1–18, 2024, doi: 10.1080/10447318.2024.2366015.

[6] N. L. Rane, M. Paramesha, S. P. Choudhary, and J. Rane, "Machine Learning and Deep Learning for Big Data Analytics : A Review of Methods and Applications Partners Universal International Innovation Journal (PUIIJ)," no. June, pp. 172–197, 2024, doi: 10.5281/zenodo.12271006.

[7] B. M. Ndlovu, B. Chipangura, and S. Singh, "The readiness to use quantified self-technology : A case of diabetic patients from a hospital in Bulawayo , Zimbabwe," *Digit. Heal.*, vol. 11, pp. 1–13, 2025, doi: 10.1177/20552076251376286.

[8] J. K. Paul, M. Azmal, O. F. Talukder, A. N. M. S. N. B. Haque, M. Meem, and A. Ghosh, "Harnessing machine learning for improved diagnosis, drug discovery, and patient care," *Comput. Struct. Biotechnol. Reports*, vol. 2, p. 100051, 2025, doi: https://doi.org/10.1016/j.csbr.2025.100051.

[9] N. H. Alhumaidi, D. Dermawan, H. F. Kamaruzaman, and N. Alotaiq, "The Use of Machine Learning for Analyzing Real-World Data in Disease Prediction and Management: Systematic Review.," *JMIR Med. informatics*, vol. 13, p. e68898, Jun. 2025, doi: 10.2196/68898.

[10] I. Murere, B. Ndlovu, S. Dube, and M. Muduva, "Comparative Analysis of Machine Learning Techniques for Predicting Diabetes," vol. 2030, no. Mozaffarian 2020, 2024, doi: 10.46254/EU07.20240073.

[11] B. Ndlovu, "A Personalized Generative AI Model for Diabetes Drug Discovery:

Integrating Molecular and Clinical Data Using Variational Autoencoders (VAE)," *Indones. J. Comput. Sci.*, vol. 15, no. 1, pp. 79–100, 2026.
[12] P. Murwirapachena, B. Ndlovu, and K. Maguraushe, "Generative Artificial Intelligence in Drug Discovery for Diabetes : A Systematic Literature," in *Smart Computing Paradigms: Intelligence and Network Applications. SCI 2025.*, 2026, vol. 1, pp. 111–123, doi: 10.1007/978-3-032-08240-4.
[13] S. Hadebe, B. Ndlovu, and K. Maguraushe, "Managing Diabetes Using Machine Learning and Digital Twins," *Indones. J. Innov. Appl. Sci.*, vol. 5, no. 2, pp. 145–162, 2025, doi: 10.47540/ijias.v5i2.1981.
[14] S. Dihan *et al.*, "Heliyon Digital twin : Data exploration , architecture , implementation and future," *Heliyon*, vol. 10, no. 5, p. e26503, 2024, doi: 10.1016/j.heliyon.2024.e26503.
[15] M. Iliu, "applied sciences Digital Twin — A Review of the Evolution from Concept to Technology and Its Analytical Perspectives on Applications in Various Fields," 2024.
[16] J. Corral-acero *et al.*, "The ' Digital Twin ' to enable the vision of precision cardiology," pp. 4556–4564, 2020, doi: 10.1093/eurheartj/ehaa159.
[17] L. Moreno and J. L. Peck, "Nurse Practitioner–Led Telehealth to Improve Outpatient Pediatric Tracheostomy Management in South Texas," *J. Pediatr. Heal. Care*, vol. 34, no. 3, pp. 246–255, 2020, doi: 10.1016/j.pedhc.2019.11.008.
[18] H. K. Rudsari *et al.*, "Digital twins in healthcare : a comprehensive review and future directions," no. November, pp. 1–25, 2025, doi: 10.3389/fdgth.2025.1633539.
[19] C. Mosquera-Lopez and P. G. Jacobs, "Digital twins and artificial intelligence in metabolic disease research.," *Trends Endocrinol. Metab.*, vol. 35, no. 6, pp. 549–557, Jun. 2024, doi: 10.1016/j.tem.2024.04.019.
[20] G. Menon, B. Malave, M. Mhaske, S. Parjane, and H. Mhaismale, "Digital twin technologies in medicine: The innovations, barriers, and future directions," *Intell. Hosp.*, p. 100043, 2025, doi: https://doi.org/10.1016/j.inhs.2025.100043.
[21] F. S. Rad, R. Hendawi, and X. Yang, "Personalized Diabetes Management with Digital Twins : A Patient-Centric Knowledge Graph Approach," 2024.
[22] C. Meijer, H. Uh, and S. Bouhaddani, "Digital Twins in Healthcare : Methodological Challenges and Opportunities," 2023.
[23] J. W. Syed Hasib Akhter Faruqui, Adel Alaeddini, Yan Du, Shiyu Li, Kumar Sharma, "Nurse-in-the-Loop Artificial Intelligence for Precision Management of Type 2 Diabetes in a Clinical Trial Utilizing Transfer-Learned Predictive Digital Twin," *J. Med. Internet Res.*, pp. 1–15, 2024.
[24] P. Thamotharan *et al.*, "Human Digital Twin for Personalized Elderly Type 2 Diabetes Management.," *J. Clin. Med.*, vol. 12, no. 6, Mar. 2023, doi: 10.3390/jcm12062094.
[25] R. Wirth, "CRISP-DM : Towards a Standard Process Model for Data Mining," in *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 2000, no. 24959, pp. 29–39, [Online]. Available: https://api.semanticscholar.org/CorpusID:1211505.

[26] D. B. Olawade *et al.*, “Diabetes Research and Clinical Practice Digital twin paradigm in diabetes prediction and management,” *Diabetes Res. Clin. Pract.*, vol. 231, no. December 2025, p. 113075, 2026, doi: 10.1016/j.diabres.2025.113075.

[27] R. Wijaya, F. Saeed, P. Samimi, and A. M. Albarrak, “An Ensemble Machine Learning and Data Mining Approach to Enhance Stroke Prediction,” 2024.

[28] R. Saha, S. Saif, T. Khan, and A. Alabdultif, “A critical systematic review of ensemble learning methods for predictive healthcare,” 2026.

[29] N. Malandrino *et al.*, “Body Mass Index and Diabetes Incidence Across the Adult Lifespan : The Baltimore Longitudinal Study of Aging,” *J. Endocr. Soc.*, vol. 8, no. 11, pp. 1–10, 2024, doi: 10.1210/jendso/bvae156.

[30] M. G. Samuelsson *et al.*, “Evaluation of Type 2 Diabetes Risk in Individuals With or Without Metabolically Healthy Obesity,” vol. 6, pp. 1–17, 2025.

[31] O. Mabikwa, B. Ndlovu, and K. Maguraushe, “A Comparative Analysis of Machine Learning Techniques and Explainable AI on Voice Biomarkers for Effective Parkinson ’ s Disease Prediction,” vol. 7, no. 3, pp. 2196–2228, 2025, doi: 10.51519/journalisi.v7i3.1172.

[32] T. Ngwazi and B. Ndlovu, “Early Detection of Diabetic Retinopathy Through Explainable AI Models : A Systematic Review,” *Int. J. Informatics Dev.*, vol. 14, no. 2, pp. 616–628, 2025, doi: 10.14421/ijid.2025.5200.

[33] S. Harrison, S. U. Islam, and G. Epiphaniou, “Digital health interventions use cases : Classi fi cation and taxonomy development , a scoping review,” vol. 11, pp. 1–26, 2025, doi: 10.1177/20552076251375958.