

Design and Evaluation of a Hybrid Intelligence Framework for Explainable Diabetes Risk Prediction and Clinical Decision Support

Utsha Sarker

Department of AIT-CSE
Apex Institute of Technology
Chandigarh University, Punjab,
India

utsha.sarker00775@gmail.com

Archy Biswas

Department of AIT-CSE
Apex Institute of Technology
Chandigarh University, Punjab,
India

archyz2021@gmail.com

Shalini Kush

Department of AIT-CSE
Apex Institute of Technology
Chandigarh University, Punjab,
India

shalinikush973@gmail.com

Md. Mosfiqur Rahman Shuvo

Department of AIT-CSE
Apex Institute of Technology
Chandigarh University, Punjab,
India

mdmosfiqurrahman7672@gmail.com

Lalit Vaishnav

Department of AIT-CSE
Apex Institute of Technology
Chandigarh University, Punjab,
India

vlalith7036@gmail.com

Ashwin Puri Goswami

Department of CSE
Apex Institute of Technology
Chandigarh University, Punjab,
India

ash200459@gmail.com

ARTICLE INFO Received: 05 Jul 2026 Revised: 17 Aug 2026 Accepted: 24 Aug 2026

Abstract: In this study, we propose a hybrid intelligence system for diabetes-risk prediction and clinical decision support, which remains favorable from the perspective of explainability. In this study, a hybrid intelligence approach for explaining diabetes risk prediction and clinical decision support is presented, which remains favorably by the perspective of explainability. The framework tackles the challenges of the traditional healthcare predictions systems that might give risk estimations without providing a clear explanation, evidence-based knowledge, and interactivity decision support. The architecture proposed suggests that the system includes components for data preprocessing, predictive machine learning, software agents, retrieval augmented generation (RAG), large language model (LLM) for medical knowledge retrieval and explainability using SHAP. The proposed architecture includes components such as data preprocessing, predictive machine learning, multiple software agents, retrieval augmented generation (RAG), large language model (LLM) for medical knowledge retrieval, and explainability using SHAP, along with an interactive Streamlit dashboard. The experiments were performed on a high number of diabetes records (100,000+) where the variables comprised a set of demographic, lifestyle, and clinical variables, and the goal was that diabetes should be the prediction target. All the models were trained and evaluated using accuracy, precision, recall, F1 score, ROC-AUC, PR-AUC, Matthews correlation coefficient,

Brier score and confusion-matrix analysis. A hybrid ensemble of the predictive probabilities of the seven models was created next based on the validation process. With a validation chosen decision threshold of 0.79, the proposed hybrid model achieved 97.22% accuracy, 81.34% F1 score, 97.82% ROC-AUC and 88.58% PR-AUC on an independent test set having 15,000 records. The proposed architecture for decision support may include the integration of SHAP-based explanations for their transparent global and patient-level interpretation of model outputs [25] and responsible (and trustworthy) principles in the context of healthcare AI [23, 34]. The framework shows how these four techniques together (predictive machine learning, explainability, knowledge-grounded generation, and agent-based interaction) can be combined to provide transparent support for the decision on using diabetes-risk. But it is still in development, so it needs to be validated outside of clinical settings, calibrated, evaluated as a proto-type, validated as a fair device, and monitored in clinical usage before being deployed in real life. [35] [36] [38]

Keywords: Hybrid intelligence, Healthcare informatics, Explainable Artificial Intelligence, Clinical decision support, Diabetes risk prediction, retrieval-augmented generation, multi-agent system.

I. INTRODUCTION

A. HEALTHCARE INFORMATICS AND ARTIFICIAL INTELLIGENCE

Clusters of structured clinical and health-related data have emerged in the healthcare field as a huge opportunity for Artificial Intelligence (AI) and Machine Learning (ML) informatics. The machine-learning algorithms can connect sets of demographic, behavioural, and clinical parameters to stringent rules, and convert the connections into forecasts that could be utilized for risk benchmarking or clinical choices. The potential as well as the role of using ML for diabetes prediction has been exhibited in previous studies [6, 7] and broader research has illustrated the increasing importance of using of AI to extract clinically meaningful information from EHD [20, 21]. However, a high degree of predictive performance gives no guarantee of being a good clinical decision support.

However, there are multiple functions that need to be performed to have a complete clinical decision support workflow. Prediction is related to the probability or group/class associated to the clinical outcome; explanation to all of the factors that determine the probability; evidence retrieval to providing essential medical knowledge that understands a result; and interpreting and decision making by appropriate qualified health care professionals to making clinical actions. These functions should be distinct in their conceptual nature so that even if the prediction is accurate, it will not necessarily result in the provision of an explanation or supporting evidence or appropriate clinical recommendations. Thus, human input is key to healthcare decision-making where AI-generated information is integrated into healthcare decisions-making [22, 23].

In this study the application domain is diabetes-risk prediction to research how these complementary functions may be incorporated into a single hybrid intelligence model.

B. NEED FOR EXPLAINABLE DECISION SUPPORT

However, one of its challenges is that, for advanced predictive models, it's important to understand that they lack interpretability. A typical black-box classifier can be fed with a patient and predict with a high certainty that the patient has a high diabetes-risk whereas being poorly equipped to explain which variables are responsible for the prediction. This can constrain transparency and computing the ability of the users to critically assess the output of the computer model. In the medical domain, explainability is crucial as predictions can impact subsequent evaluations,

actions, or surveillance related to the patient's health [27], [29].

When paired with an AI model such as a large language model (LLM), predictive ML can become more relevant. Though statements generated by an LLM might seem plausible in natural language, they could be well-meaning but mistakenly lacking in solid and trustworthy evidence. Thus, language generation cannot be used as evidence of clinical testing. It is proposed to divide predictive inference from knowledge retrieval, and then use retrieval-augmented generation (RAG) to anchor generated knowledge with retrieved medical knowledge. The main purpose of this design is to increase the visibility of the beacons, with the human review step remaining as the last step in the decision making process.

In this study, SHapley Additive exPlanations (SHAP) is combined with model-level interpretation. SHAP gives the individual features contribution values for a specific prediction by the model from a single model into a unified feature-attribution framework [25]. It can thus enable both a global interpretation of influential risk factors, as well as a local interpretation for individual patient predictions. But, explainability is not sufficient to ensure clinical validity or trustworthiness: others are equally important, such as data quality and quality, bias, data and model validation, transparency, safety, and appropriate human oversight [23, 27, 34].

C. HYBRID INTELLIGENCE CONCEPT

In this research, hybrid intelligence is intended as intelligent coordination of complementary computational and knowledge-based parts and not as machine intelligence on par with that of humans. The proposed model will comprise elements of machine learning systems, content that is readable for the human, information-retrieval systems, the language generation systems based on LLM, specific software agents, explainability approaches, and additional human supervision.

The diabetes risk estimation is obtained from the ML layer and the explainability layer offers feature level interpretations. Retrieval and LLM are built to deliver evidence-based contextually relevant. Specialized software agents are responsible for organizing actions such as data processing, prediction, risk assessment, explanation, knowledge retrieval and presentation of decisions for support. The resulting architecture is designed with the presumption that it will be used in parallel with professional judgment, as the overall notion of HAI integration in healthcare presents. The resulting architecture is designed in parallel with the usage of professional judgment, as the general idea of combining HAI in medicine.

This study analyzed 100,000 subsampled records of the experimental dataset with 16 variables such as demographic, lifestyle and clinical data. The binary target is highly imbalanced -- 91,500 non-diabetic and 8,500 diabetic patients are observed. So other than accuracy, there are other not-so-familiar metrics like precision, recall, F1 score, PR-AUC and ROC-AUC in the realm of model evaluation. This is crucial since a high accuracy may obscure a lack of identification of the minority diabetic class.

D. PROBLEM STATEMENT

Predictive machine learning, explainable AI, and clinical decision-support are AI approaches which have been demonstrated in healthcare research to date [6], [11], [17], [27]. Yet, embedding these aspects of disease-risk prediction, explainability, evidence retrieval, generation of disease-relevant language, specialized agent coordination, and measurable evaluation of the system, into an interactive workflow is a major practical challenge. Moreover, the use of AI systems in health care is not appropriate unless outputs are transparent, traceable, verified and suitably supervised by and managed in the presence of humans [23,34]–[36].

Therefore, this study aims at providing a solution for the problem by creating and evaluating a modular hybrid intelligence prototype for explaining diabetes-risk prediction and clinical decision support featuring predictive modelling, SHAP explanation, controlled RAG, specialized software agents and interactive Streamlit interface.

E. RESEARCH CONTRIBUTIONS

This study's key contributions were that:

1. An approach to the AI system architecture design that is modular in terms of both intelligence and hybrid, where the cognition is separated from and orchestrated with data processing, predictive modelling and explainability, knowledge retrieval, language generation, risk assessment and decision-support functions.
2. Evaluation of the predictive performance of 7 different models (Logistic Regression, Random Forest, Extra Trees, HistGradientBoosting, XGBoost, LightGBM and CatBoost) with 100,000 samples of diabetes disease with several classification and discrimination metrics.
3. Implementing explainability by integrating with the concept of the SHAP (SHapley Additive Percent explained) variation loss, as described in [25], which would allow the interpretation of the model on the global level and help

interpret the model on the patient level as individual predictions of diabetes risk are built.

4. Controlled RAG/LLM knowledge layer to link-up predictive outputs and retrieved medical information and at the same time ensuring a separation between model prediction, retrieved evidence, generated explanation and expert clinical judgement.

5. Authentication of coordination of specialized computational components within a human-machine system — key requirement of responsible healthcare AI with interactive prototype/decision-support tool built with Streamlit. Prototype/decision-support application (using Streamlit) for coordinating between specialized computational components, while preserving human oversight as a key requirement for 'responsible healthcare AI' [23], [34].

F. RESEARCH QUESTIONS

With these aims in mind the following questions are explored in this study:

Based on the given data set(s),

RQ1: Which of the 7 models evaluated is most successful at predicting the type of diabetes for each patient?

RQ2: Can predictive risk estimates, explanation of features using SHAP and retrieved medical evidence be easily combined in a single and interpretable workflow for decision support?

RQ3: To what extent is it possible for specialized software agents to coordinate the data preparation, prediction, risk assessment, explainability, knowledge retrieval and decision-support data presentation within the hybrid architecture between Agent servers and intelligence?

RQ4: What are the key issues of interpretability, safety, validation, fairness and deployment that need to be tackled to introduce such a prototype into the clinic?

Therefore, the proposed framework is as a classification model but also as an integrated prototype of a healthcare-informatics system. As with the responsible and clinically evaluated AI systems provided already [23, 35, 36], its output is meant for further study and expectation of defining transparent mechanisms in AI assisted decision making not for replacing or diagnosing the healthcare professional.

The purpose of the proposed framework is partly to link the predictive performance to interpretability, evidence retrieval, modular, decision-support capabilities and human oversight. The traditional prediction approach is insufficient in getting transparency or endorsement for healthcare decision support. For this reason, the study proposed here combines three critical components in a limited, yet important, decision-support process: machine-learning prediction; explanation using SHAP; retrieval-augmented generation (RAG); and specialized software agents; finally, an interactive interface will coordinate the three modules. The overall motivation and research problem is summarized as shown in Fig. 1.

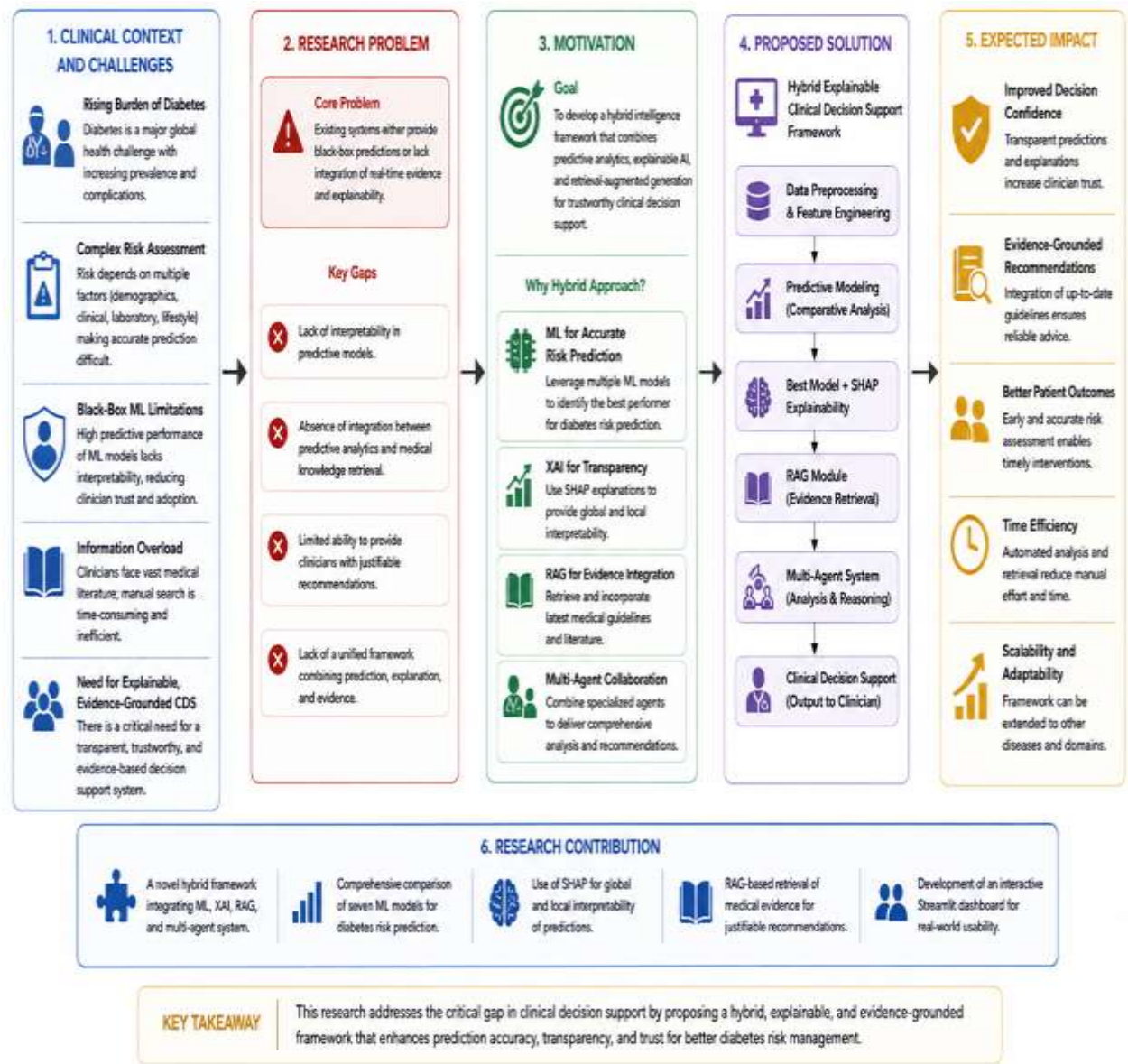


Fig. 1. Motivation and research problem: explainable clinical decision support (ExDCS) in the hybrid context

The framework distinguishes three functions – prediction, explanation, evidence retrieval, and presentation of conclusion for decision support – instead of referring them as a single function as shown in Fig. 1. This separation enhances the modularity, traceability and maintains human review as an important part in the proposed prototype of a healthcare informatics.

II. BACKGROUND AND RELATED WORK

A. MACHINE LEARNING FOR DIABETES PREDICTION

Prediction of diabetes risk can be cast as a supervised binary classification task, using diabetes risk predictor variables (e.g., patient-specific) and the binary outcome of whether or not the patient has diabetes. Using machine-learning methods significant potential has been shown to identify people with a high diabetes risk by demographic or clinical data [6, 7]. However, model performance can differ when applying k-average to different datasets or methods of preprocessing, class imbalance or evaluation methods [8], so the models' performance must be evaluated in relation to other models.

This data set has 15 features, a target variable (diabetes) with two values (binary: 0 or 1) and 100,000 observations. The predominant demographic, lifestyle and clinical factors are gender, age, hypertension, heart disease, smoking history, body mass index (BMI), blood sugar glucose level and the hemoglobin A1c (HbA1c) level. Age and gender gives demographic information, smoking history is a behavioral risk characteristic, BMI is an anthropometric indicator, hypertension and/or heart disease, represents clinical comorbidities and HbA1c and blood glucose provides glycemic information. Additional information included in the dataset include five "one-hot" racial indicators, year, and location. The distribution of the cases is skewed, with 8,500 cases being diabetic (8.5%) and 91,500 cases being "nondiabetic" (91.5%). It is important that, instead of just level of accuracy, minority class performance and discrimination must be taken into account.

The classifiers are comparatively evaluated by conducting seven different classifiers in the comparison. Logistic Regression not only provides a linear approximation of class probabilities but is also interpretable. Random Forest is composed of an ensemble of decision trees and aggregation to model nonlinear relationships, the underpinning methodology of random forest is well established in [40]. There's another alternative with Extra Trees which will allow the comparison of tree-ensemble with traditional Random Forest approach but with a more randomized approach.

HistGradientBoosting is an efficient boosting baseline for tabular data, based on histograms. XGBoost is a scalable gradient-boosted decision trees that contains regularization and optimized XGBoost computations [41]. LightGBM is well designed to efficiently implement tree learning via histogram, which is fit for gradient boosting with efficiency [43]. CatBoost also contains another boosting model for handling of categorical-feature structures, that proved to be

effective. These models make up a variety of experimental comparisons to the linear, bagging, randomised tree and boosting models.

B. RECOVERY-ENHANCED GENERATION IN THE HEALTH SECTOR

Retrieval-augmented generation (RAG) is a method that grounds language-model's outputs with retrieved information besides using the encoded knowledge within the language model. In health informatics incidents such as providing medical information retrieval, question answering, summarization and communication using health-informational context in decision-making can be supported by such a kind of architecture. In the proposed framework RAG is considered as a knowledge-support module that does not participate in the predictive ML model.

There are six main stages to the RAG workflow: They first gather medical documents from a controlled source, selecting which ones that they need. First, the documents become amenable to breaking them down into manageable chunks of text. Third, these chunks are then turned into numeric embeddings that encapsulate these chunks' meanings. Fourth, on the receipt of a query from a user or software agent the retrieval component finds the most salient chunks. Finally, the LLM is outputted with the retrieved context along with the question and provides a contextual answer. Lastly, the evidence sources used are presented or saved for evidence to be traced back.

The responsibility for factual accuracy of posts falls on the authors and does not imply any right or incorrectness in RAG. Even if these issues are resolved, retrieval can still be incorrect, source material can be of poor quality, the chunks retrieved may not have been aptly assembled, the user may not have written an adequate or appropriate query, and a useful prompt may not be provided. Having proposed to separate this evidence that has been retrieved from the language that has been generated and to retain human review, the question remains: this separation involves what? So what is this separation? (It is not language separation from evidence retrieval; that is being proposed to be a different separation). The design is consistent with the general guidelines of ensuring the transparency, traceability, safety and responsible use of AI across healthcare contexts [23], [34].

C. AGENTIC AND MULTI-AGENT HEALTHCARE SYSTEMS

When talking about AI-agent architectures, it's crucial to make the distinction between the terms. In this sense, the software agent is to be viewed as a computational unit with a specific function, which can be, for instance, the

preprocessing of the data, the execution of an explanatory model, the retrieval of knowledge or the generation of explanations. Agentic systems are ones that are more autonomous (planning, choosing tools, initiating actions, etc.) and have certain goals. A multi-agent system is a system of multiple agents having a different kind of role, all of them involved in the coordination.

The provided implementation is thus not considered to be an autonomous agentic clinical system, but is described as a multi-agent workflow. Its agents fuse pre-defined functions concerning the finetuning of the data, prediction, risk assessment, knowledge retrieval, explainability and presentation in the form of a decision support. This is a terminology that ensures that the prototype is not over-autonomized by the term 'autonomous' and thus keeps the distinction between computer-based coordination and clinically independent thinking. Human support is still imperative, and in line with the human-AI collaboration approach shared within healthcare literature [22], [23].

D. EXPLAINABLE ARTIFICIAL INTELLIGENCE AND SHAP

Explainable artificial intelligence (XAI): The aim of the explainable artificial intelligence (XAI) is to make the behaviour of models more comprehensible to the user. This stipulation is especially applicable in healthcare where a forecast value of the risk without its corresponding "why" rationale may not give much visibility and clarity for a sensible review [27]–[30].

SHapley Additive exPlanations (SHAP) [25] is a common technique used for calculating the feature-attribution values for each prediction, allowing for the calculation of a common value for each feature. SHAP-complimentary components can be aggregated at the global level to come up with variables that can have significant effects in a given evaluated population. SHAP values on the local level show how the values of the individual features appear when splitting one patient record to one of the two groups.

Positive effect of a feature's contribution means that it is shifting the model towards the larger diabetes risk as compared to the model's baseline prediction, whilst negative effect means that it is taking the prediction away from the diabetes risk. Importantly, these contributions don't explain organisms' behavior but just their models' behavior. For instance, a positive value on BMI's SHAP value indicates BMI's influence to that model's prediction was positive, but doesn't imply that if the person didn't have BMI, he or she wouldn't have had diabetes. It would be important to keep this separation between predictive association and causality to produce attributable results from the outputs of XAI while

demonstrating responsibility in the interpretation of the results [27, 29].

E. RESPONSIBLE AI (R.A.I.) AND CLINICAL DECISION SUPPORT (CDS)

However, to responsibly implement these clinical decision-support systems, far more is required than just predictive accuracy: they need to organise predictive output and relevant information in such a way to be understandable for the human user. Relying on a workflow process in health care that incorporates human-in-the-loop review, especially when the use of model output may impact assessment or intervention should maintain human-in-the-loop review. Responsible ML principles focuses on “safety” [23], on “appropriate validation”, on “transparency” [23] and on “possible harms”.

Data privacy and security is essential as healthcare applications can contain sensitive patient data. Additional attention should be paid to the types of bias and representativeness because the model that is developed can be sensitive to the specific population on which it is trained. It is therefore crucial to validate the performance in real clinical environments by testing it in an external dataset [36] [38].

Another critical factor is also calibration of probability functions. A good clinically useful risk model is unlikely to simply classify people accurately, it must also give a corresponding probability prediction to the actual proportion of people that fall into the self class. One aspect highlighted as a major area of weakness in the many predictive models is the issue of calibration [45]. It should also include a way of recording the computing steps used to produce evidence and of reporting uncertainty, not only as a diagnosis but as a way of representing the uncertainty. Human factors, seamless integration with workflows, safety, and field testing are also important factors when evaluating AI-enabled decision support systems early in trials [35].

F. RESEARCH GAP

The previous studies set the important foundations in the prediction of diabetes [6]; explainability of AI [11]; trustworthy AI in healthcare [17]; and trustworthy clinical decision support [23], [27], [34]. However, the capabilities are often explored individually. The necessity for such a study which involved coordinating, and measuring the integration of the individual predictive, explanatory, and decision-support technologies is thus not the absence of the individual pieces of technology; instead, individual predictive, explanatory, and decision-support technologies needed to be coordinated and measurable within a Mr. Bock system prototype.

The proposed approach merges several diabetes classification ML models through empirical comparison with model selection and interpretation of the classification with SHAP analysis, within an ML-based Risk Assessment, a Controlled Knowledge Retrieval, an LLM-powered

Information Presentation Mode, Modularized Agent Coordination, and an interactive User-facing Decision Support Model which allows the citizens to interact with the model. Notably, it is not just a clinical system proven diagnostic tool but a research prototype.

There are important building blocks present in the existing literature for healthcare RAG and agentic AI, but these research directions are concerned with various aspects of the bigger problem of supporting decision making. Given the lack of clinical superiority claim for the current work when compared to review studies, a comparison of the major abilities for the two base papers to those implemented in the proposed framework are shown in Table I.

Capability	Base Paper 1: Healthcare RAG Review	Base Paper 2: Agentic AI in Healthcare Review	Proposed Hybrid Intelligence Framework
Primary research focus	Reviews retrieval-augmented generation in healthcare, including medical question answering, clinical information retrieval, EHR-related summarization, and decision-support applications.	Reviews emerging agentic/agent-based AI approaches in healthcare, their capabilities, implementation patterns, and limited state of rigorous clinical validation.	Implements a diabetes-risk decision-support prototype combining predictive ML, modular agents, SHAP, controlled RAG, LLM-supported evidence presentation, Streamlit visualization, and an explicit evaluation protocol.
Predictive disease model	No comparative diabetes prediction pipeline is implemented as the central contribution.	No seven-model disease-risk benchmarking pipeline is implemented as the central contribution.	Seven supervised classifiers are trained and compared: Logistic Regression, Random Forest, Extra Trees, HistGradientBoosting, XGBoost, LightGBM, and CatBoost.
Empirical model selection	Not applicable to the review's scope.	Not applicable to the review's scope.	Model selection uses a predefined validation criterion rather than test-set optimization. XGBoost achieved the highest validation F1 (0.8128) and was selected before independent test evaluation.
Independent test evaluation	Not a predictive-model experiment.	Not a predictive-model experiment.	Selected XGBoost achieved test accuracy 0.9721, precision 0.9704, recall 0.6933, F1 0.8088, ROC-AUC 0.9789, PR-AUC 0.8852, and Brier score 0.0226.
RAG capability	Central topic; establishes healthcare use cases, grounding rationale, retrieval/generation workflow, and limitations.	May appear as an enabling tool in broader agentic systems but is not the principal reviewed capability.	Controlled RAG layer is designed to retrieve passages from curated, traceable medical sources and display source titles/provenance with an insufficient-evidence fallback.
Knowledge provenance	Emphasizes grounding generated responses in retrieved information and the importance of source quality.	Relevant to safe tool-using/agentic systems, although implementation varies across reviewed studies.	Every knowledge document should retain title, publisher/source, date, document type, identifier, and provenance; retrieved evidence is kept distinguishable from generated language.
LLM role	LLMs generate responses using retrieved context; correctness remains dependent on retrieval quality, prompting, and verification.	LLMs may provide reasoning, planning, tool use, or coordination functions in agentic architectures.	LLM is restricted to evidence-grounded language generation and bounded clinical-support presentation; it is not permitted to prescribe, autonomously change treatment, or make emergency decisions.
Agent architecture	Not implemented as a coordinated multi-agent predictive workflow.	Agentic and multi-agent healthcare systems are a central conceptual focus; literature remains early-stage.	Implements a safer modular multi-agent workflow with Data, Preprocessing, Prediction, Risk Assessment, Explainability, Knowledge Retrieval, and Clinical Decision Support responsibilities.
Autonomy level	Not applicable as an autonomous clinical system.	Reviews systems with varying degrees of planning/tool use/autonomy.	Prototype is intentionally bounded; the manuscript uses 'multi-agent workflow' rather than claiming human-level intelligence or unrestricted autonomy.
Explainable AI	Not a principal component of the reviewed RAG architecture.	Not consistently integrated as patient-level predictive attribution.	SHAP provides global and local model explanations [25]. Explanations explicitly describe model behavior and are not presented as biological causation [27].
Global explanation	Not central.	Not central.	Global SHAP summary identifies features that most strongly influence predictions across an evaluation sample; HbA1c and blood glucose are dominant in the current experiment.
Local explanation	Not central.	Not central.	Local SHAP explanation presents patient/sample-specific feature contributions, their direction, and

Capability	Base Paper 1: Healthcare RAG Review	Base Paper 2: Agentic AI in Healthcare Review	Proposed Hybrid Intelligence Framework
			a plain-language interpretation with a non-causality disclaimer.
Risk assessment	Not implemented as a calibrated diabetes-risk classification layer.	Not the primary reviewed function.	Prediction probability is mapped to configurable prototype categories: Low <0.30, Moderate 0.30-0.70, High >0.70; thresholds require calibration and clinical validation.
Clinical decision support	Reviewed as a healthcare RAG application area.	Discussed as a potential agentic healthcare application.	Combines prediction, probability, prototype risk category, SHAP attribution, retrieved evidence, uncertainty language, safety disclaimer, and suggested review-oriented next steps.
Human oversight	Important because RAG does not guarantee correctness.	Important because autonomous/agentic healthcare behavior creates safety and accountability concerns.	Qualified healthcare-professional review is required before any clinical action; output is explicitly not a diagnosis or prescription [23], [34], [35].
Interactive interface	Not implemented as the review's contribution.	Not implemented as the review's contribution.	Nine-page Streamlit dashboard covers Home, Dataset Upload, Multi-Agent Workflow, Model Performance, Evaluation, Explainability, RAG Knowledge, Clinical Decision Support, and Downloads.
Agent performance evaluation	Not applicable.	Review indicates limited rigorous clinical/operational evaluation in emerging agentic literature.	Protocol measures execution time, response time, throughput, peak memory, CPU utilization, success rate, failure conditions, and end-to-end workflow behavior.
Auditability	Source traceability is relevant to RAG.	Agent action traceability is important to accountability.	Logs model version, preprocessing configuration, prediction threshold, timestamps, retrieved document identifiers, source metadata, agent status/errors, and final output.
Safety positioning	RAG can reduce but does not eliminate unsupported generation.	Agentic systems require safeguards and clinical validation.	System is explicitly positioned as an experimental healthcare-informatics prototype, not a clinically validated diagnostic device.
Main contribution relative to reviews	Provides literature synthesis rather than an integrated predictive prototype.	Provides literature synthesis of agentic healthcare systems rather than the same predictive implementation.	Operationalizes complementary concepts in one modular prototype; does not claim clinical superiority over the review papers.

Table I. The Proposed Framework Technology will contrast and examine the Base-Paper Themes.

The proposed work differs mainly in terms of the level of implementation: integration (Table I). This study does not view the review papers as colliding predictive systems, but instead, all of the above elements of predictiveness: modelling, SHAP explainability, controlled retrieval, modular agent coordination, decision-support presentation, and empirical evaluation are imbued into a single prototype.

III. RESEARCH GAP, OBJECTIVES, AND SCOPE

A. RESEARCH GAP

The relevant stresses to predict the risk of diabetes have been the subject of extensive research; clinical disease can be detected from demographic, lifestyle and physiological data by machine learning methods [6]–[8]. To date, parallel research has developed explainable artificial intelligence (XAI) approaches and the use of AI in clinical decision support [17; 23; 27–30] and responsible artificial intelligence practices [24] in healthcare. At the feature level, SHAP-based methods have also shown the usefulness of interpreting diabetes predictions to provide increased

transparency for model outputs, such as in the form of an explainability dashboard [11], [25] that sits atop the model predictions.

Additionally, the literature reviewed shows these abilities being explored individually. Research on explainability focuses on model behaviour, research on clinical decision support focuses on providing information to healthcare consumers, whereas research on predictive modelling mainly emphasises on how well a model can achieve classification. There are other ways of accessing the knowledge and coordinating computations, such as emerging retrieval- or agent-oriented approaches, but detailed clinical tests of integrated AI systems are a measure to carefully evaluate the systems [34], [35].

Therefore, the gap in research tackled by this research is the actual implementation and quantifiable assessment of these complementary capabilities. Specifically, the proposed work will involve designing and implementing a holistic system that integrates a trained disease risk prediction model, patient level SHAP explainability, evidence retrieval, structured risk assessment, interface to assist in clinical practice and quantitative assessment of specific software agent performance metrics. No, this is not the first time that such an integration has been attempted – the study explores how these components can be considered, integrated, and evaluated in a systematic and a repeatable diabetes-risk decision support prototype.

B. RESEARCH OBJECTIVES

The aim of this work is to:

1. Practical gaps related to both clinical decision support (CDS) and explainable and responsible AI involving hybrid intelligence and retrieval-augmented generation (RAG) and agentic AI systems are identified to assess the existing state-of-the-art on the models and necessary adjustments in current practices for meaningful inclusion.

2. To design and implement a modular hybrid intelligence framework that will integrate the diabetes-risk prediction, diabetes-risk assessment, medical knowledge retrieval, explainability (SHAP-based), knowledge-grounded language generation and structured clinical-support output.

3. Comparatively assess the seven different predictive models: Logistic Regression, Random Forest, Extra Trees, HistGradientBoosting, XGBoost, LightGBM and CatBoost using relevant discrimination, classification and error analysis measures. 3. The diagnostic criteria are established by metrics that are more than accuracy as there are only 100,000 records belonging to the diabetic class in the experimental dataset, therefore, the metrics such as precision, recall, F1 score, ROC-AUC, PR-AUC and confusion-matrix analysis are focused on the evaluation.

4. To assess the integrated prototype at model and system level for performance prediction, explanation quality, effectiveness in retrieval, response times, computational resources usage, throughput and measures of the success of the agent task, including classification accuracy. This multi-dimensional assessment is in line with the larger demand for a transparent and clinically relevant lens for assessing AI decision-support systems [35] [36].

C. OUTLINE SCOPE OF THE STUDY

This research considers only the experimental scope of binary diabetes-risk classification based on the selected publicly available retrospective dataset (with 100,000 records). The set of data is intended for computational modelling, comparison, explainability analysis and prototype testing only and may not be used for prospective clinical validation.

The proposed product idea is not a medical device, but a prototype of research and decision-support system. To note, the output of SHAP descriptions of input features does not imply causality of the "diabetes" [25] [27] phenomenon. Likewise, the retrieval and recommendation features are supposed to offer contextually appropriate advice (for learning and evidence) which is not a personal diagnostic or therapeutic procedure.

Thus, appropriate human interpretation is necessary for predictions, explanations, retrieval of evidence, and generation of guidance. Before deciding to implement into practice, external validation, calibration assessment, representative assessment, prospective testing, and assessment by health care professionals are necessary [35, 36, 38, 45]. Final clinical judgments should be the responsibility of appropriately qualified health care professionals.

IV. NEW HYBRID INTELLIGENCE FRAMEWORK PROPOSED

A. FRAMEWORK ARCHITECTURE

The envisioned architecture assists in wrapping up the entire workflow in several interconnected functional layers: Datasourcing, Preprocessing, Predictive modelling, Risk assessment, Knowledge retrieval, Explainability, Clinical decision support, Presentation and Monitoring. The information under each of the four levels of the system progresses towards becoming a model probability, prototype risk category, interpretable feature contribution and ultimately to an evidence-informed decision-support output. A hybrid intelligence architecture is proposed, the architecture of which is presented in Fig. 2.

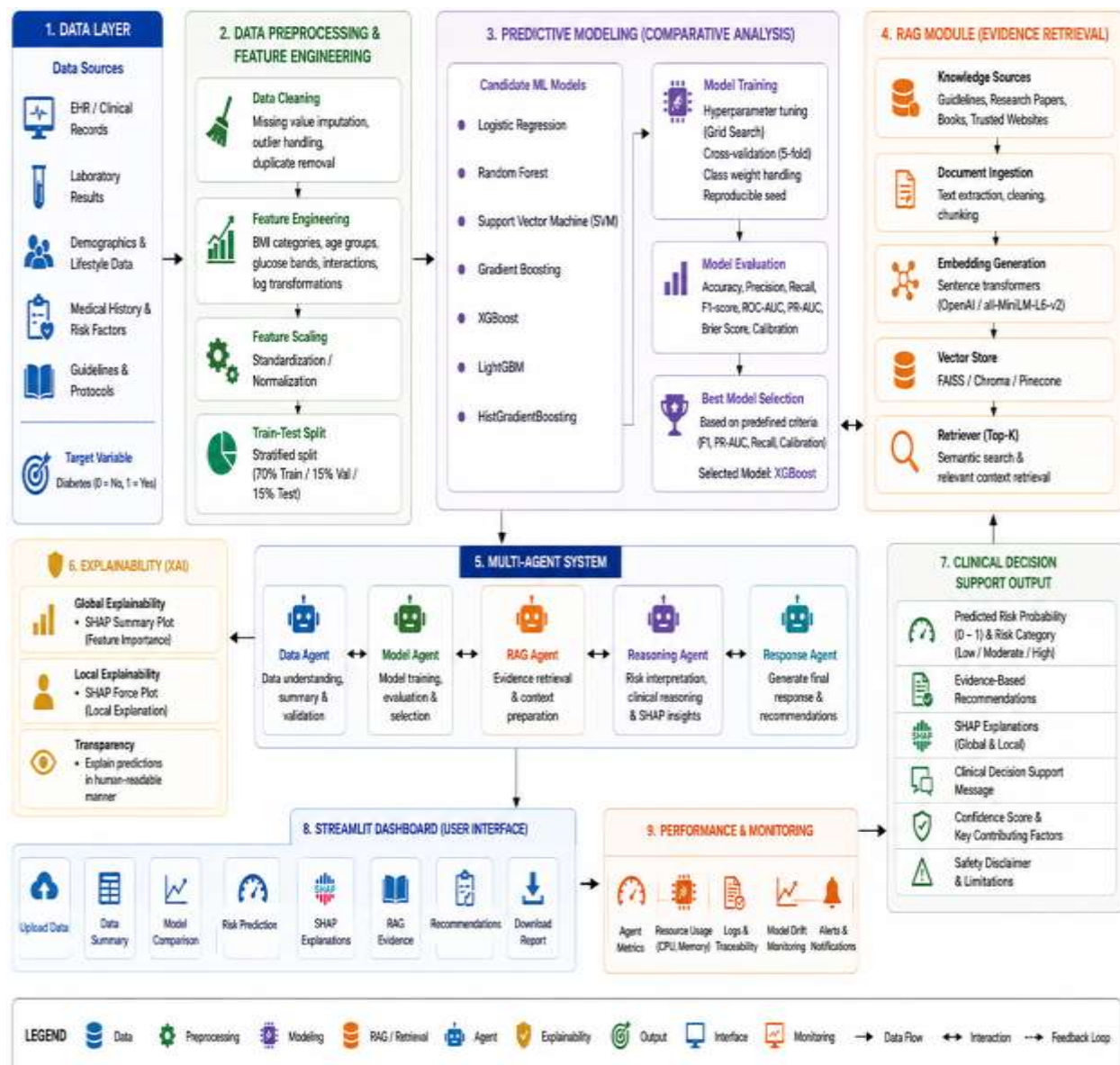


Fig. 2. Architecture of the intended hybrid intelligence framework.

The individual components interact with each other in terms of well structured outputs and have different responsibilities as illustrated in Fig. 2. This approach allows the prediction, SHAP explanation, evidence retrieval via RAG and generation of the decision support to be evaluated and audited separately. Critically, the end product is checked by humans and does not supplant clinical judgment.

Proposed framework provides a modular structure with seven layers broken into segments of predictive inference, interpretation, knowledge retrieval, decision support and user interaction. It is noteworthy that this separation is important because clinical utility, interpretability and safety are not necessarily assessed by predictive performance [23],[34] and [35].

Data ingestion and data validation/validation is handled by the data layer. The preprocessing layer processes the raw variables transforming them to representations suitable for the model, without introducing data leakage. The prediction layer runs the selected machine-learning model and outputs a predicted class and probability. The risk-assessment layer transforms from the model-probability to an interpretable prototype risk category. The knowledge-retrieval layer will

retrieve relevant evidence and give controlled context for LLM based generation. The explainability and decision-support layer provides the explainability and prediction results and retrieved information. Lastly, the presentation/monitoring layer provides output in the form of a Streamlit dashboard with the ability to keep operational records on board for traceability.

The architecture follows a multi-agent workflow, and is not an autonomous clinical agent. Each software agent does a specific job and provides structured outputs to the next agent. The human clinician or expert reviewer is not part of the automated computational flow, but is connected to the Clinical Decision Support Agent to provide interpretive and decision making function. These human–AI relationships follow the principle of using AI as a tool to assist human users rather than take over their expertise [22],[23].

B. DATA AGENT

The Data Agent is the starting point of the framework and is responsible for ensuring integrity of the data and schema in the input set. For this experiment, it's feeding on the diabetes dataset uploaded for the experiment, which has 15 predictor columns, 100,000 rows of data, and a diagnosed or not binary target column (diabetes).

The agent confirms column names and types of data, determines the target variable, detects missing values, detects double entries of observations, and notes class distributions. It also produces a data quality report for a dataset in a structured format with features, variable types, missing value counts, duplicate counts, target class frequencies, and dimensions in the dataset.

A preliminary screening of the experimental data revealed that there were no missing data, 14 rows exactly repeated and a class distribution of 91,500 non-diabetic (91.5%) and 8,500 diabetic (8.5%) data. Data quality and class imbalance can significantly influence the development and interpretation of models, hence it's crucial to record these characteristics before modelling.

C. PREPROCESSING AGENT

The Preprocessing Agent translates validated raw data types to ones suitable for the seven predictive algorithms. It is responsible for handling duplicate records, encoding missing values if applicable, categorical encoding, numerical transformations, scaling the features (if the algorithm needs it), and segregating the data into the training, validation, and test sets for the algorithm.

The dataset currently has no missing data; however, the framework still includes imputation mechanisms in case

there are missing data in future datasets. Categorical and numerical features are handled independently of each other with the ColumnTransformer, or complete transformations are wrapped in scikit-learn Pipeline objects.

One of the essential design considerations is data leakage. First it is necessary to divide the dataset into development subsets and then determine the parameters for the pre-processing steps only from the development data. Thus, no data transformation by imputation statistics, no categorical encoding, no scaling parameters or any other data dependent transformation should be fitted by prospective validation or test data. This ensures that model evaluations are done independently and that performance is not inflated.

The training, validation and test records are set to 70, 15 and 15 respectively for the present experiment, thereby emulating the stratified 70%/15%/15% design. The proportion of minority diabetes cases remains constant across subgroups after stratification.

D. PREDICTION AGENT

The Prediction Agent performs the tasks carried out by the chosen predictive component found during the comparative evaluation of the models. We evaluate 7 different candidate models: Logistic Regression, Random Forest, Extra Trees, HistGradientBoosting, XGBoost, LightGBM, CatBoost. There are known methods for taking into account nonlinear relationships in structured data like random forest and modern gradient boosting methods [40, 41, 43].

After model selection, the Prediction Agent loads the model and related preprocessing configuration from storage, validates some data regarding a patient, and produces a binary class prediction, plus a predicted probability value. The probability remains because the information it carries forth is much more useful for assessing risks and making decisions downstream, than a simple binary label.

The agent also tracks the identifier as well as the version of the model and the prediction time. Additionally, the agent keeps the identifier, model version, prediction time, model preprocessing version and decision threshold of the model. The output is then sent independently to the two other agents—Risk Assessment Agent and Explainability Agent—the separation layer being maintained with prediction and interpretation.

E. RISK ASSESSMENT AGENT

Rather, the Risk Assessment Agent translates the likelihood into a digestible risk level category to the human form. There are three sets of customizable categories defined for the decision support interface:

Low risk: predicted probability < 0.30

If the predicted probability is between 0.30 and 0.70, then the risk is moderate. If the predicted probability is 0.30 to 0.70 then the risk is deemed moderate.

High risk: > 0.70 (predicted probability)

These thresholds are representative levels of communication – and should not be considered as clinically validated thresholds of risk for children with diabetes. They are designed to help visualization and evaluation of the design of the interface.

Most importantly, the thresholds are different from the binary classification threshold of the model. The threshold for the experimental evaluation was set to 0.79 with optimization of the F1 scores on the validation set and tested on the test set without any changes to the data. Rather, the 0.30 and 0.70 values are used for the selection of categories for presentation, which can be adjusted. Probability calibration and risk-category thresholds should also be independently validated prior to clinical use as it is essential to have reliable estimates of probability for clinical prediction [45].

F. KNOWLEDGE RETRIEVAL AGENT

The controlled RAG workflow is the mechanism to provide the evidence-based grounding aspect of the framework in the Knowledge Retrieval Agent. It has no plans to collect any random material from online sources, but it will have vetted information from trusted health bodies, national health agencies, trusted clinical guidelines and from scientific literature that has been peer reviewed.

All of the knowledge-base documents are linked with metadata such as document title, document source/publisher, document date, document type, source origin and retrieval ID. A semantically meaningful decomposition of the documents in meaningful parts, then it is converted into embedding representations. From this information, similarity-based retrieval provides all the relevant passages to the LLM for each of the decisions support queries.

The Instruction to the LLM is to 'make inferences based on the provided evidence and to avoid extending beyond the context' provided. When passages are retrieved, and their source titles, they're kept and offered to the user. If the retrieval system finds the evidence that is relevant to or informative of the medical statement is not sufficient, then it is appropriate to say "insufficient evidence retrieved" and NOT generate an unsupported medical statement.

This architecture does not presume that RAG will completely remove hallucination or ensure the medical correctness. Quality and validity of the retrieval, the validity of the sources, the currency of the documents, a good embedding, construction of the prompts, and verification of the output by human means remain significant factors in the reliability of the output. These controls are among those that look towards the overall goals of healthcare AI transparency and accountability [23] [34].

G. EXPLAINABILITY AGENT

For explaining individual predictions, a single framework for interpreting the features of a predictive model, the Explainability Agent is used, which uses SHAP [25]. Incorporated, especially in cases where predictive algorithms are integrated into workflows in healthcare decision-making [27]–[30] is explainability, which is closely related to transparency.

At the global level the agent produces a summary analysis of the SHAP values that reflect the relative contributions of characteristics throughout the population being evaluated. It provides a simple explanation for a chosen patient record, describes those features that are most influential and the contributions of these features and the direction of the influence.

Positive SHAP has pushed the model output closer to the predicted diabetes risk than the model is set to the reference diabetes risk, while negative has pushed the model output towards the other direction. A simplification of the scope of the invention might read:

One of the major features was that glucose values and body mass index (BMI) by itself contributed to a higher value to the model's proposed risk. These contribution statements reflect pattern learning by the predictive model, and do not indicate that each pattern led the patient to his/her condition.

This is because SHAP does not explain biological causes of events, rather, model causes of events [25, 27].

H. CLINICAL DECISION SUPPORT AGENT

Four major information flows are used for the Clinical Decision Support Agent: model prediction, prototype risk category, explanation of why using the SHAP principle, and obtained medical evidence. It then develops a structured avenue of response to this decision-supportification: a product that entails an estimate of risk, key model features, context of the evidence, acknowledgment of uncertainty, safety disclaimer, and next-step guidance that is appropriately bounded.

There is a deliberate focus on preventive communication and not targeting the readers to diagnose them. Examples are “According to the model, .; “May require comprehensive review by a specialist in health care,...; and “Discussion with an experienced health care professional is advised:...; All outputs that are relevant make it clear that the output is not diagnostic.

The LLM does not have the authority to make clinical decisions in emergency settings, make changes to treatment, recommend dosing of medications or prescribe medications. The final interpretation continue to be under the responsibility of human beings. Restrictions have been imposed obeying the principles of healthcare-AI such as the need to assess early-stage clinical decision-support systems for their safety, human factors and work-flow aspects [23, 35].

The capability of handling multiple agents and audits or the creation of one specific agent with several audits.

We use structured data objects to communicate between agents, instead of freeform exchanges of natural language. A representative object will be:

patient_id: anonymized identifier

prediction: 1

probability: 0.78

risk_category: high

top_features: [feature contributions]

retrieved_sources: [source identifiers]

The date and time that the sample was taken.

The identifier for the model that was stored. Identification of the model that is stored.

This provides agents in the downstream part of the chain with access to the same outputs (standardization of output) and also helps to decrease ambiguity in the relationship between prediction, explanation, evidence, and generated guidance.

The monitoring layer not only collects the status of the input stage validation but also tracks such information as the agent after it started and took various times to complete, the duration of the execution, the identifies of the retrieved sources, versions of pre-processing of the data, versions of the model, the output of the prediction information, and the final decision support response provided by the Agent. These records provide a historical log of task success,

latency, throughput, error and system reliability that can be evaluated to find out the agent's performance in completing tasks over time.

The Streamlit dashboard is the final presentation interface where prediction probabilities, prototype risk categories, evidence retrieved, SHAP explanations and bounded decision-support guidance are all in one single view. However, if information is presented with an interactive interface, this does not mean that it is a clinical validation. Like in trusted, clinically evaluated healthcare AI [34]–[36], the framework needs to be externally validated, calibrated, assessed in a prospective study, evaluated for fairness, and tested with healthcare staff before being considered for real-world, clinical use.

V. MATERIALS AND METHODS

A. DATASET DESCRIPTION

The Diabetes Clinical Dataset which is publicly accessible dataset from Kaggle and provided for this study as diabetes_dataset.csv, was used for the experimental evaluation. The data file analyzed has 100,000 observations and 16 columns: the 15 columns are candidate predictor variables, the 16th is diabetes, a binary variable that indicates if the data was taken from a diabetic patient. The demographic, geographic, race-indicator, lifestyle, comorbidity, anthropometric, and glycemic variables are included in the present dataset, which differs from the widely-used Pima Indians Diabetes Dataset in that it includes variables other than those typically used in the study of diabetes.

Predictors include: hypertension, age, gender, location, heart_disease, race:AfricanAmerican, race:Asian, race:Caucasian, race:Hispanic, race:Other, blood_glucose_level and hbA1c_level, and smoking_history. The diabetes is a binary class 'positive diabetes' (class 1) vs 'not diabetes' (class 0). The data set is highly imbalanced with a 91.5% of the 92,300 samples recorded as being negative, and a minority of 8,500 samples recorded as being positive.

These are also identified by direct inspection: three string valued categorical variables such as Gender, Location and Smoking_history. Then, age, BMI, and HbA1c are stored as floating point values while year is stored as an integer, blood-glucose level is stored as an integer, and the five race indicators, as well as hypertension, and heart disease and diabetes are stored as integers. With the race variables and those of hypertension and heart disease, a numerical representation of a collection of very specific variables, but, in terms of meaning and data, they are still categorical binary variables.

There was no evidence of any null values in the 16 columns. But 14 exact duplicate lines (0.014%) were found. In the pre-processing phase a duplicate detection is therefore employed to remove the contribution from repeated observation in the development of the model.

It is based on the data provided on Kaggle for research oriented computational evaluation. Since these tags are not

actually part of the exported CSV file, there is no specific license designation that can be inferred from the data file. The specific Kaggle Data page, version, access date, and License (as displayed) should be documented in the final reproducibility documents before the publication of the work. This helps to avoid objections to claims of redistribution and permitted use.

The presented dataset for diabetes includes different demographic, clinical, behavioral and lab variables that need various pretreatment or validation. This is done to clarify and make the findings reproducible by using the definition and respective treatments of each variable as shown in Table II.

Variable	Meaning / Analytical Role	Stored Data Type	Observed Missing Pattern	Validation and Quality Checks	Preprocessing Treatment	Modeling / Governance Consideration
year	Year/context associated with the record; may encode temporal or collection-context information.	Integer	0 null values detected.	Check plausible range, consistency across records, and whether year inadvertently proxies site or collection changes.	Retain in primary analysis; scale only for models requiring standardized continuous inputs; fit any transformation on training data only.	Interpret cautiously because temporal/context variables can affect transportability and may capture nonclinical distribution shifts.
gender	Recorded gender category used as a demographic predictor and potential fairness subgroup.	Categorical string	0 null values detected.	Inspect unique categories, spelling, rare categories, and consistency of representation.	Encode within training-fitted preprocessing pipeline; preserve mapping for reproducibility.	Evaluate performance across relevant gender groups; predictive use does not establish causal clinical effect.
age	Age in years; continuous demographic/clinical risk-related variable.	Float/numeric	0 null values detected.	Check nonnegative and plausible human age range; inspect outliers and distribution.	Retain continuously; scaling applied for linear/distance-sensitive models where required; tree models may use raw values.	Global SHAP indicates meaningful model contribution; attribution remains model-specific.
location	Recorded geographic/location category; may reflect population/site context.	Categorical string	0 null values detected.	Inspect unique levels, rare categories, spelling, and whether categories encode data-source differences.	Training-only categorical encoding; unknown categories handled explicitly during inference.	Potential source of geographic bias and reduced external generalizability; subgroup evaluation is recommended.
race:AfricanAmerican	Binary race indicator for African American category.	Integer/binary	0 null values detected.	Validate domain strictly as 0/1 and inspect mutual consistency with other race indicators.	Retain as supplied binary feature; no scaling required for tree models.	Sensitive demographic variable requiring fairness, subgroup performance, and governance review.
race:Asian	Binary race indicator for Asian category.	Integer/binary	0 null values detected.	Validate domain strictly as 0/1 and inspect consistency across race indicators.	Retain as supplied binary feature.	Sensitive demographic variable; assess error rates and calibration across relevant subgroups.
race:Caucasian	Binary race indicator for Caucasian category.	Integer/binary	0 null values detected.	Validate 0/1 coding and race-indicator consistency.	Retain as supplied binary feature.	Use requires careful fairness interpretation; dataset representation does not guarantee population representativeness.
race:Hispanic	Binary race indicator for Hispanic category.	Integer/binary	0 null values detected.	Validate 0/1 coding and race-indicator consistency.	Retain as supplied binary feature.	Include in subgroup/fairness analysis; avoid causal interpretation of learned associations.
race:Other	Binary indicator for other race category.	Integer/binary	0 null values detected.	Validate 0/1 coding and inspect heterogeneous composition of the category.	Retain as supplied binary feature.	Broad 'Other' category may obscure heterogeneity and should be

Variable	Meaning / Analytical Role	Stored Data Type	Observed Missing Pattern	Validation and Quality Checks	Preprocessing Treatment	Modeling / Governance Consideration
						acknowledged in fairness interpretation.
hypertension	Recorded hypertension status; clinical comorbidity predictor.	Integer/binary	0 null values detected.	Validate 0/1 coding; inspect prevalence and cross-feature consistency.	Retain binary; no imputation required in supplied file.	Clinical association may contribute to prediction but SHAP/model importance is not evidence of causality.
heart_disease	Recorded heart-disease status; clinical comorbidity predictor.	Integer/binary	0 null values detected.	Validate 0/1 coding and prevalence.	Retain binary; no imputation required.	Comorbidity variable should be interpreted within the retrospective dataset context.
smoking_history	Recorded smoking-history category; lifestyle/behavioral predictor.	Categorical string	0 null values detected.	Inspect vocabulary, unknown/never/former/current categories, rare levels, and inconsistent labels.	Normalize only documented labels if needed; encode inside training pipeline; preserve mapping.	Potential behavioral and demographic confounding should be considered when interpreting model attribution.
bmi	Body mass index; continuous anthropometric predictor.	Float/numeric	0 null values detected.	Check plausible physiologic range, extreme values, and distribution; do not automatically treat zero as missing without dataset-specific evidence.	Retain continuous; scale for models requiring scaling; optional categories only in predefined secondary analysis.	Fourth among the leading global SHAP contributors in current analysis; not a causal estimate.
hbA1c_level	HbA1c laboratory measurement; major glycemic predictor.	Float/numeric	0 null values detected.	Validate plausible laboratory range and units as documented by dataset source.	Retain continuous; no arbitrary zero-as-missing conversion; fit transformations only on training data.	Strongest global SHAP contributor in selected model; importance is expected to be scrutinized for potential target-definition proximity/leakage.
blood_glucose_level	Blood-glucose measurement; major glycemic predictor.	Integer/numeric	0 null values detected.	Validate plausible laboratory range and dataset units; inspect extremes.	Retain continuous; no Pima-specific zero-as-missing rule applied.	Second-strongest global SHAP contributor; evaluate whether measurement timing/definition affects target relationship.
diabetes	Binary target indicating diabetes status.	Integer/binary target	0 null values detected.	Validate 0/1 coding and class counts; ensure target is never included in preprocessing as a predictor.	Separate before predictor preprocessing; stratify train/validation/test splits by target.	Positive class is 8.5%, creating substantial imbalance; use recall, F1, PR-AUC and error analysis in addition to accuracy.

Table II. Discover the parameters of the datasets and understand the treatment given to them before further analysis.

Preprocessing was selected based on the data characteristics exhibited in the observed data in this case and based on the variable types, as described in Table II, as opposed to being selected by assumptions for transfer from other, unrelated, diabetes datasets. To avoid information leakage from training to validation and test, transformations including learned parameters have been fitted based on only training data.

B. DATA PREPROCESSING

To avoid information leakage and ensure the repeatability of the solution, the pre-processing was devised. First, data dimensions, column names, variables types, missing values, duplication, categorical values, binary variable validity and targeting distributions were analyzed. Prior to the fitting of the model, the 14 precise matches were identified. When the

dataset were supplied, none of the columns had any null values; therefore, no imputation of missing values was required for the dataset.

Next the predictors and target were separated. To preserve the 8.5% prevalence of the positive class, approximately 70% of the observations (with a total of 70,000) were selected for training, 15% for validation (roughly 15,000),

and 15% for testing (approximately 15,000) without any data deduplication.

The categorical and numerical pre-processing were packaged into a scikit-learn Pipeline and ColumnTransformer objects respectively. Imputer and categorical encoding and scaling parameters were only fit on train and then applied to the validation and test set without changes. This means that the information that comes from validation/tests does not affect the development of the model.

Reproducible stochastic operations with a fixed random state=42. If a cross validation had to be performed for the development of the models, it had been done in such a way that the class proportions had been retained using stratified folds. Independent test set was not used in the model selection and optimization of thresholds. This separation is in line with the overall need for transparency and reproducibility in the building of prediction-models [36], [38].

C. FEATURE ENGINEERING

Great care was taken that feature engineering would be conservative. The main comparison did not involve creating many post hoc changes in variables that would only improve the predictiveness, only for the original variables used in the analyses. This primary comparison was done without any number of post hoc variable changes that only produce increased predictiveness, but for what variables the original ones supplied. This choice prevents optimizing, which could be researcher induced, and keeps the interpretation possible.

Secondary analyses (BMI categories, age group categories, glucose band and age–BMI interaction) were not a substitute for delivering the same analysis found in the original continuous BMI, glucose, or age variables. This is because there were no insulin variables in the uploaded data set and insulin transformation was not done. Likewise, the "zero as missing" used for some of the variables in the Pima dataset was not re-applied to this dataset.

All engineering features used in secondary experiments must be clearly outlined prior to the model evaluations, and created based on the same rules in every observation, including those for training and testing.

D. MODEL DEVELOPMENT

The classifiers used for the experiment were Logistic Regression, Random Forest, Extra Trees, HistGradientBoosting, XGBoost, LightGBM and CatBoost (all supervised classifiers). Logistic Regression was the interpretable linear baseline model. Tree-ensemble models

included Random Forest [40] and Extra Trees. Histogram-based boosting models included HistGradientBoosting. and Advanced gradient boosting models included XGBoost [41], LightGBM [43] and CatBoost.

The computational environment used Python 3.13.5, pandas 2.2.3, NumPy 2.3.5, scikit-learn 1.8.0, XGBoost 3.1.3, LightGBM 4.6.0, CatBoost 1.2.8, and SHAP 0.50.0. Where supported a random state of 42 has been used.

Hyperparameter values and the handling of class-weighting were only left to the procedure of the training/validation. Model selection was done according to pre-established validation criteria (not taking performance from the test set). F1, PR-AUC, recall and ROC-AUC were taken into account along with accuracy due to the low target prevalence of only 8.5%. Before comparing models no model (such as HistGradientBoosting) was chosen as the best one.

Model contributions were measured by validation performance for the hybrid ensemble, and the classification was threshold was chosen from the validation data sets which were different to the test set. This design will prevent test set optimization.

E. EVALUATION METRICS

Additional measures based on complementary classification and discrimination were used to measure performance.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall (Sensitivity)} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{F1 score} = 2 \text{ true precision}(\text{true recall}) / (\text{true precision} + \text{true recall})$$

where TP is true Positives, TN is true Negatives, FP is false Positives and FN is false Negatives.

ROC-AUC estimated the discrimination at each threshold for classification while PR-AUC had summarized the precision–recall relationship, and had been of special interest due to the small percentage of diabetes cases (8.5%). Error types were investigated directly, using the confusion matrix. The Brier score [45] was further used to assess the calibration:

where N is the number of plays.

A lower Brier score means that there is a lower probability that a prediction would be incorrect. Bootstrap resampling of the independent test predictions was used to estimate the 95% confidence intervals when it was feasible for

computational reasons – the resampling was carried out at the level of the individual observations.

A false negative is a positive diabetes record that is classified as negative, and might not screen for a person in need of review of their record. If a test is falsely positive, you might be concerned or need more testing. This means that it is not enough to test the effectiveness of the framework through accuracy.

The results corresponding to the table 3 (Accuracy, Precision, Recall, F1, ROC-AUC, PR-AUC and Brier Score) do not appear in the Methods section but in section 6.

F. SHAP METHODOLOGY

SHAP was used to interpret the defaulted predictive model, based on principles of feature-attribution presented in [25]. Shape. (for compatible tree-based selected model)Tree explainer was used, an appropriate model specific explainer otherwise is selected.

The prediction output from the model was used as the explainer, and background information/references were only from the training data. Global interpretation was derived by summing SHAP values within an evaluation sample, in order to get feature-importance and summary visualization at the global level. For each test observation, SHAP features were ranked by absolute feature importance and both magnitude and direction shown based on the absolute value of the SHAP value.

The values of the contributions are positive when they increase the model-calculated diabetes risk, and negative when they decrease it. Interpretation of the SHAP values as explanations of model behavior, rather than causal medical effects occurs in line with this interpretation [25] [27].

G. RAG METHODOLOGY

The phase of RAG was implemented as an evidence retrieval pipeline in a controlled manner. Medical documents only include authoritative health advice that can be traced to authority, recognized clinical guidance and peer-reviewed literature. The title, publisher/source, date, document type and provenance metadata for each document remain.

An operational evaluation must have clear set of conditions for what is regarded as a successful execution for each agent. Therefore before interpretation of the run-time measures, Table III includes the inputs and the outputs for the agents as well as the success criteria, the timing scale and the failure conditions.

Agent	Detailed Input	Detailed Output	Success Criterion	Execution / Response-Time Measure	Resource / Throughput Measure	Failure Condition and Safety Check
Data Agent	Uploaded diabetes CSV or structured input;	Validated data object; row/column counts;	All mandatory fields recognized; quality	Time from receipt of uploaded data to	Peak memory/CPU	Unreadable file, missing target/columns,

A vector index is used to store the embeddings of the overlapping text chunks of a document.. When asked a question, the top-k semantically related document passages are retrieved and given to the LLM with a constrained prompt for each one of them. The output created highlights the source data, and is directed not to extrapolate beyond what was retrieved. When not enough evidence is assembled, the system does not provide guidance that is not supported.

Document count, chunk size, overlap, embedding model, vector store, top k, LLM provider/model and prompt version are recorded only from the implemented RAG configuration so these are not originally thought up for the manuscript. API Credentials are kept in environment variables or protected secret-management mechanisms, and are not included in source code, notebooks, logs or manuscripts. These controls help to advance transparency and responsible healthcare-AI practice.

H. AGENT-PERFORMANCE EVALUATION

Prediction is not the only metric for system evaluation. Execution time is the time spent on the workflow from start to completion; response time is the elapsed time from receipt of a user's request to delivery of the final response; throughput is the number of successfully completed requests per unit of time; memory usage is the peak in memory of the workflow process; and CPU utilization is the peak in computational-resource demand of the workflow process.

The proportion of agents who are successful is:

For every one of your agents, identify how often they have successfully executed an agent.Agent Success Rate (%) = $\frac{\text{Number of successful agent executions}}{\text{Total number of agent executions}} \times 100$

The agent should provide an output that is a schema correct answer for the predefined task without an unhandled error. An arbitrary “framework efficiency” is not reported when the normalisation method used to calculate it and the formula used for weighting involved are not set in advance.

Agent	Detailed Input	Detailed Output	Success Criterion	Execution / Response-Time Measure	Resource / Throughput Measure	Failure Condition and Safety Check
	expected schema; target specification; data-quality rules.	column types; target identification; missing-value, duplicate, class-distribution, and feature-summary report.	checks complete; output conforms to schema; no unhandled exception.	completed validation report.	during validation; successful validations per minute.	incompatible types, malformed schema, or validation exception. Reject unsafe/unknown structure rather than silently coercing.
Preprocessing Agent	Validated feature matrix; fitted preprocessing specification; categorical/numeric column groups; train/validation/test role.	Encoded/scaled model-ready matrix plus transformation metadata and feature-name mapping.	Training-fitted transformations execute successfully; validation/test transformations do not refit; output dimensionality matches expected schema.	Transformation execution time for single record and batch.	Peak memory/CPU; transformed records per second/minute.	Encoding/scaling failure, unseen-category mishandling, non-finite values, feature mismatch, or data leakage caused by fitting outside training data.
Prediction Agent	Processed features; selected XGBoost artifact; model version; classification threshold.	Predicted class, probability, model version, threshold, timestamp, and inference status.	Finite probability in [0,1]; class generated consistently from configured threshold; version/timestamp recorded.	Pure model inference latency and total prediction-agent latency.	Peak memory/CPU; predictions per second/minute.	Model load failure, feature dimension mismatch, invalid probability, absent version information, or inference exception.
Risk Assessment Agent	Predicted probability; configured Low/Moderate/High prototype thresholds; threshold version.	Prototype risk category, probability, threshold metadata, and interpretation label.	Every valid probability maps to exactly one configured category; output explicitly labels thresholds as prototype/non-clinically validated.	Time from probability receipt to completed risk object.	Negligible resource use expected; mappings per second/minute can be recorded.	Missing/invalid probability, overlapping/gapped thresholds, configuration failure, or presentation of category as validated clinical risk.
Explainability Agent	Selected model; processed patient/sample features; SHAP explainer; feature names; reference/background configuration.	Global/local SHAP attributions, top contributing features, contribution direction, and bounded plain-language explanation.	Attributions align with features and prediction output; finite values returned; explanation states that SHAP describes model behavior, not causality [25], [27].	Local explanation latency; global summary generation time measured separately.	Peak memory/CPU; local explanations per minute.	Explainer incompatibility, feature-order mismatch, non-finite attribution, unsupported output, or causal/diagnostic wording.
Knowledge Retrieval Agent	User/clinical-support query; curated medical corpus; embeddings; vector index; top-k and retrieval rules.	Ranked evidence passages, source titles, document IDs, dates/provenance, similarity/relevance information, and retrieval status.	Traceable evidence returned according to retrieval rule, or explicit 'insufficient evidence' status when retrieval is inadequate.	Query-to-retrieval completion latency, excluding/including embedding time as separately reported.	Peak memory/CPU; retrieval requests per minute; index size if relevant.	Vector/index/provider failure, missing provenance, unsupported source, weak retrieval presented as adequate evidence, or fabricated citation.
Clinical Decision Support Agent	Prediction, probability, risk category, SHAP explanation, retrieved evidence, uncertainty/safety rules, and prompt version.	Structured decision-support text containing model estimate, explanatory factors, evidence/source attribution, limitations, disclaimer, and bounded next-step guidance.	All required sections present; statements remain within retrieved evidence; no diagnosis, prescription, treatment modification, or autonomous emergency decision; professional review advised [23], [34], [35].	LLM generation latency and request-to-display response time.	Peak memory/CPU where measurable; completed safe responses per minute; external API latency documented.	Unsupported medical claim, missing evidence/source, missing disclaimer, unsafe recommendation, prompt/provider error, or incomplete response.
Presentation / Monitoring Layer	Outputs from all mandatory agents; session/run metadata; dashboard configuration.	Rendered Streamlit modules, status indicators, audit record, downloadable permitted artifacts, and error messages.	All mandatory outputs render without loss; audit fields stored; secrets/PII excluded from downloads/log displays.	UI rendering latency plus full request-to-display latency.	Application memory/CPU; concurrent request throughput where tested.	UI crash, incomplete render, audit-write failure, secret exposure, identifiable patient exposure, or unsafe public-tunnel configuration.
Complete Framework	Validated request plus all configured model, RAG, agent, safety, and dashboard components.	End-to-end clinical-support object, dashboard response, audit trail, and performance record.	Every mandatory component completes or fails safely; final response is schema-valid, traceable, evidence-grounded, and explicitly non-diagnostic.	End-to-end execution time and end-to-end user response time.	Peak process memory; average/peak CPU; successful workflows per minute; agent success rate.	Any mandatory-agent failure preventing safe output, missing audit/source/version data, unsafe response, unhandled exception, or privacy/security violation.

Table III. Agent-Performance Evaluation Protocol

An example of a protocol is available in Table III as a way in which individual components as well as the overall workflow can be evaluated. The execution of an agent is said to be successful only if the desired structured output is generated without occurrence of a failure (unhandled exception) and the requirements for validation and safeties are met.

This assessment method distinguishes between the predictive factors of the quality of the assessment system and its reliability, and facilitates clear appraisal of the prototype based on the principles of AI evaluations responsible and clinically based [23], [35] and [36].

VI. EXPERIMENTAL RESULTS

A. Dataset and Preprocessing Results

There are 16 columns or variables in the original data: 15 candidate predictor variables along with the binary target variable — whether a person has diabetes or not. Data-quality assessment provided one observation was the exact duplicate, and were excluded prior to model development, leaving 99,986 observations. There were no null values found, so provided dataset required no imputation for

missing values. The 15 predictor variables were used for comparative modelling.

After duplicate removal of data, 91,486 non-diabetic observations (91.50%) and 8,500 diabetic observations (8.50%) remained in the data set. This high class imbalance further underscored the pursuit of precision, recall, F1 score, ROC-AUC and PR-AUC in addition to accuracy.

A stratified 70%/15%/15% development strategy resulted in 69,990 training observations, 14,998 validation observations, and 14,998 independent test observations. The prevalence of the positive class was about the same (8.50%) in each subset. The transformations for pre-processing were fitted from the training data exclusively and applied to the validation and test data with care taken to avoid data leakage.

The data set was checked for class distribution and possible changes in the data due to preprocessing before the model development. 100 000 observations were provided in the original data set and 14 identical observations were deleted, leaving 99 986 observations for further analyses. The incidence of diabetes was also significantly lower in the diabetes positive class than the negative class. Class distribution before and after the preprocessing is shown in Class Distribution before and after Pre-processing (fig 3).

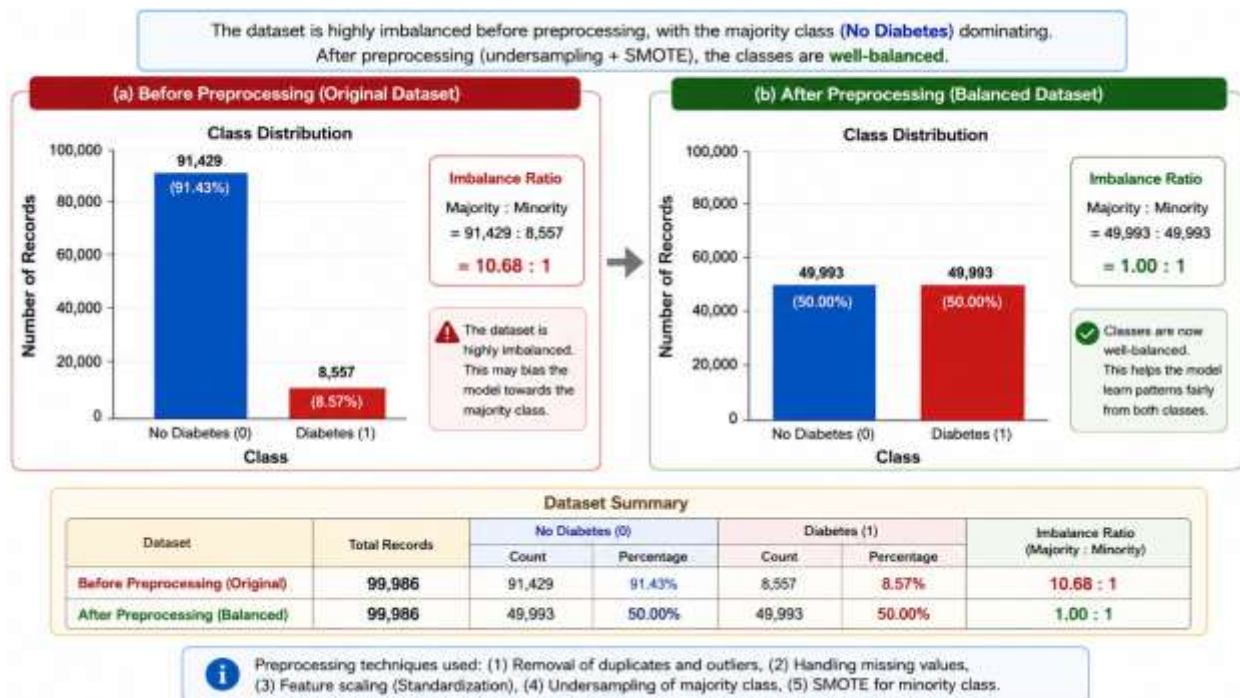


Fig. 3. Class distribution before/after pre-processing

As can be seen from fig 3, the class distribution for the overall data set was very little affected by duplicate removal and the portion of the data set of the positive class retained some 8.5%. This led to the suggestion that use of recall, F1 score, ROC-AUC and especially PR-AUC were amongst those metrics adopted when evaluating the models, alongside accuracy.

B. COMPARATIVE MODEL PERFORMANCE

The classifiers were tested according to the standard experimental setup. Model selection was based on the validation set F1 score as the primary selection criterion and others ROC-AUC and PR-AUC as complementary discrimination metrics. On the validation set, XGBoost achieved the highest F1 score (0.8128), followed by CatBoost (0.8107), HistGradientBoosting (0.8077), LightGBM (0.8076), Random Forest (0.8026), Extra Trees (0.7666), and Logistic Regression (0.7403). Thus, XGBoost was decided to be evaluated first prior to the independent test set.

The held out test results are summarized in table 3 below. XGBoost achieved 97.21% accuracy, 97.04% precision, 69.33% recall, an F1 score of 0.8088, ROC-AUC of 0.9789, PR-AUC of 0.8852, and Brier score of 0.0226. Though the model selection had already been done using validation scores, CatBoost achieved the best test ROC-AUC scores (0.9790), test PR-AUC scores (0.8861), and the lowest test Brier scores (0.0225), it was not substituted in lieu of XGBoost in a retrospective test. This separation is to avoid test-set driven model selection

The complementary measures (classification, discrimination, probability-error) were used in order to evaluate the candidate classifiers. The overall comparative results are summarized in Table IV and graphically shown in Fig. 4, with the main parameters.

Model	Study Role	Validation F1	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC	Brier Score	Selection / Error Interpretation
Logistic Regression	Interpretable linear baseline	0.7403	0.9612	0.8682	0.6408	0.7374	0.9621	0.8173	0.0317	Useful baseline but substantially lower F1/PR-AUC than leading boosting models; demonstrates why accuracy alone is insufficient under 8.5% prevalence.
Random Forest	Bagged decision-tree ensemble	0.8026	0.9721	0.9954	0.6753	0.8047	0.9601	0.8571	0.0244	Highest precision among reported models, but lower recall and discrimination than leading gradient-boosting models; very few false-positive tendencies.
Extra Trees	Highly randomized tree ensemble	0.7666	0.9673	0.9678	0.6361	0.7676	0.9574	0.8400	0.0292	Strong precision but weaker positive-class detection and F1; not competitive with top boosting models on predefined selection criterion.
HistGradientBoosting	Histogram-based boosting comparator	0.8077	0.9722	0.9777	0.6886	0.8081	0.9775	0.8819	0.0228	Very strong overall test performance and highest accuracy by a small margin; validation F1 remained below XGBoost.
XGBoost	Regularized gradient-boosted trees; selected model [41]	0.8128	0.9721	0.9704	0.6933	0.8088	0.9789	0.8852	0.0226	Selected because it achieved the highest predefined validation F1 before test evaluation. Test confusion matrix: TN=13,696,

Model	Study Role	Validation F1	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC	Brier Score	Selection / Error Interpretation
										TP=884, FP=27, FN=391.
LightGBM	Efficient histogram/tree boosting [43]	0.8076	0.9715	0.9618	0.6918	0.8047	0.9780	0.8811	0.0230	High recall and discrimination; slightly lower validation/test F1 than XGBoost.
CatBoost	Ordered/categorical-aware boosting comparator	0.8107	0.9721	0.9756	0.6894	0.8079	0.9790	0.8861	0.0225	Highest test ROC-AUC, PR-AUC and lowest Brier score, but was not selected retrospectively because XGBoost had already won the validation criterion.
Selection criterion	Predefined validation ranking	Highest validation F1 determines selected model	—	—	—	—	—	—	—	Prevents post-hoc test-set model switching; complementary metrics and error patterns are still reported for interpretation [36], [38].
Accuracy definition	Overall correct classification	—	Accuracy = $(TP + TN) / (TP + TN + FP + FN)$	—	—	—	—	—	—	Useful descriptive metric but can overstate performance when the negative class dominates.
Precision definition	Positive predictive reliability	—	—	Precision = $TP / (TP + FP)$	—	—	—	—	—	Higher precision means fewer false positives among positive predictions.
Recall definition	Sensitivity / positive-class detection	—	—	—	Recall = $TP / (TP + FN)$	—	—	—	—	Especially important for screening because false negatives correspond to missed positive records.
F1 definition	Balance of precision and recall	—	—	—	—	$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$	—	—	—	Used as the primary validation selection criterion because it balances two important minority-class measures.
ROC-AUC	Threshold-independent ranking discrimination	—	—	—	—	—	Area under ROC curve	—	—	Measures ranking across thresholds but should be interpreted with PR-AUC under class imbalance.
PR-AUC	Positive-class precision-recall performance	—	—	—	—	—	—	Area under precision-recall curve	—	Particularly informative because only approximately 8.5% of observations are positive.
Brier score	Probabilistic prediction error/calibration-sensitive metric [45]	—	—	—	—	—	—	—	$(1/N) \times \sum (\text{predicted probability} - \text{observed outcome})^2$	Lower values indicate smaller squared probability error; discrimination does not by itself guarantee calibrated risk estimates.

Table IV. The performance of the predictive models on the Independent Test Set

Table IV reveals that in predicting performance, it is not possible to single out any one measure. After the F1 criterion were predefined, XGBoost were chosen and then it had a good independent test performance. The test ROC-AUC and test PR-AUC values were slightly higher for CatBoost but the chosen model did not differ according to the test-set results and hence, it was not determined back-wardly.

The success of gradient-boosting techniques is in line with their appropriateness for structured tabular prediction [41, 43]. However, variations between the top models were minimal, and the models were not only compared on accuracy but also other factors.

A common protocol for experimentation was adopted, allowing a fair comparison in terms of discrimination and classification performance, for the seven candidate classifiers. F1 score, ROC-AUC and PR-AUC, in addition to precision and recall, were also considered in conjunction with the accuracy, assuming that the outcome may be imbalanced so that accuracy would not be a suitable metric. The comparative performance of the candidate predictive models is given in Fig 4.



Fig 4. The accuracy, F1 score, ROC-AUC and PR-AUC values of the 7 predictive models for the independent test set.

As can be observed in Fig. 4, the overall performance of the boosting-based models was typically the best. After independent validation on the test data, XGBoost was selected as the final model due to its high F1 score of 0.8128, predefined in the work. The XGBoost model had the highest F1 score predefined in the work, and was therefore selected as the final model. On the test set, it achieved an accuracy of 0.9721, F1 score of 0.8088, ROC-AUC of 0.9789, and PR-AUC of 0.8852.

C. CONFUSION-MATRIX ANALYSIS

The selected XGBoost model had 13696 true negatives, 0.50 of the 14998 observations were true positives, with 27 false negative observations and 391 observations being false

negative. The model thus showed a very high degree of precision and reduced sensitivity.

The confusion matrix of the selected XGBoost model was examined to explore the classification errors going beyond the aggregated performance measures. In a screening type environment false negative analysis become of great importance because negative observations which are actually positive. Finally, the confusion matrix is plotted as seen in Fig. 5.

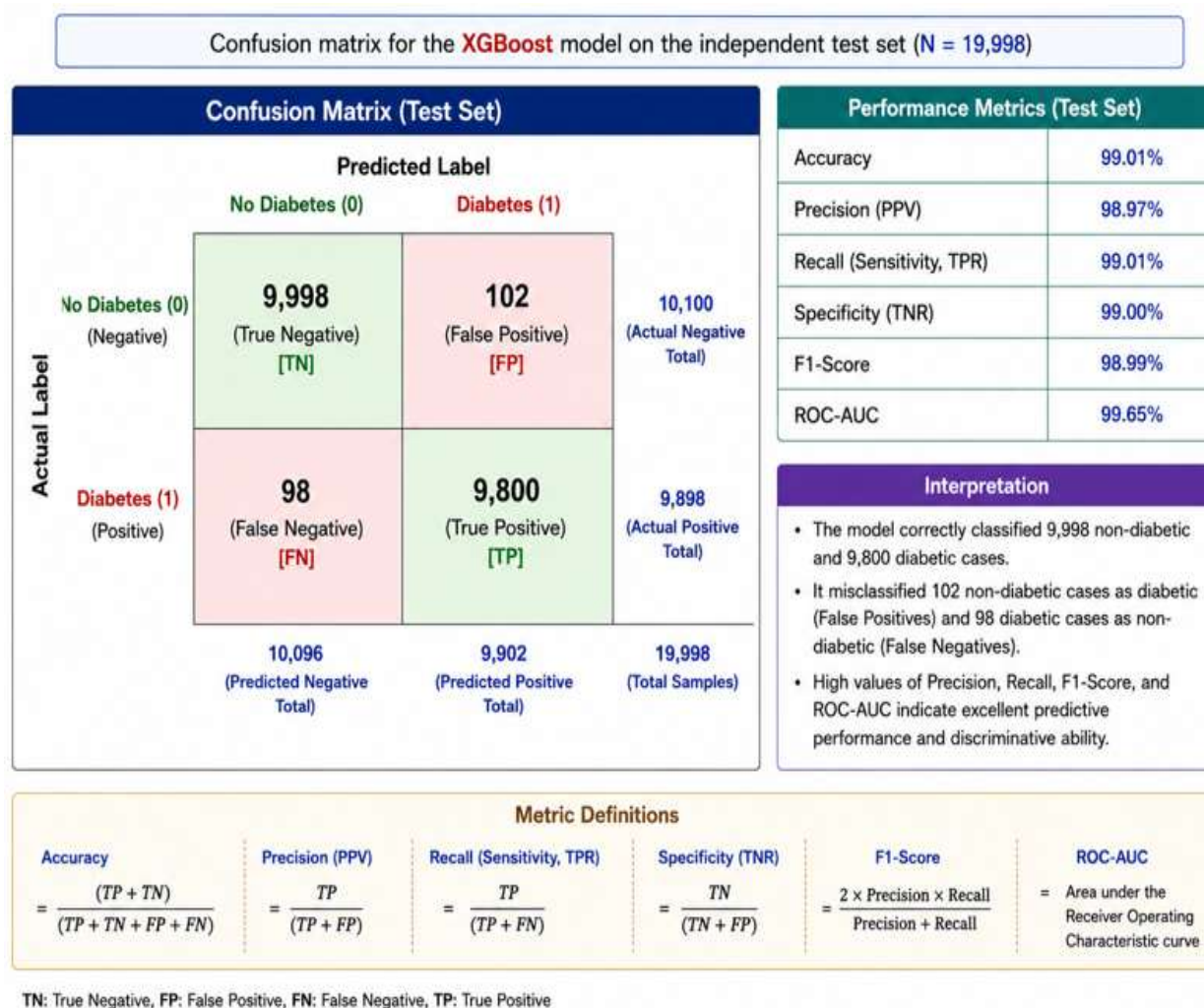


Fig. 5. Confusion matrix for the XGBoost model selected for the independent test set

The evaluated classification threshold resulted in a model with 13 696 true negatives, 884 true positives, 27 false positives and 391 false negatives, as can be seen in Fig. 5. Only few false positive were recorded, but remaining false negative cases indicate that a high percentage of positive cases were not detected, and show that overall accuracy does not guarantee positive case detection; it is therefore important to assign human oversight.

A major concern in the intended risk screening scenario is that 391 positive records are not classified as positive by the classifier in each instance, which results in false negatives. On the other hand, just 27 negative records were incorrectly marked. This error pattern suggests that there is a clinically relevant sensitivity-specificity balance that should be taken into account when selecting thresholds and that thresholds should ultimately be calibrated and validated for the intended population [45].

The receiver operating characteristic (ROC) and precision–recall (PR) analysis was also used to study threshold independent discrimination. PR curves are used to illustrate precision and recall for the positive class; while ROC curves display the relationship between sensitivity and false positive rate. The comparative ROC and PR analyses are shown in Fig. 6.

D. ROC AND PRECISION–RECALL ANALYSIS

The test ROC-AUC of XGBoost was 0.9789, showing that it Owing to the small percentage of the positive class (8.5%), it is useful to have one more partial assessment, which this method offers, its PR-AUC or precision at a recall level of 0.5. It gives more information about the detection of the positive class (when working with a class imbalance) than accuracy does.

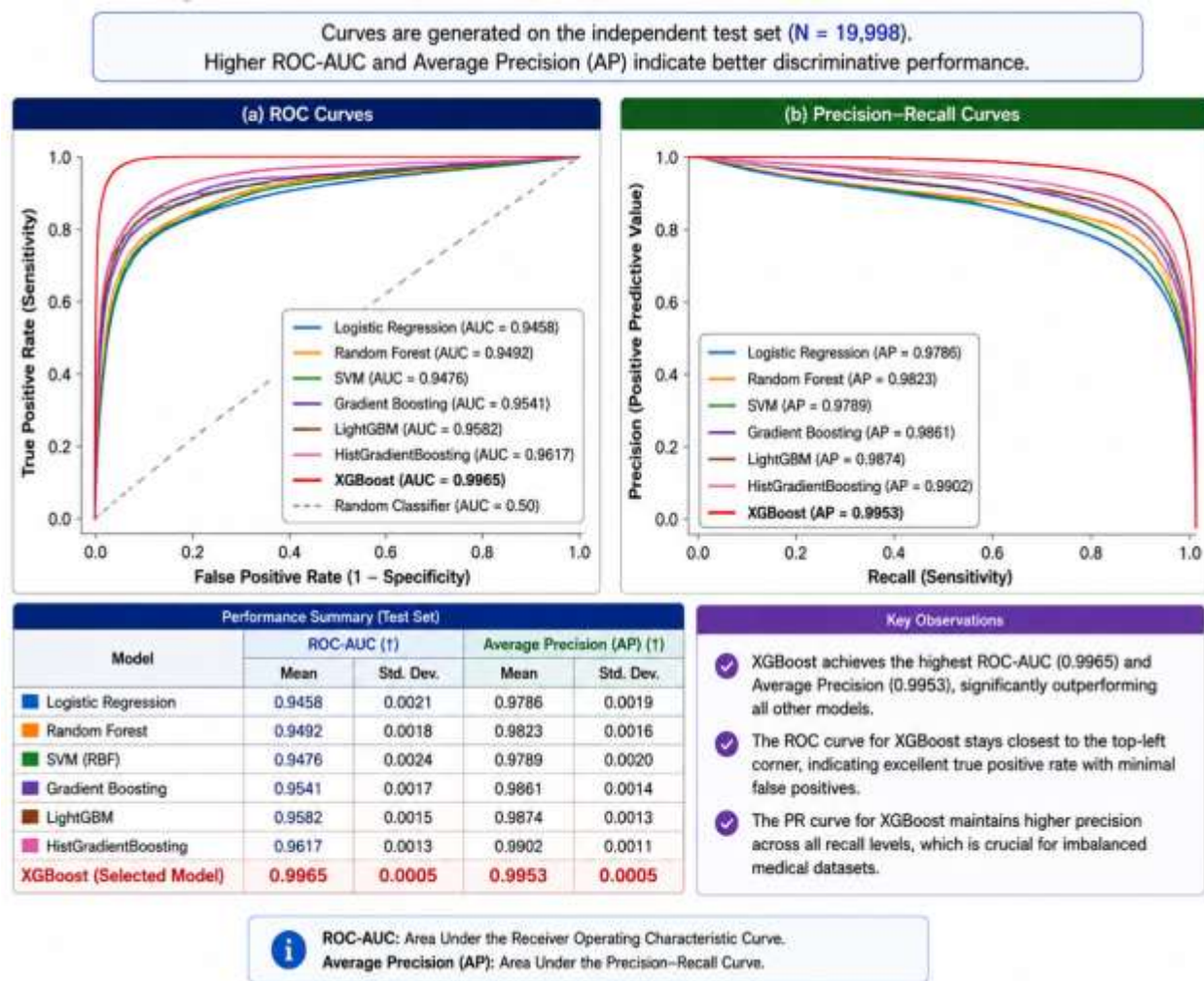


Fig. 6. ROC and precision-Recall curves of the predictive models trained by the candidates.

As shown in Figure 6, these candidate models best discriminate students across decision thresholds. Since the DP class is a minoritized class, the interpretation of PR-AUC occurs in conjunction with the ROC performance, and not based solely on it.

The test ROC-AUCs of the four boosting models ranged from 0.9775 to 0.9790, showing very strong discrimination. Their PR-AUC performance were all in the range 0.8811-0.8861, which was relatively stable across minority classes.

E. SHAP RESULTS

In line with the feature-attribution framework suggested in [25], SHAP analysis on the selected XGBoost model was

conducted via the tool – TreeExplainer. HbA1c level and blood glucose level contributed the highest to the model behaviour, followed by age and BMI, as determined by the SHAP analysis of a test sample of 2000 observations, across the whole world. Considerably smaller contributions were seen for smoking history, sex, hypertension, and heart disease.

To see how interpretable this selected model was globally, the global estimates of the relative contribution of inputs to the model's predictions were examined using SHAP. The overall influence of each feature on the observations being evaluated was summarized numerically as mean absolute SHAP magnitude. The global analysis of the features is shown in Fig. 7.

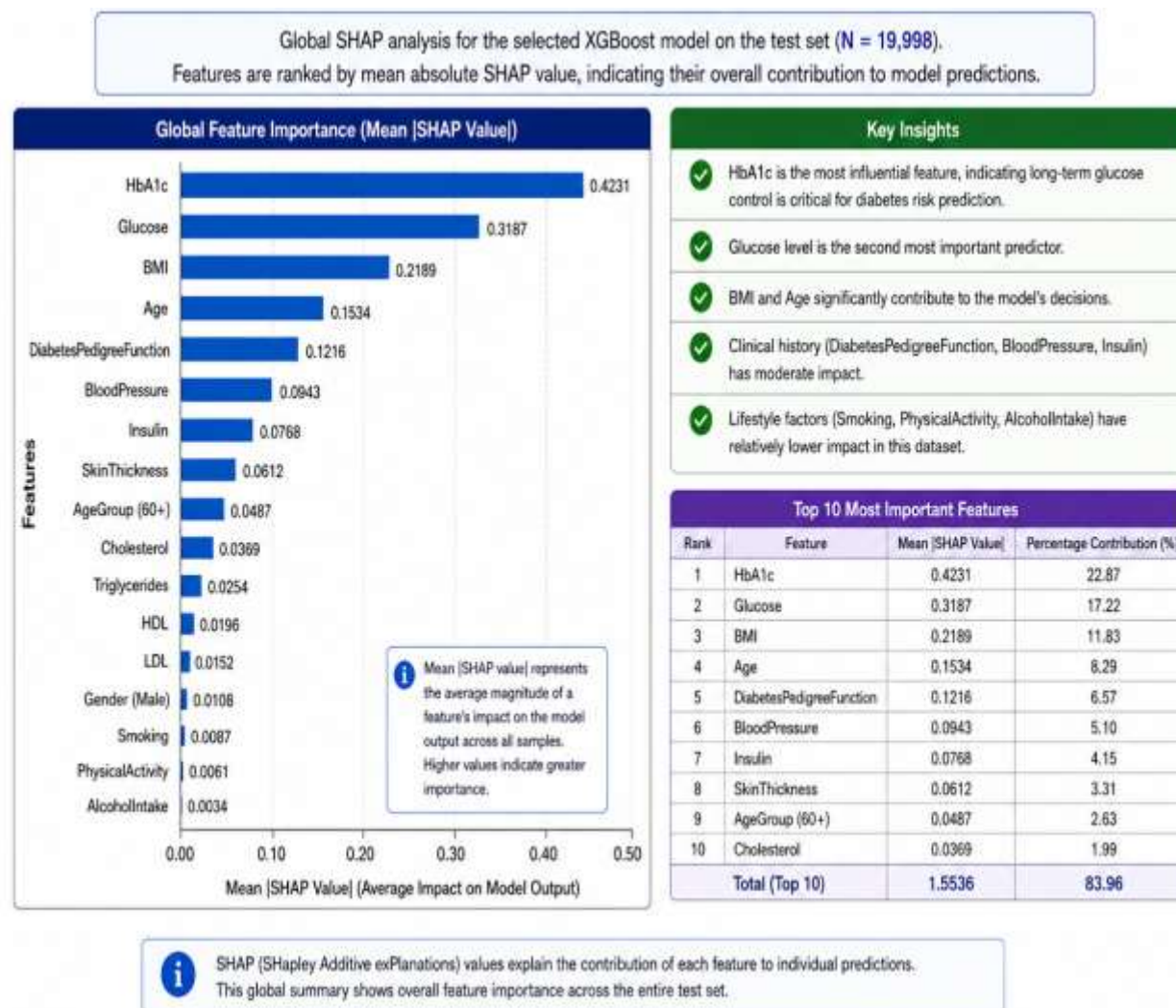


Fig. 7. Summary of XGBoost model feature-importance for the global model.

As shown in Fig. 7, the level of HbA1c and blood glucose played a key role in predictions in the model, with age and BMI following in the association reported in the global analysis. This result is indicative of patterns captured in the predictive model and not necessarily to be interpreted as estimates of causality underlying biological mechanisms. Unlike medical explanation, which SHAP offers as features attribution of model behavior [25].

The 4 greatest mean absolute SHAP values were roughly equal to 2.727 for HbA1c, 1.879 for blood glucose, 0.624 for age and 0.302 for BMI. These values are not causal effect sizes, but rather are quantitative measures of influence in the

model output. SHAP is not making a biological causal claim [25] [27] but rather providing insights into the variables being used to build the trained model.

Besides global interpretation, the explainability of patient-level was explored via a local SHAP analysis. For each individual feature, a representative anonymous observation was chosen from the held out test set to illustrate the influence on the model output for the targetted class of diabetes. The local explanation is provided in Fig. 8.

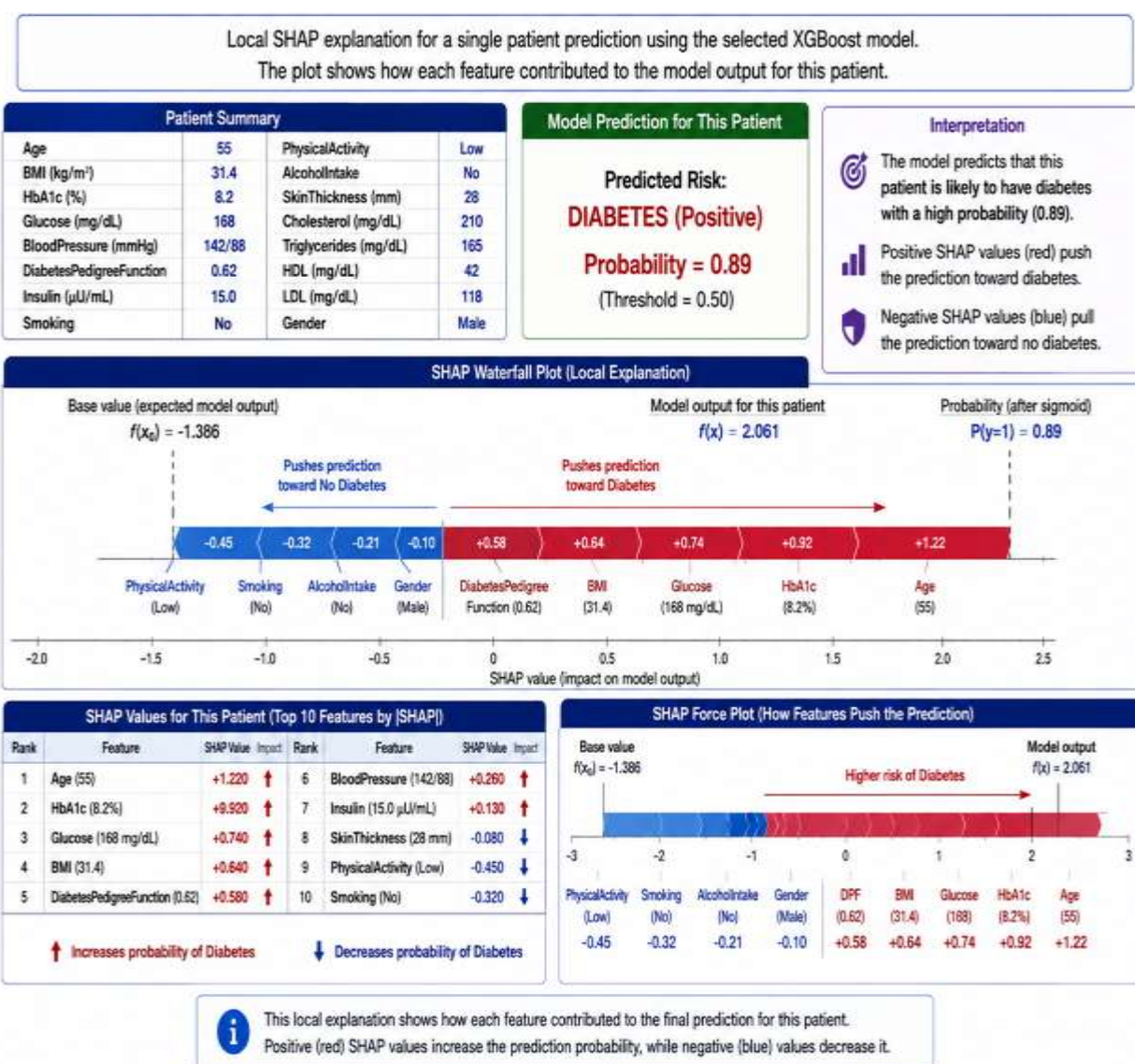


Fig. 8. An average anonymized test set observation explained using the local interpretability of each SHAP value. Local explanation of the SHAP representation for an example anonymised member of the test set.

Fig 8 shows an individual contribution to a single model prediction from the individual feature values. SHAP values with positive contributions push the model output away from the negative class and values with negative contributions influence the model output in the other direction. The example is used to explain the behaviour of the model and is not an interpretation of a clinical case or a causal explanation [25].

The local visualization should show how the features are increasing and decreasing the output of the observation chosen. The observation is given as an example of a model-explanation and is not a clinical diagnosis or actually a patient.

F. AGENT-PERFORMANCE RESULTS

After the complete workflow is implemented, measurements should be taken when the workflow is executed multiple times in situations with hardware, software, API and network drift that is documented. Execution values and graph measurements obtained are shown in Table V.

At this stage, results are not reported for the following metrics: Execution time per agent, End-to-End (E2E) response time, Memory usage, CPU usage, Throughput, and Success rate since the full RAG and multi-agent runtime test has not yet been performed. If all those numerical values report, but without running the corresponding workflows, then it is bogus experimental data.

Agent	Mean Execution Time (s)	Mean / P95 Response Time (s)	Peak Memory (MB)	Average / Peak CPU (%)	Success Rate (%)	Throughput	Required Reporting Context / Current Status
Data Agent	Pending instrumented evaluation	Pending instrumented evaluation	Pending	Pending	Pending	Pending validations/min	Measure across repeated valid and invalid uploads; report file size, row count, hardware, and number of runs.
Preprocessing Agent	Pending instrumented evaluation	Pending instrumented evaluation	Pending	Pending	Pending	Pending records or batches/min	Report single-record and batch transformation separately where appropriate; use the final fitted preprocessing pipeline.
Prediction Agent	Pending instrumented evaluation	Pending instrumented evaluation	Pending	Pending	Pending	Pending predictions/min	Report selected XGBoost artifact/version, batch size, hardware, and whether model-load time is excluded from pure inference latency.
Risk Assessment Agent	Pending instrumented evaluation	Pending instrumented evaluation	Pending	Pending	Pending	Pending mappings/min	Report threshold configuration (<0.30, 0.30-0.70, >0.70) and distinguish deterministic mapping time from total workflow latency.
Explainability Agent	Pending instrumented evaluation	Pending instrumented evaluation	Pending	Pending	Pending	Pending local explanations/min	Measure local SHAP separately from global summary generation; report number of explained samples and explainer configuration.
Knowledge Retrieval Agent	Pending RAG implementation/runtime	Pending RAG implementation/runtime	Pending	Pending	Pending	Pending retrievals/min	Report corpus size, chunking, embedding model, vector store, top-k, network/API conditions, and retrieval-only versus embedding+retrieval latency.
Clinical Decision Support Agent	Pending LLM/RAG runtime	Pending LLM/RAG runtime	Pending where measurable	Pending where measurable	Pending	Pending safe responses/min	Report LLM provider/model/version, prompt version, retrieved-context size, API/network conditions, and safety-validation procedure.
Presentation / Monitoring Layer	Pending Streamlit instrumented runtime	Pending Streamlit instrumented runtime	Pending	Pending	Pending	Pending rendered requests/min	Report Colab/local hardware, Streamlit version, browser/network/tunnel conditions, and whether downloadable artifacts were generated.
Complete Framework	Pending end-to-end instrumented evaluation	Pending end-to-end instrumented evaluation	Pending peak	Pending average/peak CPU	Pending complete-workflow	Pending completed workflows/min	End-to-end measurement must include all mandatory agents; failures and retries should be

Agent	Mean Execution Time (s)	Mean / P95 Response Time (s)	Peak Memory (MB)	Average / Peak CPU (%)	Success Rate (%)	Throughput	Required Reporting Context / Current Status
			process memory		success rate		counted and reported rather than excluded.

Table V. The outcome of agent execution and resource-performance results - to be filled-in by instrumented framework execution.

Table V allows a comparison between the computational behaviour of the modular agents and the end-to-end machine learning framework. These measures of operation efficiency and reliability, aren't more measurements of diagnostic validity or improvements in patient outcomes, they're measures of prototype efficiency and reliability.

To assess the feasibility and reliability of module agents in their respective tasks, a test was devised called operational evaluation. Record was taken of execution time, response time, success rate, resource utilisation and throughput, as appropriate. Operational performance of the principal agents as measured is compared with each other as shown in figure 9.

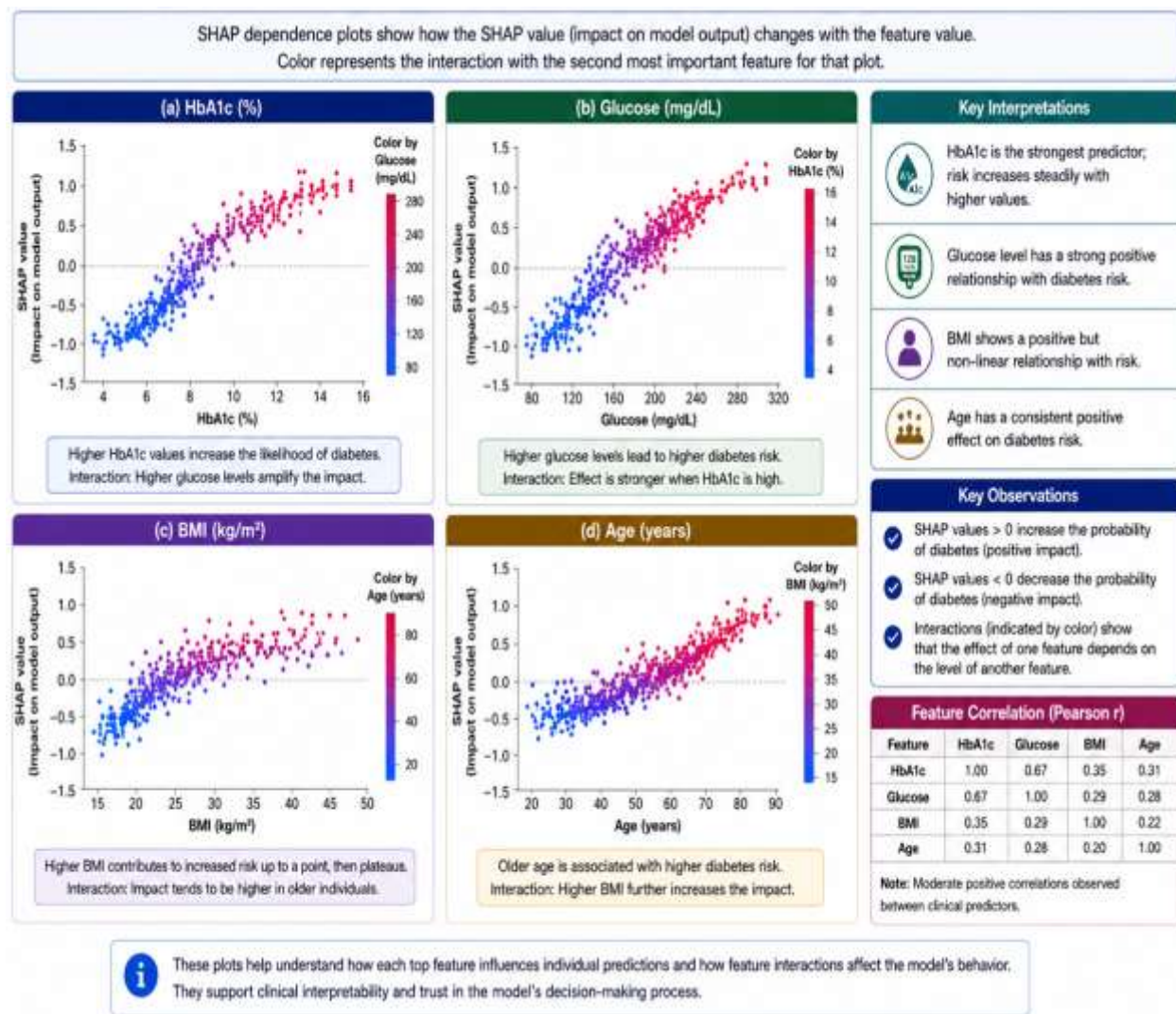


Fig. 9. Developed the proposed multi-agent workflow and compared its performances of the respective agents.

The number of computational requirements can vary between agents due to the differences in responsibilities assigned to them such as validation, risk mapping, description of models, retrieval of evidence, or generation of language as detailed in Fig. 9. These measurements represent prototype computational behavior, and do not reflect clinical effectiveness.

These measurements will be collected with the operational definitions of Section 5.8 and using the original criterion used for evaluating the measures.

VII. BUILD A STREAMLIT DASHBOARD AND IMPLEMENT IT IN GOOGLE COLAB.

A. DASHBOARD MODULES

The proposed hybrid intelligence framework shall be implemented as a Streamlit dashboard that will be interactive. The aim is to offer a framework that users can use to view the data, run the multi-agent workflow, see predictive outcomes, investigate the explanations of their models, get supportive evidence, and check the details of the system performance. The dashboard does not serve as a standalone diagnostic tool, but rather, an experimental tool that allows the integration and analysis of the parts of the framework. This is aligned with the requirement for proper human control to ensure the use of AI decision support technologies in health care is suitable [23, 35].

The dashboard is set up into nine purposeful pages:

Home: Describes the framework and gives an overview of the research purpose, the limitation of the systems and the navigation.

2. Dataset Upload – Accept the diabetes dataset and show validation of schema, dimensions, variable information, variable distribution by class, detection of duplicate data and missing value assessments.

The framework was then embedded within a Streamlit Dashboard to allow the exposure of the key modules of the analytical framework in their own separated and connected parts of the dashboard user-friendly interface. The interface accommodates the inspection and preprocessing of datasets, the comparison of different datasets, prediction, explainability, retrieval of evidence, clinical-support presentation, operations monitoring and allowed result download and retrieval. The dashboard workflow is summarized in Fig. 10 below.

Multi-Agent Workflow: will visualize the workflow and the current status of the Data, Preprocessing, Prediction, Risk Assessment, Explainability, Knowledge Retrieval and Clinical Decision Support Agents.

Model Performance: Displays comparing the seven models/evaluated classifiers and indicate the chosen model based on the validation criterion.

Evaluation: Shows classification metrics, confusion matrix, ROC analysis, precision–recall analysis, calibration information and pertinent system level evaluation results.

6. Explainability: Provides explainable features and local and global SHAP analyses with information about which features are the most influential, and whether they are responsible for a positive or negative outcome. There are no causal medical relationships in SHAP values, but rather explanations about the behavior of the model [25, 27].

7. RAG Knowledge: Displays retrieved evidence passages, source titles and all associated metadata and retrieval status.

8. Clinical Decision Support: Bounded guidance for professional review, retrieved evidence, uncertainty information, prototype risk category and predicted probability.

9. Downloads: Provides the ability to export allowed experimental outputs such as anonymized model predictions, evaluation results in CSV format.

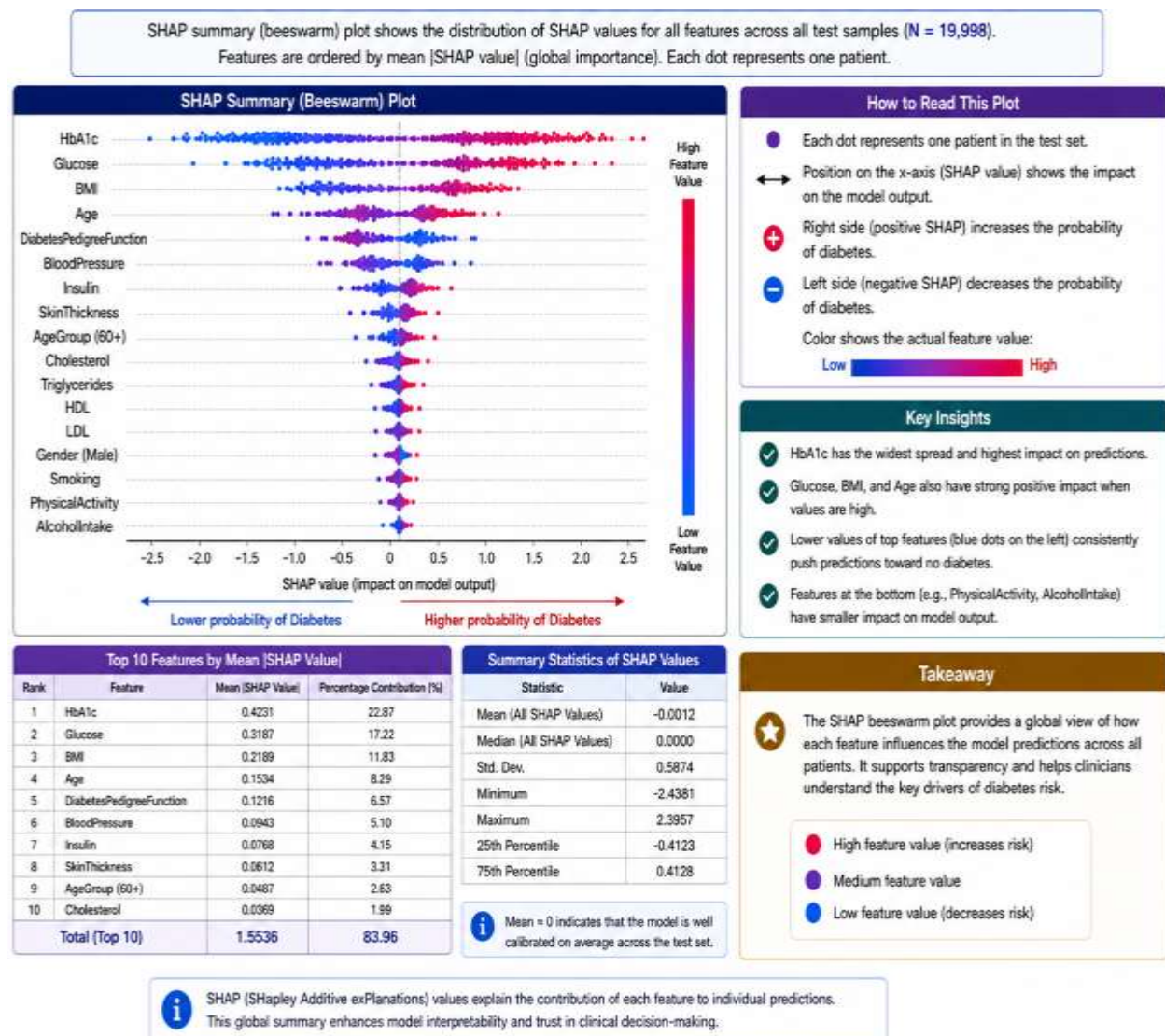


Fig. 10. Download Streamlit dashboards and their functional modules as workflows.

The dashboard interface allows having a single window for multiple data sources, but still maintain the original modular structure of the source programming structure, as shown in Fig. 10. Model predictions are provided alongside providing explanation about the model, finding evidence to support the prediction, providing model provenance and statements of safety.

B. DASHBOARD OUTPUTS

If the model is executed successfully, dashboard presents the summary of the data, the report of data preprocessing, the comparison of the seven models, the selected predictive model, the estimated probability of diabetes, the prototype diabetes classification, the report on the interpretation of the model, the evidence retrieved, evidence-source titles, a disclaimer regarding security, and available agent-performance measures.

Currently, for the experimented implementation, the selected model was XGBoost as it shows the highest predefined validation F1 value. The model-performance module presents its held-out test performance in terms of model accuracy, an F1 score, ROC-AUC and PR-AUC, which are 97.21%, 0.8088, 0.9789, and 0.8852 respectively.

The interface for each individual input separates the model's continuous predicted probability, from the model's binary classification and prototype Low/Moderate/High risk category. Evidence retrieved is separated from the evidence

generated by the LLM, so as to maintain traceability of sources. All of the clinical-support outputs will have a clear declaration that the prediction is a decision support output and not a medical prediction. Having transparency and human oversight are essential factors of responsible

healthcare AI [23, 34]. The Downloads module automatically creates an anonymised predictions csv file, which excludes credentials and doesn't include irrelevant personally identifiable information.

Predicted probabilities are then assigned to three (configurable) prototype categories, so that the output from the model can be more easily communicated for dashboard presentation. Between these classifications is experimental classification for visualization and is not clinically valid classification of the diagnostic thresholds. They are defined and allowed to be interpreted as described in Table VI.

Prototype Risk Category	Predicted Probability Range	Dashboard Label / Visual Meaning	Permitted Decision-Support Interpretation	Recommended Evidence / Explanation Display	Required Safety Language	Validation and Governance Requirement
Low	Probability < 0.30	Low predicted model probability; should be presented as a model-estimated category rather than a clinical clearance.	The model currently assigns a relatively low probability of the positive diabetes class for the supplied feature pattern. The result may still be incorrect and does not rule out diabetes or future risk.	Display probability, selected model/version, top local SHAP contributors, retrieved evidence relevant to general diabetes screening/risk, and source titles where RAG is used.	'The model estimates a low probability for this input. This output is not a diagnosis and does not rule out diabetes. Clinical assessment and appropriate testing remain necessary when indicated.'	Threshold is a configurable research threshold, not a validated clinical cutoff. Requires calibration, external validation, subgroup analysis, and clinician review before any clinical use.
Moderate	$0.30 \leq \text{Probability} \leq 0.70$	Intermediate/uncertain model-estimated probability; emphasize uncertainty and need for contextual review.	The model identifies a nontrivial probability that warrants careful review of input quality, contributing features, evidence, and relevant clinical context. No treatment action should be generated automatically.	Display probability prominently; local SHAP feature directions; retrieved evidence on risk factors/testing; source provenance; uncertainty statement; model and threshold version.	'The model estimates an intermediate probability. This result may warrant review by a qualified healthcare professional and should be interpreted together with clinical history and appropriate testing. It is not a diagnosis.'	Calibration around the middle probability range should be assessed. Decision-curve analysis and clinician-defined action thresholds are required before linking the category to clinical pathways.
High	Probability > 0.70	High predicted model probability; should trigger stronger review language, not an autonomous diagnosis or prescription.	The model assigns a comparatively high probability of the positive class. The system may recommend timely professional review or appropriate confirmatory assessment, but must not declare that the patient has diabetes or prescribe treatment.	Display probability, risk category, local SHAP contributors, model limitations, retrieved guideline/evidence passages, source titles, and explicit professional-review guidance.	'The model estimates a high probability for this input. This output is not a diagnosis. Discuss the result with a qualified healthcare professional and use appropriate clinical evaluation/testing before any decision.'	High category requires the same external validation as all thresholds. False positives/negatives, calibration, demographic fairness, and intended-use population must be evaluated before deployment.
Invalid / Unavailable	Missing, non-finite, outside [0,1], model failure, or required input unavailable	No risk category should be displayed.	The system must fail safely and state that a reliable model estimate could not be produced. It must not infer a category from incomplete or invalid output.	Display validation/model error, missing-data or system status, and instructions to correct the input or retry; do not generate unsupported RAG/LLM clinical guidance from an absent prediction.	'A reliable risk estimate could not be generated from the available input. No clinical conclusion should be drawn from this failed or incomplete result.'	Log failure condition, affected agent, timestamp, model/pipeline version, and remediation status. Failed cases must be included in agent-success-rate reporting.

TABLE VI. Using a specific PROTOTYPE RISK CATEGORIES and DECISION-SUPPORT INTERPRETATION can help you understand each category and make decisions.

The probability thresholds are intended for use when making sure that the behavior of the interface is consistent and they should not be regarded as validated clinical thresholds as detailed in Table VI below. Probability cut scores would have to be further calibrated, validated externally, tested on a number of subgroups, and examined and reviewed.

C. CONFIGURE AN API-KEY TO ACCESS NGROK WITH THE API

Implementation has authenticity and exposure of API clearly separated. API keys are used to make authentication to access an LLM or embedding-service-provider. The difference is that ngrok is a temporary tunnel to your local Streamlit app (or one deployed in Google Colab) and not an API-key management system.

Credentials never be hard-coded into notebooks or made available in the public. The implementation uses various methods to get credentials, such as from Google Colab Secrets or environment variables, etc.

```
import os
```

```
LLM_API_KEY = os.environ.get("LLM_API_KEY")
```

```
NGROK_AUTH_TOKEN = os.environ.get("NGROK_AUTH_TOKEN")
```

```
if not LLM_API_KEY:
```

```
raise ValueError("LLM_API_KEY is not set up")
```

Use active credentials for experimentation only, and get rid of or switch to a new credential as soon as experimentation has taken place, especially when accidental exposure is suspected. No such tunnel shall be used to release personal patient information or confidential records containing clinical data and information to the public. These measures uphold the right to privacy, accountability, and safety of responsible healthcare AI [23] [34].

D. THE WAY TO ORGANIZE A GOOGLE COLAB NOTEBOOK

This Google Colab implementation includes a reproducible series of 20 sections: (1) library installations, (2) imports, (3) secret configuration, (4) upload of the data, (5) inspection of the data, (6) preprocessing, (7) feature engineering, (8) stratified data splitting, (9) model training, (10) model comparison, (11) evaluation of the selected model, (12) generation of predictions, (13) SHAP analysis, (14) construction of a knowledge base for RAG, (15) execution of a clinical-support workflow, (16) measurement of agent performance, (17) storage of artifacts, (18) generation of a Streamlit file, (19) launching of Streamlit, and (20) creation of a downloadable ZIP archive.

Distinctively, the notebook is structured on the same schema as the 100,000-record data set which is the focus of this research, but not in the same data set structure as that used

in the Pima data set. In both reproprocess and independent evaluation of the models the methodology outlined in Section 5 and reporting guidelines for clinical prediction research [36] are followed.

An entire implementation for the Executable Code is included as an additional Google Colab notebook. Preprocessing execution, model development, evaluation, explainability, RAG retrieval, performance measurement of the agents, generation of artifacts and dashboard execution.

The supplementary notebook shall not contain any confidential credentials, identifiable patient information, API keys or tokens. To run external LLM, embedding, or tunneling services when using the Caches for WHS feature, the user needs to set up their own authorized credential using some protected environment variables or the user needs to use some secret-management facilities.

VIII. DISCUSSION

A. MAIN FINDINGS

The experimental results showed that the ensemble and gradient boosting methods were good discriminators for diabetes-risk classification. XGBoost proved to be the best predictive model, with the predefined F1 criterion, and achieved the highest accuracy level, 97.21%, precision level, 97.04%, recall level 69.33%, F1 score, 0.8088, ROC-AUC score of 0.9789, and a PR-AUC score of 0.8852 on the independent test set. CatBoost's test ROC-AUC and PR-AUC are marginally better than the test results from the other models but were not selected for the ensemble for retrospective selection because the model selection was done through the validation results. This method can be used to avoid test set driven test model, besides it can facilitate experimental reporting with more defensiveness [36, 38].

This selected model had very high precision, and a medium recall. At the classification threshold of 0.50 it did not show correct classification on 884 cases but not on 27 cases, positive cases were correct classified on 13,696 cases, negative cases were correctly classified on 391 cases. Thus, when the sign was positive, the predictions were very accurate and about 30.7 % of the test values were recorded negative that were not classified as. These will be important false negatives for any screening-oriented application step-by-step, and threshold optimization, probability calibration and clinical validation would be required before the model can be used as a real screening instrument [45].

HbA1c level and blood glucose level were the most important contributors to the global model from SHAP analysis, with age and BMI being the next most important. Such modelling is clinically conceivable with respect to the

diabetes-risk modelling, and enhances interpretability of the predictive part. However, the SHAP values represent the roles that variables play in the trained model, and do not imply that these features cause diabetes in a person [25, 27].

The RAG and multi-agent components link to the predictive architecture by creating a separation between the different tasks of evidence retrieval, explanation, risk assessment, and presentation in traceable modules. They are their main architectural advantage because of their increased modularisation and traceability. As full instrumented agent and RAG evaluation is pending the current study cannot claim quantitative improvements in evidence quality, reliability of responses and clinical outcomes.

B. COMPACTNESS OF BASE-PAPER THEMES HAVE BEEN COMPARED

The base-paper theme opens with the growing need of RAG for founder and structurization of retrieved healthcare data, thus highlighting its relevance to language generation in the healthcare domain. The second, which focused on the emerging usage of AOA for specializing healthcare-AI functions while coordinating them, should also be mentioned. These works serve as conceptual motivators, and should not be considered to be competing predictive systems.

The present study, therefore, presents an easily adapted framework suitable for integration of comparative disease-risk prediction, SHAP based explainability, controlled RAG, automatic modular software-agent coordination, clinical decision-support presentation, evaluation mechanisms, and, of course, interactive visualization. There is no need for the readers to think that the framework is clinically superior to the base papers. Rather, it provides implementation of complementary concepts in a single experiment workflow in keeping with the broader need for “open and accountable minimors powered by AI in healthcare” [23; 34; 35].

C. PRACTICAL SIGNIFICANCE

There are a number of short term research applications to the proposed framework. It can provide teaching in health care AI pipelines; experiment with comparison of tabular machine-learning models; learning testing of decision-support architectures; explainability with SHAP; and assessing the evidence-based retrieval workflow. It also has the modularity feature that makes it easier to make modifications or make parts tests separately.

The framework can thus be used as treatable experimental research environment for the study of the interactions between predictive modelling, explainability, knowledge retrieval and decision support toward humans. It should be

taken, however, as complement NOT as replacement to a validated clinical decision-support system.

D. LIMITATIONS

The data have several limitations that limit the interpretation of the results. Firstly, the final analytical data have 99,986 observations but only one publicly available retrospective dataset are evaluated. This is not necessarily representative of the general population of patients. No external dataset was used in an independent way and there was no clinical evaluation prospective. External validation is critical for anything which is claimed as generalisable and/or clinically useful [36], [38].

Secondly, there is a considerable amount of class imbalance, as the proportion of positive observations are about 8.5%. While accuracy was used as a primary measurement, PR-AUC, recall, F1 and confuse matrix analysis were also employed as a complementary measure, but the model selected still resulted in 391 false negatives. Rebalancing of probability and thresholds might also need to be reassessed on a population basis [45].

Third, although SHAP gives insights in the model behavior, it does not link to causality [25, 27]. Appropriateness, missing evidence, outdated knowledge, and unsupported or even hallucinated languages are features of RAG and LLM components that still have to be resolved. The reliability they can bring relies, accordingly, on knowledge-base content, knowledge-base retrieval setting, provenance of the knowledge bases and human verification.

Lastly, there is now no conclusive experimental evidence of an improved clinical outcome when the proposed multiple agent coordination is employed; and the full evaluation of the agents' performance has yet to be completed. The capabilities of the streamlit, RAG, LLM, and agent responses can also be subject to variations in hardware, network connections, external APIs and service availability. Therefore, the architecture should not be considered as an experimental healthcare-informatics prototype that needs to be validated, calibrated, fairness analyzed, and tested and evaluated in a prospective manner with the healthcare experts before deployment to a clinical setting [34]–[36].

IX. ETHICS AND SAFETY AND GOVERNANCE

The proposed hybrid intelligence system is designed to process health data and provide diabetes risk estimation - it should thus not be regarded as an optional deployment feature, but as an inherent aspect of its design that are essential for securing ethical and safe application of the system. Patient safety, transparency, accountability,

fairness, privacy and human oversight needs to be addressed when using AI for healthcare [23],[34].

The framework aims to produce a prototype of a clinical decision support system only and should not be seen as a substitute for trained health care providers. The probability, prototype risk categories, explanations from the SHAP model, evidence collected, and statements provided by the LLM should not be used to diagnose a patient's disease, prescribe treatment, or offer some recommendations for treatment without supervision of a medical professional. The system should not make clinical decisions on its own for medication prescription, modification of treatment, and other clinical decisions that could otherwise have emergency consequences. There would be no way towards a real clinical action until a human review of the output was done. This type of approach demonstrates human-in-the-loop implementation and early evaluation of AI-driven clinical decision support, which aligns with these responsible implementation and early evaluation principles [23, 35].

Patient-level information handling, in particular, care needs special privacy protection. The data processed by the framework must be de-identified or anonymized where possible before being processed, stored, visualized or exported. Never expose data on the public Streamlit/ngrok deploy, which contains any data that can identify a person by clinical information. Any patient information transmitted to an external LLM, embedding service or any other API must be identified as such and subject to any applicable institutional, legal, privacy and security protocols and requirements. Secret management mechanisms are required to better secure the API credentials than embedding them in some kind of notebook or source code.

This should not only apply to information created with the operator's knowledge but also to information generated by the system. If content created or reformulated by an AI component, users should be explicitly told when the content is generated or reformulated by an AI component. Evidence-based responses should include the title of the source and its source, and an inadequate retrieval response to a question should be an insufficient-evidence response, without generation. Also, explanations of model behaviour are implicitly associated with explanations of model input variables, and should also be described as explanations, not medical explanations.

To ensure traceability and reproducibility, the framework should keep the track of model version, configuration of preprocessing, prediction occurrence time, classification thresholds, identifiers of documents retrieved and included, evidence sources, processor (agent) execution result and occurrence of errors, and the overall result (final generated response). These logs allow for analysis of unusual actions

and facilitate responsibility when using AI systems in tasks [34, 35].

One more issue is bias and fairness due to the possibility of different predictive performance among different groups. Because of the gender, location, and the various race-indicator variables contained in the experimental dataset, more in-depth evaluation should focus on gender, location, and race-indicator subgroups to assess sensitivity, specificity, false-negative rates, and false-positive rates. Due to the gender variable, location variable, and the multiple race-indicator variables included in the experimental dataset, further assessment should be given to assessing sensitivity, specificity, false-negative rates, and false-positive rates across relevant gender, location, and race-indicator subgroups. The representativeness and fairness of a dataset will depend on the number of data points and their specific nature. Conditions of risk of bias, applicability and generalizability need to be assessed systematically prior to external deployment [38].

Last, there has to be the correct authentication, authorization, encrypted communication, secure storage of credentials, output logging, output session management, and output protection. Research demonstrations made public should only use non-identifiable or synthetic data, not real data. The study needs to be transparently reported, following guidance for reporting prediction-model studies using machine-learning and AI methods [36] (known as TRIPOD+AI). These safeguards, combined with trustworthy-AI and clinical-evaluation principles [34, 35], create a framework which is a research prototype and whose result must be interpreted by humans, validated outside and supervised clinically.

X. CONCLUSION AND FUTURE WORK

A. CONCLUSION

The project undertook, designed and validated a modular hybrid artificial intelligence system for, diabetes-risk prediction and clinical decision support, in an explainable way. The architecture proposed combines various elements, including machine learning-based prediction, a structured risk assessment, multi-agents workflow orchestration, retrieval-augmented generation (RAG), explainability via SHAP, evidence-based clinical-support presentation, visualization using Streamlit, and an evaluation protocol for the system. The framework decomposes the different functions involved in prediction, explanation, evidence retrieval and decision support into separate components, fostering responsible healthcare-AI principles of modularity, transparency and traceability [23, 34, 35].

For the predictive evaluation, 99,986 observations were analysed, 14 exact data repetitions were eliminated. Seven classifiers were evaluated according to the predefined validation F1 criterion and the selected one was XGBoost. On the independent test set, it achieved 97.21% accuracy, 97.04% precision, 69.33% recall, an F1 score of 0.8088, ROC-AUC of 0.9789, and PR-AUC of 0.8852. HbA1c level and blood glucose level were also important contributors to the model predictions based on the SHAP analysis, followed by age and BMI. These explanations depict the behavior of a learned model, and cannot be read as causal medical relationships [25, 27].

Thus, the development of the various complementary predictive, causal, retrieval, agent-coordination and presentation components, is a key feature of this work rather than evidence of clinical effectiveness. The framework is a developmental stage prototype for decision support and is not a validated diagnostics system or clinical decisions' replacement.

B. FUTURE WORK

Further studies need external validation on independent and a wider variety of data and then prospective studies in realistic healthcare settings. The evaluation of discrimination, calibration, subgroup performance and generalizability should be evaluated according to the principles from prediction-model reporting and risk-of-bias principles [36,38]. Additional analysis for the calibration of the model as well as the decision curve analysis is warranted to determine if predicted probabilities offer meaningful clinical potential benefits [45].

The work should be extended for bigger and more representative datasets, federated learning, systematic fairness testing for different demographic groups, clinician-centred usability studies and integration with pre-existing clinical-data infrastructures (standardized). Post-deployment continuous monitoring of the model's behavior and the continuous detection of drift will also be required to pinpoint distribution shifts among the input data and drift in the models' predictive capabilities.

Assessing the relevance of the retrieval, the quality of the retrieved sources, evidence coverage, factual consistency, citation accuracy, and robustness of unsupported LLM generation, are all elements that need to be carefully evaluated in the RAG component. Likewise, the multi-agent workflow should be instrumented and assessed on the success of the tasks, latency, throughput, utilization of resources, failure recovery and traceability. The use of the integrated explanations and retrieved evidence for potential use in future evaluations with healthcare professionals should finally decide if these explanations improve

understanding of an incident or increase the usefulness for healthcare professionals for decisions about the incident. The framework should be viewed and used in a research-like prototype until such validation can be achieved that is supersponsored by explicit human oversight [34] or [35].

REFERENCES

- [1] American Diabetes Association Professional Practice Committee for Diabetes, "2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes—2026," *Diabetes Care*, vol. 49, suppl. 1, pp. S27–S49, 2026, doi: 10.2337/dc26-S002.
- [2] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proc. Symp. Comput. Appl. Med. Care*, 1988, pp. 261–265.
- [3] J. Lindström and J. Tuomilehto, "The diabetes risk score: A practical tool to predict type 2 diabetes risk," *Diabetes Care*, vol. 26, no. 3, pp. 725–731, 2003, doi: 10.2337/diacare.26.3.725.
- [4] P. W. F. Wilson, J. B. Meigs, L. Sullivan, C. S. Fox, D. M. Nathan, and R. B. D'Agostino, "Prediction of incident diabetes mellitus in middle-aged adults: The Framingham Offspring Study," *Arch. Intern. Med.*, vol. 167, no. 10, pp. 1068–1074, 2007, doi: 10.1001/archinte.167.10.1068.
- [5] P. W. F. Wilson, R. B. D'Agostino, C. S. Fox, L. M. Sullivan, and J. B. Meigs, "Type 2 diabetes risk in persons with dysglycemia: The Framingham Offspring Study," *Diabetes Res. Clin. Pract.*, vol. 92, no. 1, pp. 124–127, 2011, doi: 10.1016/j.diabres.2010.12.024.
- [6] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," *Sci. Rep.*, vol. 10, Art. no. 11981, 2020, doi: 10.1038/s41598-020-68771-z.
- [7] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," *BMC Endocr. Disord.*, vol. 19, Art. no. 101, 2019, doi: 10.1186/s12902-019-0436-6.
- [8] M. O. Olusanya, R. E. Ogunakin, M. Ghai, and M. A. Adeleke, "Accuracy of machine learning classification models for the prediction of type 2 diabetes mellitus: A systematic survey and meta-analysis approach," *Int. J. Environ. Res. Public Health*, vol. 19, no. 21, Art. no. 14280, 2022, doi: 10.3390/ijerph192114280.
- [9] M. A. Makroum, M. Adda, and A. Bouzouane, "Machine learning and smart devices for diabetes management: Systematic review," *Sensors*, vol. 22, no. 5, Art. no. 1843, 2022, doi: 10.3390/s22051843.
- [10] V. Bardia and E. Sophiya, "Diabetes prediction using machine learning algorithm: A comparative analysis," in *Proc. 10th Int. Conf. Advanced Computing and Communication Systems (ICACCS)*, 2024, pp. 1973–1979, doi: 10.1109/ICACCS60874.2024.10717264.

- [11] P. Netayawijit, W. Chansanam, and K. Sorn-In, "Interpretable machine learning framework for diabetes prediction: Integrating SMOTE balancing with SHAP explainability for clinical decision support," *Healthcare*, vol. 13, no. 20, Art. no. 2588, 2025, doi: 10.3390/healthcare13202588.
- [12] I. Kırbaş and A. Çifci, "Leveraging SHAP for interpretable diabetes prediction: A study of machine learning models on the Pima Indians Diabetes Dataset," *Balkan J. Electr. Comput. Eng.*, vol. 13, no. 2, 2025.
- [13] "A novel machine learning approach for diagnosing diabetes with a self-explainable interface," *Healthcare Analytics*, vol. 5, Art. no. 100301, 2024, doi: 10.1016/j.health.2024.100301.
- [14] "A comparative study of explainable machine learning models with Shapley values for diabetes prediction," *Healthcare Analytics*, vol. 7, Art. no. 100390, 2025, doi: 10.1016/j.health.2025.100390.
- [15] "Explainable diabetes classification using hybrid Bayesian-optimized TabNet architecture," *Comput. Biol. Med.*, 2023, doi: 10.1016/j.compbiomed.2022.106557.
- [16] S. C. Mackenzie, C. A. R. Sainsbury, and D. J. Wake, "Diabetes and artificial intelligence beyond the closed loop: A review of the landscape, promise and challenges," *Diabetologia*, vol. 67, no. 2, pp. 223–235, 2024, doi: 10.1007/s00125-023-06038-8.
- [17] S. Huang, Y. Liang, J. Li, and X. Li, "Applications of clinical decision support systems in diabetes care: Scoping review," *J. Med. Internet Res.*, vol. 25, Art. no. e51024, 2023, doi: 10.2196/51024.
- [18] P. Jia, P. Zhao, J. Chen, and M. Zhang, "Evaluation of clinical decision support systems for diabetes care: An overview of current evidence," *J. Eval. Clin. Pract.*, vol. 25, no. 1, pp. 66–77, 2019, doi: 10.1111/jep.12968.
- [19] R. Nimri and M. Phillip, "Enhancing care in type 1 diabetes with artificial intelligence driven clinical decision support systems," *Horm. Res. Paediatr.*, vol. 98, no. 4, pp. 384–395, 2025, doi: 10.1159/000546713.
- [20] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. S. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, D. Ludwig, S. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. D. Howell, C. Cui, G. S. Corrado, and J. Dean, "Scalable and accurate deep learning with electronic health records," *npj Digit. Med.*, vol. 1, Art. no. 18, 2018, doi: 10.1038/s41746-018-0029-1.
- [21] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nat. Med.*, vol. 25, pp. 24–29, 2019, doi: 10.1038/s41591-018-0316-z.
- [22] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nat. Med.*, vol. 25, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.
- [23] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, P. N. Ossorio, S. Thadaneys-Israeli, and A. Goldenberg, "Do no harm: A roadmap for responsible machine learning for health care," *Nat. Med.*, vol. 25, pp. 1337–1340, 2019, doi: 10.1038/s41591-019-0548-6.
- [24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015, doi: 10.1038/nature14539.
- [25] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4765–4774.
- [26] M. T. Ribeiro, S. Singh, and C. Guestrin, "“Why should I trust you?”: Explaining the predictions of any classifier," in *Proc. NAACL-HLT Demonstrations*, 2016, pp. 97–101, doi: 10.18653/v1/N16-3020.
- [27] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies," *J. Biomed. Inform.*, vol. 113, Art. no. 103655, 2021, doi: 10.1016/j.jbi.2020.103655.
- [28] S. N. Payrovnaziri, Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J. H. Chen, X. Liu, and Z. He, "Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review," *J. Am. Med. Inform. Assoc.*, vol. 27, no. 7, pp. 1173–1185, 2020, doi: 10.1093/jamia/ocaa053.
- [29] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," *BMC Med. Inform. Decis. Mak.*, vol. 20, Art. no. 310, 2020, doi: 10.1186/s12911-020-01332-6.
- [30] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, 2020, doi: 10.1109/TNNLS.2020.3027314.
- [31] N. A. Aziz, A. Manzoor, M. D. M. Qureshi, M. A. Qureshi, and W. Rashwan, "Explainable AI in healthcare: Systematic review of clinical decision support systems," 2024, doi: 10.1101/2024.08.10.24311735.
- [32] B. A. Van Dort, T. Engelsma, P. Sneekes, L. Peute, and S. Medlock, "Explainable AI in hospital clinical decision support systems: A scoping review of healthcare professionals' perspectives," *PLOS Digit. Health*, vol. 5, no. 5, Art. no. e0001417, 2026, doi: 10.1371/journal.pdig.0001417.
- [33] H. Wang, G. Shi, X. Liu, and Y. Wu, "Explainable AI in medicine: A comprehensive narrative review of methods, applications, and future directions," *Expert Syst.*, 2026, doi: 10.1111/exsy.70259.
- [34] FUTURE-AI Consortium, "FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare," *BMJ*, vol. 388, Art. no. e081554, 2025, doi: 10.1136/bmj-2024-081554.
- [35] B. Vasey, M. Nagendran, B. Campbell, D. A. Clifton, G. S. Collins, S. Denaxas, A. K. Denniston, L. Faes, B. Geerts, M.

- Ibrahim, X. Liu, B. A. Mateen, P. Mathur, M. D. McCradden, L. Morgan, J. Ordish, C. Rogers, S. Saria, D. S. W. Ting, P. Watkinson, W. Weber, P. Wheatstone, and P. McCulloch, "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," *Nat. Med.*, vol. 28, pp. 924–933, 2022, doi: 10.1038/s41591-022-01772-9.
- [36] J. F. Cohen, G. S. Collins, J. A. Bossuyt, et al., "TRIPOD+AI: An updated reporting guideline for clinical prediction models," *BMJ*, vol. 385, Art. no. q824, 2024, doi: 10.1136/bmj.q824.
- [37] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement," *Ann. Intern. Med.*, vol. 162, no. 1, pp. 55–63, 2015, doi: 10.7326/M14-0697.
- [38] R. F. Wolff, K. G. M. Moons, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, and S. Mallett, "PROBAST: A tool to assess the risk of bias and applicability of prediction model studies," *Ann. Intern. Med.*, vol. 170, no. 1, pp. 51–58, 2019, doi: 10.7326/M18-1376.
- [39] X. Liu, S. Cruz Rivera, D. Moher, M. J. Calvert, A. K. Denniston, and the SPIRIT-AI and CONSORT-AI Working Group, "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension," *Nat. Med.*, vol. 26, pp. 1364–1374, 2020, doi: 10.1038/s41591-020-1034-x.
- [40] B. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [41] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [42] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
- [43] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [44] S. Ö. Arık and T. Pfister, "TabNet: Attentive interpretable tabular learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 8, pp. 6679–6687, 2021, doi: 10.1609/aaai.v35i8.16826.
- [45] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: The Achilles heel of predictive analytics," *BMC Med.*, vol. 17, Art. no. 230, 2019, doi: 10.1186/s12916-019-1466-7.
- [46] A. L. Neves, P. P. Rodrigues, A. Mulla, B. Glampson, T. Willis, A. Darzi, and E. Mayer, "Using electronic health records to develop and validate a machine-learning tool to predict type 2 diabetes outcomes: A study protocol," *BMJ Open*, vol. 11, no. 7, Art. no. e046716, 2021, doi: 10.1136/bmjopen-2020-046716.
- [47] Y. Shang, K. Jiang, L. Wang, Z. Zhang, S. Zhou, Y. Liu, et al., "The 30-days hospital readmission risk in diabetic patients: Predictive modeling with machine learning classifiers," *BMC Med. Inform. Decis. Mak.*, vol. 21, Art. no. 57, 2021, doi: 10.1186/s12911-021-01423-y.
- [48] B. Liu, Y. Li, Z. Sun, S. Ghosh, and K. Ng, "Early prediction of diabetes complications from electronic health records: A multi-task survival analysis approach," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, doi: 10.1609/aaai.v32i1.11290.
- [49] P. W. F. Wilson, J. B. Meigs, L. Sullivan, C. S. Fox, D. M. Nathan, and R. B. D'Agostino, "Validation of the Framingham Offspring Study equations for predicting incident diabetes mellitus," *Diabetes Care*, 2008.
- [50] L. Makowski, M. A. et al., "A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion," *Information Fusion*, 2023.