Artificial Intelligence Applications And Innovations Vol. 1 No. 1 (2024)

Article history:
Received 09 October 2023
Revised 11 December 2023
Accepted 21 December 2023
Published online 01 January 2024

Intelligent Prediction of Cardiovascular Disease Mortality Using Machine Learning Techniques

Soha. Parto[1], Ali Akbar. Safavi[1*], Shiva. Naghsh[1], Mahsa. Keikha[2], Amir. Sharafkhaneh[3]

[1] School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran
[2] Department of Mechatronic Systems Engineering, Simon Fraser University, Burnaby, British Columbia, Canada
[3] Department of Medicine, Baylor College of Medicine, Houston, TX, USA

*** Corresponding author email address**: safavi@shirazu.ac.ir

A r t i c l e I n f o

Article type:
Original Research

How to cite this article:
Parto, S., Safavi, A.A., Naghsh, S., Keikha, M., & Sharafkhaneh, A. (2024). Intelligent Prediction of Cardiovascular Disease Mortality Using Machine Learning Techniques. Artificial Intelligence Applications and Innovations, *1*(1), 78-86.
https://doi.org/10.61838/jaiai.1.1.6

A B S T R A C T

This study focuses on predicting cardiovascular disease (CVD) mortality using various machine learning (ML) techniques. A diverse set of parameters from different categories within the Sleep Heart Health Study (SHHS) dataset is leveraged, and ML techniques including LR, KNN, SVM, RF, ETC, and SGD, are employed. To ensure the reliability of these techniques, 10-fold cross-validation is applied. Furthermore, the mutual information technique with K-fold stratified cross-validation is used to determine feature importance, enhancing the model's interpretability. The proposed approach predicts CVD mortality over a 10 to 15-year period and aims to identify influential parameters to facilitate timely interventions and lifestyle improvements for patients, ultimately contributing to an increased lifespan. Among the algorithms, KNN outperforms others, achieving an accuracy of 77%, an F1-score of 77%, an AUC of 79%, a sensitivity of 77.34%, and a specificity of 76.56%.

Keywords: *CVD mortality, ML, SHHS, Cross-Validation.*

1. Introduction

According to the World Health Organization (WHO), cardiovascular diseases (CVDs) are the leading cause of death worldwide. These conditions include various disorders that affect the heart and blood vessels, such as coronary artery disease (CAD), stroke, peripheral artery disease (PAD), rheumatic heart disease (RHM), congenital heart defects, and heart failure (HF). In addition to being the primary cause of mortality, CVDs impose a significant financial burden on healthcare systems globally due to the high costs associated with treatment and long-term care [1]. Therefore, CVD diagnosis and prediction as well as CVD

mortality prediction is necessary for timely intervention and medical work-up.

Monitoring key health parameters can provide valuable insights into predicting CVD mortality. Various studies have investigated parameters contributing to disease progression and mortality prediction. These factors include smoking [2], diabetes [3], demographic characteristics [4], CVD history [5], hypertension [6], the use of medications such as aspirin [7], and sleep-related parameters like sleep efficiency and duration [8, 9].

Given the advancements in artificial intelligence (AI) for problem-solving, there have been significant developments in its application. This is especially true in the medical field for disease prediction and diagnosis. Predicting CVD mortality is crucial, as a result, many studies have employed AI, particularly ML, for predicting disease and mortality. The following sections highlight some of the related works on CVD mortality studies.

Martin-Morales et al. (2023) aimed to identify risk factors for CVD mortality using the National Health and Nutrition Examination Survey (NHANES) dataset. They developed three models based on dietary data, non-diet-related health data, and a combination of both. The ML models, especially Random Forest (RF), consistently predicted CVD mortality across these categories. Shapley Additive Explanations (SHAP) values highlighted age, systolic blood pressure, and various health factors, while fiber, calcium, and vitamin E were significant nutritional variables. Their findings underscore the importance of integrating health and dietary data to improve CVD mortality predictions [10].

Li et al. (2022) aimed to assess the 10-year cardiovascular disease (CVD) mortality risk in individuals with obstructive sleep apnea (OSA) using an ML approach. They analyzed data from 2,464 patients from the Sleep Heart Health Study (SHHS), identifying the top 9 predictive features through mutual information analysis. A random forest model was then developed and evaluated on a test set of 493 patients, achieving an area under the receiver operating curve (AUC) of 0.84, with sensitivity and specificity of 81.82% and 73.94%, respectively. The study found that individuals over 63 years old and those with severe OSA were at higher risk of CVD mortality. Their findings suggest that the random forest model could provide a quick and informative assessment of future CVD mortality risk in OSA patients [5].

Sajeev et al. (2021) aimed to enhance the prediction of CVD mortality risk in the Australian population using ML models and compared their performance to the traditional Framingham model. The dataset was derived from three Australian cohort studies: the North West Adelaide Health Study (NWAHS), the Australian Diabetes, Obesity, and Lifestyle Study (AusDiab), and the Melbourne Collaborative Cohort Study (MCCS). The research involved developing four ML models to predict 15-year CVD mortality risk, with these models achieving a 2.7% to 5.2% improvement in prediction accuracy compared to the Framingham model. In the aggregated cohort, the ML models achieved an area under the curve (AUC) of 0.852, representing a 5.1% enhancement in prediction accuracy. Additionally, the models demonstrated a net reclassification improvement of up to 26%, with better performance observed when stratified by sex and diabetes status [3].

The SHHS dataset is a rich resource and has been widely used in numerous studies. It includes a broad range of parameters, such as health metrics, demographic information, and other personal and health-related data. We will now review three studies that have used this dataset to predict CVD.

Park et al. (2021) predicted CVD within ten years in patients with sleep-disordered breathing (SDB) using an ML algorithm. The model was trained on data from the SHHS, incorporating ECG features, clinical risk factors, and AI-based features. Feature selection was performed using statistical analysis and SVM-RFE. The SVM model effectively predicted CVD, coronary heart disease (CHD), HF, and stroke, achieving high recall and precision, particularly in distinguishing CVD-free cases. The results demonstrated the model's potential in predicting CVD development in SDB patients [11].

Zhang et al. (2020) investigated the link between sleep heart rate variability (HRV) and long-term CVD outcomes, aiming to improve automatic CVD prediction. They analyzed PSG data from 2111 participants over a median follow-up of 11.8 years, finding that decreased HRV, especially high-frequency components, independently predicted CVD outcomes. Using HRV and clinical features, they trained a model with the eXtreme Gradient Boosting algorithm, achieving 75.3% accuracy. Their findings suggest that changes in sleep HRV may precede CVD onset, and combining HRV with other factors enhances early prediction [12].

Zhang and Xu (2023) aimed to predict angina pectoris events in middle-aged and elderly individuals by analyzing RR interval time series in the resting state. Using data from the SHHS involving 2,977 participants over a 15-year

follow-up, they developed a Bi-directional Long Short-Term Memory (Bi-LSTM) model with an Attention layer. The RR intervals from ECG signals were used as inputs, and participants were split into training (n = 2,680) and testing (n = 297) sets. The model demonstrated strong predictive performance, with an accuracy of 0.913. The study suggests that RR intervals could serve as valuable predictors for angina pectoris, supported by the increasing accessibility of heart rate data through wearable devices.

Studies utilizing the SHHS dataset have focused predominantly on predicting the onset of CVDs. However, there is relatively less research aimed at predicting CVD mortality. One such study [5] aimed to predict CVD mortality in patients with sleep apnea using the SHHS dataset. This research was limited to individuals with sleep apnea and employed ML techniques for mortality prediction. In contrast, our study includes a broader population, not restricted to sleep apnea patients. We examine a significantly wider range of parameters compared to Li et al.'s (2022) study, including sleep metrics, disease history, various cardiovascular and respiratory medications, and more. Our comprehensive study seeks to predict CVD mortality over a 10 to 15-year period and assess the importance of these parameters in mortality prediction.

In addition to the aforementioned studies, there has been limited research focused on predicting CVD mortality, with most studies aiming for predictions within a ten-year timeframe. However, in our research, we extend the prediction period to 10 to 15 years. Unlike the majority of studies in this field, we examine a relatively large number of mortality-related parameters across various categories. Our goal is to predict CVD mortality using ML techniques and to identify the key factors influencing this outcome.

2. Materials and Methods

2.1. Data

In our research, we utilize the Sleep Heart Health Study (SHHS) dataset, a prospective cohort study conducted by the National Heart, Lung, and Blood Institute aimed at identifying sleep-related breathing disorders in individuals over 40. The SHHS1 dataset was collected from 6,441 participants between November 1, 1995, and January 31, 1998, while SHHS2 involved 3,295 participants from January 2001 to June 2003, with CVD outcomes gathered until 2011. The study primarily investigated the relationships between sleep-disordered breathing and CVD outcomes [13].

Our focus is on the SHHS1 dataset, which comprises a rich array of categories. The parameters that are utilized in this research are detailed in Table 1. The SHHS1 dataset includes data from 5,804 individuals. ML algorithms are used for binary classification, categorizing individuals into either deceased or surviving classes. Some parameters have missing values for certain individuals; after excluding those with missing data, 3,516 individuals remain for analysis. Among these, only 256 individuals died due to CVD, while the rest have survived, resulting in imbalanced labels. To balance the labels, we employ under-sampling, selecting the data of 256 deceased individuals along with 256 randomly selected surviving individuals for further analysis.

Table 1

Description of categories and parameters.

Categories	Parameters
Demographics	Age, sex, race, ethnicity
Anthropometric measurements	BMI, neck circumference, hip circumference, waist circumference
Medical history of general condition	Alcohol, heavy smoker, cigarette pack-years
Medical history of disease	CVD History, stroke history, Congestive heart failure, history of sleep apnea, Diabetes history
Medical history of sleep parameters	Sleep efficiency, total sleep time, WASO, sleep latency, total time in bed
Laboratory test results	Triglycerides level, blood cholesterol, HDL
Lung Function measurements	FVC, FEV1
Polysomnography measurements	Obstructive apnea-hypopnea (Oahi)
Medications	ALPHA1, ANAR1A1, ANAR31, Aspirin, CCB1, CCBIR1, CCBSR1, DIG1, DIURET1, ESTRGN1, HCTZ1, HCTZK1, Insulins, ISTRD1, LIPID1, LOOP1, NIAC1, NSAID1, NTCA1, Nitrates, PDEI1, Premarin, Progestins, SYMPTH1, TCA1, Thyroid agents, Warfarin
Physical examination test	Hypertension, systolic and diastolic blood pressure

The parameters in Table 1 are defined as follows. BMI refers to body mass index. Heavy smoker refers to moking at least 20 packs in a life time. WASO refers to wake after sleep onset. HDL refers to high-density lipoprotein. FVC

refers to forced vital capacity. FEV1 refers to forced expiratory volume. ALPHA1 refers to Alpha-Blockers without Diuretics. ANAR1A1 refers to Anti-Arrhythmics, class 1A. ANAR31 refers to Anti-Arrhythmics, class 3. CCB1 refers to any calcium-channel blocker (CCIR or CCBSR or CCBT). CCBIR1 refers to Immediate-release CCBS = NFIR or DIHIR or VERIR or DLTIR. CCBSR1 refers to slow-release calcium channel blockers (DLTSR/ AMLOD drug) for hypertension and angina treatment. DIG1 refers to Digitalis preparations for HF. DIURET1 refers to diuretics used to treat hypertension. ESTRGN1 refers to Estrogens, excluding vaginal creams. HCTZ1 refers to Thiazide diuretics without Potassium-sparing agents. HCTZK1 refers to Thiazide diuretics with k-sparing agents. ISTRD1 refers to Inhaled steroids for asthma. LIPID1 refers to any Lipid-Lowering Medication. LOOP1 refers to Loop diuretics used for HF. NIAC1 refers to Niacin and nicotinic acid. NSAID1 refers to non-steroidal anti-inflammatory agents. NTCA1 refers to non-tricyclic antidepressants other than monoamine oxidase inhibitor. Nitrates used for treatment of Angina. PDEI1 refers to Phosphodiesterase inhibitors. Premarin used for treatment of menopausal symptoms. Progestins refers to Synthetic progesterone. SYMPTH1 refers to sympathomimetics, oral and inhaled, for treatment of asthma. TCA1 refers to Tricylic anti-depressants. Thyroid agents used to treat thyroid-related conditions. Warfarin used for treatment of Thromboembolic disorders.

2.2. *Methodology*

Our goal is to predict CVD mortality through the development of a predictive model. For this purpose, we use various parameters from different categories, as listed in Table 1, treating them as features for ML algorithms to predict mortality [2, 4-9, 13]. Additionally, we analyze their impact on CVD mortality.

In this study, we use the mutual information technique with 5-fold stratified cross-validation to determine the importance of each parameter. Mutual information measures the amount of information each feature provides about the target label, helping to identify which features are most relevant for predicting mortality [14, 15]. After identifying the most important features through mutual information (2, 4, 6, 8, 10 features), we applied various ML techniques listed in Table 2 (including KNN, LR, SVM, RF, ETC, and SGD) to predict CVD mortality. We used Grid Search to identify the best hyper-parameters for the ML techniques. Grid Search systematically explores different combinations of hyper-parameters and uses cross-validation to determine which set of values provides the best performance for the model [16].

Table 2

Description of ML techniques

Method	Description
KNN	K-Nearest Neighbors (KNN) is an instance-based algorithm that classifies a point based on the majority class of its k nearest neighbors using distance metrics [17].
LR	Logistic Regression (LR) is a statistical method for predicting binary outcomes by fitting a logistic curve using independent variables [18].
SVM	Support Vector Machine (SVM) is a supervised algorithm that identifies the optimal hyperplane to separate classes while maximizing the margin between them [19].
RF	Random Forest (RF) is an ensemble technique that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting [20].
ETC	Similar to RF, Extra Trees Classifier (ETC) introduces more randomness in tree splitting, enhancing diversity and performance [21].
SGD	Stochastic Gradient Descent (SGD) is an optimization algorithm that incrementally updates model parameters based on the gradient of the loss function from a data subset [22].

3. Results

Table 3 lists the 57 parameters used to assess their impact on predicting CVD mortality. Figure 1 shows the feature importance rankings obtained using the mutual information technique. Figure 2 illustrates the performance of various ML techniques in predicting CVD mortality. The accuracy varies between 70% and 79% depending on the number of input features, while the F1-score ranges from 68% to 80%. Table 4 compares the results of the ML algorithms to identify the best-performing technique. The comparison indicates that the KNN algorithm outperforms others when using the two most important features, age and forced

expiratory volume. Figure 3 presents the Receiver Operating Characteristic (ROC) curve, and Figure 4 displays the confusion matrix, both plotted for the KNN algorithm, which demonstrates the best performance.

Table 3

Description of parameters

Parameters	Alive (256 subject)	Dead (256 subject)
Age; (Mean ± SE)	64.84 ± 0.64	76.24 ± 0.46
SEX; (Male or Female)	Female; 139 (54.30%), Male; 117(45.70%)	Female; 116(45.31%), Male; 140 (54.69%)
Race; (White, Black or Other)	White; 233 (91.01%), Black; 12 (4.69%), Others; 11 (4.30%)	White; 226 (88.28%), Black; 29 (11.33%), Others; 1 (0.39%)
Ethnicity; (Hispanic or Latino)	Hispanic; 246 (96.09%), Latino; 10 (3.91%)	Hispanic; 255 (99.61%), Latino; 1 (0.39%)
BMI; (Mean ± SE) (in kilograms per square meter)	28.36 ± 0.28	27.50 ± 0.30
Hip circumference; (Mean ± SE) (in centimeters)	105.61 ± 0.58	103.33 ± 0.66
Neck circumference; (Mean ± SE) (in centimeters)	37.75 ± 0.25	38.12 ± 0.23
Waist circumference; (Mean ± SE) (in centimeters)	98.27 ± 0.84	98.52 ± 0.81
Cholesterol; (Mean ± SE) (in milligrams per deciliter)	207.70 ± 2.26	206.89 ± 2.57
HDL; (Mean ± SE) (in milligrams per deciliter)	51.26 ± 1.03	50.28 ± 0.91
Triglycerides; (Mean ± SE) (in milligrams per deciliter)	153.26 ± 6.64	159.61 ± 6.34
Systolic BP; (Mean ± SE) (in millimeters of mercury)	123.74 ± 1.10	132.69 ± 1.34
Diastolic BP; (Mean ± SE) (in millimeters of mercury)	71.27 ± 0.65	66.99 ± 0.75
FEV1; (Mean ± SE) (in Liters)	2.71 ± 0.05	2.10 ± 0.04
FVC; (Mean ± SE) (in Liters)	3.63 ± 0.06	2.80 ± 0.06
Alcohol; (drinks per day) (yes or no)	Yes; 112 (43.75%)	Yes; 66 (25.78%)
Cigarette pack-years; (Mean ± SE)	13.84 ± 1.35	16.40 ± 1.42
Heavy smoker; (yes or no)	Yes; 148 (57.81%)	Yes; 141 (55.08%)
History of sleep apnea; (yes or no)	Yes; 2 (0.78%)	Yes; 1 (0.39%)
Hypertension; (yes or no)	Yes; 102 (39.84%)	Yes; 187 (73.05%)
Diabetes; (yes or no)	Yes; 16 (6.25%)	Yes; 56 (21.87%)
Stroke; (yes or no)	Yes; 4 (1.56%)	Yes; 23 (8.98%)
CHD; (yes or no)	Yes; 3 (1.17%)	Yes; 34 (13.28%)
CVD; (yes or no)	Yes; 27 (10.55%)	Yes; 83 (32.42%)
Sleep efficiency; (Mean ± SE)	85.05 ± 0.58	77.93 ± 0.77
WASO; (Mean ± SE) (in minutes)	55.02 ± 2.53	83.17 ± 3.23
Sleep time; (Mean ± SE) (in minutes)	604.18 ± 5.97	569.84 ± 7.17
Time in bed; (Mean ± SE) (in minutes)	438.00 ± 3.35	441.44 ± 3.66
Sleep latency; (Mean ± SE) (in minutes)	11.22 ± 1.15	13.83 ± 1.12
Oahi; (OAHI at >=4%) (Mean ± SE)	10.69 ± 0.83	11.56 ± 0.86
ALPHA1; (yes or no)	Yes; 14 (5.47%)	Yes; 15 (5.86%)
ANAR1A1; (yes or no)	Yes; 2 (0.78%)	Yes; 4 (1.56%)
ANAR31; (yes or no)	Yes; 1 (0.39%)	Yes; 1 (0.39%)
Aspirin; (yes or no)	Yes; 82 (32.03%)	Yes; 119 (46.48%)
CCB1; (yes or no)	Yes; 32 (12.50%)	Yes; 63 (24.61%)
CCBIR1; (yes or no)	Yes; 5 (1.95%)	Yes; 14 (5.47%)
CCBSR1; (yes or no)	Yes; 27 (10.55%)	Yes; 50 (19.53%)
DIG1; (yes or no)	Yes; 4 (1.56%)	Yes; 42 (16.40%)
DIURET1; (yes or no)	Yes; 50 (19.53%)	Yes; 88 (34.37%)
ESTRGN1; (yes or no)	Yes; 48 (18.75%)	Yes; 23 (8.98%)
HCTZ1	Yes; 25 (9.76%)	Yes; 25 (9.76%)
HCTZK1	Yes; 13 (5.08%)	Yes; 17 (6.64%)
Insulins	Yes; 6 (2.34%)	Yes; 10 (3.91%)
ISTRD1	Yes; 4 (1.56%)	Yes; 5 (1.95%)
LIPID1	Yes; 44 (17.19%)	Yes; 33 (12.89%)
LOOP1; (yes or no)	Yes; 11 (4.30%)	Yes; 49 (19.14%)
NIAC1; (yes or no)	Yes; 5 (1.95%)	Yes; 4 (1.56%)
NSAID1; (yes or no)	Yes; 57 (22.26%)	Yes; 41 (16.01%)
NTCA1	Yes; 9 (3.51%)	Yes; 13 (5.08%)
Nitrates	Yes; 7 (2.73%)	Yes; 31 (12.11%)
PDEI1	Yes; 5 (1.95%)	Yes; 9 (3.51%)
Premarin; (yes or no)	Yes; 37 (14.45%)	Yes; 20 (7.81%)
Progestins; (yes or no)	Yes; 24 (9.37%)	Yes; 8 (3.12%)
SYMPTH1; (yes or no)	Yes; 7 (2.73%)	Yes; 11 (4.30%)
TCA1	Yes; 7 (2.73%)	Yes; 6 (2.34%)
Thyroid agents	Yes; 25 (9.76%)	Yes; 25 (9.76%)
Warfarin; (yes or no)	Yes; 2 (0.78%)	Yes; 24 (9.37%)

Alive refers to individuals who were alive throughout the SHHS dataset collection period. Dead refers to individuals who died from CVD. SE refers to standard error.

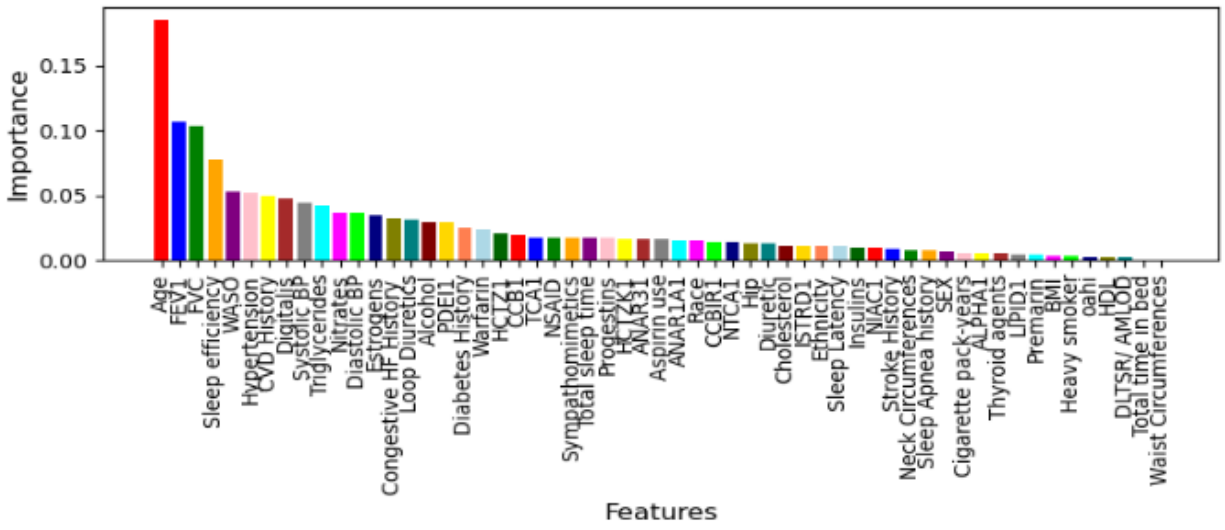

Figure 1

Feature Importance Using the Mutual Information technique

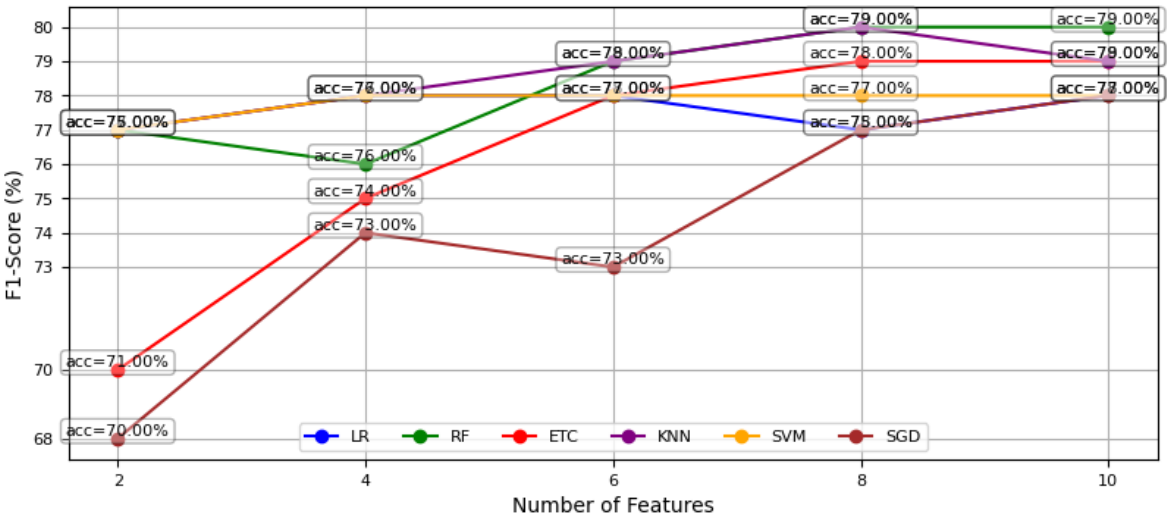

Figure 2

Performance of ML Algorithms in predicting CVD mortality

Table 4

Comparison of ML algorithms' results

	Precision (%)		Recall (%)		F1-Score (%)		Performance (%)		n
ML	*C0*	*C1*	*C0*	*C1*	*C0*	*C1*	*Acc*	*AUC*	*n*
LR	79	74	71	81	74	77	76	82	2
KNN	77	77	76.56	77.34	76	77	77	79	2
RF	77	75	74	79	75	77	76	81	2
SVM	80	72	68	83	73	77	75	80	2

*C*0 refers to class 0. *C*1 refers to class 1. Acc refers to accuracy. AUC refers to the area under the receiver operating curve. n refers to number of features used for training the ML algorithms**.**

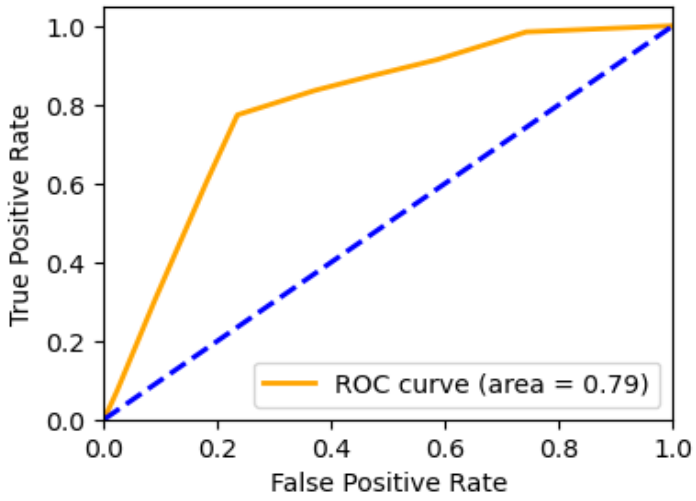

Figure 3

ROC Curve of KNN method for predicting CVD mortality

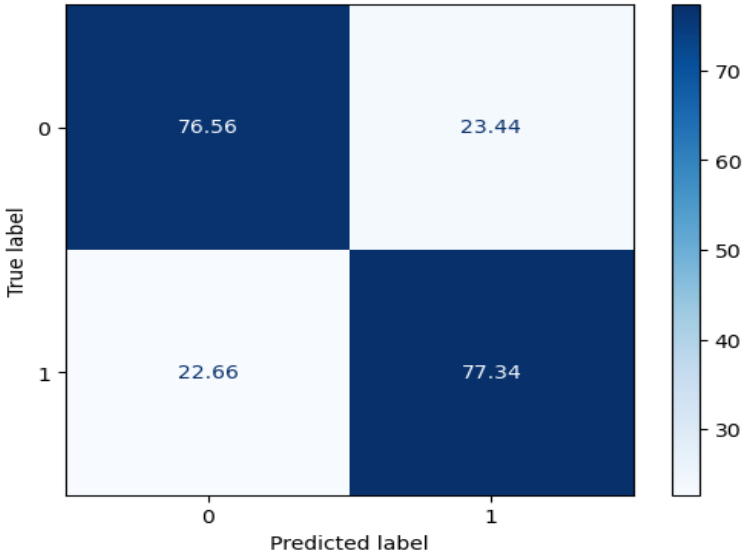

Figure 4

Confusion Matrix of KNN for CVD mortality prediction

4. Discussion

Given that CVD mortality is the leading cause of death globally, effective monitoring of CVD patients is critical. To facilitate proper monitoring, it is essential to first identify the key factors contributing to CVD mortality.

Using ML techniques, we predict CVD mortality over a 10 to 15-year period to determine these influential factors. We train various ML algorithms, including LR, SVM, KNN, RF, SGD, and ETC, using different sets of 2, 4, 6, 8, and 10 most significant features through mutual information. To ensure accuracy, we employ 10-fold cross-validation.

The performance of these algorithms, based on F1-score for different numbers of input features, is depicted in Figure 2. Our analysis reveals that LR, RF, KNN, and SVM produced relatively similar results across varying feature counts. Notably, all algorithms perform comparably well with the top 10 features (age, FEV1, FVC, sleep efficiency, WASO, hypertension, CVD history, digitalis use, systolic blood pressure, triglycerides).

Given the similarity in overall performance, we focus on the minimal number of features that yield comparable results to the best-performing algorithms. As shown in Figure 2, using the top 2 features results in an F1-score of 77%, which slightly increases to 80% with 10 features. Therefore, we compare the algorithms using just the top 2 features in Table 4. The results indicate that the KNN algorithm outperforms the others, achieving an F1-score of 77%, an accuracy of 77%, an AUC of 79%, a sensitivity of 77.34%, and a specificity of 76.56% with age and forced expiratory volume as inputs. The optimal value of k, determined via grid search for KNN with these two features, is 8.

In this study, we encounter several limitations, including a significant number of missing values across various parameters, an average participant age above 64 years, and highly imbalanced labels. Future research should focus on analyzing data from younger populations with fewer missing values to achieve a more comprehensive and broader prediction across a larger population.

Despite these limitations, our study provides a comprehensive examination of various factors influencing long-term CVD mortality over a period of ten to fifteen years. We thoroughly analyze different parameters to predict mortality across individuals of varying ages. Importantly, we identify the key factors contributing to early CVD mortality and determine the most effective ML technique for prediction.

Timely and intelligent prediction of CVD mortality and the identification of influencing parameters are crucial for preventing premature deaths in populations. Such predictions allow individuals to consider lifestyle changes towards healthier living and enable physicians to focus on reducing the impact of or monitoring the key parameters affecting CVD mortality.

5. Conclusion

This study develops an intelligent approach for predicting CVD mortality over a 10 to 15-year period. Various ML techniques are employed to investigate the relationship between multiple parameters and mortality prediction. A relatively large number of parameters from different categories are examined, and their importance is determined using mutual information. These parameters are then used as inputs for ML models. Model robustness is ensured through k-fold cross-validation (with k = 10), and grid search is employed for hyper-parameter optimization.

Among the tested algorithms, KNN demonstrates the best performance across different numbers of important parameters, particularly with the key features of age and forced expiratory volume, achieving an optimal k = 8. KNN achieves an accuracy of 77%, an F1-score of 77%, an AUC of 79%, a sensitivity of 77.34%, and a specificity of 76.56%.

As mentioned earlier, the primary goal is to identify the influential parameters in predicting CVD mortality. By closely monitoring these key parameters, healthcare providers can better manage CVD patients, improve their health conditions, and ultimately contribute to extending their lifespan and enhancing their quality of life.

Authors' Contributions

The contribution of the authors is according to the names of the authors in the article.

Declaration

Transparency Statement

Data are available for research purposes upon reasonable request to the corresponding author.

Acknowledgments

The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

Declaration of Interest

The authors report no conflict of interest.

Funding

This work is partially supported upon research funded by Iran National Science Foundation (INSF) under project No.4024604.

Ethical Considerations

Not applicable.

References

[1] W. World Health Organization, "Cardiovascular diseases (CVDs)," 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] A. Rawat and et al., "Smoking And Coronary Heart Disease Impact," *Journal of Pharmaceutical Negative Results*, vol. 2023, pp. 1737-1742, 2023.

[3] S. Sajeev *et al.*, "Predicting Australian adults at high risk of cardiovascular disease mortality using standard risk factors and machine learning," *International Journal of Environmental Research and Public Health*, vol. 18, no. 6, p. 3187, 2021, doi: 10.3390/ijerph18063187.

[4] R. Nakanishi *et al.*, "Machine Learning Adds to Clinical and CAC Assessments in Predicting 10-Year CHD and CVD Deaths," *JACC Cardiovascular Imaging*, vol. 14, no. 3, pp. 615-625, 2021, doi: 10.1016/j.jcmg.2020.08.024.

[5] A. Li, J. M. Roveda, L. S. Powers, and S. F. Quan, "Obstructive sleep apnea predicts 10-year cardiovascular disease-related mortality in the Sleep Heart Health Study: a machine learning approach," *Journal of Clinical Sleep Medicine*, vol. 18, no. 2, pp. 497-504, 2022, doi: 10.5664/jcsm.9630.

[6] A. S. Vaughan and et al., "Country-level trends in hypertension-related cardiovascular disease mortality—United States, 2000 to 2019," *Journal of the American Heart Association*, vol. 11, no. 7, p. e024785, 2022, doi: 10.1161/JAHA.122.027832.

[7] E. G. Ross, N. H. Shah, R. L. Dalman, K. T. Nead, J. P. Cooke, and N. J. Leeper, "The use of machine learning for the identification of peripheral artery disease and future mortality risk," *Journal of Vascular Surgery*, vol. 64, no. 5, pp. 1515-1522.e3, 2016, doi: 10.1016/j.jvs.2016.04.026.

[8] H. Han *et al.*, "Sleep Duration and Risks of Incident Cardiovascular Disease and Mortality Among People With Type 2 Diabetes," *Diabetes Care*, vol. 46, no. 1, pp. 101-110, 2023, doi: 10.2337/dc22-1127.

[9] B. Zhao *et al.*, "Association of objective and self-reported sleep duration with all-cause and cardiovascular disease mortality: A community-based study," *Journal of the American Heart Association*, vol. 12, no. 6, p. e027832, 2023, doi: 10.1161/JAHA.122.027832.

[10] A. Martin-Morales, M. Yamamoto, M. Inoue, T. Vu, R. Dawadi, and M. Araki, "Predicting cardiovascular disease mortality: Leveraging machine learning for comprehensive assessment of health and nutrition variables," *Nutrients*, vol. 15, no. 18, p. 3937, 2023, doi: 10.3390/nu15183937.

[11] J. U. Park, E. Urtnasan, S. H. Kim, and K. J. Lee, "A Prediction Model of Incident Cardiovascular Disease in Patients with Sleep-Disordered Breathing," *Diagnostics (Basel)*, vol. 11, no. 12, p. 2212, 2021, doi: 10.3390/diagnostics11122212.

[12] L. Zhang, H. Wu, X. Zhang, X. Wei, F. Hou, and Y. Ma, "Sleep heart rate variability assists the automatic prediction of long-term cardiovascular outcomes," *Sleep Medicine*, vol. 67, pp. 217-224, 2020, doi: 10.1016/j.sleep.2019.11.1259.

[13] Sleepdata.org. "Sleep Heart Health Study (SHHS)." https://sleepdata.org/datasets/shhs (accessed.

[14] T. M. Cover and J. A. Thomas, *Elements of International Theory*. Hoboken, NJ, USA: Wiley, 2006.

[15] R. Kohavi, "A study of Cross-Validation and Boostrap for Accuracy Estimation and Model Selection," presented at the Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), 1995.

[16] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 281-305, 2012.

[17] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100-108, 1979.

[18] D. W. Hosmer, Jr. and S. Lemeshow, *Applied Logistic Regression*. Hoboken, NJ, USA: Wiley, 2013.

[19] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge University Press, 2000.

[20] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[21] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3-42, 2006, doi: 10.1007/s10994-006-6226-1.

[22] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," presented at the Proceedings of the 2nd International Conference on Learning Representations (ICLR), 2014.