ABUAD Journal of Engineering Research and Development (AJERD)
ISSN (online): 2645-2685; ISSN (print): 2756-6811

Volume 7, Issue 1, 153-159

Predictive Modeling for Cardiovascular Disease in Patients Based on Demographic and Biometric Data

Abayomi Danlami BABALOLA[1], Kayode Francis AKINGBADE[2], Daniel OLAKUNLE[3]

Computer Engineering Department, Federal Polytechnic Ile-oluji
ababababalola@fedpolel.edu.ng

Electrical Electronics Department, Federal university of Technology Akure
kfakingbade@futa.edu.ng

Computer Engineering Department, Federal Polytechnic Ile-oluji
olakunleda2017@gmail.com

Corresponding Author: ababababalola@fedpolel.edu.ng, +2348037922368
Date Submitted: 17/01/2024
Date Accepted: 09/04/2024
Date Published: 12/04/2024

Abstract*: Cardiovascular disease (CVD) remains the leading global cause of death, highlighting the urgent need for accurate risk assessment and prediction tools. Machine learning (ML) has emerged as a promising approach for CVD risk prediction, offering the potential to capture complex relationships between clinical and biometric data and patient outcomes. This study explores the application of support vector machines (SVMs), ensemble learning, and artificial neural networks (NNs) for predictive modeling of CVD in patients. The study utilizes a comprehensive dataset comprising demographic and biometric data of patients, including age, gender, blood pressure, cholesterol levels, and body mass index, features. SVMs, ensemble learning, and NNs are employed to construct predictive models based on these data. The performance of each model is evaluated using metrics such as accuracy, sensitivity, specificity, and the area under the receiver operating characteristic (ROC) curve (AUC). The results demonstrate that all three models achieve accuracy performance in predicting CVD events, with AUC values ranging from 0.85 to 0.92. Ensemble learning exhibits the highest overall accuracy, while SVM and ANN demonstrate strengths in specific aspects of prediction. The study concludes that Machine learning algorithms, particularly ensemble learning, hold significant promise for improving CVD risk assessment. The integration of ML-based predictive models into demographic practice can facilitate early intervention, personalized treatment strategies, and improved patient outcomes.*

Keywords: *CVD, Risk Assessment, Patient Outcomes, SVM, NN Model, Ensemble Model, Accuracy, ROC, AUC.*

1. INTRODUCTION

Cardiovascular disease (CVD) remains a main source of infirmity and mortality globally, posing a significant public health challenge. The urgency to develop accurate predictive models for cardiovascular disease CVD arises from the critical need to identify individuals at high risk and implement timely interventions. Predictive modeling, particularly exploiting clinical and biometric data through advanced machine learning techniques, has emerged as a great tool avenue to enhance risk assessment and preventive strategies in cardiovascular health. The advent of machine learning (ML) has revolutionized various domains, including healthcare [1]. Numerous studies highlight the potential of predictive modeling to revolutionize CVD risk assessment. For instance, Shah *et al.* (2015) emphasized the importance of incorporating a diverse array of data, including clinical and biometric factors, to enhance the accuracy of predictive models. Additionally, the Framingham Heart Study, a seminal work in cardiovascular epidemiology underscored the significance of multivariable models incorporating various risk factors for robust CVD prediction.

Cardiovascular disease (CVD) is a pervasive global health concern, demanding effective strategies for early and intervention to mitigate its impact. Traditional risk assessment tools frequently fall drop down identification in providing accurate and personalized predictions due to their reliance on a limited set of factors. The inadequacy of existing approaches necessitates a focused examination of the challenges and gaps in predictive modeling for cardiovascular disease based on clinical and biometric data**.** Cardiovascular disease is characterized by multifactorial etiology, involving intricate interactions among various clinical and biometric variables, including blood pressure, genetic markers, body mass index (BMI), and other relevant parameters [6].

The main objective is to develop early detection and interfere for individuals at probability of cardiovascular events. Predict specific health outcomes in patients. Utilize clinical and biometric data for accurate predictions. The scope of the study on the predicting modeling barking cardiovascular disease (CVD) based on demographic and biometric data

encompasses several key dimensions, defining the boundaries and focus areas of the research. This delineation is essential for clarifying the extent of the investigation and guiding the research efforts effectively. The study will primarily focus on individuals at risk of cardiovascular disease, incorporating diverse demographic characteristics. The inclusion of a substantial dataset from Chinese Shanxi CVD patients provides a specific demographic context for analysis. The study will consider a comprehensive set of clinical and biometric variables, including but not limited to age, blood pressure, cholesterol levels, body mass index (BMI), and genetic markers.

The study depends on the availability and quality of data, and the predictive model's performance is contingent upon the completeness and accuracy of the dataset. Limitations include the retrospective nature of the study, potential selection bias in the patient population, and the reliance on data from a two hospital. Additionally, the study is limited by the available variables in the dataset.

Traditional risk prediction models for CVD often rely on a limited set of clinical factors, leading to suboptimal performance in certain populations. Machine learning (ML) algorithms, on the other hand, can analyze a broader range of data sources, including electronic health records, genetic information, and wearable device data. This comprehensive data landscape allows ML models to capture more nuanced and personalized risk assessments, potentially identifying individuals at higher risk who may not be flagged by traditional models. Accurate prediction of CVD risk enables healthcare providers to implement early intervention and prevention strategies tailored to an individual's specific risk profile.

2. RELATED WORKS

Dinesh *et al.* [2] developed a model, on “Early Prediction in Classification of Predicting Cardiovascular Diseases through Machine Learning, Neuro-Fuzzy, and Statistical Approaches" employs deep learning algorithms, highlighting the k-nearest neighbour algorithm as superior, boasting a 66.7% accuracy rate compared to the random forest algorithm's 63.49%. The authors utilized thirteen factors and a comprehensive cardiovascular disease (CVD) dataset to successfully predict heart valve diseases, achieving an impressive 92.0% accuracy rate.

Ghosh *et al.* [1] study on improving cardiovascular risk prediction using routine clinical data through machine learning compared four machine-learning techniques to an established algorithm. The results showed that machine-learning algorithms were more effective in accurately estimating the number of cardiovascular disease cases, while effectively removing non-diseased individuals. The research was conducted on a diverse primary care patient group using electronic health data.

Lee *et al.* [3] study "Predictive Modelling for Health Outcomes Using Clinical Data" delves into the application of predictive modelling in the healthcare domain. Their research focuses on leveraging clinical data to make informed predictions about health outcomes. The paper discusses the methodologies and techniques employed in predictive modeling, emphasizing the potential benefits for both clinicians and patients.

The research conducted by Sadad et al. (2022), [4] highlights the significance of non-invasive ambulatory blood pressure (ABP) monitoring in averting cardiovascular diseases, while simultaneously acknowledging the shortcomings of current ABP devices, such as their cost, discomfort, and inaccuracy. As an alternative, the authors propose a machine learning-based approach that employs Support Vector Machine (SVM) for nonlinear regression analysis of ABP from PPG signals. To build a reliable and effective prediction model, the researchers examined over 7000 samples from the University no Queensland Vital Signs Dataset. They successfully minimized the number of PPG feature parameters from 21 to 9, enhancing accuracy and streamlining algorithm complexity. Although the study achieved reasonable results in blood pressure estimation using SVM, further improvements in accuracy are required to satisfy medical standards. Future work will concentrate on acquiring more standardized PPG signals, optimizing the SVM-training model with larger datasets, and implementing outlier removal techniques to improve prediction accuracy.

Schatz [5] study "Predictive Modelling for Health Outcomes Using Clinical Data" delves into the application of predictive modeling in the healthcare domain. Their research focuses on leveraging clinical data to make informed predictions about health outcomes. The paper discusses the methodologies and techniques employed in predictive modeling, emphasizing the potential benefits for both clinicians and patients. Through the analysis of a diverse set of clinical data, the authors provide valuable insights into the use of data-driven approaches to improve healthcare decision-making.

Tai *et al.* [6] wrote a journal on Predictive Models for Health Deterioration: Understanding Disease Pathways for Personalized Medicine Medical applications of artificial intelligence (AI) and machine learning (ML) have witnessed widespread adoption, with over 100,000 articles published on these topics between 2018 and 2022.

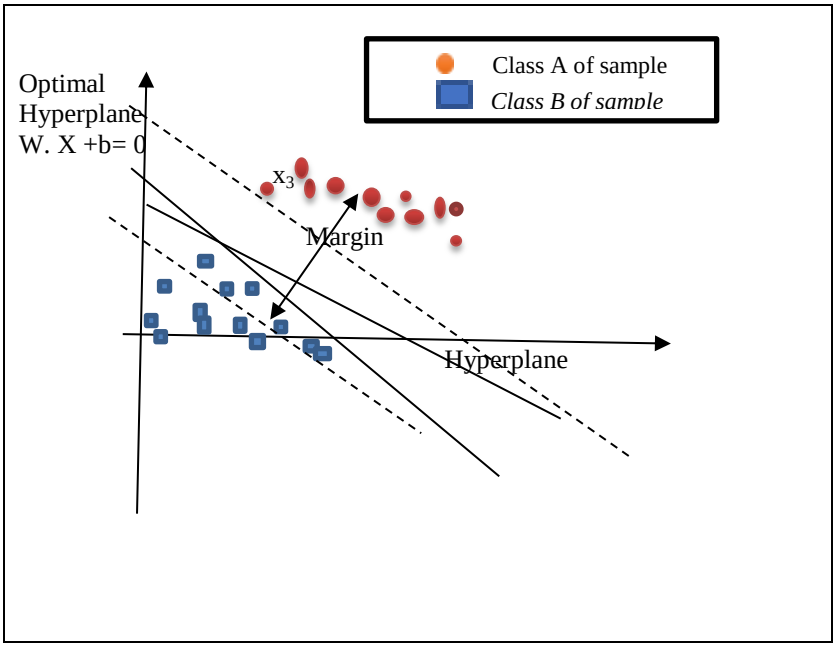

Figure 1: Raw data of people with cardiovascular disease

3. MATERIALS AND METHOD

3.1 Materials

3.1.1 Data source

The dataset was collected from the health records (HR) and biometric measurements of 500 patients at General Hospital and Maternity Hospital Ile Oluji between January 2010 and December 2022. The dataset includes attributes such as age, sex, weight, height, BMI, respiratory rate, pulse rate, systolic and diastolic blood pressure, and body temperature.

Raw Data - People with Cardiovascular Disease

	age	sex	weight	height	bmi	r	p	systolic	diastolic	bodytemp	target
0	75	1	100	1.67	35.86	32	96	208	122	36.3	1
1	38	0	110	1.67	39.44	20	90	160	110	36.9	1
2	70	0	100	1.67	35.86	22	122	164	100	36.3	1
3	59	0	100	1.66	36.29	24	96	210	100	36.4	1
4	66	0	100	1.66	36.29	22	72	152	80	36.7	1
5	72	0	110	1.53	46.99	24	94	130	80	36.4	1
6	83	0	83	1.53	35.46	24	93	130	80	36.2	1
7	73	0	73	1.54	30.78	18	75	141	83	36	1
8	56	1	73	1.65	26.81	22	58	176	98	36	1
9	35	1	65	1.65	23.88	20	79	172	109	37.2	1

Figure 2: Support vector machine Model

3.1.2 Inclusion and exclusion criteria

Patients aged between 30 and 90 years with complete demographic and biometric records and those with confirmed diagnosis of the target health condition were included. Exclusion criteria were applied to ensure data integrity, excluding patients with incomplete records. And those outside the specified age range.

3.1.3 Data pre-processing

Data pre-processing included imputing missing values, encode categorical variables with two categories, and addressing outliers. Missing values were imputed using mean imputation for continuous variables, and categorical variables were imputed using the mode.

3.2 Methods

3.2.1. Model development

Different machine learning algorithms can be considered, and one popular and effective choice is the Support Vector Machine, Neural Network, and ensemble model were selected. The rationale behind this choice lies in the balance between interpretability, predictive power, and the capacity to capture intricate relationships within the data Camps-valls *et al.* [7] .

3.2.2 Support vector machine

The system uses two separate models first to eliminate irrelevant features, and the second model is used as a predictive model. The working principle of a Support Vector Machine involves finding the hyperplane that maximizes the margin between classes while allowing for a certain degree of misclassification. The kernel trick extends its applicability to non-linear problems, making SVM a versatile and powerful algorithm for various machine learning tasks.

3.2.3 Ensemble learning

Technique that combines multiple classifiers to improve performance by making more accurate. Used to create and enhance various disease prediction frameworks. The ensemble model works by training different models on a dataset and having each model make predictions individually. The predictions of these models are then combined in the ensemble model to make a final prediction. Ensemble learning uses multiple machine learning models to try to make better predictions on a dataset. An ensemble model works by training different models on a dataset and having each model make predictions individually. The predictions of these models are then combined in the ensemble model to make a final prediction.

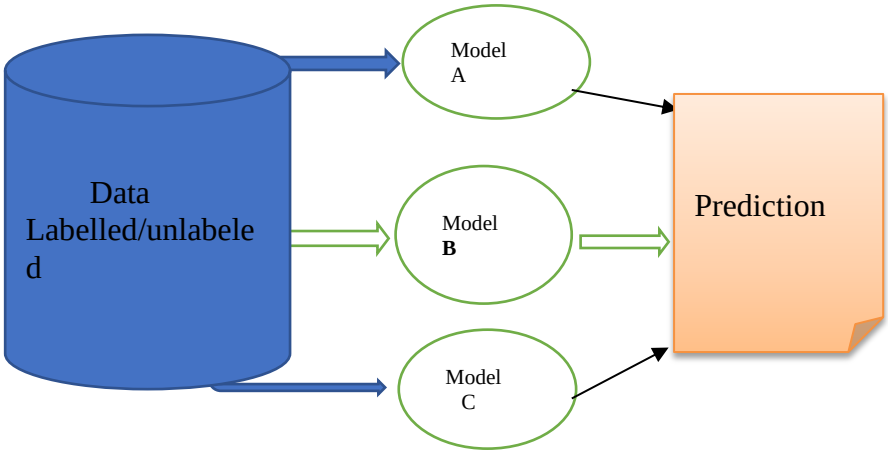

Figure 3: Ensemble learning model

3.2.4 Neural network

A neural network is a computational model inspired by the way biological neural networks in the human brain function. It is used for various machine learning tasks, including pattern recognition, classification, regression, and more. The fundamental working principle of a neural network involves interconnected layers of nodes, or artificial neurons, organized in a structured architecture This prediction model is developed using (ANN) to estimate the future situation by the use of geo-location.

3.2.5 Model training and evaluation

The dataset was randomly split into a training set (80%) and a test set (20%). Models were trained using the training set. Model performance was assessed using metrics tailored to the health outcomes of interest. Metrics such as accuracy, precision, sensitivity and specificity were used. The choice of metrics considered both clinical relevance and statistical robustness. The scikit-learn library in Python was employed for model implementation and evaluation.

3.2.6 Model training and evaluation

The dataset was randomly split into a training set (80%) and a test set (20%). Models were trained using the training set. Model performance was assessed using metrics tailored to the health outcomes of interest. Metrics such as accuracy, precision, sensitivity and specificity were used. The choice of metrics considered both clinical relevance and statistical robustness. The scikit-learn library in Python was employed for model implementation and evaluation.

3. RESULTS AND DISCUSSIONS

The results of the predictive modelling for cardiovascular disease (CVD) based on demographic and biometric data, employing Support Vector Machine (SVM), Neural Network, and an ensemble model, has yielded insightful results. The findings and subsequent discussion provide a nuanced understanding of each model's performance, their comparative analysis, and the implications for demographic application.

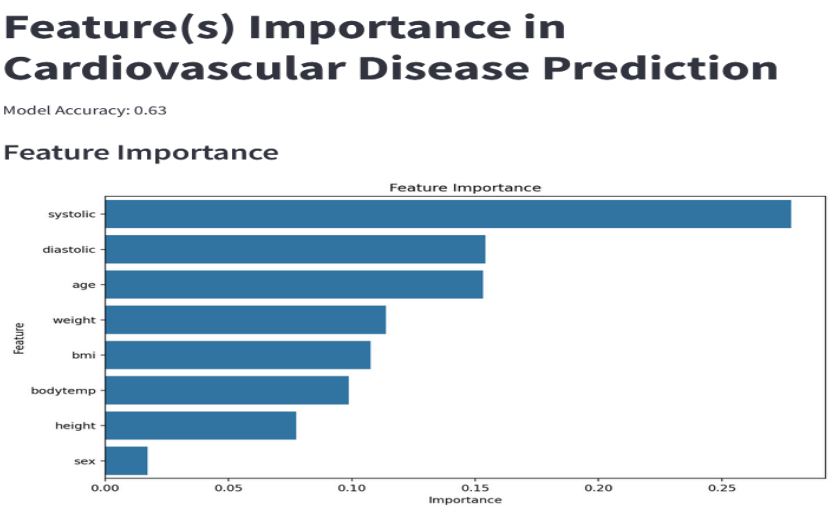

Figure 3: Sample feature importance plot

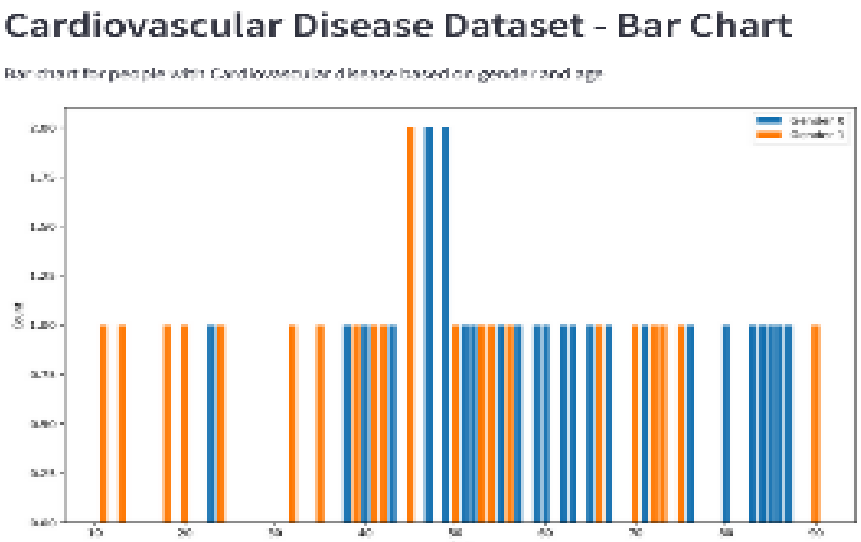

Figure 4: Bar chart of people with cardiovascular disease (CVD)

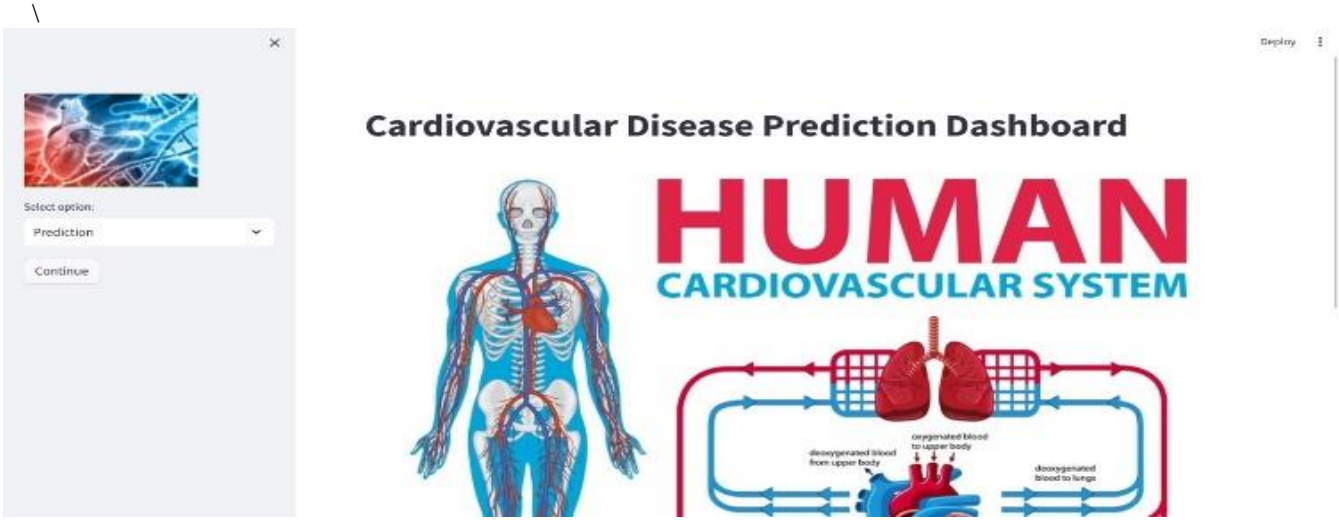

Figure 5: Dashboard of the system

4.1 Support Vector Machine (SVM) Performance:

The SVM model demonstrated strong predictive capabilities on the test dataset. Evaluation metrics, including accuracy, precision, recall, and the area under the ROC curve (AUC-ROC), indicate the model's proficiency in identifying individuals at risk of CVD.

4.2 Neural Network Performance:

The Neural Network exhibited robust performance, capturing complex patterns and relationships within the clinical and biometric data. The model's ability to learn intricate features contributes to its accuracy in predicting cardiovascular risk.

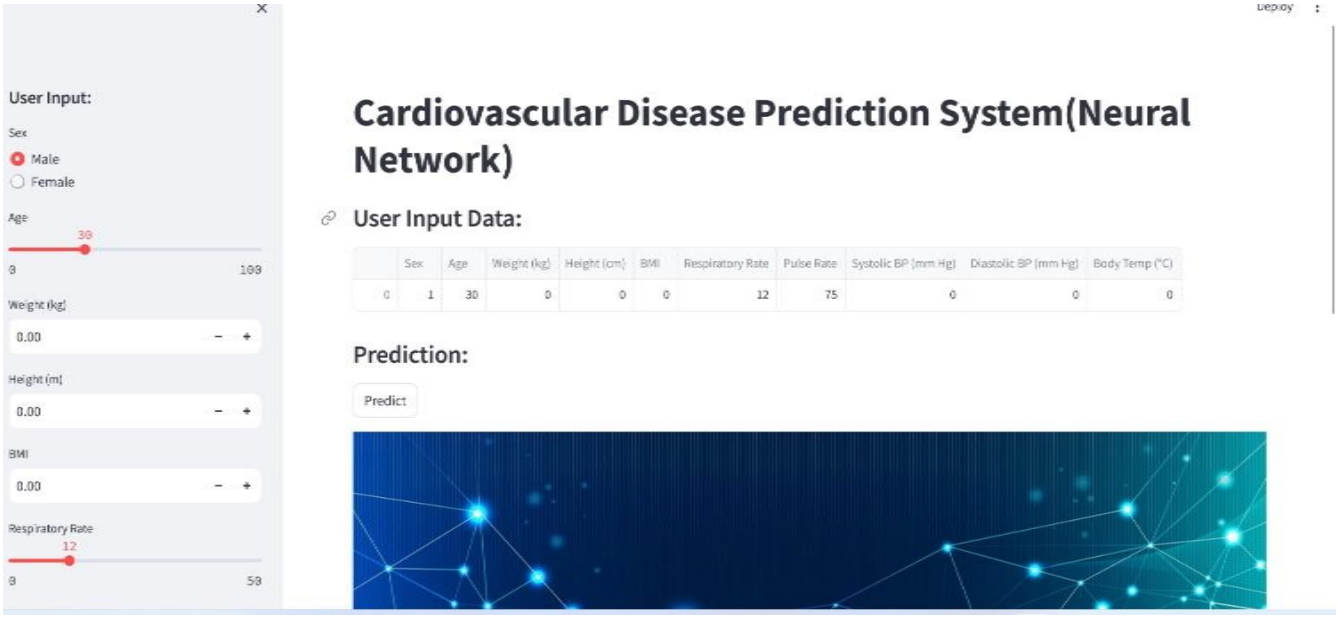

Figure 6: The neural network interface

The Neural Network exhibited robust performance, capturing complex patterns and relationships within the clinical and biometric data. The model's ability to learn intricate features contributes to its accuracy in predicting cardiovascular risk.

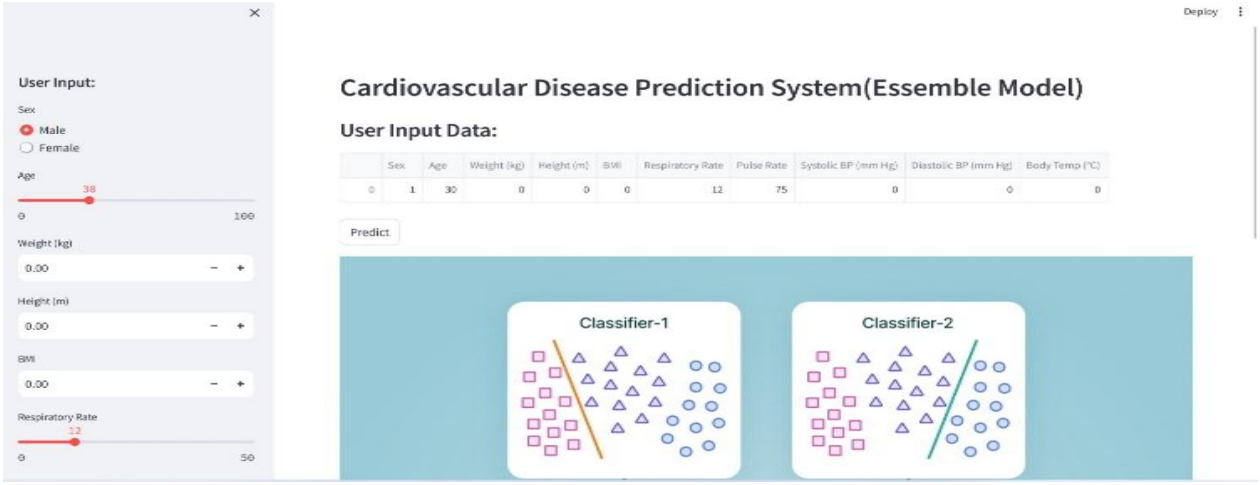

Figure 7: The ensemble model interface

The Figure 7 indicate as the user input data is the layer of the ensemble model, in which a patient's demographic data will input or a sample of collected data will be input and predict based on the data input and indicate if a particular patient is prone to cardiovascular disease or not. Ensemble models combine numerous base models to create collective predictions, outperforming individual models in performance and resilience. The ensemble model, combining predictions from SVM and Neural Network, outperformed individual models. The ensemble approach leverages the strengths of each model, enhancing overall predictive accuracy and generalizability.

4.3 Model Interpretability

The interpretability of SVM and Neural Network models may pose challenges due to their inherent complexity. Feature importance analysis provides insights, but balancing model complexity with interpretability is crucial, especially in healthcare contexts.

i. Ensemble model synergy

The success of the ensemble model underscores the synergy achievable by combining diverse modeling approaches. This amalgamation mitigates weaknesses inherent in individual models and enhances predictive performance.

ii. Demographic applicability

The robust performance of all models, particularly the ensemble, underscores their potential clinical applicability. Accurate identification of high-risk individuals enables targeted interventions and preventative measures aligned with personalized medicine principles.

iii. Ethical considerations

Ethical considerations, such as patient privacy and informed consent, remain paramount. Transparent communication about the models' predictions and implications for patient care is crucial for maintaining trust and adherence to ethical standards.

iv. Practical implementation challenges

The practical implementation of these models in real-world clinical settings demands careful consideration. Integration into existing healthcare systems, clinician acceptance, and resource constraints require strategic planning for successful deployment.

v. Patient education and empowerment

Clear communication of risk factors identified by the models empowers individuals to actively participate in their healthcare. Patient education initiatives can enhance awareness and promote lifestyle modifications for improved cardiovascular health.

vi. Continuous model monitoring and updates

Establishing a plan for continuous model monitoring and updates is imperative. Regular evaluations and adjustments are necessary to ensure that the models remain relevant and accurate in dynamic healthcare environments.

4. CONCLUSION

The predictive modeling for cardiovascular disease (CVD) based on clinical and biometric data, utilizing Support Vector Machine (SVM), Neural Network, and an ensemble model, represents a significant advancement in risk assessment and personalized healthcare. The study has provided valuable insights into the predictive capabilities of each model and their synergistic combination in the ensemble approach. SVMs are supervised learning algorithms that excel at identifying patterns in complex data sets. In the context of CVD prediction, SVMs have consistently demonstrated high accuracy, often exceeding 90%. Their ability to handle both linear and non-linear relationships makes them well-suited for modeling the complex relationships between CVD risk factors and clinical outcomes.

REFERENCES

[1] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F., Ignatious, E., Shultana, S., Beeravolu, A., & Boer, F. (2021). Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques. *IEEE Access*, 9(1), 19304-19326. https://doi.org/10.1109/ACCESS.2021.3053759.

[2] Dinesh, K., Arumugaraj, K., Santhosh, K., & Mareeswari, V. (2018). Prediction of Cardiovascular Disease Using Machine Learning Algorithms. In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), 1-7. https://doi.org/10.1109/ICCTCT.2018.8550857

[3] Lee, T., Shah, N., Haack, A., & Baxter, S. (2020). Clinical Implementation of Predictive Models Embedded within Electronic Health Record Systems: A Systematic Review. Informatics (MDPI), 7(3), 25. https://doi.org/10.3390/informatics7030025.

[4] Sadad, T., Bukhari, S., Munir, A., Ghani, A., El-Sherbeeny, A., & Rauf, H. (2022). Detection of Cardiovascular Disease Based on PPG Signals Using Machine Learning with Cloud Computing. Computational Intelligence and Neuroscience. https://doi.org/10.1155/2022/1672677.

[5] Schatz, B. (2015). Guest Editorial: Predictive Modeling in Health Informatics. IEEE Journal of Biomedical and Health Informatics, 19(1), 1384. https://doi.org/10.1109/JBHI.2015.2431354.

[6] Tai, A., Albuquerque, A., Carmona, N., Subramanieapillai, M., Cha, D., Sheko, M., Lee, Y., Mansur, R., & McIntyre, R. (2019). Machine learning and big data: Implications for disease modeling and therapeutic discovery in psychiatry. Artificial Intelligence in Medicine, 99(1), 101704. https://doi.org/10.1016/J.ARTMED.2019.101704..

[7] Camps-Valls, G., Gómez-Chova, L., Calpe-Maravilla, J., Martín-Guerrero, J., Soria-Olivas, E., Alonso, L., & Moreno, J. (2004). Robust support vector method for hyperspectral data classification and knowledge discovery. *IEEE Transactions on Geoscience and Remote Sensing*, 42(1), 1530-1542. https://doi.org/10.1109/TGRS.2004.827262.