Grenze International Journal of Engineering and Technology, June Issue

Transformative Approaches in Early Cardiovascular Risk Evaluation through AI-Enhanced Modeling Techniques

Vijendra Rai[1], Vikrant[2], Sonam Dubey[3], Subodh Rastogi[4], Neeraj[5] and Arjun Singh[6]
[1]I.T.S Engineering College, CSE, Greater Noida, India
Email: vijendrarail@gmail.com
[2]COER University, Roorkee, AI&ML, Uttarakhand, India
Email: vikrantrana.in@gmail.com
[3]Pranveer Singh Institute of Technology, Data Science, Kanpur, India
Email: sonamdubey728@gmail.com
[4]Moradabad Institute of Technology, MCA, Moradabad, India
Email: subodhrastogimit@gmail.com
[5]Galgotias University, AI, Greater Noida, India
Email: neer3069@gmail.com
[6]Greater Noida Institute of Technology, CSE, Greater Noida, India
Email: innovativearjunsingh@gmail.com

Abstract*— The prevalence of cardiovascular disease (CVD) around the globe is steadily increasing. Thus, there is an increasing need to implement ways to assess and identify Individual Risk for CVD in Proactive ways. To assess this risk, The Behavioral Risk Factor Surveillance System (BRFSS) 2022 data from available Public Health Surveys have been utilized to develop The Artificial Intelligence (AI) Heart Disease Risk Assessment System, in order to better predict the Future Cardiovascular Disease Risk through the identification of Cardiovascular Risk factors early on, in order to improve patient outcomes and delay disease progression. The research program utilizes AI to identify those that are at Risk of developing heart disease, through the use of (but not limited to) supervised machine learning algorithms, (Random Forests, Light Gradient Boosting Machines and Logistic Regression etc.), and UAA (Universal Applicability of Algorithms). The Entire Data Science Pipeline (Data Wrangling, Distribution-Based Imputation, EDA (Exploratory Data Analysis), Advanced Feature Selection & Robustly Developed Machine Learning Models) has, also, been implemented into this study and the proposed system is a good example of how explainable Artificial Intelligence and accessible Digital Platforms can work together to help early Identify Cardiovascular Health Risks and Support Preventive Health Behaviors.[1] Also, to increase prediction accuracy, the framework utilizes Feature Engineering, Distribution-Based Imputation and Thorough Preprocessing of Data. Overall, the results indicated that The Easy Ensemble Classifier (with Light) provides the best and most reliable model for large, real-time screening of cardiovascular risk by providing the most balanced level of Sensitivity and Specificity.

***Keywords*— Cardiovascular risk, artificial intelligence, LightGBM, Easy Ensemble, predictive modeling, model training, SHAP.**

Grenze ID: 01.GIJET.12.2.74_20

I. Introduction

Health diagnostics have advanced more in cardiovascular illness than any other type. Additionally, there is now a need for Smart Data Technology to perform better and more often at risk assessment for cardiovascular disease. Mitigating risk through the early detection of at-risk groups will allow health systems to intervene before a disease occurs, thereby reducing mortality rates, costs of treatment, and health system overload. This research presents an Artificial Intelligence-driven predictive model for cardiovascular risk based on advanced Machine Learning and the availability of population data. The platform integrates data preprocessing, exploratory data. The global burden of cardiovascular disease (CVD) continues to be the leading cause of morbidity and mortality, even after many years of clinical advancements in medicines, diagnostics, and preventive measures. One of the most significant problems for cardiovascular practitioners is that traditional cardiovascular risk assessments (although they have clinical value) typically identify individuals only after the progression of subclinical disease has started. The existing risk scoring systems used (such as the Framingham Risk Score and pooled cohort equations) use a limited number of population-based risk factors that can only establish a linear relationship between risk factor(s) and CVD outcome(s) and, because of this, are at a high risk of underestimating the risk for younger adults, females, and patients that have atypical or emerging risk factors. This gap has resulted in the international healthcare systems to shift towards approaching risk assessment through precision prevention, which focuses on detecting vulnerability to developing CVD at a much earlier time and with much more specificity [2][5].The tremendous advances of artificial intelligence (AI) technology, particularly through machine learning (ML) and deep learning (DL), allow for an innovation in how we assess cardiovascular risk. AI-based risk prediction models have the capability of integrating enormous amounts of heterogeneous data and use multiple forms of data including clinical parameters, laboratory-derived biomarkers, imaging characteristics, genetic and molecular data, longitudinal data obtained from wearable health devices, and social determinants of health. In addition, predictive models based on artificial intelligence are able to account for more complicated associations that result from interactions between nonlinear patterns that are not normally taken into account using traditional statistical methods. Because of this, AI based models are capable of developing risk stratification on an individual basis and identifying patients with a level of detail not previously possible. The goal of this study is to help overcome the barriers associated with predicting cardiovascular disease risk in the early stages of the disease process by utilizing cutting edge technologies based on artificial intelligence (AI) for prediction. In doing so, the researcher hopes to change the way in which cardiovascular disease prevention is viewed, which has typically focused on the estimation of risk based on aggregate data, to utilize the combination of all available data sources in order to create models that will allow for individualized predictions of risk. Ultimately, the intent of this research is to help create a new paradigm for the use of early risk assessment of patients by employing advancements in AI technology for improved outcomes and the prevention of cardiovascular disease.

II. Literature Review

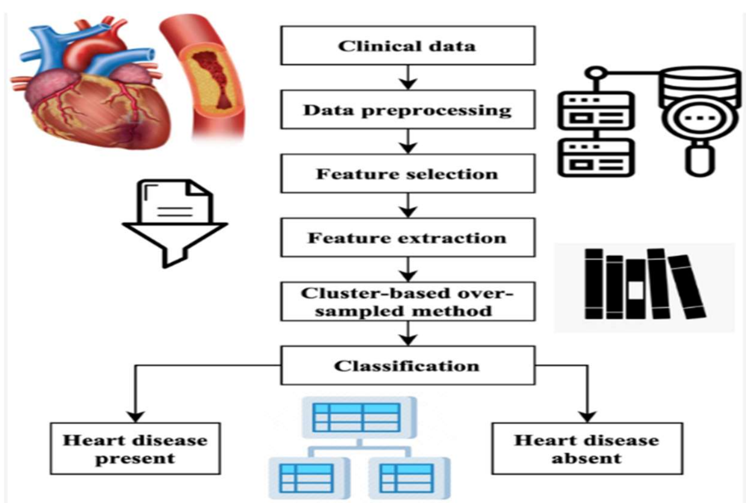

Fig 1. Work Flow Process

Many recent machine learning models (SVMs, Random Forests, Neural Networks) offer greater predictive power than those used in previous studies. Fine et, al. (2016) created a random forest model for electronic health record (EHR) data that predicted coronary artery disease with 83% accuracy. Similarly, Haq et, al. (2017) found that Gradient Boosting algorithms outperformed traditional methods when predicting heart disease from behavioral and physiological traits. Ensemble learning techniques like EasyEnsemble and Balanced Random Forest are particularly well-suited for use with imbalanced datasets, which are common in medical applications; they combine multiple weak learners to improve Recall while minimizing Bias towards the majority class. The BRFSS (Behavioral Risk Factor Surveillance System) provides evidence that large volumes of data on behavior aggregated from multiple demographics can provide a good estimate of Risk. However, the majority of research so far has been limited by a lack of Interpretability, Substandard Feature Engineering, and Limited Generalizability. This paper's goal is to develop an end-to-end pipeline utilizing AI, with a strong focus on Interpretability, balancing Training and Testing datasets, and implementing the pipeline for clinical use. Both LightGBM and EasyEnsemble will be incorporated into the developed pipeline.[6]

III. Dataset Description

The study will focus on collecting and analyzing data related to heart disease as defined in the BRFSS from the CDC. Researchers will be using the BRFSS (2022) database to collect information about more than 400,000 respondents who reported on their health behaviours, chronic illnesses, sociodemographic characteristics, and other healthcare access-related aspects throughout the U.S.A.

A. Demographic and Health Characteristics

The demographic variables are sex, age, race, and employment income.
The health behaviours that contribute to the risk of developing heart disease are: alcohol use, tobacco use, physical activity, and sleep hours.
The chronic illness categories that were identified in this project include diabetes, stroke, hypertension, chronic kidney disease, and mental illness.
Medical histories that indicate potential risk for heart disease are: body mass index (BMI), walking difficulties, and self-rated health.[7]

B. Dependent Variable

Heart disease is a dichotomous outcome variable that indicates whether or not a physician has diagnosed the respondent as having either angina or coronary artery disease.

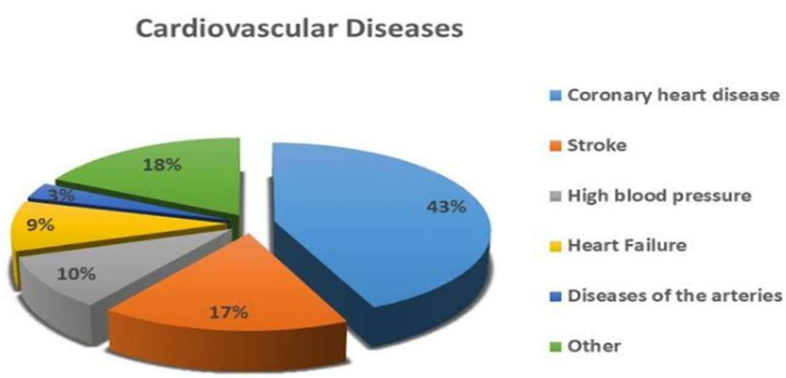

Fig 2. Summary Report for heart disease

IV. Methodologies

In this part, explaining all the algorithms and methodology used to detect early cardiovascular risks related to heart.

- Feature Selection: Features most major for cardiovascular wellness--physical activity, BMI smoking and various diseases--are retained by reviewing correlations and what experts understand about this topic.
- Data Correction: Entries not matching correctly or with extreme values were either eliminated from data or corrected so accuracy would be achieved.

➢ Column Naming: Short codes or unclear variable names were replaced with description types for easier understanding.[8]
➢ Missing Value Treatment: For filling missing values, distribution-based method is applied which gives an opportunity to keep the original categories similar [8].
➢ Model Development
seven machine learning algorithms-
- Logistic Regression
- Random Forest
- XGBoost
- LightGBM
- Balanced Bagging
- Balanced Random Forest
- EasyEnsemble

V. EXPERIMENTAL EVALUATION

We are going to separate the Dataset into 8:2 ratio as 80% for training and 20% for testing. That is important particularly in imbalanced data sets like this one for cardiovascular risk because you don't want the model to only learn from the majority class and ignore the minority. For checking the real performance of the model, we used a number of standard metrics Accuracy, Precision, Recall, F1-Score and ROC-AUC. 9.
Accuracy (A): We can say this is the ratio of total correct predictions to all predictions Although widely used, accuracy by itself is inadequate for imbalanced datasets since it may be deceptive if the model only predicts the majority class.
Precision (P): Quantifies the ratio of true positive cases correctly predicted out of the total predicted positives. High precision implies low false positives.
Recall (R): It defines the ratio of actual positive detection by the model.

A. Ensemble Techniques

Ensemble learning techniques were also investigated to address the class imbalance issue. Ensemble approaches combine many weak learners to obtain better prediction accuracy, lower variance, and a more general model. Two main approaches- Balanced Bagging and Easy Ensemble-were used. The Easy Ensemble coupled with LightGBM as a base estimator was the most promising of them. It attained 81% recall and an ROC-AUC of 0.886, surpassing all other models.
Easy Ensemble successfully learned minority-class patterns by bootstrapping the majority class and training multiple LightGBM classifiers. Bias towards majority class was avoided in this manner and we are still credibly able to generalize learning to unseen test data. Ensemble learning is powerful and important in healthcare, where the impact of false negatives can be severe [9][11].

VI. RESULTS

It is shown that due to class imbalance, conventional machine-learning methods do not perform very effectively on identifying heart disease cases. While all three baseline predictive models (logistic regression, random forest, and standalone gradient boosting) provided good overall accuracy in balanced datasets, they perform poorly in recognizing heart disease in the minority class (false negatives). Solutions for improving the models' predictive capabilities include the use of ensemble methods for managing imbalanced datasets. In particular, combining the Easy Ensemble method with LightGBM as the base learner increased the detection rate for the minority class The ultimate EasyEnsemble–LightGBM hybrid model recorded the following metrics on the test set show in Table 1.

TABLE I. MODEL PERFORMANCE

Metric	Value
Accuracy	79.35%
Recall	80.7%
Precision	20.1%
F1-score	0.32
ROC-AUC	0.885

VII. APPLICATION FRAMEWORK AND DEPLOYMENT

To create the AI healthcare assessment tool, we have built an advanced artificial intelligence method, which provides users with the ability to evaluate their healthcare in a fast, easy and efficient manner through the interpretation of their results. The process of using the AI tool will begin with the following steps.

We will construct the backend data pipeline using Catboost (also known as "Categorical Boosting") for the input data, coding the Categorical variables in Catboost format, and creating an easy ensemble model at each step of the backend pipeline to provide users with real-time results for their patient assessments.Python is by far the most common programming language used by developers to construct a personal healthcare assessment tool, and this is primarily due to the richness of the libraries available for machine learning and data analysis that are compatible with Python, offering developers an enormous array of tools to utilize.[12]

Developers will also build a user-friendly and easy to use web interface, which will allow users to quickly enter their personal and demographic information through a Streamlit application built specifically for this purpose by developers.

VIII. CONCLUSION

An AI-based system for assessing heart disease risk has been designed, developed and used in this study. In order to provide more accurate estimates of risk, the authors combined the strengths of Machine Learning and Explainable Artificial Intelligence techniques with Ensemble Learning methods to analyze large sample sizes of public health data representing diverse populations. The use of Ensemble Learning helped to overcome class imbalance. Therefore, high risk individuals are identified more accurately.The use of SHAP based interpretability helps build trust and use of the system. Trust and usability are particularly important in the area of healthcare. The development of an Interactive Web Application makes access to the system easier and provides an opportunity for end-users to connect the research of data scientists with practical applications. This application also allows users to see their own personal cardiovascular risk. Overall, this research demonstrates the potential impact that AI enabled decision support systems will have on the delivery of preventive medicine and early disease detection.

REFERENCES

[1] World Health Organization, “Cardiovascular Diseases (CVDs) Fact Sheet,” WHO, 2023.

[2] J. Krittanawong et al., “Artificial Intelligence in Precision Cardiovascular Medicine,” Nature Reviews Cardiology, vol. 18, no. 8, pp. 446–457,2021.

[3] G. D. Gupta and A. Mehta, “Heart Disease Prediction Using Machine Learning Algorithms,” IJCA, vol. 181, no. 3, pp. 39–44, 2020.

[4] S. Dey et al., “Machine Learning Approaches for Predicting Coronary Artery Disease,” IEEE Access, vol. 9, pp. 18223–18232, 2021.

[5] A. Haq et al., “Performance Analysis of Different Machine Learning Algorithms for Heart Disease Prediction,” Biomedical Signal Processing and Control, vol. 59, 2020.

[6] H. Liu and R. Zhou, “Balanced Ensemble Learning for Imbalanced Health Datasets,” Pattern Recognition Letters, vol. 139, pp. 52–59, 2020.

[7] Centers for Disease Control and Prevention, “Behavioral Risk Factor Surveillance System (BRFSS) 2022,” CDC, 2023.

[8] S. Saar-Tsechansky and F. Provost, “Handling Missing Values When Applying Classification Models,” Journal of Machine Learning Research, vol. 8, pp. 1625–1657, 2007.

[9] L. Prokhorenkova et al., “CatBoost: Unbiased Boosting with Categorical Features,” NeurIPS, 2018.

[10] T. Akiba et al., “Optuna: A Next-Generation Hyperparameter Optimization Framework,” ACM SIGKDD, 2019.

[11] Data camp. (2023). An Introduction to SHAP Values and Machine Learning Interpretability. Retrieved from https://www.datacamp.com/community/tutorials/introduction-shap-values-machine-learning

[12] Lundberg, S. M., Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. arXiv preprint arXiv:1705.07874. Retrieved from https://arxiv.org/abs/1705.07874

[13] pandas.pydata.org. (2023). Categorical data — pandas 2.2.2 documentation. Retrieved from https://pandas.pydata.org/pandas-docs/stable/userg uide/categorical.html

[14] Saar-Tsechansky, M., Provost, F. (2007). Handling missing values when applying classification models. Journal of Machine Learning Research, 8, 1625.

[15] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, “Can machine learning improve cardiovascular risk prediction using routine clinical data?” PLOS ONE, vol. 12, no. 4, Art. no. e0174944, Apr. 2017.

[16] C. Krittanawong, K. W. Johnson, R. S. Rosenson, Z. Wang, M. Aydar, U. Baber, and S. M. Narayan, “Deep learning for cardiovascular medicine: A practical primer,” European Heart Journal, vol. 38, no. 25, pp. 1837–1847, 2017.