Machine Learning Models for Cardiovascular Disease: A Holistic Review on Early Diagnosis Performance

Anirudh Chintaluri, Thomas Jefferson High School for Science and Technology
Matthew Radzihovsky, Massachusetts Institute of Technology

ABSTRACT

Cardiovascular disease (CVD) poses a significant threat to global health, responsible for approximately one-third of all deaths worldwide. Early diagnosis plays a crucial role in preventing severe complications, yet traditional methods often require expensive tests and equipment. Machine learning offers an innovative and more cost-efficient alternative, through pattern recognition of more lightweight data to accurately predict CVD presence and severity. In this paper, we look past traditional binary approaches to CVD diagnosis and look towards quantitative ordinal labels, where a higher number represents more severity. Using the UC Irvine Heart Disease dataset, we test six lightweight models — Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), Support Vector Machines (SVC), and Decision Trees (DT) — alongside a custom neural network model to evaluate their performance on this multi-class classification task. Each model's accuracy, precision, recall, and F1 score are analyzed to assess robustness in clinical settings. The RF, KNN, and custom models achieved testing accuracies above 86%, with the KNN leading in accuracy, precision, and recall, going beyond the baseline LR model's accuracy of 52%. Our findings demonstrate that lightweight models are highly effective for CVD prediction, offering potential for future deployment in AI-assisted tasks in hospitals. However, these models lack explainability behind diagnosis, which is essential for real-world deployment. Future directions include exploring wearable technology for continuous CVD monitoring and the inclusion of worldwide data to account for diverse risk factors.

KEYWORDS:

Cardiovascular disease, Machine learning, Random Forest, K-Nearest Neighbors, Skip connections, Deep learning

INTRODUCTION

The heart, one of the most important and essential organs in the human body, pumps blood out and into the rest of the body to provide oxygen necessary for the body to function and stay alive. Because the heart is such an integral part of the human body, any disease that affects the cardiovascular system, commonly referred to a cardiovascular disease (CVD), puts one's life at risk, with complications resulting in long-term health conditions, as well as fatality (3).

In the year 2022, according to the U.S. Centers for Disease Control and Prevention, over 700,000 deaths have been reported in the United States alone with one of many CVDs as the cause, including coronary artery disease, arrhythmia, and cardiac arrests (3). Furthermore, for the past 30 years, cases of CVD have been on the rise worldwide (1). Overall, this type of disease is responsible for as much as one-third of the world's deaths. Not only that, but when the COVID-19 pandemic spread worldwide in 2020 and 2021, CVD was identified as one of the biggest risk factors for severe complications from contracting the virus (1).

Historically, early detection of CVD has played a key role in preventing severe complications, and has therefore resulted in more cost-effective healthcare to treat it (10). However, detecting CVD itself can require the use of a multitude of equipment, tests, and analyses to properly diagnose one with CVD, with tests such as echocardiograms, stress tests, and coronary angiograms having median costs of up to $2588, $3230, and $9203 respectively in 2022 (9).

To provide for a more cost-effective solution for patients, machine learning can be utilized to both distinguish and learn from patterns in CVD data. Additionally, machine learning models are capable of detecting CVD at an early stage in particular (7).

Many previous studies have proposed machine learning techniques to accomplish this task of early detection. Some techniques used by studies in the past include support vector machines (SVC), multilayer perceptrons (MLP), decision trees (DT), and regression models such as logistic regression (LR) (7). Pal et al. (2022) have used machine learning model types such as MLP and K-Nearest neighbor classifiers (KNN) to predict binary classification of CVD using data from the UC Irvine machine learning repository, with the MLP model achieving an accuracy score of 82.47% (11). Korial et al. (2024), while also using the same dataset, focused on an ensemble-based approach using models such as LR, KNN, RF, and a voting ensemble model for binary classification, and achieved an accuracy of 92.11% on their voting ensemble model (6).

Unlike many existing studies that focus primarily on binary classification, this paper focuses primarily on five different classifications depending on CVD severity, and our aim is to more holistically review many different types of supervised learning models, from ensemble models to deep learning models — with the latter including a custom-built neural network model that utilizes the skip connection as data is fed into the network — using a variety of metrics, such as accuracy, F1 score, precision, and recall. The purpose of the diversity of metrics is to provide

an idea of model robustness and relevant performance metrics pertaining to a clinical task such as CVD detection, as well as to further give patients a better idea of what factors can result in CVD, and how it can be made more or less severe. All in all, we hypothesize that machine learning models can effectively move beyond binary classification approaches, and accurately predict both presence and severity of CVD in patients.

METHODS

1. **Dataset Usage:** To obtain data for training, we looked to the Heart Disease dataset available to the public in the UC Irvine Machine Learning repository. The dataset contains 14 features pertaining to a patient and labels the presence of heart disease, with 303 instances. Its attributes consist of factors that either do not require tests or can be much more easily attained and affordable than current CVD tests, such as age, sex, blood pressure, cholesterol, and electrocardiogram (ECG) results (4). The attributes consist of the following as labeled in Table 1.

Table 1

List of features from UC Irvine's heart disease dataset.

Feature	Description
age	Quantitative measurement of the patient's age in years.
sex	Binary (0 = Female, 1 = Male).
cp	Chest pain type (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
trestbps	Resting blood pressure, in mmHg.
chol	Serum cholesterol, in mg/dl.
fbs	Binary, whether or not fasting blood sugar > 120 mg/dl (0-1).
restecg	Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy).
thalach	Maximum heart rate achieved, in beats/min.
exang	Exercise induced angina (0 = No, 1 = Yes).
oldpeak	ST depression induced by exercise relative to rest.
slope	The slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping).
ca	Number of major vessels (0-3) colored by fluoroscopy.
thal	3 = normal; 6 = fixed defect; 7 = reversible defect.
label	Categorical, ordinal (0-4). Signifies presence of heart disease of patient, with 0 indicating absence and 4 indicating most severe presence.

2. **Preprocessing:** All code that was done to preprocess and train our machine learning models was built on a Jupyter Notebook with Python 3.10. This was done to better organize our code into smaller, individual cells for each step of model development. Indeed, the dataset has missing values, and this issue was addressed using a K-Nearest-Neighbors (KNN) approach to determine which missing value best aligned with the non-empty data. Through hyperparameter tuning, we were able to determine which values are to be filled in depending on the KNN model with the highest accuracy with k=10 for **ca** and k=5 for **thal**.

   Furthermore, we normalized our data into either a 0-1 range or a z-score, depending on whether each feature was quantitative or not. Machine learning models more easily learn and converge when data is normalized within a specific range, due to the use of smaller numbers while also maintaining the distinction between categories and numbers. As such, normalization avoids bias of a particular feature, while also improving model outcomes when normalized by z-score and a 0-1 range (12).

   Lastly, we had to mitigate the issue of class distribution, as the negative case had about as many values as each of the four positive cases combined. To balance out the class distribution, we turned to the Synthetic Minority Over-Sampling Technique (SMOTE), an oversampling technique that synthetically creates data points from minority classes to match the number of samples from the majority class. As a result, SMOTE reduces bias towards the majority class and places more even emphasis on the minority CVD classes (2).

   Tables 2 and 3 show the first 5 data points of the set before and after preprocessing the data respectively. For reference, “true-label” refers to the **label** category in Table 1, while “binary-label” refers to the binary presence of heart disease, with 0 indicating absence and 1 indicating presence.

Table 2

List of the first 5 data points of the dataset before preprocessing.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	true-label	binary-label
0	63	1	1	145	233	1	2	150	0	2.3	3	0.0	6.0	0	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3.0	3.0	2	1
2	67	1	4	120	229	0	2	129	1	2.6	2	2.0	7.0	1	1
3	37	1	3	130	250	0	0	187	0	3.5	3	0.0	3.0	0	0
4	41	0	2	130	204	0	2	172	0	1.4	1	0.0	3.0	0	0

Table 3

List of the first 5 data points of the dataset after preprocessing.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	true-label	binary-label
0	0.947160	1	0.25	0.756274	-0.264463	1	1.0	0.017169	0	1.085542	1.000000	0.000000	0.857143	0	0
1	1.389703	1	1.00	1.608559	0.759159	0	1.0	-1.818896	1	0.396526	0.666667	1.000000	0.428571	2	1
2	1.389703	1	1.00	-0.664201	-0.341717	0	1.0	-0.900864	1	1.343924	0.666667	0.666667	1.000000	1	1
3	-1.929372	1	0.75	-0.096011	0.063869	0	0.0	1.634655	0	2.119067	1.000000	0.000000	0.428571	0	0
4	-1.486829	0	0.50	-0.096011	-0.824558	0	1.0	0.978917	0	0.310399	0.333333	0.000000	0.428571	0	0

3. **Model development:** To analyze our model performance, the dataset was split into three categories: 70% of the data for training, 15% for validation, and 15% for testing, with the training set intended to train, the validation data intended for hyperparameter tuning, and the testing dataset for an unbiased evaluation of model performance.

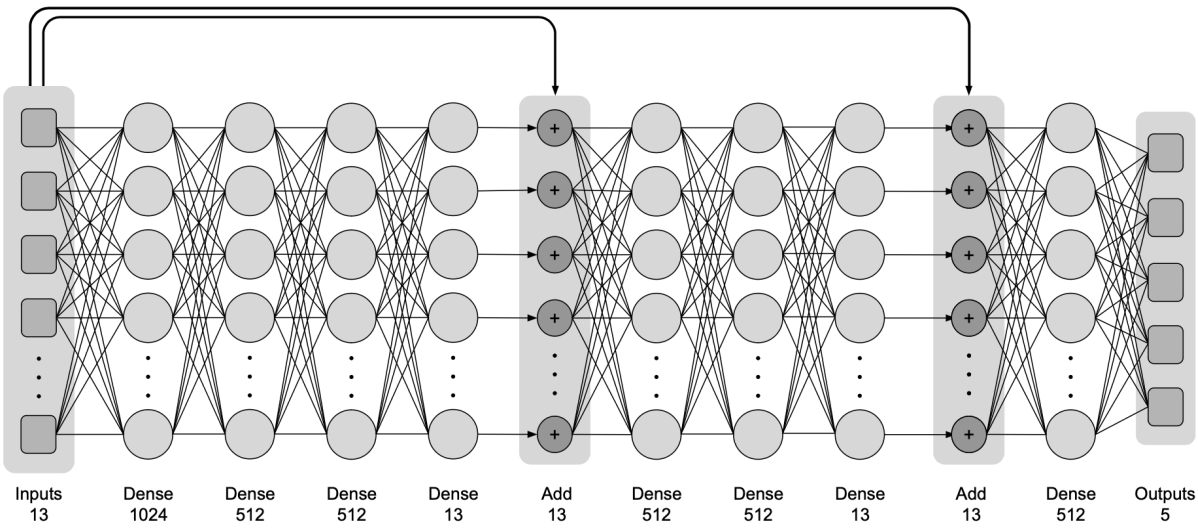
   The models used in this study include six different lightweight models in addition to a custom deep learning model. The six lightweight models are readily available in the scikit-learn library, and include the following: Logistic Regression (LR), Random Forest (RF), K-Nearest-Neighbors (KNN), Multilayer Perceptron (MLP), Support Vector Machines Classifier (SVC), and Decision Trees (DT).

   The custom neural network model used in this study was built using Keras, made with 12 layers, 10 hidden layers, and 2 skip connections, for a total of 1,094,687 weights and biases. Skip connections in neural network models have been shown to improve performance by eliminating singularities, where node overlap and elimination can cause the model to become more degenerate (8). The model was trained using the Adam optimizer, the sparse categorical cross-entropy loss function, and with a batch size of 16 for a total of 400 epochs for training. Additionally, a callback was created in order to

automatically save the best performing model during training into a .keras file. This performance callback depended on the performance of the validation dataset. Figure 1 displays the layout of the neural network.

Figure 1

Layout of the custom neural network model tested.



4. **Data analysis:** After testing the model, we compared the models on several different metrics, as accuracy does not tell the entire story of how well a model performs. It is important, especially in clinical applications of machine learning, to minimize false negatives — that is, when a model predicts a negative value when it is in fact not. As such, we also compared our custom model with the other six lightweight models with metrics such as F1 score, recall, and precision, and averaged the scores across five trials of random train-test-splits. The recall and precision metrics used in this study were calculated through macro-averaging to take into account all five **label** categories.

RESULTS

Table 4 presents the performance of the seven different machine learning models tested on the UC Irvine Heart Disease Dataset for the testing set. The top performing models in this test are the custom Keras-built model, the RF, and the KNN, achieving a testing accuracy of 0.8634, 0.8748, and 0.8764 respectively. The KNN once again got the highest precision score with 0.8822 on average, followed closely by the RF and custom model, with scores of 0.8814 and

0.8711 respectively. For recall, the KNN performs best with a score of 0.8801, followed by scores of 0.8758 and 0.8645 by the RF and custom model respectively. The high performance of these models in precision, recall, and F1 score further corroborate the high accuracy of these models, receiving scores very similar to the testing accuracy. All models showed precision, recall, and F1 scores within a 4-point margin of the accuracy score, with the DT model showing the most variability across all metrics, as shown in Figure 2. Of the three highest-performing models, though, the custom model showed the most consistent output.

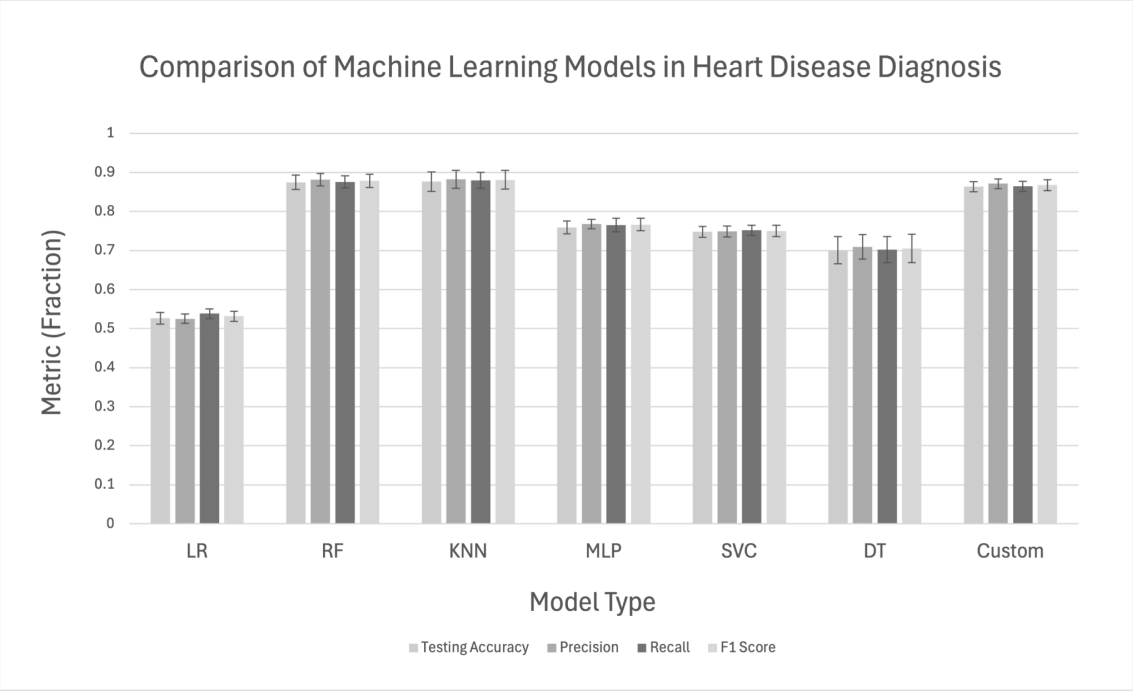
Table 4

Comparison of seven different machine learning models in average testing accuracy, precision, recall, and F1 score.

Model Type	Testing Accuracy	Precision	Recall	F1 Score
LR	0.526829268	0.525270432	0.538237895	0.531612338
RF	0.874796748	0.881410044	0.875822996	0.87859095
KNN	**0.876422764**	**0.882239259**	**0.880072836**	**0.881124205**
MLP	0.759349593	0.76767719	0.765445404	0.766434885
SVC	0.74796748	0.748949675	0.752073069	0.750497878
DT	0.700813008	0.709213101	0.702123746	0.705622519
Custom	0.863414633	0.871186808	0.864527283	0.867782041

Figure 2

A bar chart visualizing the difference in model performance through testing accuracy, precision, recall, and F1 score.



Note: Error bars indicate standard error.

DISCUSSION

The purpose of this study was to provide a holistic review of machine learning efficacy in identifying risk factors of and accurately diagnosing cardiovascular disease in patients, marking a large leap forward. The results of this study demonstrate the potential of machine learning models in accurately classifying the severity of CVD in patients. For the RF, KNN, and custom models in particular — with accuracy scores above 86% for each — the results pose a significant improvement over the baseline LR model, which saw an average testing accuracy score of 53.78%, as well as the benchmark standard of 20% for a random model that predicts one of five categories. This suggests that overall, lightweight models such as RF and KNN are generally some of the most effective models in this type of medical diagnosis. Additionally, the inclusion of other metrics, including precision, recall, and F1 score makes way for a more holistic review of what factors are especially emphasized in machine learning approaches to clinical tasks, when minimizing false negatives are especially heavily emphasized.

Our findings align with prior research in that machine learning can be effective and accurate in detecting cardiovascular disease. However, our approach moves beyond the binary classification used in prior studies, and instead focuses on the spectrum of severity, marking a major step forward for health professionals and doctors who wish to understand a more detailed review of the patient's condition. As a result, this type of model not only determines if the patient has CVD or not, but it also provides doctors a better idea of what medications should be prescribed to treat the patient's CVD specific to the patient's condition.

Despite the robustness and high accuracy of these models, they are not ready for full self-deployment in hospitals. For a clinical-based machine learning model to truly provide useful information about a patient's medical diagnosis, it must be able to reason and explain why that patient has CVD (5). It is quite unreasonable for a deployed model to not give an explanation as to what exactly led to the positive test of CVD, and in more specific scenarios, the severity of the CVD. At this point in time, models like this should instead serve as an AI-assisted tool to avoid bias and provide an objective prediction rather than a fully independent diagnosis tool.

In the future, a direction worth exploring is how these models can be incorporated in wearable technology. With existing equipment in hospitals, a medical diagnosis still requires high fees, not to mention the already high costs of healthcare and insurance. As wearable technology such as smart watches becomes more prevalent in our society, there is potential for what is already an active health monitor to monitor risk of mild or even severe cardiovascular disease. Furthermore, with active monitoring of a patient's data on a smartwatch, it may be even more capable of responding to early signs of cardiovascular disease, preventing further complications and severe symptoms from emerging in the patient, and these can be lifesaving.

Another future direction includes obtaining data from patients all over the world. In this study, we discuss relationships among several factors in Cleveland alone; however, there are also other factors not mentioned in this study that are key to cardiovascular disease diagnoses in other foreign countries, such as air quality and immunity. This can provide us with an even more holistic review of the most important risk factors that people should watch out for to avoid said CVD symptoms.

CONCLUSION

Overall, this study demonstrates that machine learning can be highly effective at both saving costs of expensive tests, and saving money with its ability to detect CVD early. We have developed models that expand on current methods and predict based on severity of CVD as opposed to simply whether or not CVD is present, and noticed that lightweight models such as KNN and RF, as well as our deep custom neural network model, stand out with that regard, producing both accurate and precise results which are particularly useful within the context of clinical diagnosis. Though these models may not be ready for standalone real-world deployment, they certainly mark a step forward with machine learning technology and can be used for AI-assisted CVD diagnosis by a doctor.

REPRODUCIBILITY

All code used for the development of our machine learning models, including steps towards building them, is publicly available in the following GitHub repository: https://github.com/anirudhc5/cvd-machine-learning-notebook. The dataset used for building this model is available to the public on the UC Irvine Machine Learning Repository: https://doi.org/10.24432/C52P4X.

ABBREVIATIONS

CVD: Cardiovascular Disease; LR: Logistic Regression; RF: Random Forest; KNN: K-Nearest Neighbors; MLP: Multilayer Perceptron; SVC: Support Vector Machines; DT: Decision Trees

REFERENCES

1. "Cardiovascular Disease Is on the Rise, but We Know How to Curb It. We've Done It before." *National Heart, Lung, and Blood Institute*, 3 Feb. 2021, www.nhlbi.nih.gov/news/2021/cardiovascular-disease-rise-we-know-how-curb-it-weve-done-it#:~:text=Over%20the%20last%2030%20years,across%20the%20globe%20external%20link%20. Accessed 15 Sept. 2024.
2. Chawla, N. V., et al. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research*, vol. 16, 1 June 2002, pp. 321-57, https://doi.org/10.1613/jair.953. Accessed 15 Sept. 2024.
3. "Heart Disease." *U.S. Centers for Disease Control and Prevention*, 15 may 2024, www.cdc.gov/heart-disease/about/index.html#:~:text=What%20is%20heart%20disease%3F,can%20cause%20a%20heart%20attack. Accessed 15 Sept. 2024.
4. Janosi, Andras, et al. "Heart Disease." UCI Machine Learning Repository, 1989, https://doi.org/10.24432/C52P4X. Accessed 15 September 2024.
5. Kononenko, Igor. "Machine Learning for Medical Diagnosis: History, State of the Art and Perspective." *Artificial Intelligence in Medicine*, vol. 23, no. 1, Aug. 2001, pp. 89-109, https://doi.org/10.1016/s0933-3657(01)00077-x. Accessed 15 Sept. 2024.
6. Korial, Ayad E., et al. "An Improved Ensemble-Based Cardiovascular Disease Detection System with Chi-Square Feature Selection." *Computers*, vol. 13, no. 6, 22 May 2024, p. 126, https://doi.org/10.3390/computers13060126. Accessed 15 Sept. 2024.
7. Latha, C. Beulah Christalin, and S. Carolin Jeeva. "Improving the Accuracy of Prediction of Heart Disease Risk Based on Ensemble Classification Techniques." *Informatics in*

Medicine Unlocked, vol. 16, 2019, https://doi.org/10.1016/j.imu.2019.100203. Accessed 15 Sept. 2024.

8. Orhan, A. Emin, and Xaq Pitkow. "Skip Connections Eliminate Singularities." arXiv preprint arXiv:1701.09175 (2017). Accessed 22 Sept. 2024.
9. Oseran, Andrew S et al. “Assessment of Prices for Cardiovascular Tests and Procedures at Top-Ranked US Hospitals.” JAMA internal medicine vol. 182,9 (2022): 996-999. doi:10.1001/jamainternmed.2022.2602. Accessed 22 September 2024.
10. Oude Wolcherink, Martijn J et al. “Health Economic Research Assessing the Value of Early Detection of Cardiovascular Disease: A Systematic Review.” PharmacoEconomics vol. 41,10 (2023): 1183-1203. doi:10.1007/s40273-023-01287-2. Accessed 15 Sept. 2024.
11. Pal, Madhumita, Parija, Smita, Panda, Ganapati, Dhama, Kuldeep and Mohapatra, Ranjan K.. "Risk prediction of cardiovascular disease using machine learning classifiers" Open Medicine, vol. 17, no. 1, 2022, pp. 1100-1113. https://doi.org/10.1515/med-2022-0508. Accessed 15 Sept. 2024.
12. Singh, Dalwinder, and Birmohan Singh. "Investigating the Impact of Data Normalization on Classification Performance." *Applied Soft Computing*, vol. 97, Dec. 2020, https://doi.org/10.1016/j.asoc.2019.105524. Accessed 15 Sept. 2024.