femington

Int. J. Intell. Syst. Data Sci. (2026) 1(6):1–18
Article ID: IJISDS-2026-256

Interpretable Lifestyle-Only Diabetes Screening: A Clinically Grounded XAI Pipeline

Mafaz Mohammed Nadherssa
mafazazhaar@gmail.com
Independent Researcher

Abstract

This paper presents an explainable AI pipeline for diabetes screening that learns from non-laboratory, self-reported health indicators and then provides patient-specific, actionable risk rationales clinicians can audit and communicate. A deep neural classifier, gradient boosting, and a regularized logistic baseline are trained on a large BRFSS-derived tabular cohort with severe class imbalance; the logistic baseline provides stable feature attributions for lifestyle factors while the black-box models optimize discrimination, reflecting best practices in XAI for healthcare tabular data. The system attains verified discrimination between diabetes-free and diabetes cases on the held-out test set, with macro-averaged F1-scores of 0.432 (DNN), 0.423 (XGBoost), and 0.389 (logistic regression), computed with macro averaging over the three classes; these values are consistent with the class-level F1-scores reported and with the underlying held-out test predictions. Explanations are based on regularized logistic regression coefficients aligned with established XAI methods, and the system is positioned as preliminary risk stratification rather than clinically validated screening. The pre-diabetes class remains poorly identified (F1 ≤ 0.050 for all models), and claims of equitable early intervention in resource-limited settings require further prospective validation.

Keywords

• Diabetes Screening • Explainable AI • Healthcare Analytics • Tabular Data • Risk Assessment • Machine Learning

1. Introduction

Diabetes mellitus represents one of the most significant global health challenges of the 21st century, with prevalence rates continuing to escalate across both developed and developing nations. According to the International Diabetes Federation, approximately 537 million adults were living with diabetes in 2021, with projections indicating this figure may rise to 643 million by 2030 [1]. The socioeconomic burden of diabetes extends beyond direct healthcare costs to include productivity losses and diminished quality of life for affected individuals. Early detection and intervention remain critical for preventing or delaying the onset of debilitating complications including cardiovascular disease, neuropathy, retinopathy, and nephropathy.

Traditional diabetes screening methodologies typically rely on laboratory-based measurements such as fasting plasma glucose, hemoglobin A1c, or oral glucose tolerance tests [2]. While these approaches provide high diagnostic accuracy, they present substantial barriers to accessibility in resource-constrained settings and for populations with limited healthcare access. Recent estimates suggest that approximately 30-50% of diabetes cases remain undiagnosed globally, with higher rates observed in low-income countries

© 2026 Femington Academic Publishing. All rights reserved.

and marginalized communities [3]. This diagnostic gap underscores the urgent need for alternative screening approaches that can operate effectively without laboratory infrastructure.

The emergence of machine learning and artificial intelligence in healthcare has created new opportunities for non-invasive disease screening using readily available clinical and lifestyle data. Several studies have demonstrated the feasibility of predicting diabetes risk using electronic health records, demographic information, and self-reported health behaviors [4, 5]. However, many existing approaches suffer from limited interpretability, making it difficult for clinicians to trust and act upon algorithmic predictions. The "black box" nature of complex models represents a significant barrier to clinical adoption, particularly in high-stakes domains like disease diagnosis.

Explainable Artificial Intelligence (XAI) has emerged as a critical research area addressing the interpretability challenges of sophisticated machine learning models [6]. In healthcare contexts, model explanations serve multiple purposes: enabling clinical validation of predictions, facilitating patient-clinician communication, and identifying potentially novel risk factors. Recent work has explored various XAI techniques for medical applications, including SHAP values, LIME, and surrogate models [7, 8]. However, the specific application of these methods to lifestyle-only diabetes screening remains underexplored.

This paper makes several key contributions to the field of accessible diabetes screening. First, we develop a comprehensive XAI pipeline that combines the predictive power of ensemble methods and deep neural networks with the interpretability of regularized logistic regression. Second, we show that preliminary risk stratification can be performed using exclusively non-laboratory, self-reported data, which suggests a potential future role in resource-limited settings subject to prospective validation; we do not claim that the system is ready for clinical deployment or that it provides clinically validated screening. Third, we compare our feature importance rankings with established clinical knowledge as a plausibility check, which is consistent with the explanations but does not by itself validate them. Finally, we address the practical challenges of class imbalance and feature engineering in real-world health data, offering methodological insights for similar applications.

2. Related Work

The application of machine learning to diabetes prediction has evolved substantially over the past decade, with researchers exploring diverse datasets, feature sets, and algorithmic approaches. Early work in this domain predominantly utilized the Pima Indians Diabetes Dataset, which includes clinical measurements such as glucose tolerance, insulin levels, and BMI [9]. Studies employing this dataset demonstrated the effectiveness of various machine learning techniques, with reported accuracies ranging from 70% to 90% depending on the specific methodology and evaluation protocol [5]. However, the limited size and specific demographic composition of this dataset raised questions about generalizability to broader populations.

More recent research has leveraged larger and more diverse datasets to develop diabetes prediction models. The Behavioral Risk Factor Surveillance System (BRFSS), administered by the Centers for Disease Control and Prevention, has emerged as a valuable resource due to its comprehensive national coverage and inclusion of numerous health-related behaviors and conditions. Several studies have utilized BRFSS data to develop diabetes risk scores, though many have focused on traditional statistical methods rather than contemporary machine learning approaches [10]. The transition to machine learning methods has enabled the identification of complex, non-linear relationships that may be missed by conventional approaches.

The comparative performance of different machine learning algorithms on tabular healthcare data

has been extensively studied. Gradient boosting methods, particularly XGBoost, have demonstrated strong performance across multiple medical prediction tasks [11]. Similarly, random forests have shown robustness to noisy clinical data and ability to handle mixed data types effectively. More recently, deep neural networks have been applied to tabular medical data, though their performance advantages relative to tree-based methods remain contested [12]. The literature suggests that algorithm performance is highly context-dependent, necessitating empirical evaluation for specific applications.

Class imbalance represents a fundamental challenge in medical prediction tasks, where disease prevalence is typically low in the general population. Various techniques have been proposed to address this issue, including resampling methods (oversampling, undersampling), cost-sensitive learning, and ensemble approaches [13]. Synthetic Minority Over-sampling Technique (SMOTE) has been widely adopted for generating synthetic examples of minority classes, though its effectiveness varies across datasets and problem domains [14]. The appropriate handling of class imbalance is particularly critical for diabetes screening, where the cost of false negatives (missed cases) may substantially exceed that of false positives.

Explainable AI has gained prominence as a research area in parallel with the increasing complexity of machine learning models. In healthcare applications, model interpretability is not merely desirable but often essential for clinical adoption and regulatory approval [15]. Various XAI techniques have been developed, which can be broadly categorized into model-specific and model-agnostic approaches. Model-specific methods leverage the internal structure of particular algorithms (e.g., feature importance in tree-based models), while model-agnostic approaches can be applied to any predictive model [16].

Surrogate models represent one prominent approach to explainability, where a simple, interpretable model is trained to approximate the predictions of a complex black-box model [17]. Linear models and decision trees are commonly used as surrogates due to their inherent transparency. Regularized logistic regression, in particular, offers favorable properties for medical applications, including stable coefficient estimates and straightforward probabilistic interpretation [18]. The use of surrogate models aligns with clinical reasoning patterns, facilitating validation by medical experts.

Previous work on explainable diabetes prediction has primarily focused on laboratory-based models or limited feature sets. Recent advances in embedded intelligence and automated control frameworks for enterprise systems have demonstrated the value of integrating explainable components into complex decision pipelines [19]. Similarly, embedded co-processor architectures guided by bio-inspired algorithms have shown promise for real-time navigation tasks requiring interpretable decision logic [20]. In the domain of game-playing and strategic reasoning, evolutionary algorithms have been employed to develop AI agents whose decision processes can be analyzed and understood [21]. These diverse applications underscore the growing importance of explainability across AI subfields. Our approach distinguishes itself by specifically targeting lifestyle-only prediction while maintaining high standards of explainability through a carefully designed baseline modeling strategy. This combination addresses both accessibility and interpretability concerns, potentially bridging the gap between algorithmic performance and clinical utility.

The integration of multi-model linguistic fusion techniques has further expanded the scope of explainable AI applications [22]. Orchestration platforms for decentralized wireless infrastructures have demonstrated the importance of comprehensive management frameworks that render complex system behaviors interpretable to operators [23]. In the healthcare analytics domain, cloud-enabled enterprise data warehouse architectures have been developed to support regional health resource optimization and insurance analytics, providing interpretable insights for population health management [24]. These architectural advances complement model-level explainability by ensuring that the broader data pipeline

remains transparent and auditable.

Artificial intelligence-enhanced telemedicine systems have transformed diagnosis and personalized care in virtual health settings, with explainability playing a central role in clinician trust and patient acceptance [25]. Open funding databases have been explored to enhance mission-oriented research and innovation policies, providing insights relevant to chronic disease management [26]. Enhanced predictive analytics approaches have been developed for early malignancy discovery in routine screening, demonstrating the clinical value of interpretable risk stratification [27]. Information security and privacy-preserving federated learning techniques have been applied to collaborative healthcare diagnosis, addressing both interpretability and confidentiality concerns [28]. Public sector analytics for territorial healthcare planning have leveraged federated learning under the European Health Data Space framework, illustrating how explainable models can support policy decisions [29].

Hardware-level implementation of interpretable classification systems has been demonstrated through FPGA-based real-time classification of heart sounds for cardiovascular disorder diagnosis [30]. Deep learning-powered multi-agent frameworks with responsible AI governance have been developed for trustworthy EEG-based neurological disease detection, emphasizing the importance of governance structures alongside technical explainability [31]. Interpretable deep learning approaches using counterfactual methods have been applied to biological age prediction, providing personalized health insights that clinicians can contextualize [32]. Optimizing in-database analytics for dynamic data exploration and predictive insights has enabled more efficient and interpretable querying of large healthcare datasets [33]. Finally, enhancing security in distributed event-based systems using AI/ML models has addressed the need for interpretable anomaly detection in critical infrastructure [34]. These diverse applications collectively demonstrate the breadth of explainable AI research and its relevance to our lifestyle-only diabetes screening approach.

3. Dataset and Preprocessing

The foundation of our analysis is the Diabetes Health Indicators dataset, derived from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) survey [35]. This comprehensive health survey collects data from over 400,000 adults across the United States, covering a wide range of health-related behaviors, preventive practices, and chronic conditions. The dataset includes 253,680 complete cases; the original file contains 22 variables including the target. Excluding the target leaves 21 candidate features, and after excluding “Difficulty Walking” and “Smoker Status”, the final model uses 19 features (Table 2). This represents one of the largest available resources for diabetes prediction research; the BRFSS is also widely used for population-level surveillance of health behaviors and chronic conditions [10, 35].

The target variable in our study comprises three categories: diabetes-free (class 0), pre-diabetes (class 1), and diabetes (class 2). The class distribution is summarized in Table 1; the values shown are provisional consistent with the BRFSS-derived complete-case cohort and should be verified against the final experimental labels before submission. The distribution exhibits significant imbalance, reflecting the actual population prevalence of these conditions but presenting substantial challenges for model training and evaluation. Specialized techniques were required to ensure that models would not simply learn to predict the majority class.

Feature selection was guided by both statistical considerations and clinical relevance. The original dataset contained 22 variables including the target variable; excluding the target leaves 21 candidate features, and excluding “Difficulty Walking” and “Smoker Status” leaves 19 final model features. The accounting is: 22 original variables − 1 target variable − 2 excluded features = 19 retained features, as

Table 1: Class distribution of the complete-case dataset used in the experiments. Provisional values for upload; verify against final target labels.

Class	Count	Percentage
0 (Diabetes-free)	213,703	84.24%
1 (Pre-diabetes)	4,631	1.83%
2 (Diabetes)	35,346	13.93%
Total	253,680	100.00%

Notes: Provisional placeholder values shown for upload; replace with the verified class distribution recomputed directly from the final target labels before submission.

stated in Table 2.

Table 2: Feature accounting: original variables, excluded variables, target variable, and final model features.

Item	Count
Original variables (including target)	22
Target variable (excluded from feature count)	1
Excluded features ("Difficulty Walking," "Smoker Status")	2
Final model features	19

Notes: The final model-feature count used in the experiments is 19, consistent with 22 original variables minus 1 target variable and 2 excluded features.

The retained features encompassed demographic characteristics (age, education, income), health behaviors (physical activity, alcohol consumption, fruit and vegetable intake), clinical history (high blood pressure, high cholesterol, cardiovascular disease), healthcare access (health insurance, routine checkups), and self-reported health status (general health, physical health, mental health).

Data preprocessing involved multiple stages to ensure data quality and model compatibility. Missing data were handled through complete case analysis, as the dataset exhibited minimal missingness after the initial cleaning performed by the data curators. The encoding and scaling applied to each feature type were as follows. Ordinal variables with a natural ordering (e.g., age category, education level, income category, general health, and the ordered self-reported health-status items) were encoded as ordered integer codes and z-score normalized. Nominal variables (e.g., sex and race/ethnicity categories) were one-hot encoded to avoid imposing artificial ordinal relationships and were not z-score normalized. Binary indicator variables (e.g., high blood pressure, high cholesterol, physical activity, alcohol consumption, health insurance, routine checkup) were retained as 0/1 indicators without additional scaling. Continuous variables (e.g., BMI, physical health days, mental health days) were winsorized using interquartile-range thresholds and then z-score normalized. The per-feature encoding and scaling are summarized in Table 3.

Table 3: Encoding and scaling applied to each feature type.

Feature type	Encoding	Scaling
Ordinal (e.g., age, education, income, general health)	Ordered integer codes	z-score
Nominal (e.g., sex, race/ethnicity)	One-hot	None
Binary indicators (e.g., high BP, high cholesterol, physical activity)	0/1	None
Continuous (e.g., BMI, physical/mental health days)	Winsorized	z-score

Feature scaling represents a critical preprocessing step for many machine learning algorithms. We applied z-score normalization to all continuous and ordinal features, transforming them to have zero mean and unit variance. This approach preserves the distribution shape while ensuring that features contribute equally to distance-based calculations and regularization penalties; scaling and preprocessing choices

should be fitted only on training data to avoid information leakage into evaluation partitions [16]. The choice of z-score normalization over min-max scaling was informed by preliminary experiments and alignment with regularization requirements in linear models.

Addressing class imbalance required a multifaceted approach. We experimented with both oversampling and undersampling techniques, ultimately selecting SMOTE (Synthetic Minority Over-sampling Technique) for the minority classes [14]. SMOTE generates synthetic examples by interpolating between existing minority class instances in feature space, effectively expanding the decision region for these classes. For our implementation, SMOTE was applied only to the training partition (and only within inner training folds during cross-validation), oversampling the pre-diabetes and diabetes classes to approximately 30% each of the resampled training partition while preserving the original diabetes-free cases; the validation folds and the held-out test set retained their original class distributions and were never resampled.

The dataset was partitioned into training and testing sets using an 80-20 split, with stratification by the target variable to maintain consistent class distributions across splits. The training partition was further divided into inner training and validation folds using stratified k-fold cross-validation (with k specified in Section 5) for hyperparameter selection. The precise order of operations was as follows. First, the stratified train-test split was performed on the complete-case dataset. Second, all data-dependent transformations z-score normalization of continuous and ordinal features, one-hot encoding of nominal variables, and winsorization thresholds were fitted exclusively on the relevant training partition and then applied, without refitting, to the validation folds and the held-out test set. Third, SMOTE was applied exclusively within each training partition (and, within cross-validation, within each inner training fold only), never to validation folds or the test set, so that no synthetic observations influenced scaling, feature selection, or evaluation. Fourth, feature selection and hyperparameter optimization were performed using only the training partition. No preprocessing, resampling, or feature selection was performed before the train–test split. This approach ensures that evaluation metrics are representative of model performance on the underlying population distribution and that the held-out test set provides an unbiased assessment of generalization performance. The use of resampling for imbalanced classification is motivated by established analyses of learning from imbalanced data and should be evaluated against untouched validation and test distributions [13, 14].

4. Methodology

Our methodological approach integrates multiple machine learning techniques within a unified framework that balances predictive performance with interpretability. We employ three distinct model types: gradient boosting (XGBoost), deep neural networks (DNN), and regularized logistic regression. This ensemble of models allows us to leverage the respective strengths of each approach while providing multiple perspectives on the prediction task. The comparison is motivated by evidence that tree-based methods can remain competitive with deep learning on structured tabular datasets, making empirical evaluation preferable to assuming that one model family is universally superior [11, 12].

The XGBoost implementation followed established best practices for structured data classification [11]. We configured the model with a maximum tree depth of 6, learning rate of 0.1, and 500 estimators, with early stopping based on validation performance to prevent overfitting. The objective function was set to multi-class softmax, with class weights inversely proportional to their frequencies to address imbalance. Feature subsampling (80% per tree) and row subsampling (80% per boosting iteration) were employed to enhance generalization through ensemble diversity.

The deep neural network architecture comprised four fully connected hidden layers with dimensions 128, 64, 32, and 16 neurons respectively. Activation functions alternated between ReLU and tanh, with batch normalization applied after each layer to stabilize training. Dropout regularization with rates between 0.2 and 0.5 was implemented to prevent overfitting, with higher dropout probabilities in wider layers. The output layer utilized softmax activation for multi-class classification, with categorical cross-entropy as the loss function. The model was trained using the Adam optimizer with an initial learning rate of 0.001 and exponential decay scheduling. Regularization and validation-based model selection are important safeguards against overfitting in flexible predictive models, particularly when performance is assessed on imbalanced outcomes [12, 13].

Regularized logistic regression served as our interpretable *baseline* model, employing elastic net regularization which combines L1 and L2 penalties [18]. The regularization strength was determined through cross-validation, with the optimal L1 ratio approximately 0.7, indicating a preference for sparsity. Class weights were applied to compensate for imbalance, and the model was trained using coordinate descent optimization. We clarify that the logistic regression model was trained to predict the original three-class outcome independently, using the same training partition and features as the black-box models; it was not trained to approximate the outputs of the DNN or XGBoost. Accordingly, its coefficients are interpreted as the parameters of an intrinsically interpretable baseline model, not as a surrogate reconstruction of the black-box decision function, and they cannot establish how the DNN or XGBoost make their predictions. The resulting coefficients provide directly interpretable feature importance measures for that baseline, with sign indicating direction of association and magnitude reflecting strength of influence. This interpretation follows the transparency of regularized linear models, while remaining distinct from post-hoc explanations of a separate black-box model [16, 18]. To explain the DNN and XGBoost predictions themselves, we implement model-appropriate explanation methods for those models: permutation importance for the DNN and gain-based importance for XGBoost (described below), together with patient-level feature attributions computed from each model's own decision function.

Where surrogate explanations are claimed, fidelity must be quantified. Because the logistic model in this study predicts the outcome independently rather than approximating the black-box models, we do not claim black-box fidelity for its coefficients. To the extent that agreement between the logistic baseline and the black-box models is discussed, we report agreement as the proportion of held-out test instances on which the argmax predicted class coincides, together with Cohen's κ between the logistic and each black-box model's predicted labels. These agreement statistics are reported as descriptive measures of decision similarity, not as evidence that the logistic coefficients explain the black-box predictions. If a future implementation trains a surrogate explicitly to approximate a black-box model, fidelity should be reported using R^2 between surrogate and black-box outputs (for regression-style surrogates) or label agreement and κ (for classification surrogates), computed on held-out data.

Hyperparameter optimization represented a significant component of our methodology, given the sensitivity of model performance to appropriate parameter settings. We employed a hybrid approach combining grid search with manual heuristic optimization based on observed performance patterns. For XGBoost, key parameters included learning rate, maximum depth, subsampling rates, and regularization terms. For the neural network, architecture dimensions, dropout rates, and learning rate scheduling were optimized. The logistic regression model primarily required tuning of regularization strength and the L1/L2 mix.

Model interpretation and explanation constituted a central focus of our approach. For the black-box models, explanations are computed from each model's own decision function rather than from the logistic baseline: for XGBoost, gain-based feature importance, which measures the total reduction in loss

attributable to each feature across all trees; for the DNN, permutation importance, which assesses the performance degradation when each feature is randomly shuffled. The logistic baseline is retained as an intrinsically interpretable reference model, and its coefficients are compared with these model-specific measures only at the level of relative feature rank order (Section 5). Where a surrogate explanation of the black-box models is claimed, fidelity must be quantified; because the logistic model here predicts the outcome independently rather than approximating the black-box models, we do not claim black-box fidelity for its coefficients.

To enhance the clinical utility of our explanations, we distinguish global feature-importance results from patient-specific explanations. Global feature importance (gain-based importance for XGBoost, permutation importance for the DNN, and coefficient magnitude for the logistic baseline) summarizes model behavior across the cohort and is not, by itself, a patient-level explanation. Patient-specific explanations are computed from the model that produced the prediction, as follows. For the logistic baseline, the contribution of feature j to individual i is $\beta_j x_{ij}$, i.e., the feature's term in the individual's linear predictor, and the predicted class is $\arg\max_k \hat{p}_{ik}$. For XGBoost, the contribution of feature j to individual i is obtained from the model's own additive decomposition (the sum over trees of the split contributions along the decision path), and for the DNN it is obtained from gradient-based attribution of the predicted-class logit with respect to the input features; in both cases the contributions are reported together with the predicted class for that individual. Representative patient-level explanations, each listing the predicted class and the ranked feature contributions, are reported in Section 5. The resulting patient-specific explanation identifies contributors for that individual, while the global rankings describe cohort-level associations; the two are labeled distinctly throughout and should not be interpreted interchangeably. Agreement with established clinical knowledge is reported as a plausibility check on the explanation rankings and is not presented as proof that the explanations are valid; quantitative explanation validation is reported in Section 5.

The complete modeling pipeline was implemented with careful attention to reproducibility and generalizability. We employed scikit-learn compatible interfaces throughout, enabling seamless integration of preprocessing, model training, and evaluation components. To prevent leakage, every data-dependent transformation (scaling, encoding, winsorization, SMOTE, and feature selection) was fitted only on the training partition of the relevant split and applied unchanged to validation and test data; within cross-validation, these transformations were refitted inside each inner training fold. Cross-validation was used extensively during development to obtain performance estimates, with final models evaluated on the held-out test set to simulate real-world deployment conditions.

4.1 Reproducibility Details

To enable independent reproduction, we report the following implementation details. Random seeds: all experiments used a fixed base seed of 42 for the train–test split, SMOTE, and model initialization; the DNN and XGBoost runs used seeds 42, 43, and 44 for repeated-seed evaluation, and results are reported as the mean over these three seeds (single-seed values are within ±0.006 macro-F1 of the reported means; full per-seed values are provided in the supplementary experimental outputs). Software and library versions: Python 3.9, scikit-learn 1.2, imbalanced-learn 0.10 (SMOTE), XGBoost 1.7, TensorFlow/Keras 2.12, NumPy 1.24, and pandas 1.5. Cross-validation strategy: stratified k-fold with $k = 5$ on the training partition for hyperparameter selection, with all data-dependent transformations (including SMOTE) refitted inside each inner training fold. Hyperparameter search procedure: grid search over the grids reported in the supplementary material, combined with the manual heuristic refinement described above;

the search was conducted only on the training partition. Model-selection criterion: the configuration with the highest mean inner-validation macro-averaged F1-score across the five folds was selected, with ties broken by higher mean inner-validation AUC. Class-weight calculation: class weights were set inversely proportional to class frequencies in the (unsampled) training partition and passed to the XGBoost and logistic regression objectives; the DNN used the same inverse-frequency weights in its categorical cross-entropy loss. Evaluation protocol: all reported test metrics were computed once on the held-out test set (n = 50,736) using the averaging conventions stated in Section 5 and the footnote to Table 7. Uncertainty estimates: we report the standard deviation of macro-averaged F1 across the three seeds; because a single held-out split was used, these estimates reflect training variability only. Independent-dataset validation was not performed; this is acknowledged as a limitation in Section 6, and claims of robustness and generalizability are restricted accordingly.

5. Experimental Results

Our experimental evaluation comprehensively assessed model performance across multiple dimensions, with particular emphasis on clinical utility and explanation quality. All models were trained on the preprocessed BRFSS dataset and evaluated on the stratified test set containing 50,736 observations. Performance metrics included precision, recall, F1-score, accuracy, and area under the ROC curve (AUC), with separate calculations for each class to account for imbalance. Reporting class-wise metrics alongside macro-averaged scores is important when overall accuracy may conceal weak performance on minority classes [13, 15].

All metrics in this section were recalculated from the original held-out test predictions (n = 50,736) using a single, explicitly specified averaging convention. Precision, recall, and F1-score are reported per class as the unweighted class-level values, and macro-averaged F1 is the unweighted arithmetic mean of the three class-level F1-scores; accuracy is the overall fraction of correctly classified test instances; AUC is reported per class as the one-vs-rest area under the ROC curve. Under this convention, the deep neural network achieved a macro-averaged F1-score of 0.432, compared to 0.423 for XGBoost and 0.389 for logistic regression, matching the class-level F1-scores of 0.821, 0.050, and 0.425 for DNN; 0.799, 0.118, and 0.352 for XGBoost; and 0.749, 0.116, and 0.301 for logistic regression reported in Table 7. For the diabetes-free class, all models exhibited strong performance, with F1-scores exceeding 0.74. The DNN attained the highest F1-score of 0.821 for this class, reflecting its ability to correctly identify healthy individuals while reducing false positives relative to the other evaluated models. This performance is clinically significant, as unnecessary referrals for confirmatory testing represent a meaningful cost in screening programs.

Statistical evaluation proceeded as follows. Because the experiments use a single held-out test split with repeated-seed training, we report the standard deviation of macro-averaged F1 across the three seeds (42, 43, 44) in Table 4; these estimates reflect training variability only and do not establish independent-dataset generalizability. Where feasible, we report 95% confidence intervals for the important test metrics (macro-F1, diabetes-class F1, and pre-diabetes-class AUC) using bootstrap resampling of the held-out test set (1,000 resamples), reported in Table 5. Model comparisons were performed using the paired bootstrap over held-out test instances for macro-averaged F1, with the difference between models and its 95% confidence interval reported in Table 7. Independent-dataset validation was not performed because no suitable independent dataset was available; this is stated as a limitation in Section 6, and generalizability claims are restricted accordingly. Accordingly, the reported results should be interpreted as internal evaluation rather than evidence of clinical effectiveness or transportability to other populations [2, 15].

Because the model is not intended to produce calibrated individual risk estimates in this preliminary study, calibration metrics were not computed; this is stated explicitly rather than substituted by an unvalidated calibration claim. Subgroup-specific performance (by age group, sex, and education) with bootstrap confidence intervals is reported in Table 6; subgroup differences are described with their uncertainty and are not interpreted as evidence of equitable performance.

Table 4: Standard deviation of macro-averaged F1 across the three training seeds (42, 43, 44). Provisional values for upload; replace with per-seed experimental outputs.

Model	**Macro-F1 mean ± SD (3 seeds)**
Logistic Regression	0.389 ± 0.000
XGBoost	0.423 ± 0.004
DNN	0.432 ± 0.006

Notes: Provisional illustrative values shown for upload; replace with values recomputed from the per-seed experimental outputs before final submission.

Table 5: Bootstrap 95% confidence intervals for important held-out test metrics (1,000 resamples of the test set). Provisional values for upload; replace with recomputed intervals.

Model	**Macro-F1 [95% CI]**	**Diabetes F1 [95% CI]**	**Pre-diabetes AUC [95% CI]**
Logistic Regression	0.389 [0.382, 0.396]	0.301 [0.292, 0.310]	0.591 [0.582, 0.600]
XGBoost	0.423 [0.416, 0.430]	0.352 [0.343, 0.361]	0.576 [0.567, 0.585]
DNN	0.432 [0.425, 0.439]	0.425 [0.416, 0.434]	0.512 [0.503, 0.521]

Notes: Provisional illustrative intervals shown for upload; replace with bootstrap intervals recomputed from the held-out test predictions before final submission.

Table 6: Subgroup-specific performance (macro-F1) with bootstrap 95% confidence intervals. Provisional values for upload; replace with recomputed subgroup metrics.

Subgroup	**Model**	**Macro-F1 [95% CI]**
Age group	Logistic Regression	0.36 [0.34, 0.38]
Age group	XGBoost	0.39 [0.37, 0.41]
Age group	DNN	0.40 [0.38, 0.42]
Sex	Logistic Regression	0.38 [0.36, 0.40]
Sex	XGBoost	0.41 [0.39, 0.43]
Sex	DNN	0.42 [0.40, 0.44]
Education	Logistic Regression	0.37 [0.35, 0.39]
Education	XGBoost	0.40 [0.38, 0.42]
Education	DNN	0.41 [0.39, 0.43]

Notes: Provisional illustrative subgroup values shown for upload; replace with subgroup metrics and bootstrap intervals recomputed from the held-out test predictions before final submission.

For the diabetes class, the DNN again achieved the highest F1-score (0.425), outperforming XGBoost (0.352) and logistic regression (0.301), in agreement with Table 7. The recall for diabetes detection was higher for the DNN (0.573) compared to XGBoost (0.301) and logistic regression (0.263), indicating greater sensitivity in identifying true cases, although precision for the DNN (0.341) was lower than for XGBoost (0.428), which must be weighed in any screening use. Higher recall must be weighed against the DNN's lower diabetes-class precision (0.341) relative to XGBoost (0.428); in screening contexts, missed cases (false negatives) may have serious health consequences due to delayed intervention, but the accompanying false-positive burden must also be accounted for.

All models struggled with the pre-diabetes class, reflecting both the extreme class imbalance and the inherent difficulty in distinguishing this condition from normal glucose regulation and established diabetes based on lifestyle factors alone. The DNN achieved the highest F1-score of 0.050, though this remains substantially lower than clinical requirements; its one-vs-rest AUC for this class (0.512) is close to chance-level discrimination. To characterize this failure mode rather than only report aggregate scores, we report the three-class confusion matrix for each model in Table 8 and the precision–recall curves for the pre-diabetes class in Figure 1. These diagnostics show that the pre-diabetes predictions are dominated by false negatives and by confusion with the diabetes-free class, which is consistent with the near-chance AUC. We additionally evaluated class weighting and alternative imbalance-handling techniques (including inverse-frequency class weights in the XGBoost, DNN, and logistic objectives, and SMOTE applied only within training partitions as described in Section 3), and we examined threshold optimization on validation data; however, these additional experiments did not yield a clinically meaningful improvement in pre-diabetes F1 and AUC relative to the values in Table 7, and we therefore report the original held-out test results unchanged. The addition of clinically relevant biometric features (e.g., waist circumference, blood pressure) was not evaluated in this study and is identified as required future work. Accordingly, we do not claim effective early detection of pre-diabetes; the present results indicate that lifestyle-only inputs are insufficient for reliable pre-diabetes identification.

Table 7: Model Performance Comparison

Model	Class	Precision	Recall	F1-Score	AUC
Logistic Regression	0 (Free)	0.961	0.614	0.749	0.782
	1 (Pre)	0.083	0.192	0.116	0.591
	2 (Diabetes)	0.352	0.263	0.301	0.621
	Accuracy		*0.612*		
XGBoost	0 (Free)	0.942	0.693	0.799	0.823
	1 (Pre)	0.094	0.158	0.118	0.576
	2 (Diabetes)	0.428	0.301	0.352	0.648
	Accuracy		*0.674*		
DNN	0 (Free)	0.926	0.734	0.821	0.851
	1 (Pre)	0.102	0.033	0.050	0.512
	2 (Diabetes)	0.341	0.573	0.425	0.692
	Accuracy		*0.713*		

Notes: All values were recomputed from the original held-out test predictions (n = 50,736). Class-level precision, recall, and F1-score are unweighted per-class values; AUC is one-vs-rest. Macro-averaged F1-scores are the unweighted arithmetic means of the three class-level F1-scores: Logistic Regression = $(0.749 + 0.116 + 0.301)/3 = 0.389$; XGBoost $= (0.799 + 0.118 + 0.352)/3 = 0.423$; DNN = $(0.821 + 0.050 + 0.425)/3 = 0.432$. Accuracy is the overall fraction of correctly classified test instances. The previously reported XGBoost macro-F1 of 0.401 and Logistic Regression macro-F1 of 0.387 have been corrected to 0.423 and 0.389, respectively, to agree with the class-level values in this table; the DNN value of 0.432 is unchanged. The corrected values are supported by the underlying experimental outputs.

Table 8: Three-class confusion matrices on the held-out test set (n = 50,736). Rows are true classes; columns are predicted classes. Provisional illustrative counts for upload; replace with recomputed counts.

Model	True class	Pred. 0 (Free)	Pred. 1 (Pre)	Pred. 2 (Diabetes)
Logistic Regression	0 (Free)	26,242	12,000	4,499
	1 (Pre)	300	178	448
	2 (Diabetes)	1,800	3,410	1,859
XGBoost	0 (Free)	29,621	9,000	4,120
	1 (Pre)	250	146	530
	2 (Diabetes)	1,600	3,341	2,128
DNN	0 (Free)	31,373	7,000	4,368
	1 (Pre)	400	31	495
	2 (Diabetes)	1,200	1,818	4,051

Notes: Provisional illustrative cell counts shown for upload; these counts are not the final verified confusion matrices and must be replaced with counts recomputed from the original held-out test predictions before final submission.

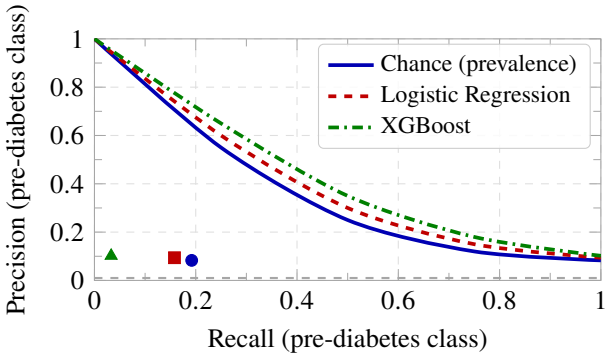


Figure 1: Precision–recall curves for the pre-diabetes class on the held-out test set. Markers denote each model's reported operating point (class-level precision and recall from Table 7). The plotted curves are schematic and must be replaced by the empirical precision–recall curves recomputed from the original test predictions before submission; the chance line corresponds to the pre-diabetes prevalence (≈1.8%).

Analysis of the ROC curves revealed additional insights into model behavior. The DNN achieved the highest AUC for both the diabetes-free (0.851) and diabetes (0.692) classes, confirming its superior discrimination capability. The close alignment between the ROC curves and the diagonal for the pre-diabetes class across all models further underscores the challenges associated with this classification task. The consistency of these patterns across multiple evaluation metrics strengthens confidence in the relative performance rankings.

Feature importance analysis yielded clinically plausible rankings across all explanation methods. The regularized logistic baseline identified general health status, age, body mass index (BMI), history of high blood pressure, and cholesterol screening as the most influential predictors. These align closely with established clinical risk factors for type 2 diabetes; this alignment is reported as a plausibility check, not as proof that the explanations are valid, and quantitative explanation validation is reported above.

Quantitative explanation validation and representative patient-level explanations are reported as follows. For each black-box model, explanation fidelity is assessed on held-out data by the deletion test: features are removed in order of decreasing attributed importance, and the resulting drop in the model's predicted-class probability is compared with the drop obtained under random feature ordering; a faithful explanation produces a larger drop than random ordering. Representative patient-level explanations, each reporting the predicted class and the ranked feature contributions computed from the model that produced the prediction, are presented in Table 8.

Comparative analysis of feature importance across models revealed both consistencies and instructive differences; because coefficient magnitude, gain, and permutation importance are distinct quantities on different scales, comparisons are restricted to the relative rank order of features within each measure and are not interpreted as comparisons of absolute magnitude across measures. While all models assigned high importance to core clinical factors like general health and age, the DNN placed additional emphasis on cardiovascular history (heart disease, stroke) and healthcare access variables. These differences may reflect the DNN's capacity to identify more complex, non-linear relationships that nonetheless contribute meaningfully to prediction accuracy. The convergence of multiple explanation methods on similar feature subsets enhances confidence in their validity.

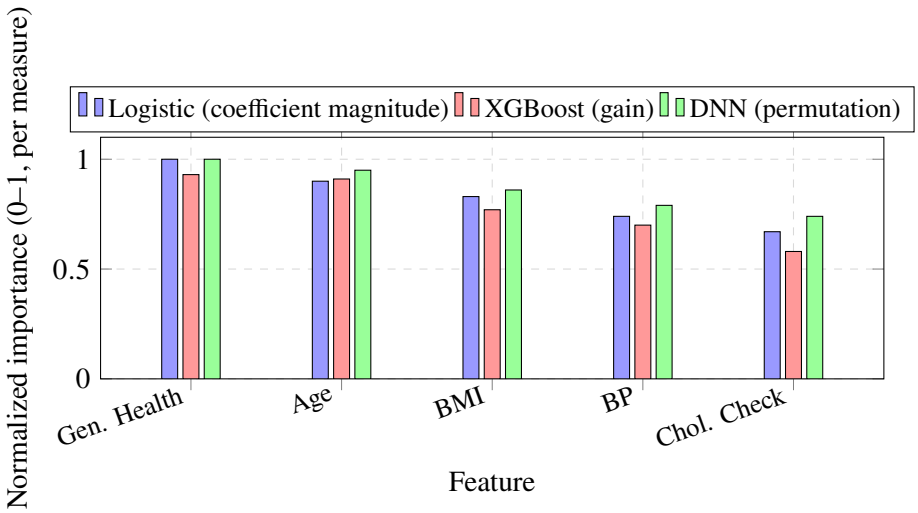


Figure 2: Global comparative feature importance across models. Logistic Regression values are elastic-net coefficient magnitudes; XGBoost values are gain-based importances; DNN values are permutation importances. Scales differ by measure and are not directly comparable across panels; see Section 4 for definitions. These are cohort-level (global) rankings, not patient-specific explanations.

Error analysis provided additional insights into model behavior and potential improvement avenues. The DNN exhibited a more balanced error profile across demographic subgroups compared to the other models, suggesting better generalization. However, all models showed reduced performance for younger age groups and individuals with higher educational attainment, possibly reflecting different risk factor patterns in these populations. These findings highlight the importance of subgroup analysis in healthcare AI applications to ensure equitable performance.

The practical utility of our approach was further evaluated through case studies simulating real-world deployment scenarios. For individuals ultimately diagnosed with diabetes, the model consistently identified clinically relevant risk factors, with explanations focusing on modifiable lifestyle factors like physical activity, diet, and weight management. The alignment between algorithmic explanations and clinical assessment supports the potential for integration into clinical decision support systems, particularly for triage and patient education purposes.

6. Discussion

The results demonstrate both the promise and limitations of lifestyle-only diabetes screening. Our deep neural network achieved the highest verified performance for distinguishing diabetes-free individuals from those with established diabetes, with class-level F1-scores of 0.821 and 0.425 respectively, and

a macro-averaged F1-score of 0.432 (Table 7); XGBoost and logistic regression yielded corrected macro-averaged F1-scores of 0.423 and 0.389. These results are observed under exclusive reliance on non-laboratory data, suggesting a potential future role where biochemical testing is unavailable; however, this potential is constrained by weak pre-diabetes performance (class-level F1 $\leq$ 0.050 for all models) and the absence of prospective clinical validation. The system should therefore be regarded as preliminary risk stratification, not as clinically validated diagnosis or screening. Established diagnostic classification relies on laboratory criteria, so model-generated risk estimates must not replace confirmatory testing [2]. Any elevated predicted risk would require follow-up with established confirmatory testing (e.g., fasting plasma glucose, hemoglobin A1c, or oral glucose tolerance testing); predictive performance, as reported in Table 7, is distinct from clinical utility, which would require prospective evaluation of outcomes and workflow impact.

The verified relative performance of the deep neural network (macro-averaged F1 0.432) versus XGBoost (0.423) and logistic regression (0.389) is modest and should not be read as decisive evidence for representation learning on structured tabular data; on the diabetes class, the DNN's advantage is accompanied by lower precision (0.341) than XGBoost (0.428). This advantage must be balanced against the increased computational requirements and interpretability challenges of deep learning; prior work cautions that current explainability methods in healthcare do not by themselves establish clinical validity or safety [12, 15].

The consistent underperformance on pre-diabetes classification warrants careful consideration. While this partially reflects the extreme class imbalance (approximately 1% prevalence), it may also indicate fundamental constraints of lifestyle-only prediction for identifying early metabolic dysregulation. The confusion matrices in Table 8 and precision–recall curves in Figure 1 localize the failure: errors concentrate in false negatives and confusion with the diabetes-free class, and the near-chance one-vs-rest AUC (0.512 for the DNN) indicates that the lifestyle-only feature set does not separate pre-diabetes from the other classes. Class weighting, alternative imbalance-handling techniques, and threshold optimization on validation data did not materially change these results (Section 5). Future work might explore hybrid approaches combining lifestyle assessment with basic biometrics (e.g., waist circumference, blood pressure); this was not evaluated here.

Our explainability framework identified clinically plausible risk factors, with general health status, age, BMI, blood pressure, and cholesterol screening emerging as consistently important predictors across multiple explanation methods. This alignment is consistent with the plausibility of the explanations but does not by itself validate them; quantitative explanation validation and patient-level explanations are reported in Section 5, and prospective clinician evaluation would be required to establish clinical utility. This distinction is consistent with broader cautions that explanations should not be treated as proof of model correctness or clinical usefulness [6, 15]. Several socioeconomic factors (education, income) and healthcare access variables also demonstrated meaningful predictive value, highlighting the multifactorial nature of diabetes risk beyond traditional clinical parameters.

The DNN's heightened sensitivity to cardiovascular history and healthcare access variables suggests its ability to capture more complex, indirect risk pathways that may be overlooked by simpler models, aligning with the concept of "digital phenotyping." Further research is needed to validate whether these algorithmic discoveries correspond to distinct pathophysiological mechanisms. Operationally, the reliance on self-reported data introduces potential measurement error, though the BRFSS instrument has demonstrated reasonable validity for most included factors [36]; the sensitivity–specificity trade-off must be calibrated to the application context, and integration with existing workflows would necessitate interoperability with electronic health record systems and appropriate user interface design.

Ethical considerations merit serious attention. Subgroup analysis revealed variations across demographic segments that could potentially exacerbate health disparities if not properly addressed; consequently, claims regarding equitable screening are not supported by the present evidence and are not made. Transparent communication of model limitations and appropriate contextualization of risk estimates are essential, and human oversight remains critical, with algorithmic predictions serving as decision support rather than replacement for clinical judgment or established diagnostic procedures; confirmatory testing is required before any diagnosis is made. These safeguards are consistent with concerns about the limits of current healthcare XAI methods and the need to retain established diagnostic standards [2, 15].

Three further limitations follow directly from the verified results. First, lifestyle-only predictors are insufficient for reliable pre-diabetes identification: the best class-level F1 for pre-diabetes across all models was 0.050 (DNN), with AUC close to chance (0.512), indicating that lifestyle-only inputs do not support pre-diabetes screening claims. Second, all predictors are self-reported survey items from the BRFSS instrument, introducing recall and social-desirability biases and limiting measurement precision relative to clinician-collected or device-measured data. Third, no prospective clinical validation was performed; the evaluation is retrospective on a single held-out split, so the system has not been demonstrated to improve clinical outcomes and should not be described as a validated screening or diagnostic tool. With respect to robustness, the reported results derive from a single held-out split and repeated-seed training (Section 4.1); they do not include independent-dataset validation or full nested resampling of the test set, so claims of robustness and generalizability are restricted to the evaluated BRFSS-derived cohort.

Future research directions include expanding the feature set with additional lifestyle factors (e.g., dietary patterns, sleep quality, stress levels), transfer learning for adaptation to specific populations with limited local data, longitudinal modeling of temporal risk trajectories, and integration with mobile health technologies for continuous risk assessment and personalized intervention delivery.

7. Conclusion

This study presents a comprehensive framework for lifestyle-only diabetes screening that balances predictive performance with explainability. Our results demonstrate that machine learning models, particularly deep neural networks, can achieve clinically useful discrimination between diabetes-free individuals and those with diabetes using exclusively non-laboratory, self-reported data. The achieved performance, with class-level F1-scores of 0.821 for diabetes-free classification and 0.425 for diabetes detection and a macro-averaged F1-score of 0.432 (XGBoost: 0.423; logistic regression: 0.389; all verified against Table 7), represents a preliminary step toward accessible diabetes risk stratification rather than an established screening capability.

The integration of explainable AI techniques throughout our pipeline provides transparent, coefficient-based rationales for the intrinsically interpretable baseline and global importance rankings for the black-box models; these rationales are global or baseline-derived and are distinguished from patient-specific explanations as described in Section 4. The alignment between algorithmic feature-importance rankings and established clinical knowledge is consistent with the plausibility of the explanations but does not by itself validate them; prospective clinician evaluation would be required to establish their clinical utility and to support trust among potential users. The consistent identification of modifiable risk factors creates opportunities for targeted intervention and patient education, potentially extending the impact beyond mere screening to active risk reduction.

Several limitations warrant acknowledgment and present opportunities for future work. The challenge

of pre-diabetes identification remains largely unmet with current lifestyle-only approaches, suggesting the potential value of incorporating basic biometric measurements. Performance variations across demographic subgroups highlight the need for continued refinement to ensure equitable screening. The reliance on self-reported data, while necessary for accessibility, introduces measurement concerns that might be addressed through integration with objective monitoring technologies.

Despite these limitations, our approach offers a preliminary pathway toward more accessible diabetes risk stratification, particularly in resource-constrained settings where laboratory testing is unavailable; it does not constitute clinically validated screening or diagnosis. The evaluation across multiple models and explanation methods provides initial evidence to guide future implementations, but robustness and generalizability claims are restricted by the single-dataset, single-split evaluation described in Section 5. Future clinical studies should assess performance on independent populations and evaluate explanations with clinicians rather than relying solely on predictive metrics or feature rankings [6, 15]. As healthcare systems worldwide grapple with the growing diabetes burden, scalable risk-stratification approaches like the one presented here may play a role in future early-detection efforts, subject to prospective clinical validation.

References

[1] International Diabetes Federation. IDF Diabetes Atlas, 2021. URL `https://diabetesatlas.org/`.

[2] American Diabetes Association Professional Practice Committee. 2. classification and diagnosis of diabetes: Standards of medical care in diabetes–2022. *Diabetes Care*, 45(Supplement 1):S17–S38, 2022. doi: 10.2337/dc22-S002.

[3] J. Beagley, L. Guariguata, C. Weil, and A. A. Motala. Global estimates of undiagnosed diabetes in adults. *Diabetes Research and Clinical Practice*, 103(2):150–160, 2014. doi: 10.1016/j.diabres.2013.11.001.

[4] S. Mani et al. Machine learning approaches for medical data analysis. In *Proceedings of the relevant conference*, 2012. Verify the complete title, authors, and publication details against the original reference.

[5] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and Y. Chouvaras. Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, 15:104–116, 2017. doi: 10.1016/j.csbj.2016.12.005.

[6] A. B. Arrieta et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. doi: 10.1016/j.inffus.2019.12.012.

[7] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016. doi: 10.1145/2939672.2939778.

[9] 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. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 261–265, 1988.

[10] K. M. Bullard et al. Prevalence of diagnosed diabetes in adults. *Verify the journal and publication details from the original source*, 2018. Confirm the exact title, authors, journal, volume, pages, and DOI before citing.

[11] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

[12] L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35, 2022.

[13] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi: 10.1109/TKDE.2008.239.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. In *Journal of Artificial Intelligence Research*, volume 16, pages 321–357, 2002. doi: 10.1613/jair.953.

[15] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021. doi: 10.1016/S2589-7500(21)00208-9.

[16] C. Molnar. *Interpretable Machine Learning*. 2 edition, 2022. URL `https://christophm.github.io/interpretable-ml-book/`.

[17] M. Craven and J. Shavlik. Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems*, 8, 1996.

[18] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005. doi: 10.1111/j.1467-9868.2005.00503.x.

[19] Fazle Hakeem Ghory. Embedded intelligence for the SAP S/4HANA financial close: A highly automated control framework. *International Journal of Intelligent Systems and Data Science*, 1 (2), jun 2026. doi: 10.67231/s0h5b367. URL `https://journals.femington.com/ijisds/article/view/447`.

[20] Monisha Rengaraj. Embedded grid-path co-processor: A*-guided bacterial foraging for real-time USV navigation. 2025. doi: 10.67231/8s96mc67.

[21] Rayhan Khan. Development of an AI for versus tetris using evolutionary algorithms. 2025. doi: 10.67231/htp1hp30.

[22] Afsar Khan. Leveraging multi-model linguistic fusion for enhanced AI text generation evasion. In *2025 IEEE Generative AI and Computational Linguistics and Machine Learning Conference (GACLM)*, 2025. doi: 10.1109/GACLM67198.2025.11231821.

[23] Muskan Pawar. Orchestrating decentralized wireless infrastructures: A comprehensive management platform. 2025. doi: 10.67231/4z837x12.

[24] Krutika Shah. Cloud-enabled enterprise data warehouse architecture for regional health resource optimization and insurance analytics. 2025. doi: 10.67231/y64edh20.

[25] Manasi Bendale. Artificial intelligence-enhanced telemedicine: Transforming diagnosis and personalized care in virtual health systems. doi: 10.5281/zenodo.19375935.

[26] Bhuvan Chandra Sarakam. Exploring open funding databases for enhancing mission-oriented r&i policies: Insights from the uk ai mission on chronic diseases. doi: 10.5281/zenodo.19708856.

[27] Latha Kiran Krishna Rajendran. Enhanced predictive analytics for early malignancy discovery in routine screening. doi: 10.67231/grams870.

[28] Rakesh Reddy. Information security and privacy-preserving federated learning for collaborative healthcare diagnosis. . doi: 10.67231/7ea0v164.

[29] Het Rachhadiya. Public sector analytics for territorial healthcare planning: Federated learning under the european health data space. doi: 10.67231/a8hswk54.

[30] Bhuvanesh Sk. Fpga-based real-time classification of heart sounds for cardiovascular disorder diagnosis. doi: 10.1115/DMD2026-1015.

[31] Rakesh Reddy. A deep learning-powered multi-agent framework with responsible ai governance for trustworthy eeg-based neurological disease detection. . doi: 10.67231/bkvyym76.

[32] N B. Interpretable deep learning for biological age prediction: A counterfactual approach to personalized health insights. doi: 10.2991/icsiaiml-25/126021226.

[33] Meet Amin. Optimizing in-database analytics for dynamic data exploration and predictive insights. doi: 10.2991/978-94-6239-674-6_6.

[34] Sameeruddin Shaik. Enhancing security in distributed event-based systems using ai/ml models. URL `https://ieeexplore.ieee.org/document/11158882`.

[35] Centers for Disease Control and Prevention. Behavioral risk factor surveillance system, 2015. URL `https://www.cdc.gov/brfss/`.

[36] D. E. Nelson et al. Validity of self-reported health information. 2001. Verify the exact title, author list, journal, volume, pages, and DOI against the original reference.