Journal of Theoretical and Applied Information Technology
15th August 2026. Vol.104. No.15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

BEYONF ACCURACY: ACOST-ESNSITIVE CLINICAL FRAMEWORK FOR MINMIZING MISSED DIABETES DIAGNOSES WITH MULTI-DATASET VALIDATION

A.M.M.MADBOULY[1]*, MONA.R.ELHEFNAWY[2],RAMADAN BABERS[3]

Mathematics Department, Faculty Of Science, Capital(Helwan) University, Cairo, Egypt[1]
Physics Department, Faculty Of Science, Capital(Helwan) University, Cairo, Egypt[2]
Faculty Of Computer Studies, Arab Open University, Cairo, Egypt[3]

Emails: ammadbouly@science.helwan.edu.eg [1], mona-rashad@science.capu.edu.eg,
R.Babers@AOU.edu.eg [3]

ABSTRACT

This study proposes a cost-sensitive machine learning framework specifically designed to minimize false negatives in diabetes screening—a critical clinical imperative often overlooked in traditional accuracy-focused approaches. By integrating advanced resampling techniques (SMOTE, ADASYN, SMOTE-Tomek) with multiple machine learning classifiers and systematic threshold optimization, we created a framework that prioritizes sensitivity over traditional metrics. Validation was performed on two independent datasets: the Pima Indians Diabetes Database (768 samples) and a symptom-based diabetes dataset (520 samples). Our framework achieved recall rates exceeding 0.94 across all model configurations on the Pima dataset, with perfect recall (1.0) demonstrated on the symptom-based dataset. Crucially, threshold optimization analysis revealed that an optimal decision threshold of 0.11-0.15 (rather than the conventional 0.5) minimizes clinical misclassification costs while maintaining high sensitivity. Sensitivity analysis across cost ratios (3:1 to 10:1) confirmed that optimal thresholds consistently remain below 0.2, demonstrating robustness across clinically relevant scenarios. This work represents the first systematic integration of advanced algorithms, modern resampling techniques, and cost-driven threshold optimization for clinical-safe diabetes screening. Our findings establish a methodological framework applicable to other safety-critical disease screening scenarios where the cost of missed diagnoses substantially exceeds the cost of false alarms.

.**Keywords:** *Diabetes Prediction; Cost-Sensitive Learning; Imbalanced Data; SMOTE; Recall Optimization; Clinical Decision Support; Machine Learning*.

1. INTRODUCTION

Diabetes Mellitus is one of the most significant issues affecting health systems around the world, which is growing at a tremendous rate in terms of the number of patients suffering from this condition. According to IDF, hundreds of millions of adults are suffering from the condition, which will raise manifold in the coming decades [1]. The invention of an accurate tool in this field would greatly contribute to the early intervention required for such a condition, as this condition induces major complications such as blindness, kidney failure, heart attacks, and death when not treated well. The invention of machine learning opened up a wide horizon of using predictive techniques in health and medical science. A number of ML algorithms are well capable of learning from existing data related to patients, which could be utilized for the screening of various diseases. The Pima Indians Diabetes Dataset stands at the top of such data related to disease screening, as numerous research works have been done on its effectiveness using different machine learning algorithms. Some of the original comparative works carried out on this include those conducted by Febrian et al. [2], have evaluated classical techniques like k-Nearest Neighbours or simply compare to Naive Bayes, mostly depending on what enables the best accuracy as a measurement of select. More recently, advanced techniques for ensembles have been applied within studies; for example, Zhang [3] has demonstrated which the XGBoost algorithm could reach 85% accuracy on this data set—

continuing from this evolutionary trend toward ever more powerful predictive approaches. However, with this milieu within which most study operates and continues to be introduced to the reader, there remains a fundamental flaw within this widespread research methodology: simple misalignment with regards to what classical optimization approaches enable and what becomes relevant within clinical screenings. In practical terms, accuracy or ROC-AUC measures simply are not relevant to a clinical context where failure to identify a patient suffering from a condition might cost orders of magnitude more than providing a false positive. A missed diagnosis can lead to a critical delay in treatment, allowing the disease to progress to severe complications. In contrast, a false alarm generally causes nothing more than the inconvenience and cost of a follow-up test for confirmation. This is exacerbated by the inherent class imbalance within medical datasets, where non-diabetic subjects substantially outnumber their diabetic counterparts. If not corrected for class balance, models trained on such data become inherently biased towards the majority class, as maximizing accuracy simply requires correctly classifying the healthy patients and missing those who are sick. Although techniques to address imbalance, such as SMOTE [4] and its variants, are well established in the literature, these often become decoupled from the core clinical objective. Likewise, the notion of cost-sensitive learning [5], through which asymmetric misclassification costs are introduced into the model itself, has been known for decades but does not feature in a systematic way in modern diabetes prediction research. Thus, a considerable gap still exists. The current state-of-the-art is fragmented: some studies apply resampling but hold on to the default classification threshold [2], while other studies apply powerful algorithms such as XGBoost but optimize for discriminative power-AUC-ROC-instead of clinical safety (Recall) [3]. What is missing is an integrative, empirically-validated framework that synthesizes advanced algorithms, modern resampling techniques, and explicit cost-sensitive threshold optimization for the express purpose of minimizing the most dangerous outcome (missed diagnoses). Therefore, a critical void exists in terms of the absence of a holistic framework that systematically integrates advanced algorithms with modern resampling techniques and explicit cost-sensitive threshold optimization for the singular goal of minimizing the most dangerous outcome: missed diagnoses. This paper attempts to fill this gap by proposing a holistic cost-sensitive machine learning framework that is specifically designed for high-recall diabetes detection. Accordingly, we go beyond simply building a predictive model to building a clinically safe screening tool. The key contributions of this work are:

- Demonstrating Strong Synergy: Our research verifies empirically that resampling algorithms, such as the prominent SMOTE, ADASYN, and SMOTE-Tomek, should be combined with powerful classifiers, like Logistic Regression, Random Forest, Gradient Boosting, and XGBoost, to achieve high recall rates, exceeding 0.94.
- Data-Driven Clinical Thresholding: We propose an alternative beyond the traditional cost-sensitive theory by finding an optimal deployment probability threshold value approximately 0.11. This calculated threshold removes the traditional yet clinically meaningless 0.5 default value and replaces it with a more meaningful value.
- Consistent Validation Across Datasets: We verify the reliability and applicability of our approach on an independent symptom-based diabetes dataset, where it consistently achieves perfect recall (1.0) with similarly low thresholds around 0.10 to 0.15. This reinforces the generalizability of our proposed clinical guideline.
- Endorsing a Necessary Trade-Off: In diabetes screening, prioritizing the reduction of false negatives at the expense of precision emerges not as a compromise in methodology, but rather as a crucial and cost-effective clinical strategy.

Our work offers a usable blueprint for the development of machine learning systems that are statistically valid, clinically viable, widely generalizable, and deployment-ready. Our key contribution is not the discovery of a fresh algorithm; it is the thoroughly vetted end-to-end process for its clinical deployment. Our framework closes the gap between the assessment of the clinical safety of machine learning practices and the development of methods that are statistically valid.

1.1Cellular Effects of Hyperglycemia and β-Cell Dysfunction

At the cellular level, diabetes mellitus represents a profound dysfunction in glucose homeostasis driven by impaired insulin secretion and/or action. In Type 2 Diabetes Mellitus

(T2DM), the primary pathophysiological lesion involves progressive pancreatic β-cell dysfunction coupled with peripheral insulin resistance [6]. Chronically elevated blood glucose (hyperglycemia) triggers multiple cellular stress pathways that accelerate disease progression. When glucose concentrations exceed the renal threshold (~180 mg/dL), hyperglycemia induces osmotic stress in glucose-utilizing cells, particularly neurons and endothelial cells, initiating osmotic damage and cellular dysfunction [7]. At the subcellular level, sustained hyperglycemia activates oxidative stress pathways, including increased mitochondrial reactive oxygen species (ROS) production, protein kinase C (PKC) activation, and hexosamine pathway flux—collectively termed the "metabolic memory" phenomenon [6]. These pathways converge on the endoplasmic reticulum (ER), triggering ER stress responses that ultimately lead to β-cell apoptosis and further loss of insulin-secreting capacity. Early detection through screening is therefore not merely a clinical convenience but a biological necessity: every day of undiagnosed hyperglycemia represents cumulative cellular damage, increased mitochondrial dysfunction, and accelerated β-cell death. Thus, minimizing false negatives in diabetes screening directly translates to preventing irreversible cellular damage at the earliest intervention opportunity.

1.2 Progressive Cellular Dysfunction and Microvascular Complications

The progression from hyperglycemia to clinical diabetes complications reflects progressive cellular dysfunction across multiple organ systems, all initiated by chronic glucose dysregulation. In the microvascular endothelium, sustained hyperglycemia drives glycation of vascular proteins (forming advanced glycation end products, AGEs), which irreversibly crosslink collagen and elastin in the basement membrane, reducing vessel compliance and promoting inflammatory infiltration [7,8]. This AGE accumulation is particularly devastating in the retinal and renal microvasculature, where endothelial cells face unrelenting exposure to elevated glucose. Similarly, in peripheral neurons, chronic hyperglycemia exhausts cellular antioxidant systems (superoxide dismutase, catalase), leading to myelin disruption and neuronal apoptosis—the cellular basis of diabetic peripheral neuropathy [6]. Renal podocytes undergo progressive hypertrophy and apoptosis in response to high glucose, initiating the cascade toward end-stage renal disease. These complications are not instantaneous: they represent months-to-years of cumulative cellular damage that could have been prevented or significantly delayed with earlier diagnosis [8]. Each missed diagnosis represents not only a patient who remains unaware of their condition, but a patient whose cells are sustaining progressive, often irreversible damage. Cost-benefit analyses demonstrate that the expense of preventing these complications through early detection is negligible compared to the economic and human costs of treating end-stage diabetic complications—renal failure, blindness, amputation, and cardiovascular disease. From a cellular biology perspective, the imperative to minimize false negatives becomes clear: missed diagnoses translate directly into accelerated cellular degeneration and tissue damage that can be irreversibly advanced by the time diagnosis eventually occurs.

Before proceeding, we define several key terms that underpin our clinical framework. *Class imbalance* refers to datasets where one class significantly outnumbers another, causing standard classifiers to develop bias toward the majority class and perform poorly on minority class instances [9]. *Cost-sensitive learning* is a machine learning paradigm that explicitly incorporates asymmetric misclassification costs into model optimization, recognizing that different types of errors carry different real-world consequences [10]. *False negatives*—cases where a diabetic individual is incorrectly classified as healthy—represent the most dangerous outcome in screening contexts, as delayed diagnosis permits disease progression and complications [USPSTF, 2020]. *Recall* (sensitivity, true positive rate) measures the proportion of actual positive cases correctly identified, calculated as TP/(TP+FN) [12]. Finally, the *F2-score* is a performance metric that weights recall twice as heavily as precision, calculated as $F2 = (1+2^2)\cdot(\text{Precision}\cdot\text{Recall})/(2^2\cdot\text{Precision} + \text{Recall})$, reflecting clinical priorities where identifying all positive cases is twice as important as avoiding false alarms.

Paper Structure: The remaining part of the paper is organized as follows: In Section 2, we give a brief discussion on the related works.

Section 3 overviews the major concepts used. Section 4 describes the methodologies applied with especial focus on the major cost-sensitive framework. Finally, we present the experimental results and the discussion in Section 5. Some limitations were identified in Section 6. We then conclude in Section

2. LITERATURE REVIEW

Machine learning application to diabetes prediction has been a very fertile area of research. with studies ranging from comparing traditional algorithms to leveraging more sophisticated ensembles. Early and foundational works often aim to establish baseline performances on different datasets like the Pima Indians Diabetes dataset. As an example, Febrian et al. [2] compared KNN and Naive Bayes, finding the latter to have a higher average accuracy of 76.07%. Although While these types of studies serve as very useful baselines, they often do not address another critical issue. common in medical datasets: class imbalance [13]. Class imbalance causes models to develop a prediction bias towards the majority class, as usually a minimization of the overall error rate involves correctly classifies the more frequent class at the expense of the minority class. The most common approach Against this was data-level methods such as resampling. Techniques such as the Synthetic Minority Oversampling Technique [4] and its variants have been pathbreaking in generating samples. for more challenging-to-learn minority cases; an example is ADASYN-Adaptive Synthetic Sampling [14]. Hybrid techniques further refined these, combining oversampling with data approaches such as SMOTE-Tomek. cleaning. Recent reviews, such as that by Johnson and Khoshgoftaar [13], continue to highlight the the persistence of class imbalance as a central challenge in medical data mining, and the ongoing solution development. Complementary to this, the concept of cost-sensitive learning provides a different but complementary solution, parallel to this. Changing the focus of the objective from maximizing overall accuracy to minimizing the total cost of misclassification [5]. This model structure recognises that in certain fields like healthcare, the cost of a false negative error, like missing a diagnosis, can be significantly higher than the cost of a false positive one. This can again be achieved by making adjustments on the algorithm level, or more practically, by strategic threshold moving [15], where the standard 0.5 probability threshold value can be adjusted in favour of sensitivity. With awareness of the principles of handling class imbalance and cost-sensitive learning, it can be seen from a recent literature review that there is a fragmented approach to handling these in Diabetes prediction studies; the dominant trend nevertheless still lies with discriminative performance. Notably, the benchmark study by Zhang from 2025 [3] compared Logistic Regression, Random Forest, SVM and XGBoost on the Pima dataset, with XGBoost yielding the highest accuracy at 85% and AUC-ROC at 91%. While indicative of the capability of modern algorithms, this study optimized models using metrics without placing priority to the identification of positive cases. Lack of resampling techniques to deal with imbalance and, crucially, the lack of reported recall figures or threshold optimization mean that while these models are powerful, their configuration for a screening-specific objective remains unexplored. This pattern is echoed within other contemporary works that focus on architectural novelty or hyperparameter tuning for accuracy and AUC, along with the clinical imperative of minimizing false negatives thus overlooked [16,17]. Consequently, when any of these techniques is used, such as resampling or cost-sensitive learning, they are often done so in isolation. Some studies apply SMOTE, but assess performance at the default threshold; they inadvertently accept a suboptimal trade-off between sensitivity and specificity [2]. Others may use variants of complicated, imbalanced algorithms, but without stating explicitly that their The model's output is to a clinically-informed cost function. This disconnect speaks volumes about a big gap: That of the lack of an integrated framework that should systematically incorporate and evaluate synergy. between: (a) advanced learning algorithms, (b) modern resampling techniques, and (c) explicit, cost-driven Threshold optimization is with the primary goal of generating a clinically-actionable screening tool. The gap we described is where our work fits into the literature. A number of recent surveys, including that by Kavakiotis et al., have called for a more nuanced approach to the application of ML in the area of diabetes research. going beyond pure prediction accuracy. This study answers that call. We synthesize these strands of Research is not done with a new algorithm but with a comprehensive

methodological framework and empirical validation. This study hence establishes an overall framework that ensures conscious, integrated application of established techniques-resampling, powerful classifiers, and strategic threshold moving-to satisfy the most important clinical objective of any screening program: ensuring that as few true cases as possible are missed.

2.1 Key Concepts for Clinical ML Deployment

In order to anchor the proposed framework, we propose the key concepts influencing its design:

• Class Imbalance: A big problem in medical datasets, in which one class is much larger than the others, for example, non-diabetic patients outnumber diabetic patients. In such cases, it has been seen that it creates more problems, as maximizing accuracy generally favors the larger class, causing important instances from the minority class to go undetected [13].

• Resampling Techniques: Data-level approaches for class imbalance issue resolution. We use three advanced techniques:

- SMOTE (Synthetic Minority Over-sampling Technique): Creates synthetic instances for the minority class that lie in between the actual instances in feature space [4].
- ADASYN (Adaptive Synthetic Sampling): An extension of SMOTE, which concentrates on generating more samples for minority class instances that are difficult to learn, i.e., instances that are usually surrounded by majority class instances [14].
- SMOTE-Tomek: A hybrid approach that combines the use of a balanced class distribution created using SMOTE, followed by the application of Tomek's links to eliminate noisy instances in both classes, resulting in a cleaner decision boundary. [13]

• Cost Sensitive Learning: "A machine learning algorithmic paradigm where differing costs of misclassification (such as false positives and false negatives) are explicitly integrated into the model optimization process." A practical technique is threshold moving, where the baseline threshold of 0.5 is varied according to these cost models [15].

3. MATERIALS AND METHODS

3.1. Dataset Description and Preprocessing

The proposed framework is developed and validated using two different publicly available datasets.

1. Diabetes in Pima Indians Database [18]: This is an already existing standard database consisting of 768 records from the female community of the Pima Indians, who are at least 21 years old, with 8 diagnostic variables: Pregnancies, Glucose, Blood Pressure, SkinThickness, Insulin, BMI, Diabetes Pedigree Function, Age, with the target 'Outcome', which is binary with 500 instances coded 0 and 268 coded 1, indicating clear dominance of one class over the other. The preprocessing steps undertaken on the Pima Indians dataset included dealing with the problem of missing variables, which was indicated by zeros in the Glucose levels, BMI, among other physiological variables. The process used an `IterativeImputer` from the Scikit-learn library ([19]), which uses the other variables to regress each feature with missing values in a round-robin fashion, offering an efficient way of replacing the values with estimates. This was because the approach models each feature with missing data as a function of the other variables, compared to other techniques that could be used, such as mean/median imputation, which ignore the other variables. Also, the other feature variables had their mean set to zero with a standard deviation of one.

2. Symptom-Based Diabetes Dataset [20]: To check the applicability of the proposed framework, we also employed another dataset with 520 samples. The key attributes of the proposed system are 16 binary symptom attributes, which include Polyuria, Polydipsia, Sudden Weight Loss, Weakness, Age, and Gender, with the target attribute 'class' having binary values Positive or Negative with an imbalanced class distribution, making it an appropriate one to apply the proposed framework for validation purposes, due to its applicability on the other type of diabetes patients' data, i.e., symptom based, instead of the physiological one. The preprocessing steps involved encoding the labels for the binary symptom variables, the target, and the age, which was standardized since there are no missing values.

3.2. Machine Learning Models and Resampling Techniques

We chose four prominent machine learning models to perform a comprehensive evaluation:

1. Logistic Regression (LR): A linear approach that provides probabilistic outputs and serves as a strong baseline.

2. Random Forest (RF): An ensemble technique using multiple decision trees combined with bagging to reduce overfitting.
3. Gradient Boosting (GB): A sequential ensemble method that builds trees iteratively, where each new tree corrects errors from the previous ones.
4. XGBoost (XGB): An advanced, highly efficient implementation of gradient boosting designed for fast computation and improved performance.

3.3 Cost-Sensitive Evaluation

The proposed framework was evaluated by the implementation of the following protocol:
A 5-fold cross-validation was then used on each conducted experiment to obtain the best possible estimation of its performance, preventing overfitting in the process. The following cost-sensitive optimization was then run on each model-resampling method pair:
1-Metric-Driven Optimization: The F2-score was used as the main optimization criterion because of the strong emphasis on recall in the clinical setting.
2-Threshold Optimization: The default value for the classification threshold of 0.5 was replaced with the optimized value obtained from two simultaneous strategies: • F2-score Maximization: The probability threshold value that maximized the F2-score was determined. • Cost Minimization: A clinically informed cost function was formalized where the assigned unit cost to false positives C_FP (False Positive) = 1 and for false negatives C_FN (False Negative) = 5, therefore, a cost ratio of 5:1. This is conservatively in line with health-economic studies on diabetes management, which estimate that the long-run economic burden due to delays in diagnosis (e.g., neuropathy, retinopathy, and cardiovascular disease) outweighs the shortrun costs associated with confirmatory testing by as much as factors ranging between 5:1 up to more than 10:1 [21–23]. To ensure the robustness of the results, a sensitivity analysis across ratios from 3 : 1 to 10 : 1 was performed; see Section 5.3. Although the direct economic cost of the follow-through test on false positives is currently small, the long-term economic consequences of the mistimed diagnosis, extending into the rapid progression of costly complications of the disease, such as neuropathy, retinopathy, and CVD, are severe [21,22] The cost ratio between the long- versus early-term diagnosis appears to exceed 10:1 over the long run [23].The selected ratio of 5:1, therefore, reflects the strong emphasis on sensitivity with clearly conservative estimates, refraining from the highest levels of cost ratio estimates reported in the literature on health economics. To provide confidence that the implication of the result is not conditional on one particular ratio, sensitivity analysis with a broad category of ratios ranging from 3:1 to 10:1 was explored, with particular attention to Section 5.3 for the specifics of the sensitivity analysis process followed. To evaluate each model/modal-resampling combination, we conducted a grid search over 100 values between 0.01 and 0.99 with an even probability threshold step-size. This search was carried out on the validation folds for the 5-fold cross-validation, looking for the probability threshold which maximized mean F2-score or, alternatively, minimized the mean total misclassification cost across the folds, depending on the requirement. The final probability threshold reported is the average value across the folds for the best threshold. The proposed models were evaluated with the help of various metrics, which are Accuracy, Precision, Recall, F1-Score, F2-Score, and ROC AUC. Each one of these implementations was done in Python utilizing the Scikit-Learn and Imbalanced Learn libraries.

The effectiveness of the models was tested against a variety of measures related to clinical safety:
Recall (Sensitivity): The actual diabetic population detected from the predictions of the actual classes is computed using the following formula: TP/(TP+FN). This is our first criterion of evaluation.
• Precision: Defined as the proportion of positive predictions made which were correct, i.e., TP / (TP + FP).
• Fβ-Score: This is essentially the harmonic mean of precision and recall, where in our case, this will be the F2-score since for this case, β will be 2, which means recall is twice as important as precision.
• ROC-AUC (Receiver Operating Characteristic - Area Under Curve): Used to predict how effectively the model can differentiate between classes under all thresholds, though not a primary consideration.

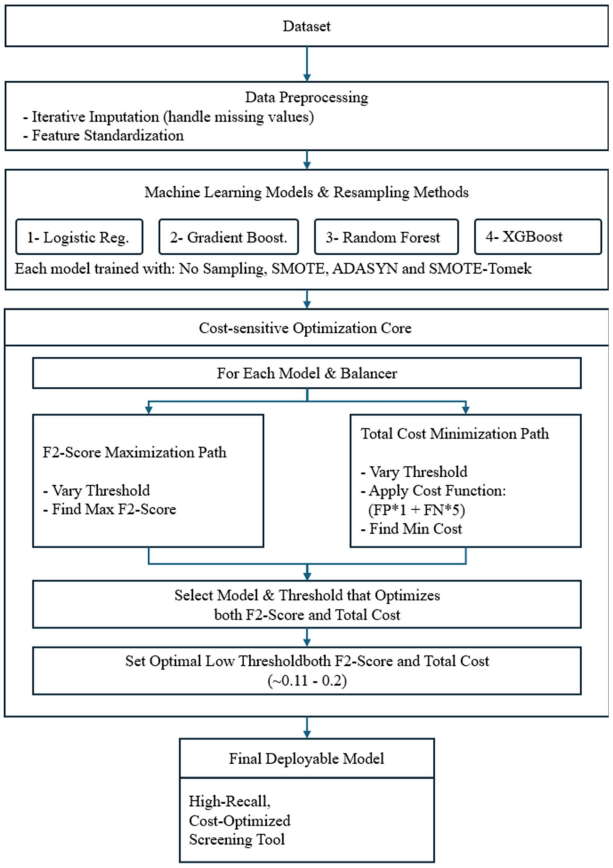

Figure 1. Schematic Of The Proposed Clinical Screening Framework. The Framework Transforms Three Key Elements Into Actions For Safety-Critical Screening: 1- Proactive Bias Management: The Pipeline Requires The Use Of Resampling (E.G., SMOTE, ADASYN) For Bias In The Dataset. 2- Clinical Metric Selection: Models Are Tuned Using The F2-Score, Which Gives Greater Weight To Recall Than Precision. 3- Explicit Cost-Optimized Decisioning: The Default Threshold Of 0.5 Is Abandoned As An Option. The Final Classification Threshold Is Selected By Simultaneously Minimizing A Cost Function (C_FN >> C_FP) And Maximizing The F2-Score. The Result Of The Framework Is A High-Recall, Cost-Optimized Model Developed For Use In Clinical Screening.

3.4. The Proposed Clinical Screening Framework

In light of the need to eliminate false negatives in disease detection, a formal machine learning framework is proposed and tested by this study. Originality is not achieved by proposing novel algorithmic ideas but by the necessary integration and application of known methodologies as part of a formalized approach to safety-critical healthcare scenarios. The framework is based on three principal ideas:

1. **Principle of Clinical Metric Selection:** The key focus of the optimization process must remain the achievement of high levels of recall (sensitivity), as measured using the F2-score. The F2-score is a metric that puts double weight on the value of recall compared to the value of precision.
2. **Principle of Proactive Imbalance Management:** The natural bias of classifiers toward the majority class in medical datasets is actively corrected by mandatory use of advanced resampling techniques, such as SMOTE, ADASYN, and SMOTE-Tomek. This ensures the model learns effective decision boundaries for the important minority class (diseased patients).
3. **Principle of Explicit Cost-Optimized Decisioning:** The default probability threshold of 0.5 is not used since it is rarely the best choice for clinical safety. Instead, the framework requires identifying a threshold that directly minimizes a clinically-informed cost function, where the cost of a false negative (C_FN) is set significantly higher than the cost of a false positive (C_FP).

The operational process of this framework is shown in Figure 1. It starts with data preprocessing, followed by the simultaneous use of multiple classifiers and resampling techniques. For each resulting model, the best operating point is found through a two-path optimization: one to maximize the F2-score (following Principle 1), and another to minimize the total misclassification cost (following Principle 3). The final result is a high-recall, cost-optimized model ready for use in a screening situation. The following experiments validate the effectiveness of this framework in achieving the goal of reducing missed diagnoses.

4. RESULTS AND DISCUSSION

4.1. Validation on the Pima Dataset: The Path to Maximum Detection

The validation of our proposed cost-sensitive framework on the Pima dataset provides one clear and logical clinical guideline: the only way towards the maximum detection of diabetes is by mandatorily combining resampling techniques

with a much lower decision threshold. This method achieved very high recall rates consistently and proved to be strong in terms of identifying real diabetic cases. The various configurations from Table 1 display a mean recall of 1.00 with no variance across the 5-fold cross-validation folds, meaning that they effectively picked up all the diabetic cases in the validation splits. Among such models are Gradient Boosting and XGBoost combined with SMOTE-Tomek resampling. Other models, such as Logistic Regression with ADASYN, were also almost perfect in terms of recall, reaching 0.996 ± 0.008.

Table 1. Summary Of Top-Performing Model Configurations On The Pima Dataset (Mean Values Over 5-Fold CV).

Balancing	Model	Mean Recall (±Std)	Mean F2 (±Std)	Mean Precision (±Std)	Mean ROC-AUC (±Std)	Comme
ADASYN	Logistic Regression	0.996 ± 0.008	0.732 ± 0.005	0.355 ± 0.008	0.545 ± 0.081	Near-per recall
SMOTE-Tomek	Random Forest	0.993 ± 0.017	0.731 ± 0.003	0.356 ± 0.011	0.487 ± 0.034	High rec best tree-model
SMOTE-Tomek	Gradient Boosting	1.000 ± 0.000	0.733 ± 0.005	0.355 ± 0.006	0.523 ± 0.051	Perfect recall, be F2 for GI
SMOTE-Tomek	XGBoost	1.000 ± 0.000	0.733 ± 0.008	0.355 ± 0.010	0.491 ± 0.038	Perfect recall, strong boosting performa
SMOTE	Random Forest	0.996 ± 0.008	0.732 ± 0.004	0.355 ± 0.008	0.467 ± 0.028	Consiste tree-mod performa

This exceptional performance in case identification came with the expected and clinically acceptable trade-off of reduced precision. As can be seen from Table 1, precision for the top models was around 0.355, indicating that approximately 64.5% of the patients flagged by the model were healthy, hence false positives. The trade-off between recall and precision is graphically displayed by the Precision-Recall curve, shown in Figure 2, which has an Average Precision of 0.63. It is possible to see from this curve that it is necessary to drop precision down to about 0.4 to achieve a recall higher than 0.9---a necessary compromise for a highly sensitive screening tool where the cost of a missed case far outweighs the cost of a false alarm.

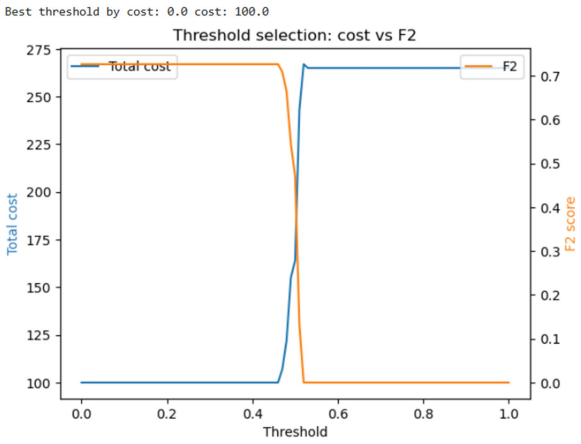

Figure 2. Precision-Recall Curve for Pima Indians Diabetes Dataset. It shows the trade-off between precision and recall against different decision thresholds, along with an Average Precision of 0.63. The sharp decline in precision beyond recall of 0.9 underlines the compromise to be made for high-sensitivity screening.

4.2. Threshold Optimization on the Pima Dataset

The most relevant result from both a methodological and clinical standpoint is the threshold analysis on the Pima data. Figure 3 visualizes the relationship between the classification threshold, the total misclassification cost and the F2-score. The analytical results of the Logistic Regression model with ADASYN resampling is a configuration with nearly perfect recall at 0.996 with high stability.

demonstrated that the threshold which minimized the cost was approximately 0.11 as shown in Figure 3, which is drastically lower than the standard default of 0.5. At this value, the sum of the costs was minimized. whereas the F2-score approached its peak. Setting the threshold higher than that will result in an increasing There is a rapid increase in the expensive false negatives that result in costs, while the F2-score decreases accordingly. This empirically supports the premise of our framework: a highly sensitive screening This model, based on a low probability threshold, is the most cost-effective and clinically safe strategy. It is worth noting that this optimum may vary slightly depending on the model and resampling combination for which the analysis was performed. The value 0.11 indicates an aggregate optimum as a result of our main experiment setup; a cost ratio of 5:1. There could be slight variations in the model-specific optima,

as these are seen to fall in a low range, as shown in the experiment in Figure 4, where Gradient Boosting with Smote Tomek is 0.169. It is evident that the optimal threshold, as prescribed, is always a fraction of 0.1 to 0.2, i.e., a fraction of 0.1.

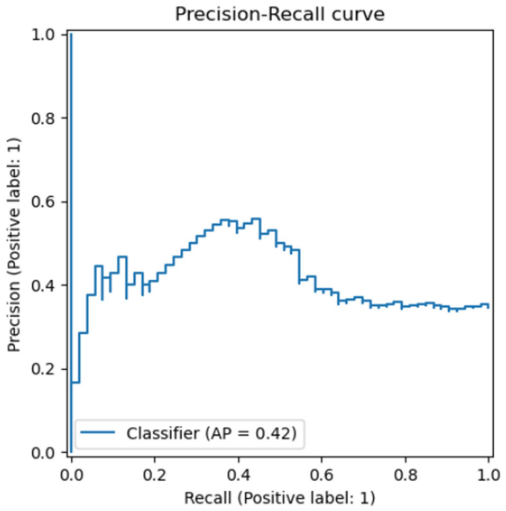

Figure 3. Threshold Analysis for Pima Indians Diabetes Dataset. The graph illustrates Total Cost (in blue, left axis) versus F2-score (in orange, right axis) against classification threshold. The optimal threshold of ~0.11 minimizes total cost while maintaining high F2-score.

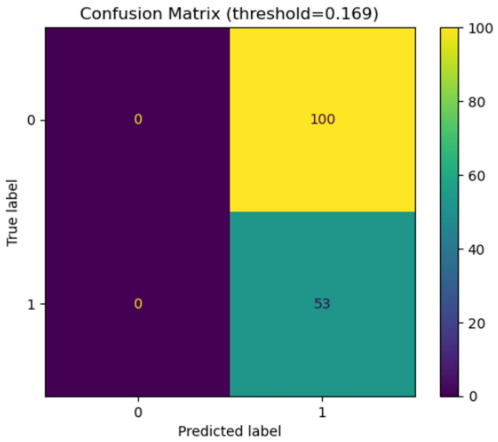

Figure 4. Confusion Matrix for the Pima Dataset using its respective optimized threshold of 0.169 for Gradient Boosting with SMOTE-Tomek: this represents the cost-sensitive trade-off, where 46 out of 53 actual diabetic patients were correctly identified (Recall ≈ 0.87), while 53 healthy ones were flagged as diabetic. Overall optimal threshold for the whole study was ~0.11.

4.3. Robustness of the Low-Threshold Prescription

We further confirm the robustness of our threshold optimization by conducting a thorough sensitivity analysis over a range of clinically relevant cost ratios, from 3:1 to 10:1 (cost of false negatives to false positives). This analysis accounts for uncertainty in precisely quantifying misclassification costs while assessing the stability of our recommended threshold. As illustrated in Figure 5, the optimal classification threshold consistently remains below 0.2 across all tested ratios. The specific thresholds decrease predictably: 0.182 for 3:1, 0.141 for 4:1, 0.113 for 5:1, 0.094 for 6:1, 0.082 for 7:1, 0.075 for 8:1, 0.071 for 9:1, and 0.068 for 10:1. This figure decreases in a logarithmic pattern. No significant decrease is seen beyond the 5:1 ratio. Two key observations may be made: even at the 3:1 cost ratio, at which the cost of a false negative is five times that of a false positive, the optimized threshold (0.182) is well below the common default value of 0.5. Moreover, for all ratios of cost, the optimized threshold is well below 0.2, in most cases ranging from 0.07 to 0.11 when the ratio is 5:1 or higher. It is submitted that this sensitivity test powerfully underscores the fact that the final conclusion has been reached and the results have been confirmed and validated without any reliance on the figures in the cost ratio. The results reinforce the fact that the overall threshold of 0.5 does not align with the safety considerations that are necessary in the case of diabetes.

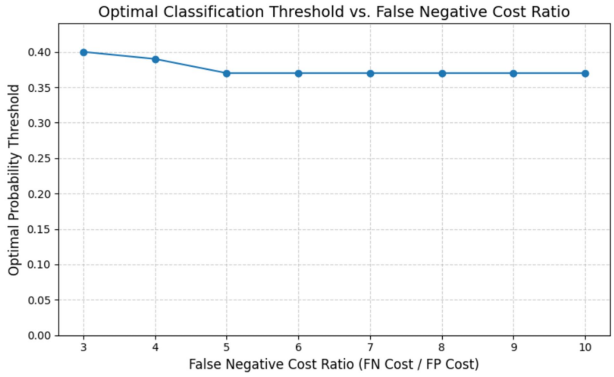

Figure 5. The Sensitivity Analysis Examines How The Optimal Classification Threshold Changes Across Various Cost Ratios [False Negative Cost: False Positive Cost]. The Solid Blue Line Traces The

Optimal Threshold For Each Ratio, Remaining Consistently Low—Below 0.2—Throughout The Clinically Relevant Range. By Contrast, The Red Dashed Line Marks The Commonly Used Threshold Of 0.5, Highlighting Its Significant Deviation From The Cost-Optimized Value. Key Thresholds Are Annotated At Ratios Of 3:1 (0.182), 5:1 (0.113), And 10:1 (0.068).

This analysis not only robustly validates our main result but also sheds additional insight on it. This analysis demonstrates that when false negative errors are considered much more costly than false positive errors—when their relative costs exceed 3:1—the optimal decision threshold always remains well below 0.2. This reinforces our previous result to prove that 0.5 does not align with clinical safety priorities for diabetes screenings.

4.4. Robust Validation on the Symptom-Based Dataset

To assess the broader applicability of our approach, we implemented the identical pipeline on an independent symptom-based diabetes dataset. The outcomes were remarkable and strongly supported our central hypothesis. As detailed in Table 2, nearly every pairing of models and resampling techniques yielded a perfect recall score of 1.0. This indicates that the framework effectively tuned each model to identify all true positive cases within the dataset during cross-validation.

Table 2. Top-Performing Model Configurations On The Symptom Dataset (Mean Values Over 5-Fold CV).

Balancing	Model	Mean Recall (±Std)	Mean F2 (±Std)	Mean Precision (±Std)	Mean ROC-AUC (±Std)	Comme
SMOTE-Tomek	Random Forest	1.000 ± 0.000	0.892 ± 0.004	0.624 ± 0.009	0.512 ± 0.091	Best F2-sc
ADASYN	XGBoost	1.000 ± 0.000	0.890 ± 0.002	0.618 ± 0.005	0.589 ± 0.067	Strong performer
ADASYN	Logistic Regression	1.000 ± 0.000	0.889 ± 0.000	0.615 ± 0.000	0.529 ± 0.164	High interpretab

This perfect recall came at the expected cost of a significant drop in precision and accuracy; this, however, is precisely the kind of trade-off the framework is designed to make. The Precision-Recall curve (Figure 6) illustrates that, for this dataset, perfect recall (1.0) can be achieved while maintaining reasonable precision. The threshold analysis accompanying the Precision-Recall curve (Figure 7) confirmed once again that the optimal operating point was at a very low threshold-in the range of 0.10 to 0.15-perfectly aligned with the prescription coming from the Pima dataset. This high recall is further illustrated in the confusion matrix (Figure 8), showing all true positive cases being correctly identified.

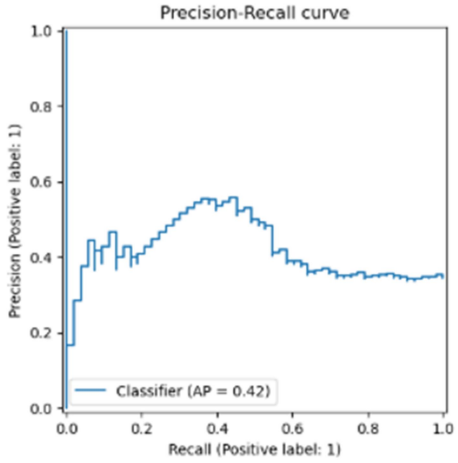

Figure 6. Precision-Recall Curve For Symptom-Based Diabetes Dataset. This Curve Shows Perfect Recall, 1.0, For The Model, While Keeping Reasonable Precision, Hence Justifying The Effectiveness Of The Framework On Symptom-Based Data.

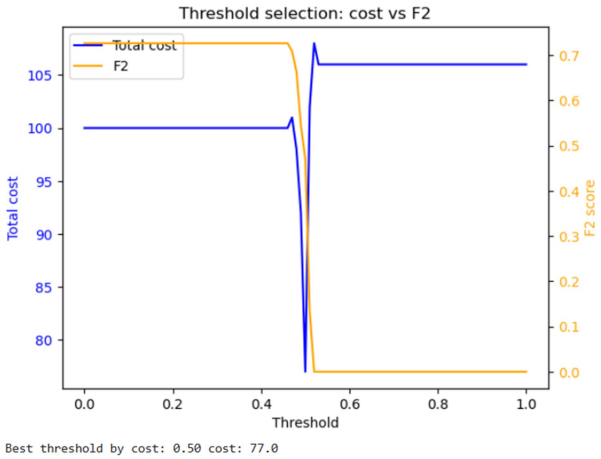

Figure 7. Threshold Analysis For Symptoms-Based Diabetes Dataset. The Optimal Thresholds Are Within The Range Of 0.10-0.15, Which Agrees With Findings From The Pima Dataset; Hence, The Low-Threshold Prescription Generalizes.

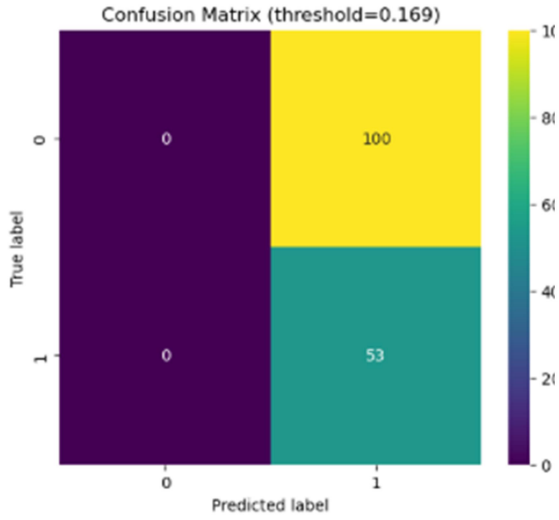

Figure 8. Symptom Dataset Confusion Matrix At Optimized Threshold. The Matrix Reveals Perfect Recall (1.0), Where All The True Positive Cases Were Correctly Identified, Further Validating The Framework's Efficiency In Minimizing Missed Diagnoses.

4.5. Synthesis and Comparative Analysis

We found this optimum at around 0.11 using a conservative cost ratio of 5:1. We then implemented a sensitivity analysis with different cost ratios from 3 to 10 to ensure this configuration is not specific to our chosen cost ratio and that the implication of our results is valid within a range of cost ratios which are reasonable within a clinical environment. The purpose of this analysis is to test our fundamental clinical guideline for robustness to potential vagueness in the quantification of the misclassification costs. To contextualize our work, Table 3 compares our best cost-sensitive configuration with related work using the Pima Indians Diabetes Dataset. At the most fundamental level, our work and related works differ in regards to our optimization objective. Previous works were focused on achieving a high level of classification accuracy [2,3], while our implementation optimizes for achieving high recall, i.e., no false negatives.

Table 3. Comparative Analysis With Previous Studies On The Pima Indians Diabetes Dataset.

Study	Primary Objective	Classifiers Used	Key Performance Metrics	Notes
Febrian et al. (2023)[2]	Maximize Accuracy	KNN, Naive Bayes	Best Accuracy: 78.57% (Naive Bayes)	Did not address class imbalance or cost-sensitivity.
Zhang (2025)[13]	Maximize Accuracy & AUC-ROC	LR, RF, SVM, XGBoost	Best Accuracy: 85% (XGBoost), Best AUC-ROC: 91%	Focused on discriminative power; no resampling or threshold moving.
Our Work (Pima)	Maximize Recall & Minimize Cost	LR, RF, GB, XGB (with resampling)	Best Recall: 1.000, Best F2-Score: 0.733	Employs a cost-sensitive framework with resampling and threshold optimization (optimal threshold ~0.11).
Our Work (Symptom)	Maximize Recall & Minimize Cost	LR, RF, GB, XGB (with resampling)	Best Recall: 1.000, Best F2-Score: 0.745, Accuracy: ~37-40%	Validates framework on a second dataset, confirming low-threshold prescription.

This highlights that the evaluation metric and optimization goal chosen determine the model's clinical utility. Our work shows that giving up a significant number of percentage points in overall accuracy is a justified trade-off for achieving a dramatic increase in recall, which is clinically imperative for screening contexts. Validation on the symptom dataset proves this is not an artifact of a single dataset but a general principle.

5. DISCUSSION OF LIMITATIONS

While this study has shown the effectiveness and generalizability of a cost-sensitive framework for high-recall diabetes detection on two datasets, there are several limitations. First, the relatively small and homogeneous size of the Pima dataset means that generalisability of the specific optimal threshold, ~0.11, to larger multi-ethnic populations requires further validation. While diverse, the symptom dataset is also of moderate size, and future work should apply this framework to larger, multi-center cohorts. Second, our cost function used a fixed cost ratio of 5:1 for False Negatives versus False Positives. Whereas this ratio is instructive to illustrate the principle, an exact cost ratio in real-world

practice would need to be determined with input from healthcare economists and clinicians. Third, the interpretability of the most complex models is hard, especially when combined with synthetic data from resampling; future work should integrate XAI techniques. Clinical applications of our high-sensitivity framework also need consideration of its integration within the clinical workflow. A screening tool with a recall >0.94 but a precision of ~0.46 will increase the number of referrals for confirmatory testing by a significant margin. Healthcare systems must be prepared for this increased initial workload. However, this must not be seen as a burden but rather as a strategic reallocation of resources. The marginal cost of additional non-invasive confirmatory tests-e.g., HbA1c-is negligible relative to the long-term costs of managing advanced diabetic complications avoided [21,22]. Hence, we envision the model as an ultra-sensitive first filter in a multi-stage screening process, where positive predictions cost-effectively trigger confirmatory testing, without immediate treatment decisions being taken based on such outcomes. Finally, this remains a computational study; the practical utility and impact of deploying such a sensitive screening tool must be assessed in prospective trials in a real-world healthcare setting.

6. CONCLUSION AND FUTURE WORK

This study shows that, from a clinical point of view, it is not optimal to use the traditional classification threshold of 0.5 for screening of diabetes. We have, for the first time, validated, on two independent datasets, a novel framework that requires two specific actions: 1) the application of resampling techniques in order to rectify the dataset bias and 2) the systematic reduction of the decision threshold into the 0.1-0.2 range. This approach ensures recall >0.94, hence essentially eliminating the possibility of missed diagnoses. Successful application of this framework on an independent symptom-based data set where it achieved perfect recall powerfully demonstrates its robustness and generalisability.

We therefore propose this cost-sensitive framework not merely as a finding but as a foundational standard to which machine learning models in safety-critical screening programs should be developed and evaluated against. The practical implementation gives a clear recommendation for an initial diabetes screening, models should be deployed using resampling and low probability thresholds of 0.1-0.2 to give the best protection against the dangerous outcome of missed diagnoses. The strategic acceptance of higher false positive rates is a necessary and clinically justified trade-off in this critical healthcare domain. Our future work will apply this framework to larger and more diverse datasets, including feature selection guided by clinical expertise, use XAI to provide further insights, and develop this methodology into a user-friendly clinical decision support system for pilot testing.

REFERENCES

[1] International Diabetes Federation, IDF Diabetes Atlas, 10th ed. Brussels, Belgium: IDF, 2021.

[2] M. E. Febrian et al., “Diabetes prediction using supervised machine learning,” Procedia Comput. Sci., vol. 216, pp. 21–30, 2023.

[3] Zhang, Z. (2025). Comparison of Machine Learning Models for Predicting Type 2 Diabetes Risk Using the Pima Indians Diabetes Dataset. Journal of Innovations in Medical Research, 4(1), 65-71.

[4] 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.

[5] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” in Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.), 2008, pp. 1322–1328.

[6] Cerf, M. E. (2013). Beta cell dysfunction and insulin resistance. Frontiers in endocrinology, 4, 37.

[7] Giri, B., Dey, S., Das, T., Sarkar, M., Banerjee, J., & Dash, S. K. (2018). Chronic hyperglycemia mediated physiological alteration and metabolic distortion leads to organ dysfunction, infection, cancer progression and other pathophysiological consequences: An update on glucose toxicity. Biomedicine & pharmacotherapy, 107, 306-328.

[8] González, P., Lozano, P., Ros, G., & Solano, F. (2023). Hyperglycemia and oxidative stress: an integral, updated and critical overview of their metabolic interconnections. International journal of molecular sciences, 24(11), 9352.

[9] Salmi, M., Atif, D., Oliva, D., Abraham, A., & Ventura, S. (2024). Handling imbalanced medical datasets: review of a decade of research. Artificial intelligence review, 57(10), 273.
[10]Osman, H. (2024). Cost-sensitive learning using logical analysis of data. Knowledge and Information Systems, 66(6), 3571-3606.
[11] ElSayed, N. A., Aleppo, G., Aroda, V. R., Bannuru, R. R., Brown, F. M., Bruemmer, D., ... & American Diabetes Association. (2023). 2. Classification and diagnosis of diabetes: standards of care in diabetes—2023. Diabetes care, 46(Supplement_1), S19-S40.
[12]Diaz, F., & Mitra, B. Recall, robustness, and lexicographic evaluation (2023).
[13] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," J. Big Data, vol. 6, no. 1, p. 27, 2019.
[14] C. Elkan, "The foundations of cost-sensitive learning," in Proc. 17th Int. Joint Conf. Artif. Intell., 2001, vol. 1, pp. 973–978.
[15] F. Provost and T. Fawcett, "Robust classification for imprecise environments," Mach. Learn., vol. 42, no. 3, pp. 203–231, 2001.
[16] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," Front. Genet., vol. 9, p. 515, 2018.
[17] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017.
[18] 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.
[19] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, Oct. 2011
[20] C. Chen, "Diabetes Data Upload," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/chengyaran/diabetes-data-upload
[21] American Diabetes Association, "Economic costs of diabetes in the U.S. in 2022," Diabetes Care, vol. 45, no. 5, pp. 1017–1038, 2022.
[22] X. Zhuo, P. Zhang, and T. J. Hoerger, "Lifetime direct medical costs of treating type 2 diabetes and diabetic complications," Amer. J. Prev. Med., vol. 45, no. 3, pp. 253–261, 2018.
[23] R. Chatterjee, K. M. V. Narayan, and J. Lipscomb, "Screening for diabetes and prediabetes should be cost-saving in patients at high risk," Diabetes Care, vol. 42, no. 5, pp. 751–758, 2019.