Identifying Health Factors Leveraging Machine Learning to Uncover Stronger Predictors of Cardiovascular Disease: Clinical Metrics vs. Self-Reports

Sandhya Dharshini Sasikumar[1*] Kavithaa Suresh Kumar[2] and Siva Sivatha Sindhu[3]

[1] West Windsor-Plainsboro High School South, Princeton Junction, NJ 08550, USA
[2] Illinois Mathematics and Science Academy, Illinois 60506, USA
[3] Nomura Securities International, NY 10019, USA
sandhya.dh0520@gmail.com

Abstract. Cardio Vascular Disease (CVD) are the major cause of death worldwide; with increasing numbers of individuals in recent days trust on internet-based health quizzes and social media for self-diagnosis. We prove that clinical examination features (blood pressure, cholesterol, glucose levels) will be significantly more predictive of cardiovascular disease than self-reported objective and subjective health information when analyzed using machine learning algorithm. To prove this, we analyzed the publicly available 70,000 clinical records of cardiovascular disease dataset using WEKA's Random Forest implementation. We segregated the dataset into three feature subsets: all features (n=11), examination-only features (n=4), and objective plus subjective features excluding examination data (n=7). Our results supported our hypothesis, showing that examination-only features achieved 72.26% accuracy in predicting CVD, significantly outperforming self-reported data alone (58% accuracy). The complete feature set accomplished the highest accuracy (83%). These findings emphasis that clinical examination features play a crucial role in CVD diagnosis. In contrast, internet-based health assessments that rely solely on self-reported data show clear limitations in predicting cardiovascular disease. Therefore, to reduce risk and support early detection, individuals are strongly advised to have an annual health check every year.

Keywords: Cardio Vascular Disease, Artificial Intelligence, Health Care.

1 Introduction

Cardiovascular disease affects almost 655,000 Americans every year and remains the foremost cause of death worldwide [1]. Despite advances in medical technology, cardiovascular disease deaths are still going up and surprisingly, the rise is most noticeable among younger people. Many of them rely on digital health tools like symptom checkers or social-media quizzes to figure out what might be wrong. However, the diagnostic accuracy of these digital tools remains questionable. Harvard

P. Johri et al. (eds.), *Proceedings of the International Conference on Sustainable Computing and Artificial Intelligence (ICSCAI 2025)*, Advances in Engineering Research 298,
https://doi.org/10.2991/978-94-6239-674-6_23

researchers reported that widely used online symptom checkers correctly identify conditions in only 34% of cases [2]. A systematic review of digital symptom checkers revealed average diagnostic accuracies below 40% with inconsistent triage guidance. This concerning trend raises big question about whether self-reported health information can meaningfully replace clinically obtained information in disease prediction.

Machine learning algorithms, particularly Random Forest classifiers, have shown promising results in many fields [3] including medical diagnosis applications. Previous studies using cardiovascular datasets have achieved higher accuracies, but few have systematically compared the predictive value of different feature categories. Understanding which types of health information are most predictive of cardiovascular disease could inform both clinical practice guidelines and the development of more reliable digital health tools.

The primary aim of this study is to determine if clinical examination features provide superior predictive accuracy for cardiovascular disease compared to self-reported health information that could be obtained through online questionnaires. We proved that clinical examination features (systolic blood pressure, diastolic blood pressure, cholesterol levels, and glucose levels) would demonstrate significantly higher predictive accuracy for cardiovascular disease than objective and subjective self-reported features (age, height, weight, gender, smoking status, alcohol intake, and physical activity level) when analyzed using Random Forest machine learning algorithms.

2 Related Work

Early CVD prediction studies have applied logistic regression, support-vector machines and ensemble trees to small cohorts, typically < 1,000 instances. Ensemble methods, especially Random Forests [4], routinely outperform linear and single-tree models, achieving accuracies up to 95% on balanced research datasets. Recent studies using Random Forest for heart disease prediction report accuracies ranging from 84% to 92%, confirming the algorithm's effectiveness for cardiovascular risk assessment[5][6][7]. Outside clinical settings, online symptom checkers deliver mixed results; Harvard researchers reported correct first-listed diagnoses in only 34% of vignettes, echoed by a BMJ meta-analysis covering 10 commercial platforms. A recent JAMA study examined five popular medical tests promoted on Tik-Tok and Instagram, finding 84% of posts omitted discussion of harms. Data preprocessing remains a pivotal but under-reported step. Normalization [8] mitigates scale disparities, accelerates convergence and can boost model stability. However, its specific effect on tree ensembles for heart-disease data has not been systematically quantified in controlled environments.

3 Methodology

This work addresses two major questions:

- Does augmenting hard-measured clinical examination data materially improve machine-learning prediction of CVD compared with self-reported objective/subjective attributes alone?
- What is the quantitative impact of feature subset selection on Random Forest performance in WEKA environment?

3.1 Dataset Acquisition and Characteristics

We obtained the CVD dataset from Kaggle [9], containing 70k health records with 11 input variables and one output variable (cardiovascular disease existence/non-existence). The dataset originated from clinical examinations conducted and includes both objective measurements and self-reported information. Data quality assessment revealed approximately 8% duplicate records and some physiologically implausible values (negative data points, extreme blood pressure readings). These anomalous entries were corrected and augmented with meaningful values. Min-max normalization was applied to continuous variables (age, height, weight, blood pressure) to scale values to the range, improving algorithm performance and numerical stability.

Min-max normalization provided measurable improvements across all configurations.

- Improved numerical stability by eliminating scale disparities
- Reduced training variance across cross-validation folds
- Enhanced convergence properties of the Random Forest ensemble
- Increased model robustness to outliers and extreme values

3.2 Feature Categorization and Subset Creation

Following the original dataset documentation, we categorized features into three types:

- Objective features (4): age (days), height (cm), weight (kg), gender (categorical)
- Examination features (4): systolic and diastolic blood pressure (ap_hi and ap_lo), cholesterol level (ordinal categories: 1–3, representing normal to well above normal) and glucose level (ordinal categories: same scale as cholesterol)
- Subjective features (3): smoking status (binary), alcohol intake (binary), physical activity (binary)

We created three experimental subsets:

- ALL: Complete feature set (11 variables)
- EXAM: Examination features only (4 variables)
- OBJ+SUBJ: Objective and subjective features, excluding examination data (7 variables)

3.3 Machine Learning Implementation

We utilized WEKA 3.8.6 machine learning workbench implemented in Java for all analyses. Random Forest algorithm configuration: Trees: 100 (ensemble size) and Random seed: 1 (for reproducibility). We loaded the preprocessed heart disease dataset in .csv format with target class as discrete value into WEKA tree models which is written in Java. WEKA provides set of hyper parameters that allow us to fine tune our model. As selecting the appropriate hyper parameter is vital in optimizing the performance of the model. The first crucial parameter we picked is the size of the forest. Generally more trees can improve performance but with increased computational time. So we started with number of trees with 50 and increased till 150 by tens. We noticed that our model performed well for 100 trees. Next, parameter we considered for tuning is the depth of the tree, as deeper trees can learn complex patterns but sometimes may result in over fitting. So we tried different experiments with value between 5 and 20 and obtained good results between 12 and 15 as the depth. The number of feature that is considered for each split is decided based on split quality measured using information gain.

3.4 Design and Validation

To achieve a stable and unbiased performance estimate, we adopted a 10-fold stratified cross-validation approach. The dataset was randomly separated into ten equal parts while preserving class balance. In each run, the model was trained on nine parts and tested on the remaining one, the process was repeated until all parts had been used for testing.

For each feature subset (ALL, EXAM, OBJ+SUBJ), we recorded: Overall accuracy and error rates, Kappa statistic (inter-rater agreement measure), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), Receiver Operating Characteristic (ROC) area, Class-specific precision, recall, and F1-scores, and Training time.

3.5 Statistical Analysis

We compared performance using different feature subsets and assessed their significance based on accuracy gains. The Kappa statistic provided insight into classification reliability beyond simple accuracy measures. ROC area analysis evaluated discriminative capability across different classification thresholds. MAE calculates how far predictions are from the actual values on average by taking the absolute difference for each data point and then averaging them.

RMSE measures prediction accuracy by squaring each error, computing their mean, and finally applying a square root, which gives more weight to larger mistakes. RAE evaluates model performance relative to a simple reference model that predicts the average of the observed values. Likewise, RSE assesses error by normalizing the model's total squared deviation against the squared deviation of the reference predictor. The results of these measures for various models are provided in Table 1. All experiments were conducted on standardized hardware to ensure fair computational comparisons. Results were verified through multiple independent runs with consistent random seeds to confirm reproducibility.

Table 1. Error rate comparison across various feature sets

4 Result Analysis

Feature Configuration	MAE	RMSE	RAE (%)	RRSE (%)
ALL Features	0.23	0.31	56.77	68.17
Examination Only	0.375	0.436	74.99	87.25
Objective + Subjective	0.428	0.527	95.05	111.04

4.1 Random Forest Performance Across Feature Subsets

Our primary hypothesis was tested by comparing Random Forest accuracy across three feature configurations using 10-fold cross-validation in WEKA. Table 2 presents the comprehensive results supporting our hypothesis. The results strongly supported our hypothesis. Examination-only features achieved 72.26% accuracy, substantially outperforming objective and subjective features (58% accuracy) by 14.26 percentage points.

Table 2. Random forest performance evaluation across various feature set

Feature Configuration	Features (n)	Accuracy (%)	Kappa Statistic	ROC Area	Training Time (sec)
ALL Features	11	83	0.66	0.897	25.5
Examination Only	4	72.26	0.445	0.77	21.31
Objective + Subjective	7	58	0.064	0.576	22

4.2 Statistical Analysis and Model Reliability

The Kappa statistic values were evaluated to further asses our findings. The eexamination-only features achieved reasonable agreement with the value of 0.445, while objective and subjective features showed minimal agreement with value of 0.064, indicating near-random performance. Also, ROC area decreased from 0.770 for examination features to 0.576 for self-reported data, showing reduced discriminative capability to distinguish between classes.

4.3 Class-Specific Performance Analysis

The performance analysis of class-specific metrics revealed that examination only features maintained unbiased performance across both CVD presence and absence predictions. Sensitivity was higher for examination-derived features across both CVD

presence/absence. For identifying CVD absence, examination features achieved 79.1% sensitivity compared with 68.2% for self-reported data. The disparity even increased for CVD presence detection, where examination features reached 65.4% sensitivity, whereas self-reported data yielded only 38.2%. The performance visualization using ROC reveals distinct patterns across different feature subsets. The eexamination-only features show robust discriminative power with ROC Area of 0.770, while objective and subjective features show poor discrimination with ROC Area of 0.576.

- **CVD Absence Detection**: Examination features achieve 79.1% sensitivity vs. 68.2% for self-reported data
- **CVD Presence Detection**: Examination features achieve 65.4% sensitivity vs. 38.2% for self-reported data
- **False Positive Control**: Examination features maintain 20.9% FPR vs. 31.8% for self-reported data

4.4 Impact of Feature Normalization

The min-max normalization was applied across all feature configurations and we observed consistent improvements on all models. This preprocessing step enhanced numerical stability and reduced variance across cross-validation folds, contributing to the model overall accuracy.

5 Discussion and Inference

The results obtained strongly support our hypothesis that clinical examination features are much better at predicting cardiovascular disease than self-reported health information. In particular, features derived from physical examinations were 14.26% more accurate than self-reported data. This has important implications for both regular health checkup and the design of digital health tools. The better performance of examination-based features highlights the significance of current clinical guidelines that recommend regular physical check-ups. Also, these results indicates measures such as blood pressure, cholesterol, and glucose levels reflect physiological conditions that patients often cannot report accurately themselves. Further, tthese objective readings provide insights into cardiovascular health that subjective symptoms may miss, especially in early stages of disease when patients may not notice any warning signs.

5.1 Machine Learning Algorithm Performance

Random Forest demonstrated better performance on the complete feature set with 83% accuracy which is consistent with previous cardiovascular prediction studies. The algorithm's ability to handle various data types and provide feature importance rankings made it particularly suitable for our comparative analysis. The 10-fold cross-validation results provide confidence in the generalizability of our findings.

5.2 Study Limitations and Future Directions

This study was based on a single dataset provided by kaggle from health screenings and limits generalizability to other populations. Future research should test these findings across different demographic groups and healthcare systems. We only evaluated Random Forest models, so comparing other machine learning approaches could provide additional insights. Using a binary classification simplified the analysis, but it does not fully capture the nuances of cardiovascular risk assessment in clinical practice. Future work could consider multi-class risk levels and incorporate temporal data to track disease progression over time..

These findings suggest that regulatory frameworks for digital health tools should require integration of objective clinical measurements where possible. Symptom checker platforms might benefit from incorporating peripheral devices (smartphone-connected blood pressure monitors, glucometers) or implementing risk stratification algorithms that direct users to professional medical evaluation when self-reported data indicates potential cardiovascular risk.

6 Conclusions

This machine learning based study quantifies the incremental diagnostic power of objective clinical examination variables over self-reported data for CVD prediction. Random Forest achieved 83% accuracy with all features, 72.26% with examination-only data, and 58% with objective + subjective attributes. The results provide evidence-based justification that annual physical examinations provide irreplaceable diagnostic information and that symptom-checker application should not substitute for professional medical assessment.

Acknowledgments. We thank the original dataset contributors for making the Cardiovascular Disease dataset publicly available through Kaggle. We acknowledge the WEKA development team for providing robust machine learning infrastructure that enabled this comparative analysis.

References

1. “Cardio Vascular Diseases.” World Health Organization. https://www.who.int/health-topics/cardiovascular-diseases.
2. https://news.harvard.edu/gazette/story/2015/07/self-diagnosis-on-internet-not-good-practice/
3. Siva S. Sivatha Sindhu, et al. “Decision tree based light weight intrusion detection using a wrapper approach,” Springer Expert Systems with Applications, vol.39, 2012, pp.129–141. doi:10.1016/j.eswa.2011.06.013.
4. Breiman, L. “Random Forests,” ACM Machine Learning, vol.45, 2001, pp.5–32. https://doi.org/10.1023/A:1010933404324
5. Xi Su, et al.” Prediction for cardiovascular diseases based on laboratory data: An analysis of random forest model,” Journal of Clinical Laboratory Analysis, vol 34, 2020. doi: 10.1002/jcla.23421.
6. Taher M. Ghazal, et al. “Heart Disease Prediction Using Machine Learning,” IEEE International Conference on Business Analytics for Technology and Security, 2023. doi: 10.1109/ICBATS57792.2023.10111368.

7. Ke Yan, et al. “Machine Learning for AI-Enhanced Healthcare and Medical Services: New Development and Promising Solution,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol.18, pp. 850 – 851, 2021. doi: 10.1109/TCBB.2021.3050935
8. https://www.deepchecks.com/question/why-is-data-normalization-necessary-for-machine-learning-models/
9. “Cardio Vascular Diseases Data Set,” Kaggle. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset?resource=download Accessed July 2025.
10. “Waikato environment for knowledge analysis,” WEKA version 3.8.6. Available on: https://ml.cms.waikato.ac.nz/weka/ Accessed July 2025.