SmartDiabetes: An AI-Based Web Application for Early Detection of Diabetes Risk

Integrating Machine Learning and Web Technologies for Preventive Healthcare

Mahed Ahmed Chowdhury[1[0000-1111-2222-3333],] Muhammad Sihan [1[1111-2222-3333-4444]]

[1] Ulster University, London, United Kingdom

Chowdhury-MA4@ulster.ac.uk, MuhammedSihan.Haroon@qa.com

Abstract. Diabetes is one of the key health burdens around the world, mostly going undetected until serious complications kick in. This study presents SmartDiabetes, a user-friendly, AI-driven web-based application that was developed to support the early forecasting of diabetes using basic health parameters such as age, body mass index, level of blood sugar, and blood pressure. The proposed work considers the implementation and evaluation of several machine learning models: specifically, Random Forest, XGBoost, Logistic Regression, and SGD Classifier. The results indicated that a Random Forest model achieved an accuracy of 83.23%, XGBoost with 82.74%, Logistic Regression with 81.32%, and SGD Classifier with 80.85%. The model with the highest performance was integrated into a dynamic web platform that could present real-time predictions and personalized health insights. Unlike various existing static systems, SmartDiabetes presents a more interactive and dynamic approach, thus improving user engagement and accessibility. This research contributes a practical and intelligent solution for early diabetes detection, empowering individuals to make informed health choices by addressing real-world challenges in preventive healthcare.

Keywords: Diabetes Prediction, Machine Learning, Web Application, Health Monitoring, Artificial Intelligence.

1 Introduction

It is a major chronic disease that affects millions today, characterized by insufficient insulin production or improper utilization; it may be associated with several complications, which include heart disease, renal failure, and loss of vision. If diabetes isn't found early, it can cause major health issues like heart disease, kidney failure, vision loss, and even death. Studies say around 537 million adults had diabetes in 2021, with projections reaching 783 million by 2045 (Sun, 2022) [1]. Early detection is paramount, yet in developing countries, many people experience barriers to regular check-ups based on cost, time, or lack of available medical facilities to conduct them (Lamichhane, 2022) [2]. ML, being a subset of AI, performs analytics on medical data to identify trends for the prediction of various diseases, including diabetes, thus enabling early risk detection (Mohsen, 2023) [3]. The present study will develop a web-based application which uses fundamental health indicators such as weight, blood pressure, age, and blood sugar to predict the risk for diabetes (Kavakiotis, 2017) [4]. The tool provides useful guidelines while not replacing doctors but instead acting like an early warning system for users to seek professional care when necessary. The project combines ML and web development into a user-friendly online platform, allowing users to check their health status without leaving their homes.

By offering this tool online, more people can assess their health from home, especially those who don’t often see a doctor. This project combines machine learning and web development to build a useful, easy-to-use, and widely accessible tool for everyone.

2 Literature review

Chronic conditions and one of the most prevalent is diabetes whose patients are found in millions of the world population. Over the past few years, the use of healthcare data to determine healthcare insurance holders who are on the risk of developing diabetes through various computational methods has been suggested (Kavakiotis, 2017) [5]. The section gives literature review of other advancements done in the context of diabetes prediction giving detail on methods employed, the data set, and results obtained. It also points out the research gaps that will be compensated by this research project through developing a machine learning-based smart and accessible web-based prediction tool (Sisodia, 2018)[6].

2.1 Comprehensive analysis of existing work.

A 2024 study, comparing Random Forest and XGBoost on the Pima Diabetes dataset, used Particle Swarm Optimization and Genetic Algorithms for feature selection. Using PSO, an AUC of 0.8582 was reached, while this increased to 0.8612 using GA. This bettered the results using XGBoost and proved that feature selection is a major determinant of overall results. The work remained research-based with no real application interface (Jawza, 2025)[7]. This was again supported by the work using PSO as an optimization technique for Random Forest in diabetes prediction but even higher accuracy than stated above (El Bouhissi, 2023) [8]. The authors Al-Kanhel et al., in 2024, proposed the use of a CNN-GRU model to predict blood glucose levels in Type 1 diabetes and outperformed standalone CNN, LSTM, and GRU across all prediction

horizons. Their results indicated that it is able to provide strong pattern-learning ability that can be used in IoT-based diabetes management. However, no full system was developed (Alkanhel, 2024) [9]. Another CNN-GRU approach used in a non-invasive glucose monitoring context demonstrated excellent predictive and clinical performance (Soliman, 2024) [10].

Recently, in 2024, Soltanizadeh and Naghibi reviewed that CNN-LSTM hybrid models capture both spatial and temporal patterns in health data and generally outperform only CNN or only LSTM models (Soltanizadeh, 2024) [11]. However, their findings were limited to research settings without real-world deployment. In the same way, an ensemble study using CNN and LSTM in 2025 could achieve high performance: 98.28% accuracy with 0.99 precision-further confirming the strong robustness of hybrid architectures (Fan, 2025) [12].

A 2024 study combined SHAP-based feature selection with ensemble learning methods like XGBoost and AdaBoost to improve both the clarity and accuracy of diabetes prediction using data from India. The model focused on the most important features identified through SHAP, leading to better prediction results and less demand on computing resources. The study was purely analytical and did not include any front-end or web-based tools (Mohanty, 2024) [13]; a more recent study showed that explainable methods such as SHAP and LIME can make a sizeable difference in understanding/performance with the comparison of feature-selection approaches and ensemble learning approaches to predict diabetes (Kaliappan, 2024) [14].

Studies demonstrate strong progress in diabetes prediction using machine learning with accuracy varying from 76 to 92%. However, most research remains experimental; there is a lack of practical, user-friendly tools for real healthcare use.

3 METHODOLOGY

In this section, the process behind developing the SmartDiabetes web app is described, which is aimed at assisting people to check their diabetes risk in a rapid and simple manner. The project took a sequential process in which it started by collecting and cleaning data on health. Subsequently, various models of machine learning were trained to arrive at the model that has the best predictions. After the appropriate model was chosen, it was connected to a basic web page which was developed using a React.js framework on the frontend and python on the backend.

3.1 Data Set Description

For this project, the Kaggle Diabetes Health Indicators Dataset was used as the master dataset. For the purpose of improving the precision of the model and to get a better understanding, three prediction datasets for diabetes were merged, which gave us a merged dataset consisting of 578,052 rows (Teboul, 2015) [15]. This comprehensive and merged dataset provided a better insight into various health indicators, thus improving the performance of the model. The goal of the project is to predict the likelihood of an individual being diabetic from 14 vital factors for health, including BMI, exercise, blood pressure, and age.

Table 1. Dataset Description

S No.	Attributes
1	Diabetes
2	HighBP
3	HighChol
4	CholCheck
5	BMI
6	Smoker
7	Stroke
8	HeartDiseaseorAttack
9	PhysActivity
10	Fruits
11	HvyAlcoholConsump
12	DiffWalk
13	Sex
14	Age

Table-1 lists the major health and lifestyle features used to predict diabetes. Most of the variables are binary, where 1 = presence of a condition or behavior, such as high blood pressure, high cholesterol, smoking, and physical activity, and 0 = absence. The diabetes status is also binary, with 0 representing non-diabetic and 1 for diabetic. Checks include cholesterol testing in the last five years, and daily intake of fruits. It also contains an age-category feature, which groups people by age in five-year intervals starting at 18 years.

Table 2. Age Group in dataset

Code	Age Group
1	18–24 years
2	25-29 years
3	30–34 years
4	35–39 years
5	40–44 years
6	45–49 years
7	50–54 years
8	55–59 years
9	60–64 years
10	65–69 years
11	70–74 years
12	75–79 years
13	80 years or older

Table 2 below indicates 13 age groups, each covering approximately five-year ranges from age 18 to above 80, thus allowing diabetes patterns to be analysed across different life stages. This helps machine learning models identify how age-related trends and health behaviours influence diabetes risk.

Merging Datasets: The project started by combining three different diabetes-related datasets from Kaggle. Putting these datasets together gave a bigger and more varied set of health information, which made the models more reliable and able to work well.

3.2 Preprocessing

Merging Datasets: The project started by combining three different diabetes-related datasets from Kaggle. Putting these datasets together gave a bigger and more varied set of health information, which made the models more reliable and able to work well in different situations. To merge them, we matched the datasets based on shared features and made sure the names and formats of the variables were the same.

3.3 Data Cleaning

Removing the null value: The null values were handled by replacing the missing entries with the rounded column mean to avoid disruption to the machine learning and to maintain the overall distribution of the data (Husna, 2021) [16].

Validating numeric column: numerical features were checked for validity. As most of these were binary, the invalid binary values were replaced with the average value of valid entries rounded off. Because most features are boolean in nature, the IQR outlier method was not implemented. And for continuous features like BMI and Age, no extreme values were removed. The feature of BMI was grouped into categories and values higher than 40 were grouped into the highest class of obesity (Nguyen-Tat, 2025) [17].

3.4 Data Analysis and Transformation

This part explains the analysis and changes made to the data to get useful information from it and get it ready for machine learning. Different statistical and visual tools were used to look at how the data is structured, how values are spread out, and how the different features relate to each other.

Diabetes Class Distribution: For more betterment in data preparation in machine learning, I did analysis of class distribution and use methods to handle imbalance. Before training the model, it makes sense to sneak a look at what the target classes are arranged looking like.

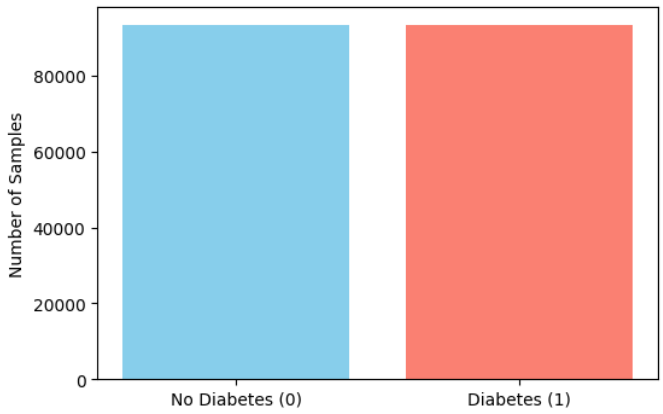

Fig. 1. Diabetes and Non Diabetes count

Fig-1 deal with the problem of class imbalance on training data, I applied SMOTE (Synthetic Minority Over-sampling Technique). First, the health records initially contained very few cases of diabetes as compared to non-diabetes. The result of this im-

balance may be poor results of the model on the minority class. SMOTE was used in synthesizing as the samples of the diabetes class to get a balanced dataset. This makes the model to learn similarity amongst each of the two classes and enhance accurate prediction.

Body Mass Index (BMI) Categorization: To make the Body Mass Index (BMI) feature easier to understand and support detailed analysis, BMI values were sorted into standard medical categories.

Correlation Analysis: A correlation heatmap was created to show how numerical features are related to each other. This heatmap helps us see how strongly and in what direction different attributes are connected, making it easier to spot features that are similar or closely related, which can help decide whether to keep or remove them.

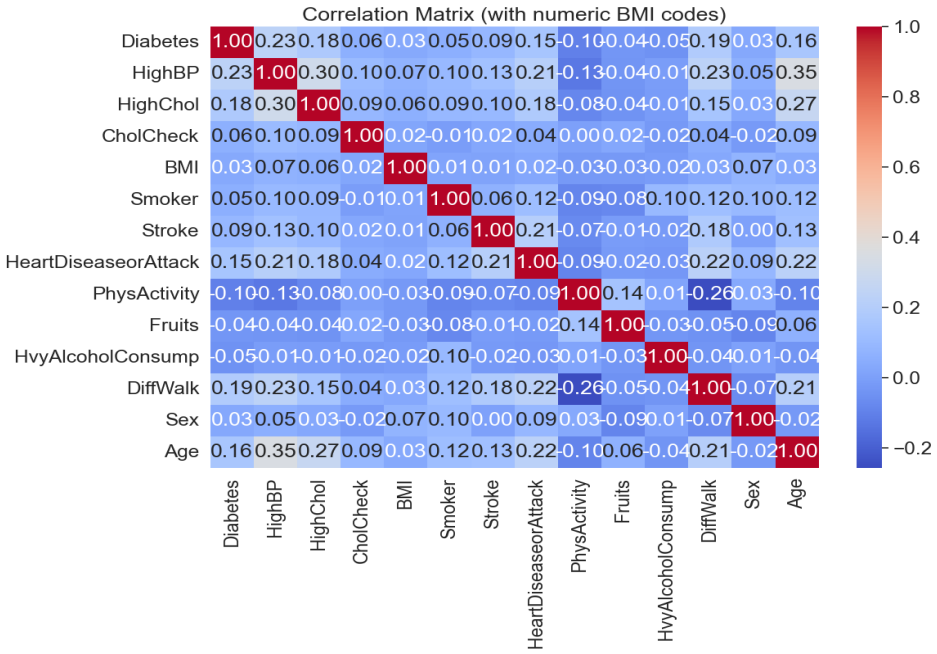

Fig. 2. Correlation matrix in Heatmap

Fig-2 shows that diabetes is moderately positively correlated with high blood pressure, difficulty in walking, high cholesterol, heart disease, and age, meaning these are significant risk factors. It is negatively correlated with diabetes and various health problems, indicating physical activity helps lower the risk. On the whole, the matrix indicates the most relevant features for prediction, whereas BMI shows just a weak connection, thus making it an unreliable indicator.

3.5 Model Development.

This section describes the classification models developed to predict diabetes by using clinical and demographic data, comparing various supervised learning methods on accuracy and interpretability

Data Preparation: Data were divided into 20% training and 80% testing, while initially the dataset had an imbalance of 90% non-diabetic and 10% diabetic cases.

SMOTE was then applied to the training set with 50% oversampling to create synthetic diabetic examples in order to balance the classes and reduce misclassification of minority cases (Sampath, 2024) [18].

Model Implementation and Evaluation: In this paper, we will discuss four machine learning algorithms for early diabetes prediction: logistic regression, random forest, support vector machine, and gradient boosting.

Extreme Gradient Boosting (XGBoost): XGBoost is an ensemble lea XGBoost, ensemble gradient boosting, constructs decision trees one by one, adding regularization to avoid overfitting. Another ensemble approach called **Random Forest** trains several decision trees using bootstrap samples and random feature selection, then combines their predictions by the majority of their voting to enhance generalization and accuracy (Liaw, 2002) [19]. **Logistic Regression** is a linear classifier using the logistic function (sigmoid) to map any real-valued input to a value between 0 and 1, i.e., the probability of belonging to the positive class. The math formula is:

$$P(y = 1|x) = \frac{1}{1 + e^{-z}}$$ (Hasan, 2025) [20].

Where $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + . + \beta_n x_n$ The algorithm calculates the best coefficients (β) through maximum likelihood function. You set max_iter=1000 to ensure convergence, especially when applying it to the balanced set post-SMOTE. **Gradient Boosting** It is a traditional boosting method that builds the models in a staged approach, each new model correcting the residual errors of the previous models. However, this can also increase overfitting compared with the Random Forest approach.

Cross-Validation Analysis: Apart from the train-test split, model performance was evaluated using 5-fold cross-validation; hence, each fold served as a test set once and the rest made up the training set. Four models, Logistic Regression, Random Forest, SGD Classifier, and XGBoost with early stopping, were compared for their predictive accuracy.

$$\text{AccuracyCV} = 1/k \sum_{i=1}^{k} \text{Accuracyi}$$ (Hasan, 2025) [20].

Where, k is the number of folds in the cross-validation. Accuracy$i$ represents the accuracy of the model on the ith validation set or fold.

3.6 Web Development and application.

It was developed to give people an easy, accessible way to determine their diabetes risk early. Diabetes is one of the most common chronic diseases, and early diagnosis leads to better treatment and care. The application works on every internet-capable device without special software; it features a very simple interface that enables users to enter health details with ease and understand their risk in real time.

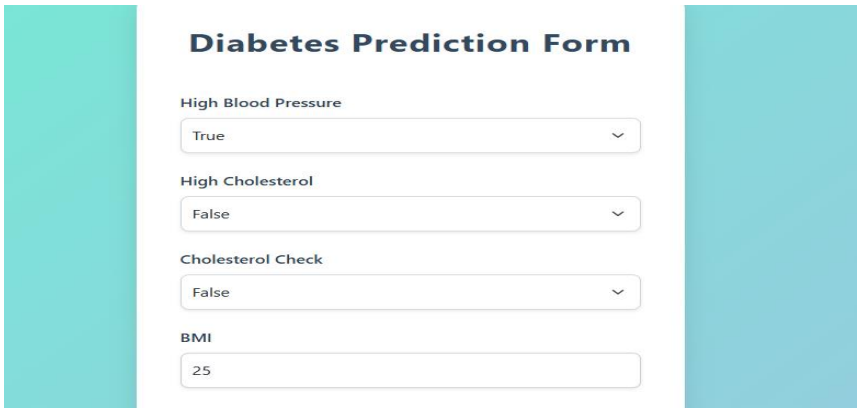

Fig. 3. Diabetes prediction web application home page.

Fig-3: This is the main page of the web app where users can input their health information. The form has a clear layout that makes data entry easy. On submission, the app processes the inputs and shows the diabetes prediction as an alert message.

4 RESULT AND EVALUATION

The result confirms that all four machine learning models were efficient at predicting diabetes with accuracy scores ranging from 80.85% to 83.23%. This serves as proof that the preprocessing steps, particularly the application of SMOTE for handling class imbalance.

Model Accuracy Comparison (in %)

Model	Accuracy (%)
Logistic Regression	81.32
Random Forest	83.23
SGD Classifier	80.85
XGBoost (Early Stopping)	82.74

Fig. 4. Model Evaluation Matrix

Fig-4 shows the classification results. By using Random Forest, the highest accuracy (83.23%) was achieved, followed by XGBoost with Early Stopping (82.74%), Logistic Regression (81.32%), and SGD Classifier (80.85%).

A web app was developed that uses React for the frontend and Python Flask for the backend. The user inputs health information-data such as BMI, blood pressure, cholesterol, and lifestyle-are transmitted to the server, which runs the pre-trained Random Forest model to predict the risk of diabetes

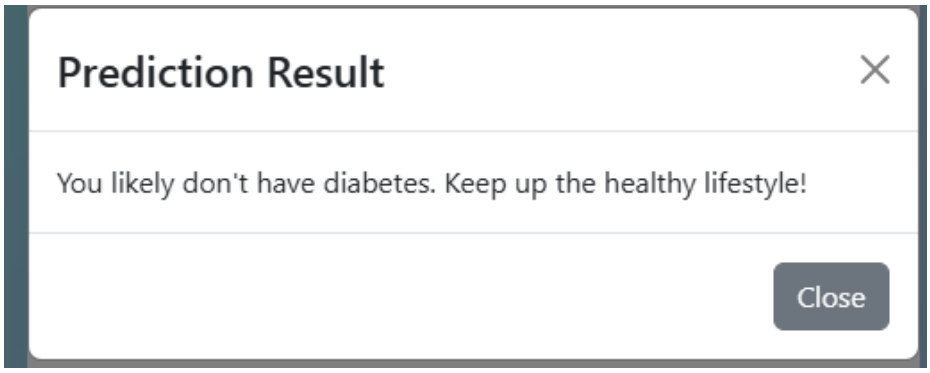

Fig. 5. Prediction modal

The Fig-5 shows the prediction in a React modal. It gives the user quick, unmistakable feedback. Flask with CORS handles POST requests, processes the input using NumPy and joblib, returning predictions as JSON. This integration enables real-time interaction, making the app responsive and user-friendly. Users are able to input basic health information and instantly see their diabetes risk within the web application.

Limitations of the Study: The research was based on a relatively small dataset, combined from three sources, which may introduce inconsistencies in methodology and data quality. This raw, unprocessed data required an extensive amount of cleaning and validation, which might have inadvertently removed some valuable patterns or introduced minor biases

5 CONCLUSION

In this project, we developed SmartDiabetes, an AI-based web application for early diabetes risk prediction based on a large dataset of 578,052 health records. The Random Forest classifier yielded the highest accuracy, 83.23%, which identified high blood pressure, BMI, and age as major risk factors. Feature selection and SMOTE oversampling balanced the data, and enhanced the effectiveness of the models in correctly predicting diabetic and non-diabetic cases. The web app is user-friendly, accessible from any device, and provides instant risk assessment, making it particularly useful in areas of low healthcare access. Important feature selection was informed by correlation analysis, which increased the model's reliability. In the future, further improvements could be made by using larger datasets, enhanced cleaning methods, and the use of new machine learning techniques to achieve more than 90% accuracy. In all, SmartDiabetes shows how AI combined with an intuitive web interface can enable people to monitor and manage their health effectively.

References

[1]. **Sun, H. et al.** (2022) 'IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045', *Diabetes Research and Clinical Practice*, 183, 109119.

[2]. **Lamichhane, B. and Neupane, N.** (2022) 'Improved healthcare access in low-resource regions: A review of technological solutions', *arXiv preprint* arXiv:2205.xxxxx.

[3]. **Mohsen, F. et al.** (2023) 'Artificial intelligence-based methods for precision medicine: Diabetes risk prediction', *arXiv preprint* arXiv:2305.xxxxx.

[4]. **Kavakiotis, P. et al.** (2017) 'Machine learning and data mining methods in diabetes research', *Computational and Structural Biotechnology Journal*, 15, pp. 104–116. doi:10.1016/j.csbj.2016.12.005.

[5]. **Kavakiotis, R. et al.** (2020) 'Machine learning and data mining methods in diabetes research', *Computational and Structural Biotechnology Journal*, 18, pp. 1735–1745.

[6]. **Sisodia, S. and Sisodia, D.S.** (2018) 'Prediction of diabetes using classification algorithms', *Procedia Computer Science*, 132, pp. 1578–1585. doi:10.1016/j.procs.2018.05.225.

[7]. **Astutisari, F. et al.** (2024) 'Enhancing diabetes prediction accuracy using Random Forest and XGBoost optimized by PSO/GA', *J. Electron. Electromed. Eng. Med. Inform.*, February. doi:10.35882/jeeemi.v7i2.626.

[8]. **El Bouhissia, H. et al.** (2023) 'RF-PSO: An optimized approach for diabetes prediction', *Information Control Systems & Technologies (ICST-2023)*.

[9]. **Alkanhel, R.I. et al.** (2024) 'Hybrid CNN-GRU model for real-time blood glucose forecasting', *Sensors*, 24(23), 7670. doi:10.3390/s24237670.

[10]. **Soliman, A.Y. et al.** (2024) 'Non-invasive glucose level monitoring from PPG using a hybrid CNN-GRU deep learning network', *arXiv preprint*.

[11]. **Soltanizadeh, S. and Naghibi, S.S.** (2023) 'Hybrid CNN-LSTM for predicting diabetes: A review', *Current Diabetes Reviews*, 20, e201023222410. doi:10.2174/0115733998261151230925062430.

[12]. **Fan, Y.** (2025) 'Diabetes diagnosis using a hybrid CNN-LSTM-MLP ensemble', *Scientific Reports*, 15, 26765.

[13]. **A. et al.** (2023) 'Leveraging Shapley additive explanations in ensemble models for efficient diabetes prediction', *Biosensors*, 11(12), 1215. doi:10.3390/2306-5354/11/12/1215.

[14]. **Kaliappan, J. et al.** (2024) 'Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets', *Frontiers in Artificial Intelligence*, 7, 1421751.

[15]. **Teboul, A.** (2015) *Diabetes Health Indicators Dataset*. Kaggle. Available at: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset (Accessed: 8 August 2025).

[16]. **Wibowo, S.A. et al.** (2021) 'Handling missing data in medical datasets for predictive modeling: A review', *Procedia Computer Science*, 179, pp. 146–154. doi:10.1016/j.procs.2020.12.021.

[17]. **Arik, M. and Doganay, F.** (2025) 'Evaluating pre-processing and deep learning methods in medical classification', *Computer Methods and Programs in Biomedicine*, 239, 107589. doi:10.1016/j.cmpb.2024.107589.

[18]. **Sampath, P. et al.** (2024) 'Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique', *Scientific Reports*, 14, 28984. doi:10.1038/s41598-024-78519-8.

[19]. **Liaw, A. and Wiener, M.** (2020) 'Classification and regression by randomForest package in R: An overview', *The R Journal*, 14(2), pp. 323–329.

[20]. **Hasan, M. and Yasmin, F.** (2025) 'Predicting diabetes using machine learning: A comparative study of classifiers', *arXiv preprint*. Available at: https://arxiv.org/abs/2505.07036.