International Journal of Advanced and Applied Sciences, 13(5) 2026, Pages: 235-245

Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: https://www.science-gate.com/IJAAS.html

AI-based cardiovascular risk stratification using population health data: An intelligent risk assessment agent (IRAA)

Oualid Ali *

College of Arts and Science, Applied Science University, Manama, Kingdom of Bahrain

A R T I C L E I N F O

Article history:
Received 19 January 2026
Received in revised form
20 May 2026
Accepted 24 May 2026

Keywords:
Cardiovascular risk assessment
Explainable artificial intelligence
Risk stratification
Preventive health screening
Machine learning in healthcare

A B S T R A C T

Cardiovascular disease remains one of the leading causes of death worldwide, and identifying Artificial Intelligence paradigms that can support early diagnosis and preventive measures is therefore of great importance. Although many machine learning (ML) studies report high classification accuracy using cardiovascular risk predictors, these results can be misleading because of strong class imbalance and the population-screening nature of the data. In this study, we developed an explainable AI-based IRAA that focuses on cardiovascular risk categorization rather than binary diagnosis. To systematically evaluate different ML models using an imbalance- and risk-sensitive assessment framework, we employed a large population-based health investigation dataset. The proposed system achieved a stable ROC-AUC of approximately 0.83 and a PR-AUC of around 0.31, identifying more than 63% of heart disease cases within the top 25% of risk groups and nearly 78% within the top 30%. These results demonstrate the potential of the model for early screening and case prioritization rather than final clinical decision-making. To improve transparency and user trust, SHAP-based explanations were integrated into a conversational IRAA interface, enabling doctors and users to understand how demographic, lifestyle, and comorbidity factors contribute to an individual's risk assessment. This functionality helps bridge the gap between the interpretability of complex predictive models and user understanding. The findings highlight the limitations of accuracy-focused evaluation methods and support a shift toward explainable and risk-aware AI-based cardiovascular screening at the population level.

1. Introduction

Cardiovascular disease (CVD) has remained among the leading causes of death and long-term disability in all parts of the world, hence putting significant pressure on health-care infrastructures and catalyzing the need to develop scalable measures of early screening (Gul et al., 2026). The most crucial factor is the early identification of individuals at increased cardiovascular risk, which enables the prevention of illness, the equitable distribution of clinical resources, and population-level health management programs (Kissi et al., 2025). Advances in artificial intelligence (AI) and machine learning (ML) have recently enabled modelling of cardiovascular risk based on data, using a large dataset which includes demographic, lifestyle, and comorbidity-related variables (Reátegui et al., 2025; Sen and Bhattacharya, 2025). Relative to traditional clinical risk scores, ML-based models have the power to model nonlinear relationships between risk factors and the ability to adjust to heterogeneous strata of a population (Lippert et al., 2024). Based on this, the concept of AI-based cardiovascular risk assessment systems has attracted growing research interest as a measure of population-wide screening and decision support, as opposed to a conclusive diagnostic modality. However, a majority of population-based cardiovascular data, especially those collected via lifestyle surveys, is highly imbalanced, with clinically meaningful disease states represented by a small fraction of the sample (Babicki et al., 2025). Under these conditions, common performance measures, including overall accuracy or even the area under the receiver operating characteristic curve (ROC-AUC), can be misleading, because they can have high scores

---

* Corresponding Author.
Email Address: oualid.ali@asu.edu.bh
https://doi.org/10.21833/ijaas.2026.05.022
Corresponding author's ORCID profile:
https://orcid.org/0009-0001-1433-017X
2313-626X/

and fail to pick up most true positives (Martin et al., 2025). The recent methodological studies thus suggest the use of precision-recall-based measures, including the area under the precision-recall curve (PR-AUC), to assess rare event prediction and screening systems (Zhu et al., 2025). Alongside, decision policy specification has become an essential component of responsible AI implementation in healthcare settings. Set probability cutoffs (such as a probability cut-off of 0.5) cannot be used in imbalanced screening situations. Rather, methods of evaluation based on ranking, such as recall under a restricted fraction of the population (Recall@K), have been suggested as less realistic measures of screening performance (Talukder et al., 2025). These measures are direct reflections of practical constraints of operation, including limited clinical follow-up capacity. In addition, the explainable AI (XAI) has become more urgent in healthcare applications. Shannaq et al. (2024) proved the ability of the intelligent, explicable human-computer interaction systems to tailor complex decision-support content to heterogeneous user cohorts, which is why the role of transparency and cultural sensitivity cannot be underestimated. Clear disclosure of risk determination is essential to the confidence of clinicians, regulatory compliance, and ethical responsibility (Bilal et al., 2025). Such approaches as feature-attribution methods (especially SHapley Additive exPlanations (SHAP)) have become popular to explain cardiovascular risk models, allowing both a broad view of the negative risk factors and specific explanations of the particular prediction. As a continuation of the results presented in Muneer et al. (2024), which combined data acquisition with chatbot and SHAP /LIME interpretability tools and demonstrated nearly 92 percent classification accuracy on canonical cardiac disease datasets, our research will add to this research question, implementing more advanced ensemble learning methods, a carefully balanced evaluation to class imbalance, and developing an Intelligent Risk Assessment Agent. The result of this whole process is a more resilient, generalizable, and risk-sensitive cardiovascular risk stratification system that goes beyond binary measures of accuracy.

Considering these obstacles, the current study suggests a cardiovascular risk assessment system based on XAI that is clearly designed to stratify risks and screen at risk in an imbalanced population health dataset. Instead of focusing on accuracy, such a framework preempts PR-AUC and Recall at K as the chief measures of evaluation, implicates composite risk characteristics based on clinical intuition, and provides amenable explanations that are coherent with screening judgments. This form of reframing leads to realistic, morally available, and practically applicable AI-led cardiovascular screening.

This work presents an Intelligent Risk Assessment Agent (IRAA) that can be used to provide explainable population health data for cardiovascular risk stratification. Our method is in contrast to the previous machine learning literature, which largely concerns how to enhance general classification accuracy, and users are more inclined to use it in population screening as a risk-ranking task. Accuracy can be misleading in large-scale health surveys that have a relatively low disease prevalence, since models can do well by simply predicting the majority class.

Consequently, the work focuses on ranking-based evaluation measures, especially Precision–Recall Area Under the Curve (PR-AUC) and Recall@K, which are more relevant to the role of the model to recognize individuals who are at risk in the limited mental screening resources. The point of view matches the real-world preventive healthcare situations when the aim is not simply to label the people but to prioritize those whose health is most likely to improve with the early medical examination. The proposed IRAA framework is built on imbalance-conscious machine learning, multi-faceted risk feature engineering, and explainable AI in order to offer an operationally-relevant, transparent strategy to large-scale cardiovascular risk stratification.

2. Literature review

According to recent research, in modern epidemiology, the field of machine learning that uses algorithms to address epidemiological problems, specifically logistic regression and support vector machines, has demonstrated a reliable predictive performance for premature coronary heart disease among younger adults, allowing for premature risk stratification and implementation of precision-guided secondary prevention initiatives using tools that are interpretable and clinically implementable. This has been demonstrated in medical research in recent times, where interpretable machine-learning models, most notably LightGBM, provide accurate predictions of 30-day postoperative cardiovascular risk in older adults undergoing non-cardiac surgery, thus facilitating early risk stratification and informing both clinical decision-making in the perioperative period (Al Khatib et al., 2026).

Recent research still supports the fact that ensemble learning, a subcategory of machine learning, such as boosting, stacking, and hybrid meta-learners, is still very competitive in predicting cardiovascular disease (CVD), especially when the datasets are based on demographic variables, comorbidity histories, and behavioral risk factors. Hybrid ensemble models that combine LightGBM, XGBoost, and CatBoost with meta-learners have been reported to show strong discrimination and improved stability in a wide variety of data sources; however, the new paradigm of promising screening-focused evaluation measures critical to clinical triage is still mostly ignored in reports (Shah et al., 2025).

A study by Wei et al. (2023) reaffirmed the excellent results of highly tuned CatBoost models and multi-model ensembles in predicting CVD risk, but indicates that discriminative conclusions no less than strong ones can be achieved through careful hyperparameter methods as well as complex feature manipulations. Though the reliance on aggregate statistics is also highlighted in these works and can fail to capture the realities of the practical triage decisions in settings with a relatively low rate of positive cases. According to Talaat et al. (2024), the explainable AI model, the CardioRiskNet, based on active learning and attention processes, is much more accurate at cardiovascular risk prediction than conventional approaches.

2.1. Cardiovascular risk prediction using machine learning

Modern literature is regularly characterized by ongoing interest in applying machine-learning methods to the problem of cardiovascular risk prediction, and ensemble-based algorithms often assert a leading role in published performance leaders (Almutairi and Dardouri, 2025). Gradient-boosting algorithms, in particular, XGBoost, LightGBM, and CatBoost, have been experimentally shown to outperform traditional classifiers, especially in extreme cases, where the determinants of risk are heterogeneous, and where missing data concerns (Jiang et al., 2025). A group of more current studies, up to 2024 and beyond, supports this claim, showing that, in fact, CatBoost offers very specific benefits in relation to algorithmic stability and the treatment of categorical variables in healthcare data (Hamid et al., 2025). Massive survey and lifestyle archives are increasingly being used as the foundation of preventive heart valve analytics, especially in the context of the milieu of public health (Babicki et al., 2025). Even though these datasets do not require the use of granular clinical biomarkers, their broad population base and scalability make them exceptionally appealing to initial screening programs (Bilal et al., 2025). However, numerous studies highlight inherent noise and bias associated with the self-reported variables and thus the need for careful interpretation and strict validation procedures (Lippert et al., 2024). This emerging concept illustrates the way in which predictive information systems can be used to support decision-making by estimating and ranking the critical quantities, avoiding deterministic results (Shannaq et al., 2019).

2.2. Class imbalance and evaluation metrics in screening tasks

There is a persistent issue of class imbalance as one of the most difficult methodological challenges of cardiovascular prediction. Even high accuracy rates have been reported recently to be accompanied by very low sensitivity rates in cases where only positive results are uncommon. In its turn, this has led to a growing questioning of the assessment practices that have been inclined towards accuracy in the context of the screening and early-warning systems (Segar et al., 2021). Precision-recall diagnostics have emerged as one of the popular alternatives to assess models of rare events. It has been established that the PR-AUC measure provides a more acute assessment of the minority-class behavior and associated trade-offs related to false alarms (Talukder et al., 2025). In addition, ranking-based techniques, such as Recall@K, have also been proposed to test the proficiency with which models rank genuine cases when faced with restricted screening budgets (Zhu et al., 2025). However, despite these methodological advances, a rather high percentage of cardiovascular prediction studies are still reporting PR-AUC values within underreporting and fail to state explicit decision policies (Kissi et al., 2025).

2.3. Explainable AI in cardiovascular risk assessment

Explainable AI (XAI) has been refined into a research theme topic in cardiovascular machine-learning, driven by the need to achieve transparency, equity, and clinical interpretability (Gul et al., 2026). SHAP-based explanatory models have been embraced with a significant rate of adoption to outline the most important risk factors in heart diseases and to provide personalized justifications to the prediction of risk (Banerjee and Paçal, 2025). Shannaq (2025) presented an explicable artificial-intelligence-based medical assistant to predict Alzheimer's disease, which provides insights into the future of XAI as a way to build trust and interpretability in healthcare AI systems.

Modern research testifies to the fact that XAI has the potential to promote clinician validation, explain dataset biases, and enhance the trust in AI-based decision-support systems (Wah, 2025).

Despite such benefits, there is a literature in which the explanations are often conveyed without a strong validation or an established nexus to operational decision-making (Reátegui et al., 2025). Such explanations have not been sufficiently challenged on the issue of their stability among cross-validation folds and consistency with screening thresholds, particularly when it comes to conducting population-level studies (Sen and Bhattacharya, 2025).

2.4. AI-based risk assessment systems and deployment scenarios

As of late, syntheses highlight how cardiovascular prediction systems based on AI should be theorized as the decision-support or risk evaluation tools and not autonomous diagnostic systems (Martin et al., 2025). The distinction presupposes an increased salience when the population-screening situations are considered, and AI outputs are used to screen potential follow-ups instead of replacing the opinion of a clinician.

Digital-health explorations are increasingly promoting modular artificial intelligence (AI) risk evaluation architectures suitable for seamless administrative integration into electronic-health record systems, public-health dashboards, or automated screening platforms (Kissi et al., 2025). The harmonization of the evaluation metrics, interpretability modalities, and deployment premises is currently considered to be an indispensable aspect of the responsible adoption of AI in the context of cardiovascular screening.

Modern estimates indicate that cardiovascular diseases (CVDs) will cause nearly 23 million deaths annually in the year 2030 (Coronnello and Francipane, 2022), an amount depicting the pathogenesis of such conditions as myocardial infarction, atrial fibrillation, and heart failure (Hossain et al., 2021). The risk determinants of CVD are multi-factorial, including demographic, anthropometric, and biochemical factors, including age, ethnicity, body mass index, and laboratory indices. However, conventional risk-assessment tools are limited by their low accuracy, the need to use fixed variables, population-averaged assumptions, poor representation of nonlinear relationships, and inability to withstand the fluctuating patient information (Bhatt et al., 2023; Chang et al., 2022). To that end, the explainable artificial-intelligence model below overcomes such limitations by taking advantage of mass population-survey data, pursuing imbalance-sensitive assessment measures, and using SHAP-based interpretability, which facilitates dynamic, transparent, and risk-focused cardiovascular screening instead of traditional, accuracy-oriented prognostication.

3. Methodology

The structure of the proposed approach for developing an Intelligent Risk Assessment Agent (IRAA) is presented in Fig. 1. The proposed structure is to make the proposed cardiovascular risk assessment system based on AI reasonable, transparent, and applicable to screening the population. Therefore, Restates cardiovascular prediction as risk stratification in place of diagnosis. Uses risk-sensitive and imbalance-aware learning. Provides composite characteristics of interpretability and stability.

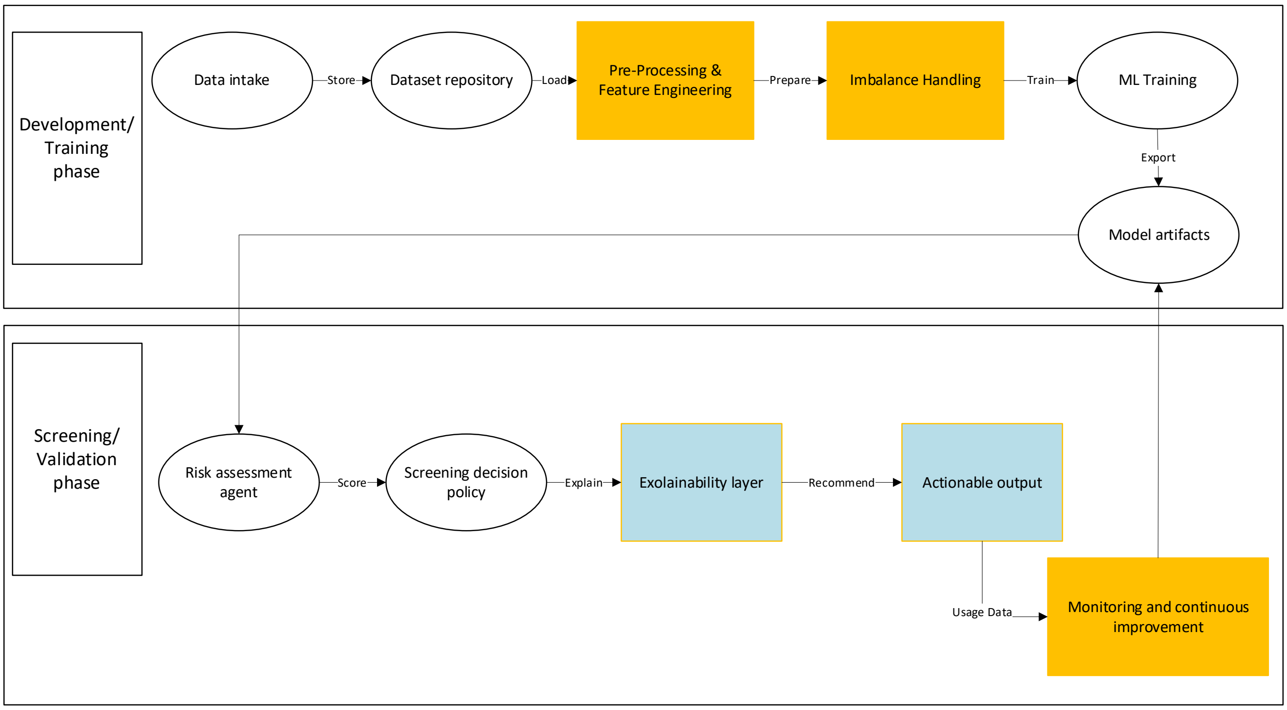


Fig. 1: Proposed intelligent risk assessment agent (IRAA)

3.1. Study design and problem formulation

This study has considered the forecast of cardiovascular disease to be an issue of risk stratification and screening, and not a mere binary diagnostic undertaking. Since population health data is highly skewed, the first objective is ranking of individuals based on the approximate cardiovascular risk and determination of the effectiveness of the true positive cases to be detected in a small sample of the population. This expression is matched by the situation in the real world of screening, whereby the clinical resources limit the number of people who can be given priority to make follow-ups. Based on this, the model performance is measured based on precision-recall and ranking-oriented metrics instead of being based on accuracy or fixed-threshold classification results.

3.2. Dataset description

The dataset used in this study was derived from the 2021 Behavioral Risk Factor Surveillance System (BRFSS), a large-scale health survey conducted annually by the Centers for Disease Control and Prevention (CDC) in the United States. The dataset was made publicly available through the Kaggle repository Cardiovascular Diseases Risk Prediction Dataset. BRFSS is one of the largest continuous health surveys worldwide and contains extensive information related to health conditions, lifestyle behaviors, and preventive healthcare practices among adults. The original dataset includes approximately 438,693 survey records and more than 300 variables covering demographic, health, and lifestyle information. Since many variables in the original dataset were not directly related to cardiovascular risk prediction, a subset of 19 relevant attributes was selected for the development of the machine learning models used in this study. These variables provide important information about demographic characteristics, physical health conditions, lifestyle behaviors, and medical history, which are commonly associated with cardiovascular disease risk.

The target variable in the dataset is Heart_Disease, which indicates whether a respondent has been diagnosed with a cardiovascular condition. The selected predictor variables include general health status, physical activity, diabetes, smoking history, alcohol consumption, dietary habits, and body mass index (BMI). Both categorical and numerical variables were included to represent different aspects of individual health and lifestyle. Demographic variables such as sex and age category were used for population grouping, while physiological indicators included height, weight, and BMI. In addition, behavioral variables such as exercise habits, smoking status, alcohol consumption, and fruit and vegetable intake were used to represent lifestyle patterns associated with cardiovascular health outcomes. The variables used in this study are presented in Table 1.

Table 1: Category and variables information

Category	Variables
Demographic variables	Sex, Age_Category
Health status variables	General_Health, Diabetes, Depression, Arthritis
Medical history variables	Skin_Cancer, Other_Cancer
Lifestyle variables	Exercise, Smoking_History, Alcohol_Consumption
Dietary behavior variables	Fruit_Consumption, Green_Vegetables_Consumption, FriedPotato_Consumption
Physical indicators	Height_(cm), Weight_(kg), BMI
Healthcare behavior	Checkup
Target variable	Heart_Disease

Applying the curated subset of variables can greatly limit the dimensionality of the dataset and retain the most predictive variables to model cardiovascular risks. The strategy of attribute selection is also useful in generating interpretable machine learning models and is more efficient in terms of computation during training and evaluation.

The resulting dataset thus represents a complete but narrow view of population health factors and thus fits well in the creation of the proposed Intelligent Risk Assessment Agent (IRAA) that is to be developed to help in explainable cardiovascular risk stratification.

3.3. Data preprocessing

Among the variables used in this study, presented in Table 1, the Heart_Disease variable is the binary target variable that is used to determine the presence of cardiovascular disease in a respondent.

1) Data Cleaning and Encoding: Label encoding was used to encode categorical variables, relying on the fact that CatBoost is capable of working with categorical variables by default. Continuous variables such as height, weight, and the BMI were left as numeric values. The missing values were treated with the built-in mechanisms of CatBoost, which was constructed to retain the information without the need to impute values explicitly.

The presented empirical investigation indicates that the choices of the experimental design, in particular, the dataset-splitting policies, can have a stronger impact on the model performance than any possible preprocessing methods in the realm of big data analytics (Shannaq, 2025).

2) The second feature is Composite Risk Feature Engineering: Composite risk features were formed by pooling together related variables to form clinically inspired indicators to further increase the interpretability and stabilize learning in the presence of noisy self-reported data. In order to improve the predictive ability of the model, a number of composite risk indicators were designed by putting together related variables in health and lifestyle. These characteristics involve an interplay between behavioral and physiological variables that are related to cardiovascular risk.

The composite indicators that were produced are presented in Table 2.

Table 2: Composite indicators

Feature	Indictor	Description
Lifestyle risk score	Lifestyle_Risk = Smoking_History + Alcohol_Consumption + FriedPotato_Consumption	This feature combines lifestyle-related behaviors known to influence cardiovascular health. Higher values indicate greater exposure to unhealthy lifestyle behaviors.
Healthy diet index	Healthy_Diet=Fruit_Consumption+Green_Vegetables_Consumption	A dietary balance indicator was constructed using fruit and vegetable consumption. Higher values indicate healthier dietary patterns.
Physical health burden	Health_Burden=Diabetes+Depression+Arthritis	This feature aggregates several chronic health conditions. This indicator reflects the cumulative burden of chronic conditions associated with cardiovascular complications.
Body risk indicator	Body_Risk=BMI	Body mass index (BMI) was used as a proxy for obesity-related cardiovascular risk. BMI is widely recognized as a major risk factor in cardiovascular epidemiology.

These composite features are not meant to add complexity to models, but rather, they are meant to give semantically meaningful representations backing both the risk stratification program and the explainability program.

3) CatBoost Model Configuration: The choice of the CatBoost gradient boosting algorithm was motivated by high results on heterogeneous tabular data and the possibility of its efficient work with categorical features. The ratio of negative to positive samples was used to weight classes to obtain the desired result of the high-class imbalance in the dataset. Primary hyperparameters of the CatBoost model are summarized in Table 3.

Table 3: Primary hyperparameters of the CatBoost model

Parameter	Value
Loss function	Logloss
Evaluation metric	AUC
Iterations	500
Learning rate	0.05
Depth	6
L2 regularization	3
Random seed	42
Scale_pos_weight	11.37
Boosting type	Ordered Boosting

These hyperparameters, presented in Table 3, were chosen to provide a tradeoff between the complexity of the model and the generalization performance and increase minority-class detection.

3.4. Handling class imbalance

Since the ratio between positive and negative cases was very high, as shown in Fig. 2, several imbalance-conscious strategies were included:

- Class-weighted learning: The scale pos weight parameter of CatBoost was determined by the proportion of negative samples to positive samples, and misclassifying the negative class was more heavily penalized in training.
- Hybrid resampling: The SMOTETomek strategy was utilized in the cross-validation fold to balance the training data, as well as to eliminate unclear samples on the boundary of the classes. This technique alleviates overfitting of a pure oversampling but retains a clear boundary of the decision.

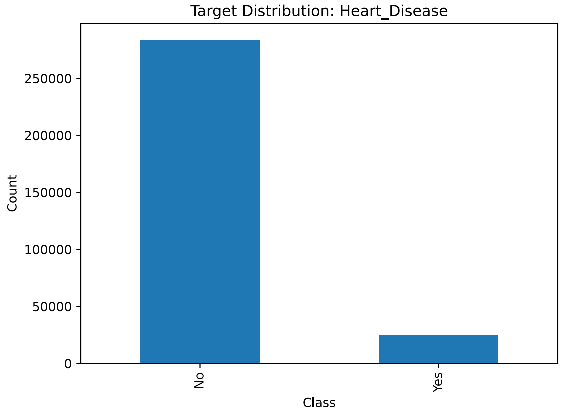


Fig. 2: Target distribution (heart disease)

Particularly, all resampling processes were conducted on training folds only, without touching the validation set or test set, so that there was no leakage of information.

3.5. Model architecture and training

1) CatBoost Classifier: The main predictive model applied in the study is the CatBoost gradient-boosting classifier, as it is highly applicable to heterogeneous tabular data, highly reacts to categorical variables, and is not prone to overfitting in imbalanced environments. The major benefits of CatBoost in this case are:

- Built-in support of categorical features.
- Boosting is ordered to minimize the target leakage.
- Fold consistency in cross-validation.

2) Cross-Validation Protocol: The stratified k -fold cross-validation plan was used to make sure that there was a uniform distribution of classes among the folds. All pre-processing tasks, such as composite feature constructions and resampling, were also performed separately in each fold to maintain a high level of experiment integrity.

3.6. Evaluation metrics

In order to support the screening-oriented nature of the study, the assessment was based on risk-conscious and imbalance-sensitive indicators.

1) Precision- Recall Area Under Curve (PR-AUC): The main global performance measure was PR-AUC, which is the direct measure of the trade-off between true positive detection and false alarms in unbalanced samples.

2) Recall@K: Recall@K was used to determine the capability of the system to capture the actual cardiovascular cases within a limited percentage of the population. Specifically:

- Recall 20% is the proportion of true positive cases found in the 20 percent of individuals who are highest-risk.
- Recall30% is the proportion of one that is captured in the top 30%.

Such measures are directly related to viable screening limits and are more informative than fixed threshold recall when doing population health.

3) Secondary Metrics: ROC-AUC and accuracy were reported as well in order to be complete and to be compared with the previous work. Nevertheless, they did not serve as metrics to select the model and draw the main conclusions.

3.7. Explainability and risk interpretation

1) SHAP-Based Explanation: SHAP (Shapley Additive Explanations) was used to interpret the trained CatBoost model in order to be transparent

and interpretable. The SHAP values are the values that measure how each feature contributes to single predictions of risk and can be interpreted globally and locally.

2) Global Risk Factor Analysis: The most influential risk factors among the population were determined using global SHAP summaries, aiding in validation of the previously accepted cardiovascular risk knowledge.

3) Risk Explanation on an individual level: SHAP force and waterfall explanations were also applied to show the individual predictions of determining which demographic, lifestyle, and comorbidity factors increase or decrease the risk of cardiovascular. Such exposition is critical in ethical implementation in screening situations where transparency and accountability are needed.

3.8. Implementation environment

The experiments were all done in Python-based scientific computing libraries. Model training, evaluation, and explainability analyses were implemented in a controlled experimental setting in order to have reproducibility. Simulations of workflow and interface of the system were then developed and tested.

4. Experiments and results

All the experiments were thoroughly planned to study the effectiveness of the suggested intelligent risk assessment agent in population-based screening conditions. The experimental structure confined its focus on risk ranking and the acquisition of examples representing the minority-class group, rather than the traditional objective of maximizing accuracy based on a threshold, to the extent it understands the strength of the class imbalance in the dataset. Table 4 presents the main experimental settings. To this end, a stratified fivefold cross-validation procedure was followed to maintain the homogenous class representation across the divided subsets. Each fold was confined with all the data preprocessing methods, composite risk feature engineering, and mitigating imbalanced processes confined to the training part, and hence preventing leakage of such information in the validation stream. Only training data was resampled with a hybrid scheme of SMOTETomek making the prevalence of the minority class equal, but excising samples that are close to decision boundaries, and leaving the validation sets untouched to accurately represent the prevalence in the real-world. The predictive machine was a CatBoost classifier, which is adjusted to the issue of class disparity by proper weighting plans based on the proportion of negative and positive samples and reinforced by ordered boosting to encourage a predictor to be stable. The metric variables of evaluation were based on the need to understand screening imperatives, and PR-AUC was selected as the key performance measure in the derivation of which were Recall@30 and Recall@20 that were used to measure the percentage of recovered authentic heart-disease cases among tightly defined high-risk groups. ROC-AUC was said to be calculated as a secondary measure of general discriminative ability.

Per-fold performance estimates were created, which were aggregated using mean and standard deviation measures, but consolidated cross-validated predictions were used to project effectiveness in screening the entire world.

Table 4: Summary of experimental settings

Component	Description
Validation strategy	Stratified 5-fold cross-validation
Class imbalance handling	Class-weighted learning + SMOTETomek (training only)
Feature design	Original features + composite risk indicators
Predictive model	CatBoost (ordered boosting, probability output)
Primary evaluation metric	PR-AUC
Screening metrics	Recall@20%, Recall@30%
Secondary metric	ROC-AUC
Data leakage prevention	All preprocessing applied within training folds only

4.1. Cross-validation performance results

Table 5 shows the performance of the proposed intelligent risk assessment agent in terms of cross-validation, which is displayed in the form of PR-AUC, ROC-AUC, and Recall at K.

Table 5: Cross-validation results of the proposed risk assessment agent

Fold	PR-AUC	ROC-AUC	Recall@20%	Recall@30%
1	0.299	0.831	0.622	0.772
2	0.300	0.832	0.622	0.770
3	0.310	0.835	0.634	0.778
4	0.310	0.838	0.631	0.785
5	0.316	0.837	0.634	0.786
Mean ± SD	0.307 ± 0.007	0.835 ± 0.003	0.629 ± 0.006	0.778 ± 0.007

4.2. Pooled screening performance

To evaluate deployment-level screening performance, pooled predictions from all validation folds were aggregated. The combined results demonstrated stable screening effectiveness, achieving a pooled PR-AUC of 0.307, a Recall@20% of 0.628, and a Recall@30% of 0.779. These findings further confirm the robustness of the proposed framework across validation folds and support its applicability to large-scale cardiovascular risk screening.

According to these findings, it would be reasonable to conclude that about 63 % of heart-disease cases are within the top 20 per cent of people with the highest risk scores, with about 78 %

of cases falling within the top 30 %. These proportions provide a solid argument in favor of the appropriateness of the system in large-scale screening programs, as well as of prioritization of resources in a population-based healthcare scenario.

The pooled precision-recall (PR) curve in Fig. 3 demonstrates the performance of the proposed Intelligent Risk Assessment Agent organized in a screening-oriented manner in all of the folds during the validation and with a total PR-AUC of 0.3066. In contrast to receiver-operating characteristic (ROC) curves, which can leave an overly optimistic impression when faced with highly skewed data, the PR curve provides a more realistic portrayal of what the model is able to do right when it comes to indicating that a particular case is of the minority-class, namely, one with heart disease, in a setting that approximates the reality in terms of population screening.

On the curve at low recall levels, the precision is relatively high and indicates that the individuals who are at the very top of the risk list have an experimentally high possibility of actually having cardiovascular disease. The ability to think so is particularly welcome in screening procedures, whereby the early identification of at-risk individuals promotes effective prioritization of individuals to undergo clinical follow-up. Recall is increasing smoothly and not suddenly, so that the effect is an increased precision as recall increases, not a high precision rate.

It is interesting to note that the systematic increase of the curve above the baseline prevalence range throughout the recall range proves that the model is dispensing significant risk discrimination significantly beyond random sorting. Since the prevalence of heart disease in the population survey data is low, the achieved PR-AUC is a significant improvement on the non-parametric performance of chance; therefore, resolving the fact that the model is a risk-ranking and prioritization tool, but not a dichotomous diagnostic tool.

The novelty perspective, in its turn, enables this figure to visually support the key research hypothesis: effective cardiovascular screening must be evaluated by precision, recall behavior, and risk concentration, but not by the overall accuracy. The PR curve clearly demonstrates the maintenance of the screening utility in spite of reaching a performance plateau period in the conventional metrics, and, therefore, provides an operationally and clinically relevant assessment of AI-based risk-stratification systems.

Fig. 3 shows how precision and recall change with screening thresholds, giving the indications of efficient stratification of risk and a good sample of the minority class in an imbalanced population health dataset with a PR-AUC of 0.3066.

4.3. Screening metrics interpretation

Although the total classification performance was comfortably higher than 91 percent, which is not reported here to save space, threshold-based analysis showed a conspicuously low recall when traditional probability cut-offs were used. This result is in line with the established literature on the inflammatory nature of accuracy-based measures on perceptions of screening algorithm effectiveness in the setting of class imbalance typical of cardiovascular data.

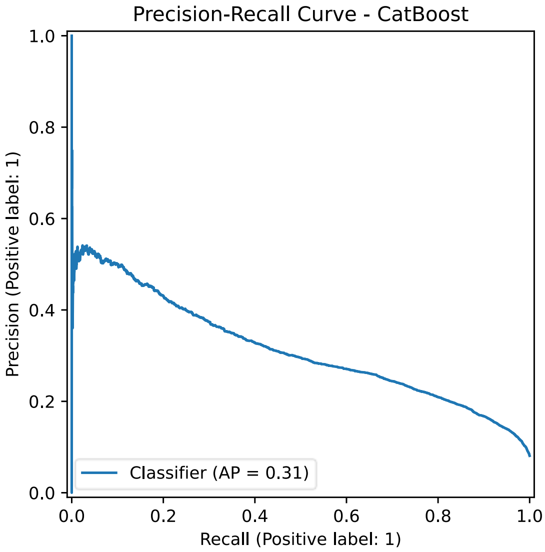


Fig. 3: Pooled precision–recall curve of the proposed intelligent risk assessment agent

In contrast, performance indicators like PR-AUC and Recall@K provide a much more valid estimation of real-world performance. The PR-AUC that was observed, of about 0.31, is three times more than the discriminative power when compared with random ranking, which is high, taking into consideration the fact that the prevalence of heart disease in the cohort is low. In addition, the agreement of Recall@K between cross-validation folds testifies to a strong risk prioritization and not artefact overfitting.

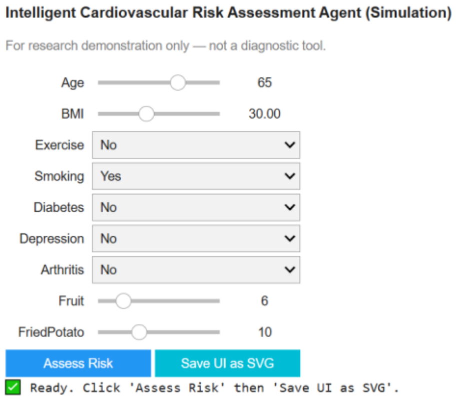
4.4. Implications on population screening

The findings demonstrate that the proposed intelligent risk assessment agent is effective as a population-level risk-ranking and screening tool rather than a definitive diagnostic system. The results further show that high overall accuracy alone is insufficient for evaluating screening performance in highly imbalanced cardiovascular datasets. Instead, PR-AUC and Recall@K provide more clinically meaningful measures because they better capture the model’s ability to prioritize high-risk individuals under limited screening resources. The proposed framework consistently identified a large proportion of true positive cardiovascular cases within restricted screening budgets, while the stability of the cross-validation results supports the robustness and generalizability of the system for preventive healthcare applications. Fig. 4 demonstrates the developed simulator of the Intelligent Risk Assessment Agent (IRAA), and Fig. 5 demonstrates the output of the IRAA simulator.

5. Discussion

Although the positive outcomes of this study have been achieved, some limitations should be mentioned. To begin with, the data used in this paper is based on the Behavioral Risk Factor Surveillance System (BRFSS), which is a source based on self-reported survey data. Although this type of data can provide important insights about populations, the use of self-reporting introduces recall biases, reporting biases, and measurement errors, especially when it comes to behavioral data, including the consumption of alcohol, diet, and exercise frequency. Such variables can undermine the validity of some predictors and, by implication, influence the performance of the model.

Second, since the proposed model achieves a pooled PR-AUC of about 0.31, which is significantly higher than random ranking in an imbalanced dataset, it is also an indicator of high levels of predictive uncertainty. The risk of cardiovascular diseases is controlled by a complex of interactions between genetic, behavioral, and physiological factors that may not be well-represented in survey-based data. Thus, the available results are to be viewed as the warning signs of risk stratification as opposed to the clinical anticipations.



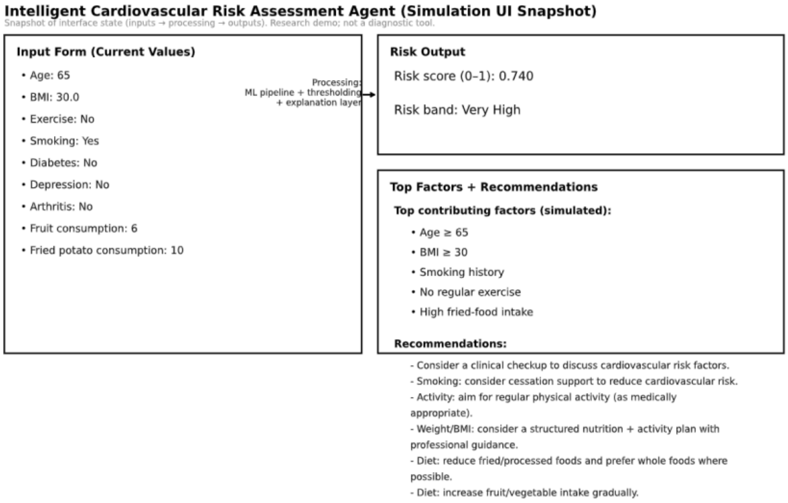
Intelligent Cardiovascular Risk Assessment Agent (Simulation)

For research demonstration only — not a diagnostic tool.

Age 65
BMI 30.00
Exercise No
Smoking Yes
Diabetes No
Depression No
Arthritis No
Fruit 6
FriedPotato 10

✓ `Ready. Click 'Assess Risk' then 'Save UI as SVG'.`

Fig. 4: IRAA simulator interface (input)



Intelligent Cardiovascular Risk Assessment Agent (Simulation UI Snapshot)

Snapshot of interface state (inputs → processing → outputs). Research demo; not a diagnostic tool.

Input Form (Current Values)

- Age: 65
- BMI: 30.0
- Exercise: No
- Smoking: Yes
- Diabetes: No
- Depression: No
- Arthritis: No
- Fruit consumption: 6
- Fried potato consumption: 10

Risk Output

Risk score (0–1): 0.740

Risk band: Very High

Top Factors + Recommendations

Top contributing factors (simulated):

- Age ≥ 65
- BMI ≥ 30
- Smoking history
- No regular exercise
- High fried-food intake

Recommendations:

- Consider a clinical checkup to discuss cardiovascular risk factors.
- Smoking: consider cessation support to reduce cardiovascular risk.
- Activity: aim for regular physical activity (as medically appropriate).
- Weight/BMI: consider a structured nutrition + activity plan with professional guidance.
- Diet: reduce fried/processed foods and prefer whole foods where possible.
- Diet: increase fruit/vegetable intake gradually.

Fig. 5: Sample of the output of the IRAA simulator

Third, the data corresponds to a population health survey as opposed to a clinical cohort, which means that there are no medical variables of interest, including laboratory biomarkers, blood pressure, or electrocardiograms. The integration of these clinical pointers might have the potential to improve the predictive performance and provide finer patient-level risk estimates.

The proposed Intelligent Risk Assessment Agent requires further development in future research directions in a number of ways. Making sure that the framework is validated by using clinically verified datasets where cardiovascular diagnoses are verified using medical examinations, and not self-reported survey data, will be a critical step. This would determine the applicability of the model to real-life healthcare settings.

Also, the incorporation of additional data modalities, such as electronic health records, laboratory tests, or wearable health-monitoring data, can significantly enhance the accuracy and reliability of risk-stratification models. Another perspective is that the suggested system can be integrated into the preventive healthcare screening process; that is, AI-based risk prioritization can support the practices of healthcare professionals prioritizing people to undergo follow-up assessment or preventive measures.

Finally, the potential synthesis of population-wide survey data and clinically validated sources of information could facilitate the construction of stronger and interpretable cardiovascular risk assessment systems, thus facilitating the establishment of large-scale preventive healthcare programs.

6. Conclusion

The key value addition of the paper is to redefine cardiovascular disease prediction as a risk-ranked population screening task instead of a typical accuracy-based classification problem. Though other past research has focused on enhancing the general predictive accuracy, these measures can be valuable only in imbalanced population health data. The suggested Intelligent Risk Assessment Agent, on the contrary, gives more priority to precision recall-based evaluation and Recall@k measures, which allows identification of individuals at high risk in limited screening budgets. This ranking-based viewpoint offers a more feasible application to preventive healthcare, in which the goal is to prioritize people to undergo further medical evaluation, instead of giving them definite diagnoses. The experiment results indicate that the given approach can provide meaningful risk stratification at the same time as it can be interpreted in the form of explainable AI methodologies.

This study explored the issue of cardiovascular disease prediction based on population health survey data, but the paradigm in use was risk stratification as opposed to the simple diagnostic paradigm. By doing this, it attempted to manipulate the methodological and practical challenges that occur in the face of extreme class imbalance and lack of clinical variables. The obtained empirical data prove that although the conventional performance measures, including the overall accuracy and ROC-AUC, are increasing at a slow rate, the screening-oriented measures, e.g., the recall-based thresholds, provide a more realistic and practical assessment of the usefulness of the model. Overall, the obtained results in this work demonstrate that the idea of population-level cardiovascular screening is best conceptualised as a risk-ranking problem, and the central aim of such screening is the effective identification of people at high risk, as opposed to the exceptional classification rates.

Though the suggested solution has shown good potential in terms of cardiovascular risk stratification of the population, more clinical applications should be allowed before such systems are implemented into regular healthcare practice.

The proposed intelligent risk assessment agent provides a real-world, ethically based avenue to AI-assisted cardiovascular screening as model design, evaluation metrics, and interpretability mechanisms are synchronized to this end.

List of abbreviations

AI	Artificial intelligence
AP	Average precision
AUC	Area under the curve
BMI	Body mass index
BRFSS	Behavioral Risk Factor Surveillance System
CatBoost	Categorical boosting
CDC	Centers for Disease Control and Prevention
cm	Centimeter
CVD	Cardiovascular disease
IRAA	Intelligent Risk Assessment Agent
kg	Kilogram
L2	Level 2 regularization
LightGBM	Light Gradient Boosting Machine
LIME	Local Interpretable Model-agnostic Explanations
ML	Machine learning
PR	Precision–recall
PR-AUC	Precision–Recall Area Under the Curve
Recall@20%	Recall within the top 20% highest-risk population
Recall@30%	Recall within the top 30% highest-risk population
Recall@K	Recall at top K ranked instances
ROC	Receiver operating characteristic
ROC-AUC	Receiver Operating Characteristic – Area Under the Curve
SD	Standard deviation
SHAP	SHapley Additive exPlanations
SMOTETomek	Synthetic Minority Oversampling Technique with Tomek Links
SVG	Scalable Vector Graphics
UI	User interface
XAI	Explainable artificial intelligence
XGBoost	Extreme Gradient Boosting

Data availability

The dataset used in this study was derived from the 2021 Behavioral Risk Factor Surveillance System (BRFSS) published by the Centers for Disease Control and Prevention (CDC). The processed dataset analyzed during the current study is publicly available through the Kaggle repository: https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset

The original BRFSS dataset is available from the CDC repository: https://www.cdc.gov/brfss/

Funding

Funding was received for this study from the Applied Science University, Manama, Kingdom of Bahrain.

Acknowledgment

The author gratefully acknowledges the support provided by Applied Science University, Manama, Kingdom of Bahrain, for funding and supporting this research work.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Al Khatib N, Chehab A, Tamim H, Dakik HA, and Tajeddine R (2026). Machine learning-based cardiovascular risk calculator for non-cardiac surgery. Open Heart, 13: e003565. https://doi.org/10.1136/openhrt-2025-003565 **PMid:41500561 PMCid:PMC12781980**

Almutairi M and Dardouri S (2025). Intelligent hybrid modeling for heart disease prediction. Information, 16(10): 869. https://doi.org/10.3390/info16100869

Babicki M, Ledwoch J, Zieliński T et al. (2025). Assessment of cardiovascular risk factors and effect of lifestyle in individuals without cardiovascular disease, diabetes or chronic kidney disease. Scientific Reports, 15: 13544. https://doi.org/10.1038/s41598-025-98215-5 **PMid:40253544 PMCid:PMC12009300**

Banerjee T and Paçal İ (2025). A systematic review of machine learning in heart disease prediction. Turkish Journal of Biology, 49(5): 600-634. https://doi.org/10.55730/1300-0152.2766 **PMid:41246228 PMCid:PMC12614364**

Bhatt CM, Patel P, Ghetia T, and Mazzeo PL (2023). Effective heart disease prediction using machine learning techniques. Algorithms, 16(2): 88. https://doi.org/10.3390/a16020088

Bilal A, Alzahrani A, Almohammadi K, Saleem M, Farooq MS, and Sarwar R (2025). Explainable AI-driven intelligent system for precision forecasting in cardiovascular disease. Frontiers in Medicine, 12: 1596335. https://doi.org/10.3389/fmed.2025.1596335 **PMid:40703259 PMCid:PMC12283576**

Chang V, Bhavani VR, Xu AQ, and Hossain MA (2022). An artificial intelligence model for heart disease detection using machine learning algorithms. Healthcare Analytics, 2: 100016. https://doi.org/10.1016/j.health.2022.100016

Coronnello C and Francipane MG (2022). Moving towards induced pluripotent stem cell-based therapies with artificial intelligence and machine learning. Stem Cell Reviews and Reports, 18: 559-569. https://doi.org/10.1007/s12015-021-10302-y **PMid:34843066 PMCid:PMC8930923**

Gul G, Korejo IA, Hakro DN, Alqahtani H, Abbasi A, Babar M, Al Rahbi O, and Ali NI (2026). Machine learning and ensemble methods for cardiovascular disease prediction: A systematic review of approaches, performance trends, and research challenges. Computers, 15(1): 25. https://doi.org/10.3390/computers15010025

Hamid M, Hajjej F, Alluhaidan AS, and bin Mannie NW (2025). Fine tuned CatBoost machine learning approach for early detection of cardiovascular disease through predictive modeling. Scientific Reports, 15: 31199. https://doi.org/10.1038/s41598-025-13790-x **PMid:40854918 PMCid:PMC12378338**

Hossain ME, Uddin S, and Khan A (2021). Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes. Expert Systems with Applications, 164: 113918. https://doi.org/10.1016/j.eswa.2020.113918

Jiang ZZ, Jiang YF, Wang Y, Zhou Y, Peng RL, Ma CY, and Liu XT (2025). Development of an explainable machine learning model for 3-year cardiovascular risk prediction in new-onset type 2 diabetes using the TyG index and ultrasound features. BMC Medical Informatics and Decision Making, 25: 409. https://doi.org/10.1186/s12911-025-03247-6 **PMid:41188894 PMCid:PMC12584365**

Kissi SA, Talukder MGM, and Iqbal MZ (2025). Data-driven predictive modelling of lifestyle risk factors for cardiovascular health. Electronics, 14(14): 2906. https://doi.org/10.3390/electronics14142906

Lippert M, Dumont KA, Birkeland S, Nainamalai V, Solvin H, Suther KR, Bendz B, Elle OJ, and Brun H (2024). Cardiac anatomic digital twins: Findings from a single national centre. European Heart Journal - Digital Health, 5(6): 725-734. https://doi.org/10.1093/ehjdh/ztae070 **PMid:39563912 PMCid:PMC11570384**

Martin B, Bennett TD, DeWitt PE, Russell S, and Sanchez-Pinto LN (2025). Use of the area under the precision-recall curve to evaluate prediction models of rare critical illness events. Pediatric Critical Care Medicine, 26(6): e855-e859. https://doi.org/10.1097/PCC.0000000000003752 **PMid:40304543 PMCid:PMC12133047**

Muneer S, Ghazal TM, Alyas T, Raza MA, Abbas S, AlZoubi O, and Ali O (2024). Explainable AI-driven chatbot system for heart disease prediction using machine learning. International Journal of Advanced Computer Science and Applications, 15(12): 249-261. https://doi.org/10.14569/IJACSA.2024.0151227

Reátegui R, Tandazo-Malla C, Suárez R, and Ramírez-Cerna L (2025). Cardiovascular risk prediction via ensemble machine learning and oversampling methods. Scientific Reports, 15: 43576. https://doi.org/10.1038/s41598-025-30895-5 **PMid:41326754 PMCid:PMC12698720**

Segar MW, Jaeger BC, Patel KV et al. (2021). Development and validation of machine learning–based race-specific models to predict 10-year risk of heart failure: A multicohort analysis. Circulation, 143(24): 2370-2383. https://doi.org/10.1161/CIRCULATIONAHA.120.053134 **PMid:33845593 PMCid:PMC9976274**

Sen J and Bhattacharya S (2025). An explainable hybrid framework for early detection of cardiovascular diseases using Categorical Boosting and Bees algorithm. Scientific Reports, 15: 45748. https://doi.org/10.1038/s41598-025-28514-4 **PMid:41390781 PMCid:PMC12756275**

Shah P, Shukla M, Dholakia NH, and Gupta H (2025). Predicting cardiovascular risk with hybrid ensemble learning and explainable AI. Scientific Reports, 15: 17927. https://doi.org/10.1038/s41598-025-01650-7 **PMid:40410273 PMCid:PMC12102235**

Shannaq B (2025). A smart medical assistant robot for explainable AI-based Alzheimer's disease prediction using big data analytics. International Journal of Advanced and Applied Sciences, 12(10): 118-128. https://doi.org/10.21833/ijaas.2025.10.014

Shannaq B, Adebiaye R, Owusu T, and Al-Zeidi A (2024). An intelligent online human-computer interaction tool for adapting educational content to diverse learning capabilities across Arab cultures: Challenges and strategies. Journal of Infrastructure, Policy and Development, 8(9): 7172. https://doi.org/10.24294/jipd.v8i9.7172

Shannaq B, Al Shamsi I, and Majeed SNA (2019). Management information system for predicting quantity martials. TEM Journal, 8(4): 1143-1149. https://doi.org/10.18421/TEM84-06

Talaat FM, Elnaggar AR, Shaban WM, Shehata M, and Elhosseini M (2024). CardioRiskNet: A hybrid AI-based model for explainable risk prediction and prognosis in cardiovascular disease. Bioengineering, 11(8): 822. https://doi.org/10.3390/bioengineering11080822 **PMid:39199780 PMCid:PMC11351968**

Talukder MA, Talaat AS, Kazi M, and Khraisat A (2025). XAI-HD: An explainable artificial intelligence framework for heart disease detection. Artificial Intelligence Review, 58: 385. https://doi.org/10.1007/s10462-025-11385-6

Wah JNK (2025). Revolutionizing e-health: The transformative role of AI-powered hybrid chatbots in healthcare solutions. Frontiers in Public Health, 13: 1530799. https://doi.org/10.3389/fpubh.2025.1530799 **PMid:40017541 PMCid:PMC11865260**

Wei X, Rao C, Xiao X, Chen L, and Goh M (2023). Risk assessment of cardiovascular disease based on SOLSSA-CatBoost model. Expert Systems with Applications, 219: 119648. https://doi.org/10.1016/j.eswa.2023.119648

Zhu H, Asiaee A, Azinfar L, Li J, Liang H, Irajizad E, Do KA, and Long JP (2025). AUPRC: A metric for evaluating the performance of in-silico perturbation methods in identifying differentially expressed genes. Briefings in Bioinformatics, 26(5): bbaf426. https://doi.org/10.1093/bib/bbaf426 **PMid:40889115 PMCid:PMC12400816**