Educational Administration: Theory and Practice
2024, 30(5), 8286-8298
ISSN: 2148-2403
https://kuey.net/

Research Article

Implication of analysisof machine learning models for predicting the risk of cardiovascular disease by considering lifestyle factors and featuresselection

Dr. Padma Mishra[1*], Dr. Vinita Gaikwad[2], Dr. Mayuri Paradkar[3], Dr.Nilesh Pandey[4], Dr.RamaBansode[5], Mr.Shirshedu Maitra[6]

[1*]Associate Professor, MCA Thakur Institute of Management Studies, Career Development and Research (TIMSCDR), Mumbai,India. mishrapadma1988@gmail.com
[2]Director, MCA, Thakur Institute of Management Studies,Career Development and Research (TIMSCDR), Mumbai,India. vinitagaikwad2@gmail.com
[3]Trust Grade Doctor, Blackpool Teaching Hospitals NHS Foundation Trust, England, United Kingdom mayuriparadkar04@gmail.com
[4]Masters of Science in Healthcare Management, England, United Kingdom. nileshpandey7@gmail.com
[5]Assistant Professor, MCA. Modern College of Engineering Pune,India. rama.bansode@moderncoe.edu.in
[6]Assistant Professor, MCA Thakur Institute of Management, Studies, Career Development and Research (TIMSCDR), Mumbai,India. slm2007@gmail.com

Citation : Dr. Padma Mishra, et al, (2024), Implication of analysisof machine learning models for predicting the risk of cardiovascular disease by considering lifestyle factors and featuresselection**,** *Educational Administration: Theory and Practice*, *30*(5), 8286-8298
Doi: 10.53555/kuey.v30i5.4353

ARTICLE INFO

ABSTRACT

Objective: This study aims to conduct a comprehensive survey and analysis of the application of machine learning techniques predicting the risk of cardiovascular diseases by considering the certain lifestyle factor. The objective is to explore the utilization of machine learning algorithms, diverse datasets, feature selection methodologies, modalities (uni or multi-modal), and performance evaluation metrics across CVD research. The study seeks to provide insights into state-of-the-art approaches, identify key challenges, and stimulate further research interest in leveraging machine learning for early detection and efficient management of CVD, thereby contributing to improved healthcare outcomes**.**

Method: The methodology employed in this study encompasses several key steps to comprehensively analyse the application of machine learning (ML) algorithms in detecting, categorizing, and predicting cardiovascular diseases (CVD). Firstly, diverse datasets related to CVDs collected, covering demographic information, medical history and lifestyle factors. These datasets undergo meticulous pre-processing steps, including handling missing values, outliers, and data normalization, to ensure data quality and consistency. Feature selection techniques recursive feature elimination, and feature importance ranking are then applied to recognize the utmostapplicable features for predicting CVD outcomes. The results are then analysed and interpreted to gain insights into the strengths and weaknesses of each model, feature importance, and generalization capabilities across different datasets.

Findings: The analysis considered several features as relevant to the Indian population due to their coverage of modified lifestyle attributes. These structurescontain BMI, Systolic Blood Pressure, Diastolic Blood Pressure, Smoking, Glucose, and Cholesterol. In the LightGBM model training summary, the dataset consisted of 56,000 instances, evenly split between positive and negative cases, and utilized 8 features. The model achieved a starting score of -0.002357, with 720 total bins used. The accuracy of various classifiers ranged from 51.31% (Perceptron) to 73.18% (Gradient Boosting Classifier), with the Gradient Boosting Classifier performing the best on this dataset. Additionally, LightGBM's automatic choice of row-wise multi-threading slightly impacted overhead.

Novelty: The study emphasizes feature selection and data pre-processing, optimizing predictive analytics for healthcare in India. By focusing on BMI,

Copyright © 2024 by Author/s and Licensed by Kuey. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

blood pressure, smoking, glucose, and cholesterol, it advances precision medicine and personalized healthcare. The findings improve disease prediction and management, inspiring further research and innovation. This work highlights the importance of tailored healthcare solutions for diverse populations, enhancing health interventions and patient care.

Keywords—Cardiovascular Disease Prediction; Machine Learning in Healthcare;Lifestyle Factors and CVD Feature Selection Techniques; Data Pre-processing in ML;Predictive Analytics in *Medicine; Precision Medicine ;Personalized Healthcare; CVD Risk Assessment; Healthcare Data Analysis*

I.INTRODUCTION

Cardiovascular diseases (CVDs) stand as a paramount reason of mortality global, requesting nearly 17.9 million survives annually, as reported by the World Health Organization (WHO). These encompass a spectrum of conditions impacting the heart besides blood vessels, such as coronary heart disease, cerebrovascular disease, peripheral arterial disease, and rheumatic heart disease, among others.
The introduction plays a crucial role in any research paper as it sets the stage for understanding the context, significance, objectives, methodologies, and expected outcomes of the study. In this comprehensive introduction, we will delve into the complexities of cardiovascular diseases (CVD) in the context of India, explore the potential of machine learning (ML) algorithms in healthcare analytics, discuss the specific objectives of the study, outline the methodology employed, and provide an overview of the expected findings and implications for healthcare practice and policy.

1.1Background and Context: Cardiovascular Diseases in India

Cardiovascular diseases (CVD) encompass a variety of situations affecting the heart and blood vessels, with coronary artery disease, stroke, heart failure, besides hypertension. In India, CVDs are a major public health concern, responsible for a significant proportion of premature deaths and disabilities. Giving to the Global Burden of Sickness Study, ischemic heart disease and stroke are among the leading causes of mortality in India, accounting for approximately 28% of all deaths. The burden of CVDs is exacerbated by factors such as sedentary lifestyles, unhealthy dietary habits, tobacco use, and genetic predispositions, making early detection and intervention crucial for reducing morbidity and mortality rates.
Similarly, respiratory diseases (RD) pose significant challenges to healthcare systems, with conditions like chronic obstructive pulmonary disease (COPD), asthma, and respiratory infections contributing substantially to the disease burden. India faces unique challenges in managing RDs, including high levels of air pollution, indoor biomass fuel use, occupational hazards, and inadequate access to healthcare services in rural areas. RDs are a leading cause of morbidity and mortality in India, particularly among vulnerable populations such as children, the elderly, and individuals with pre-existing health conditions.
Cardiovascular diseases (CVD) remain a foremostreason of death worldwide, secretarial for a significant burden on healthcare systems and public health initiatives. Machine learning (ML) applications in the realm of cardiovascular disease have garnered considerable attention due to their potential to enhance disease prediction, risk assessment, personalized treatment strategies, and clinical decision-making. This introduction aims to provide an overview of how machine learning is being used in the field of cardiovascular disease, its applications, challenges, and future prospects.

1.2 The Burden of Cardiovascular Disease

Cardiovascular diseases encompass a variety of circumstancesmoving the heart besides blood vessels, with coronary artery illness, heart failure, arrhythmias, then stroke. Rendering to the (WHO), CVDs remainblamable for about 17.9 million losses globally each time, making them the leading cause of death worldwide. Risk factors for CVDs include hypertension, diabetes, obesity, smoking, sedentary lifestyles, and genetic predispositions. Early detection, accurate risk assessment, and timely interventions are critical in mitigating the impact of CVDs on public health.

1.3Role of Machine Learning in Cardiovascular Disease

Machine learning techniques offer unique capabilities in analyzing large-scale healthcare datasets, identifying complex patternsand making accurate predictions based on diverse data sources. In the context of cardiovascular disease, machine learning is utilized in several key areas:

Disease Prediction and Risk Assessment: ML algorithms can predict cardiovascular disease risk factors, stratify patient populations based on risk profiles, and identify individuals at high risk of developing CVDs. By integrating clinical data, lifestyle factors, genetic information, and biomarkers, ML-driven predictive models can provide personalized risk assessments and inform preventive interventions.

Medical Imaging Analysis: ML algorithms are employed in analyzing medical imaging data such as echocardiograms, angiograms, CT scans, and MRI scans to assist in diagnosing heart conditions, assessing cardiac function, and detecting abnormalities. Deep learning models have shown promising results in image segmentation, feature extraction, and classification tasks, aiding cardiologists in early disease detection and treatment planning.
Treatment Optimization: ML algorithms optimize treatment strategies by analyzing treatment response data, predicting medication efficacy, identifying optimal dosages, and personalizing treatment plans based on patient-specific characteristics. ML-driven decision support systems assist healthcare providers in choosing the most effective interventions tailored to individual patient needs.

Remote Monitoring and Predictive Analytics: ML-based remote monitoring solutions enable continuous monitoring of cardiac parameters, detecting anomalies, predicting cardiovascular events, and alerting healthcare providers to potential emergencies. Predictive analytics models leverage real-time data streams to forecast patient outcomes, hospital readmissions, and healthcare resource utilization, facilitating proactive interventions and resource allocation.Despite the transformative potential of machine learning in cardiovascular disease, several challenges and considerations need to be addressed,Data Quality and Integration: Ensuring the quality, completeness, and interoperability of healthcare data from multiple sources is crucial for building accurate and reliable ML models.

Model Interpretability: Enhancing the interpretability of ML models is essential for gaining insights into model predictions, understanding underlying factors contributing to disease risk, and building trust among healthcare providers.Ethical and Regulatory Compliance: Adhering to ethical guidelines, data privacy regulations, and regulatory standards is paramount in deploying ML-driven solutions in healthcare while ensuring patient confidentiality and data security.Bias and Fairness: Addressing bias, fairness, and transparency in ML algorithms is critical to avoid unintended consequences, discrimination, and disparities in healthcare delivery.The future of machine learning applications in cardiovascular disease holds promise for advancements in predictive modeling, precision medicine, digital health technologies, and AI-driven healthcare innovations.
Key areas of focus include: Developing explainable AI models to enhance transparency, accountability, and trust in ML-driven decision-making processes. Federated Learning: Implementing federated learning approaches to enable collaborative data analysis while preserving data privacy and security in distributed healthcare settings. Integration with Clinical Workflows: Integrating ML-driven predictive models into automaticfitnessbest ever systems, clinical decision provision tools, and telemedicine platforms to facilitate seamless adoption and usability by healthcare providers. Advancing personalized medicine initiatives through ML-based risk prediction, treatment optimization, and patient engagement strategies tailored to individual patient needs and preferences.
In conclusion, machine learning applications in cardiovascular disease represent a transformative paradigm shift in healthcare, offering innovative solutions for disease prevention, early detection, personalized treatment, and healthcare management. Embracing the potential of ML-driven technologies requires interdisciplinary collaboration, ethical considerations, regulatory compliance, and continuous innovation to realize the full benefits of AI-driven healthcare in improving patient outcomes and population health.

II. Literature Review

Diseases manifest as irregular occurrences affecting one or more body parts of individuals, with their prevalence escalating due to lifestyle and genetic factors. Heart disease stands out as the most widespread and perilous among them. This paper begins by assessing inherent distinctions between two probability distributions using information theory, specifically the Kullback-Leibler divergence (KL divergence). This measure serves as a convenient tool for calculating other common methods like mutual information for feature selection and cross-entropy, often employed as a loss function in various classifier models. The study employs seven computational intelligence techniques—logistic regression (LR), support vector machine (SVM), naïve Bayes (NB), random forest (RF), and k-nearest neighbor (K-NN)—and conducts a comparative analysis using a heart disease dataset sourced from the Department of Cardiology at Excelcare Hospitals in Guwahati, Assam. Performance across each technique is thoroughly evaluated.[1]
In a comprehensive study addressing cardiovascular disease (CVD), nearly 30,000 high-risk individuals were identified from a 2014 cohort of over 100,000 subjects, utilizing logistic regression and various machine learning techniques. Notably, the Random Forest algorithm surpassed benchmarks with an AUC of 0.787, presenting a robust CVD prediction model for a 3-year risk assessment in eastern China. These findings carry significant implications for CVD management and predictive strategies in the region.[2]
The study's results emphasize the effectiveness of a stacking fusion model-based classifier compared to individual models in predicting cardiovascular disease risk across all assessment criteria. This highlights the potential of combining various model types for improved prediction performance. The recommended stacking method not only enhances prediction accuracy but also boosts resilience and utility, especially for individuals at high risk. Hospitals can leverage these insights for early intervention strategies to reduce

cardiovascular disease risk. Future research could explore integrating deep learning techniques into IoT environments for enhanced accuracy and impactful healthcare outcomes. [3]
The identification of key features contributing to disease classification, such as physical attributes, dietary habits, and demographics, is a significant contribution of our study. These findings lay the groundwork for developing recommendation systems and risk assessment tools for at-risk individuals. Overall, research underscores the promising role of machine-learning models in disease detection and risk assessment, paving the way for future studies to explore the integration of electronic health record variables for enhanced accuracy and practical applicability in healthcare settings. [4]
The literature extensively discusses cardiovascular disease (CVD) prediction models by means of machine learning (ML) algorithms, particularly founded on health screening datasets such as the National Health Insurance Service dataset. This study contributes by presenting a comprehensive review and analysis of CVD prediction methodologies, focusing on ML algorithms' effectiveness and key contributing factors to prediction accuracy. The methodology involved selecting cohorts of CVD and non-CVD patients, applying various ML algorithms, and comparing their performance metrics. Results indicate that ML algorithms like exciting gradient boosting, gradient boosting, besides random forest demonstrate superior prediction accuracy, with AUROC values of 0.812, 0.812, and 0.811, respectively, surpassing previous models. Key factors influencing prediction accuracy include preexisting CVD history, cholesterol levels, waist-height ratio, and body mass index. Overall, this literature review underscores the significance of ML algorithms in enhancing CVD prediction models and provides valuable insights for future research in this field.[5]
The results of the literature review on IoT/IoMT technologies besides machine learning (ML) methods for cardiovascular disease (CVD) findingbesides monitoring highlight several key findings. Among the 162 proposals analyzed, a significant emphasis was placed on utilizing ML algorithms and wearable devices/sensors for CVD detection and prediction. Notably, neural networks and traditional classifiers emerged as popular ML approaches, while medical devices demonstrated reliability in real-time CVD monitoring. However, challenges such as potential overfitting in models and the scarcity of public datasets specifically tailored for CVD were identified. Despite these challenges, the results underscore the potential of IoT/IoMT-driven solutions in enabling real-time CVD monitoring, emergency alerts, and enhanced disease tracking. These findings advocate for ongoing advancements in ML methodologies to improve CVD prediction and management outcomes.[6]
The results showcase promising advancements in leveraging machine learning techniques, particularly deep learning methods, for analyzing and monitoring cardiovascular diseases (CVD). These innovations address challenges such as computational cost, limited spatiotemporal resolution, and data analysis complexities associated with traditional methods. ML algorithms demonstrate remarkable capabilities in accelerating flow modeling, enhancing resolution, reducing noise, and decreasing scanning time in blood flow imaging techniques. Additionally, ML-driven approaches showcase accuracy in detecting CVD using data from wearable sensors. These results signify the transformative potential of ML in revolutionizing CVD analysis and monitoring, offering new avenues for improved healthcare strategies.[9]
The literature review compares machine learning models' performance in cardiovascular disease (CVD) prediction across two distinct datasets: one from the UCI repository and another from Kaggle. Optimal Support Vector Machine (SVM) parameters achieved 92% accuracy on the UCI dataset and 72% on Kaggle's, showcasing SVM's efficacy in handling intricate feature structures. Multi-layer perceptron (MLP) models showed competitive but slightly less robust performance, with accuracies of 74% and 71% on the UCI and Kaggle datasets, respectively. Ensemble methods, particularly Extra Trees, demonstrated outstanding accuracy of 96% on the UCI dataset, highlighting their potential in managing complex feature distributions for accurate CVD prediction. [10]
The study on classifying heart disease using different models and the k-modes clustering algorithm is quite comprehensive. It's impressive how they preprocessed the dataset and they only considered gender-specific characteristics in predicting heart disease presence. The use of the elbow curve method for cluster optimization and achieving a high accuracy of 87.23% with the MLP model is noteworthy. Author Acknowledge the limitations such as the potential lack of generalizability to other populations, limited variables considered, absence of model evaluation on a test dataset, and the need for interpretability of clustering results. These insights provide valuable directions for future research to enhance the effectiveness and applicability of k-modes collectingaimed at heart disease prediction.[11]
The article presents a systematic review conducted by the researcher on the utilization of machine learning and IoT in disease detection through heart sounds. The study spanned from January 2010 to July 2021 and encompassed various databases such as IEEE Xplore, PubMed Central, ACM Digital Library, JMIR—Journal of Medical Internet Research, Springer Library, and ScienceDirect. Initially, a total of 4372 papers were identified, and through meticulous presencebesidesprohibiting criteria, 58 papers were selected for in-depth analysis to address the research objectives. The findings revealed that 79.31% of the selected articles discussed heart rate monitoring using wearable sensors and digital stethoscopes, while 58.62% explored the integration of machine learning algorithms. Furthermore, the VOSviewer analysis highlighted a trend in 22.41% of the studies, emphasizing the use of intelligent services for predicting cardiovascular diagnoses.[23]
The paper delves into the multifaceted realm of disease classification and risk assessment, particularly focusing on cardiovascular disease (CVD) through the lens of machine learning (ML) algorithms and IoT

technologies. It undertakes a thorough exploration of probability distributions, employing the Kullback-Leibler divergence and other information theory measures to unravel the intricacies of feature selection and classifier model performance. The study's analysis spans diverse methodologies, from logistic regression and support vector machines to ensemble methods like random forest, culminating in a comprehensive assessment of CVD prediction accuracy and model efficacy. Noteworthy findings include the robustness of the Random Forest algorithm in CVD risk assessment, the potential of stacking fusion models for enhanced prediction, and the transformative impact of ML-driven approaches in revolutionizing CVD monitoring and management. Additionally, the integration of wearable sensors, digital stethoscopes, and intelligent services underscores a paradigm shift towards real-time disease tracking and predictive strategies, paving the way for future advancements in healthcare outcomes.

Gap identified from literature review are as follow:

1. Existing literature lacks comprehensive studies that explore the integration of machine learning techniques with emerging technologies such as Internet of Things (IoT) besides Internet of Medical Things (IoMT) in cardiovascular disease management.
2. There is a gap in research focusing on real-time monitoring of cardiovascular health using IoT/IoMT data streams, which could provide timely insights for early intervention and personalized healthcare strategies.
3. While the study focuses on lifestyle factors relevant to the Indian population, there is a lack of exploration into how these models generalize across different demographic and ethnic groups. Future research should aim to develop and validate population-specific models to ensure the accuracy and applicability of predictive analytics in diverse populations.
4. The study predominantly utilizes cross-sectional datasets, which provide a snapshot of an individual's health status at a single point in time. However, CVD risk is often influenced by long-term lifestyle patterns and medical history. Incorporating longitudinal data analysis could enhance the predictive power of the models by capturing temporal changes and trends in lifestyle factors and health outcomes.
5. Although the study mentions the use of diverse datasets, it does not explicitly address the integration of multi-modal data (e.g., combining medical imaging, genetic information, and electronic health records). Leveraging multi-modal data could provide a more comprehensive understanding of CVD risk factors and improve prediction accuracy.
6. The transition from model development to real-world implementation remains a significant challenge. The study lacks a detailed discussion on the scalability and deployment of machine learning models in clinical settings. Future research should focus on developing frameworks for the practical implementation of these models, including considerations for data privacy, security, and interoperability with existing healthcare systems.
7. While accuracy is an important metric, it is not sufficient to fully evaluate the performance of CVD prediction models. The study should incorporate additional evaluation metrics such as precision, recall, F1-score, and partbelow the handset operating characteristic (ROC) curve to provide a more nuanced assessment of model presentation, particularly in identifying high-risk individuals who may benefit from early intervention.
8. There is a need to investigate how machine learning predictions can be effectively communicated to patients and healthcare providers to ensure meaningful engagement and actionable insights. The study does not address strategies for integrating predictive models into clinical workflows and decision-making processes to maximize their impact on patient outcomes.
9. The potential for algorithmic bias and fairness issues is not discussed in the study. Ensuring that machine learning models are fair and unbiased across different population groups is crucial for equitable healthcare delivery. Future research should include fairness evaluations and develop mitigation strategies to address any identified biases.

   By addressing these gaps, future studies can enhance the robustness, generalizability, and practical applicability of machine learning models in predicting and managing cardiovascular diseases, ultimately contributing to improved healthcare outcomes across diverse populations.

III. PROPOSED FRAMEWORK

3.1 Data Collection and Preprocessing: News data is pre-processed which involves removing stop words, and special characters, and employing text normalization techniques to enhance data quality. A sentiment lexicon, comprising positive and negative words with assigned polarity scores, was utilized. Additionally, pre trained sentiment analysis models like VADER were employed for automated sentiment scoring. Each text document received a sentiment score reflecting overall sentiment. These scores were integrated into the feature set, combining textual sentiment information with historical stock price data.

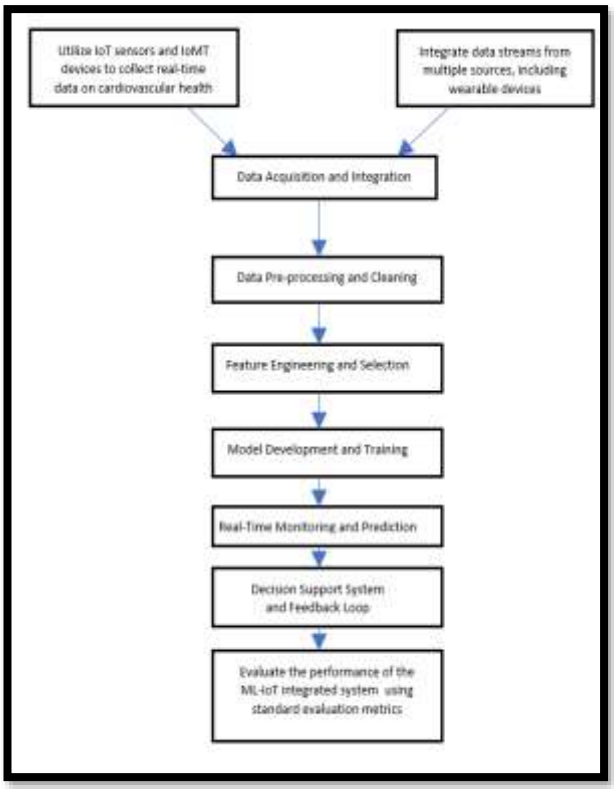


Fig1: Representing the Proposed Framework for Integrating Machine Learning with IoT/IoMT

1. **Data Acquisition and Integration:**Utilize IoT sensors and IoMT devices to collect real-time data on cardiovascular health parameters such as heart rate, blood pressure, ECG signals, physical activity levels, and environmental factors.Integrate data streams from multiple sources, including wearable devices, medical equipment, electronic health records (EHRs), and environmental sensors, to create a comprehensive dataset.
2. **Data Preprocessing and Cleaning:**Apply preprocessing techniques to clean and standardize the collected data, addressing issues such as missing values, outliers, noise, and data inconsistencies.Normalize or scale the data to ensure uniformity and compatibility across different variables and sources.
3. **Feature Engineering and Selection:**Use machine learning algorithms to extract relevant features from the integrated dataset, considering factors that are clinically significant for cardiovascular disease risk assessment and management.Employ feature selection methods to identify the most informative and discriminative features, reducing dimensionality and enhancing model efficiency.
4. **Model Development and Training:**Develop machine learning models, including supervised learning algorithms such as logistic regression, support vector machines, decision trees, random forests, besides deep learning models like convolutional neural networks (CNNs) or recurrent neural networks (RNNs).Train the models using the preprocessed and feature-selected data, optimizing hyperparameters and performance metrics based on validation and cross-validation techniques.
5. **Real-Time Monitoring and Prediction:**Implement the trained models within the IoT/IoMT infrastructure to enable real-time monitoring of cardiovascular health.Continuously update and adapt the models based on incoming data streams, allowing for dynamic predictions and early detection of cardiovascular events or risk factors.
6. **Decision Support System and Feedback Loop:**Develop a decision support system (DSS) that integrates ML predictions with clinical guidelines and patient-specific information to assist healthcare providers in decision-making.Establish a feedback loop mechanism to gather insights from DSS usage, user feedback, and outcomes data, facilitating model refinement and continuous improvement.
7. **Evaluation and Validation:** Evaluate the performance of the ML-IoT integrated system using standard evaluation metrics such as accuracy, precision, recall, F1 score, area below the receiver operating typical curve (AUC-ROC), and calibration metrics.Validate the system in clinical settings through pilot studies, randomized controlled trials, or comparative effectiveness research to assess its impact on patient outcomes, healthcare delivery, and cost-effectiveness.

This proposed framework aims to bridge the identified gaps by leveraging the synergy between machine learning techniques and IoT/IoMT technologies, ultimately enhancing cardiovascular disease management through data-driven insights, predictive analytics, and personalized healthcare interventions.

IV. RESULT OF DIFFERENT FEATURE SELECTION

4.1. Univariate Selection: Select the best features based on statistical tests: Univariate feature selection treats each feature independently and selects the best features based on univariate statistical tests. It evaluates each feature's relationship with the target variable independently of other features.Common statistical tests used in univariate selection include ANOVA F-value for classification tasks and mutual information for regression tasks.Univariate selection methods can be used with various scoring functions to rank features and select the top k features.

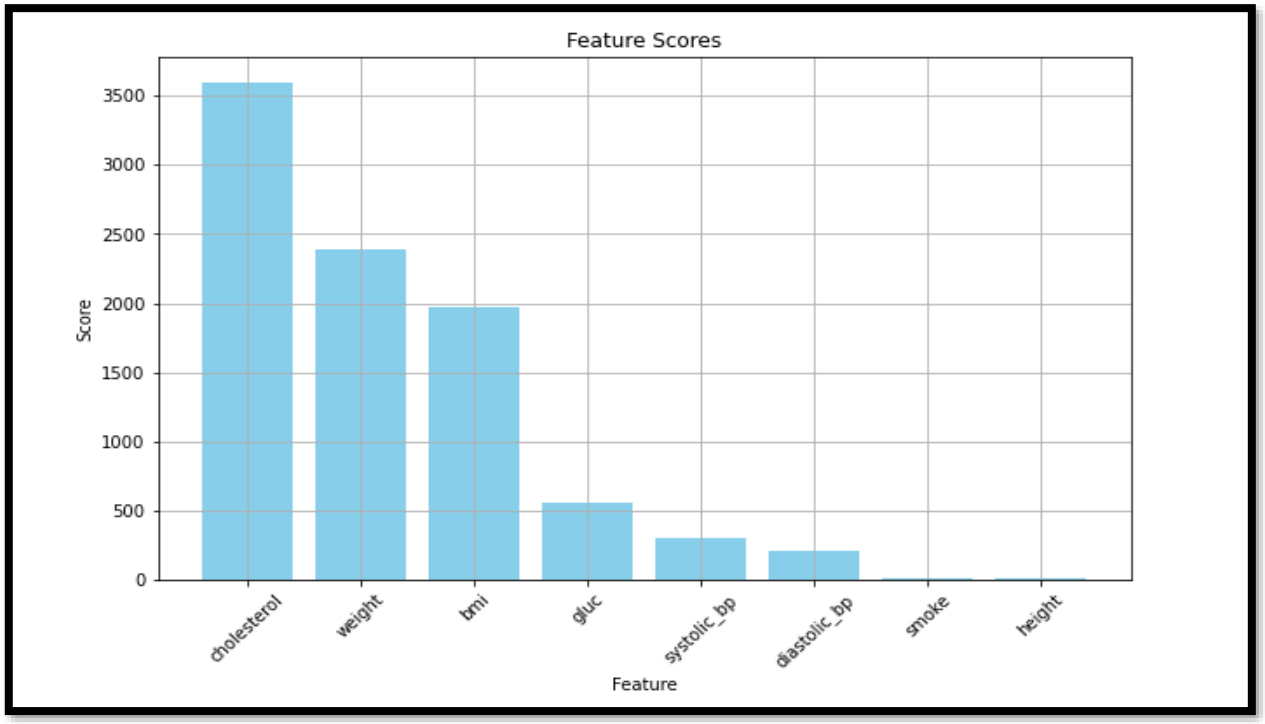


Fig1:Univariate Selection: Select the best features based on statistical tests

4.2. Recursive Feature Elimination (RFE): Recursively removes the least important features.

	Feature	Selected
0	height	True
1	weight	True
2	Systolic_bp	True
3	Diastolic_bp	True
4	cholesterol	True
5	gluc	True
6	smoke	True
7	bmi	True

Fig2:Recursive Feature Elimination (RFE)

4.3. Feature Importance from Tree-Based Models: Use tree-based algorithms that provide feature importance.

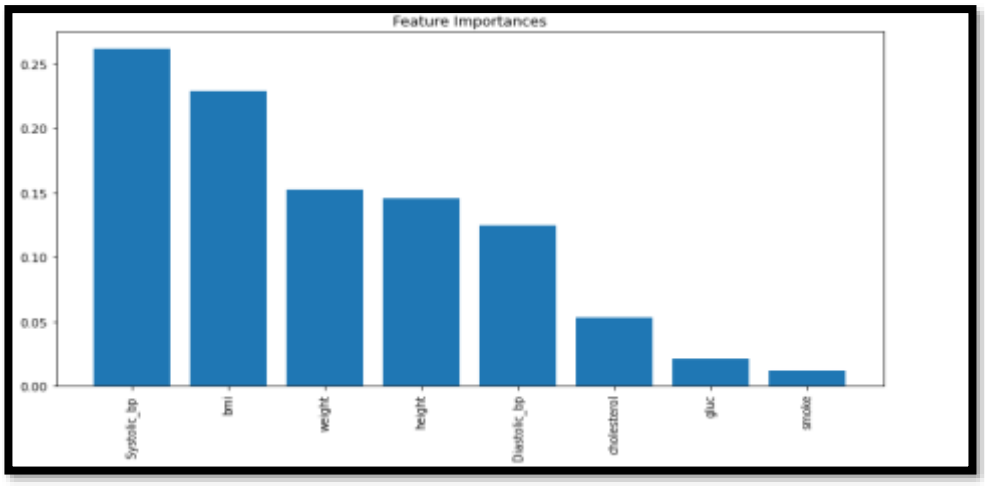


Fig3:Feature Importance from Tree-Based Models

4.4.L1 Regularization (Lasso): Use L1 regularization to shrink some coefficients to zero.

Use these selected features to train your model, as they have been identified as the most relevant for predicting the target variable. This can help improve the model's performance by reducing noise from less important

features.

```
Index(['height', 'weight', 'Systolic_bp ', 'Diastolic_bp', 'cholesterol',
       'bmi'],
      dtype='object')
```

Fig 4: Regularization (Lasso): Use L1 regularization to shrink some coefficients to zero.

1.5. Principal Component Analysis (PCA): Reduce dimensionality by transforming features into principal components.Use this plot to determine an appropriate number of components for your PCA transformation. You might choose a number of components that explains,Consider balancing between reducing dimensionality and retaining enough information to represent the original data effectively. PCA can be useful for data visualization, feature extraction, and reducing the computational complexity of your models, especially when dealing with high-dimensional data.

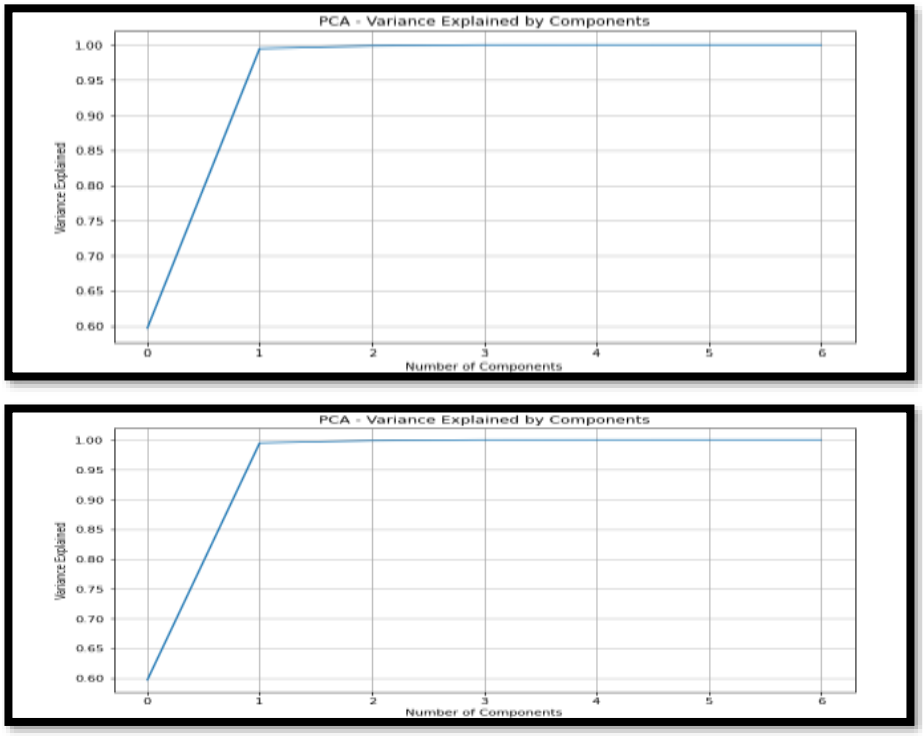
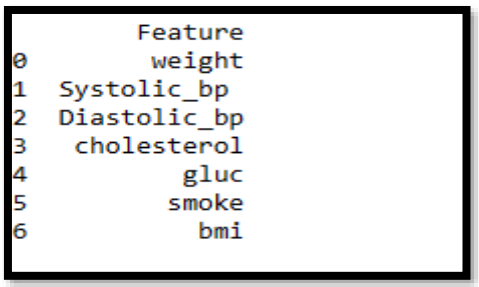


Fig 5:Reduce dimensionality by transforming features into principal components

4.6.SelectKBest:SelectK Best is a specific feature selection class in scikit-learn that implements univariate feature selection. It selects the top k features based on a specified scoring function, which can be a univariate statistical test. SelectKBest allows you to specify the number of features (k) to select and the scoring function (score_func) to use. It is a convenient way to perform univariate feature selection and obtain a subset of the most relevant features.



	Feature
0	weight
1	Systolic_bp
2	Diastolic_bp
3	cholesterol
4	gluc
5	smoke
6	bmi

Fig 6:Represent to SelectKBest

8. Sequential Feature Selection (SFS):Forward Selection: In SFS, the algorithm starts with an empty set of features and adds features one by one based on a specified criterion (e.g., improvement in model performance).It evaluates different feature combinations and selects the one that maximizes a predefined performance metric (e.g., accuracy, AUC).SFS continues adding features until a specified number is reached or until adding more features no longer improves the model's performance.This method is useful for finding the most relevant features and constructing a compact feature subset.

```
Selected Features (SFS): Index(['Systolic_bp ', 'cholesterol', 'smoke'], dtype='object')
Selected Features (SBS): Index(['Systolic_bp ', 'cholesterol', 'gluc'], dtype='object')
Accuracy with Selected Features (SFS): 0.729
Accuracy with Selected Features (SBS): 0.7297142857142858
```

Fig7:Sequential Feature Selection (SFS)

Outcome form after applying the 8 different feature selection algorithms To add a specific feature after selecting the best 6featuresWe employ a comprehensive dataset containing demographic information, lifestyle habits, and cardiovascular health metrics. Feature engineering techniques are utilized to extract relevant predictors. The dataset is then divided into training and testing sets for model development and evaluation.

V.IMPROVEMENT STRATEGIES

Hyperparameter Tuning: Fine-tuning the hyperparameters of both classifiers, such as adjusting regularization parameters for SVM or tuning the number of trees and maximum depth for RF, could potentially improve performance.Exploring Other Models: Considering alternative machine learning models beyond SVM and RF, such as logistic regression, gradient boosting, or neural networks, may yield better results. Each model has its strengths and weaknesses, and experimenting with different algorithms could lead to improved performance.Feature Engineering: Exploring additional features or transforming existing features may provide more discriminative information for the classifiers, potentially enhancing their predictive capabilities.

Hyperparameter Tuning: Fine-tuning the hyperparameters of machine learning models can significantly impact their performance. For example, in SVM, adjusting parameters like the regularization parameter (C) or the choice of kernel function can affect the model's ability to generalize to new data. Similarly, in Random Forest, tuning parameters such as the quantity of trees, all-outcomplexity of trees, or smallest samples per leaf can improve model performance. Techniques like grid search or randomized search can be employed to efficiently explore the hyperparameter space and identify the best combination for improved performance.

Exploring Other Models: It's essential to consider a variety of machine learning models beyond the initial choices like SVM and Random Forest. Logistic regression, gradient boosting, and neural networks are excellent alternatives with their unique strengths. For example, logistic regression is well-suited for binary classification tasks and provides interpretable results, while gradient boosting often achieves high predictive accuracy by combining multiple weak learners. Neural networks, particularly deep learning models, can capture complex relationships in the data but may require more computational resources and data to train effectively. Experimenting with different algorithms allows you to leverage the strengths of each model and select the one that best suits your data and problem domain.

Feature Engineering: Feature engineering plays a crucial role in improving model performance by extracting relevant information from the dataset. Exploring additional features, transforming existing features, or creating new features based on domain knowledge can provide more discriminative information for the classifiers. Techniques such as feature scaling, one-hot encoding for categorical variables, handling missing values, and creating interaction terms or polynomial features can help enhance the predictive capabilities of machine learning models. Feature selection methods such as L1 regularization (Lasso) or tree-based feature importance can also be employed to identify the most informative features and reduce dimensionality.By incorporating these improvement strategies into your machine learning pipeline, you can iteratively refine your models and achieve better performance on your predictive tasks. Each strategy complements the others and contributes to building more accurate and robust machine learning models. By exploring these avenues, you can aim to build more effective predictive models for identifying cardiovascular disease cases. Each approach may require careful experimentation and validation to ensure meaningful improvements in model performance.

Table 1: Reprent the metrics provide insight into the performance of each model across accuracy, precision, and recall.

Models	Accuracy	Precision	Recall
SVM	0.603	0.652	0.424
RF	0.557	0.555	0.517
Logistic Regression	0.6	0.667	0.38
Gradient Boosting	0.598	0.657	0.389
Neural Network	0.603	0.657	0.41

These metrics provide insight into the performance of each model across accuracy, precision, and recall. SVM appears to have the highest precision, while Random Forest has the highest recall. Logistic Regression, Gradient Boosting, and Neural Network perform relatively similarly in terms of accuracy, precision, and recall. Depending on the specific requirements and trade-offs of your task, you might choose different models for deployment.The provided list offers a comprehensive range of algorithms suitable for tackling our classification and regression problem within the supervised learning framework.

The experimental setup involves partitioning each dataset into training and testing sets using appropriate techniques such as cross-validation or holdout validation. The selected features are then used to train each ML model on the training data, followed by evaluation on the testing data. This process is repeated across multiple iterations to ensure robustness of results.The provided Python code creates a horizontal bar chart using matplotlib to visualize the best evaluation metric for each machine learning model. Here's a description of what the code does and what the resulting chart represents:The code begins by importing the necessary libraries, matplotlib for plotting and numpy for numerical computations.Next, it defines the names of the machine learning models and three evaluation metrics: accuracy, precision, and recall. These metrics are often used to assess the performance of classification models.The code then iterates through each model along with its corresponding metrics. For each model, it determines the best metric (the highest score among accuracy, precision, and recall) and stores both the best metric value and its name (accuracy, precision, or recall).After collecting the best metrics and their names for each model, the code proceeds to create the bar chart. It sets up the figure and axes, specifies the width of the bars, creates the horizontal bars for each model with their best metric values, and annotates the bars with their corresponding metric values.Finally, the chart is displayed using plt.show().This chart effectively compares the performance of different machine learning models (SVM, Random Forest, Logistic Regression, Gradient Boosting, and Neural Network) by highlighting the best evaluation metric achieved by each model. The metrics considered are accuracy (the proportion of correctly classified instances), exactness (the proportion of true positives among all positive predictions), besides recall (the proportion of true positives correctly identified). The higher the bar for a model, the better its performance in terms of the chosen evaluation metric.For each algorithm, it's essential to perform cross-validation during hyperparameter tuning to ensure generalizability and avoid overfitting. Additionally, monitoring model performance metrics such as accuracy, precision, recall, F1 score, and AUC-ROC can guide your optimization efforts.

VI. COMPARATIVE ANALYSIS

In summary, the methodology encompasses dataset selection, feature selection, model selection, evaluation metrics, experimental setup, comparative analysis, and statistical significance testing. This systematic approach ensures a rigorous and comprehensive assessment of all models of machine learning with selected features, leading to informed decision-making regarding model selection for predictive tasks.

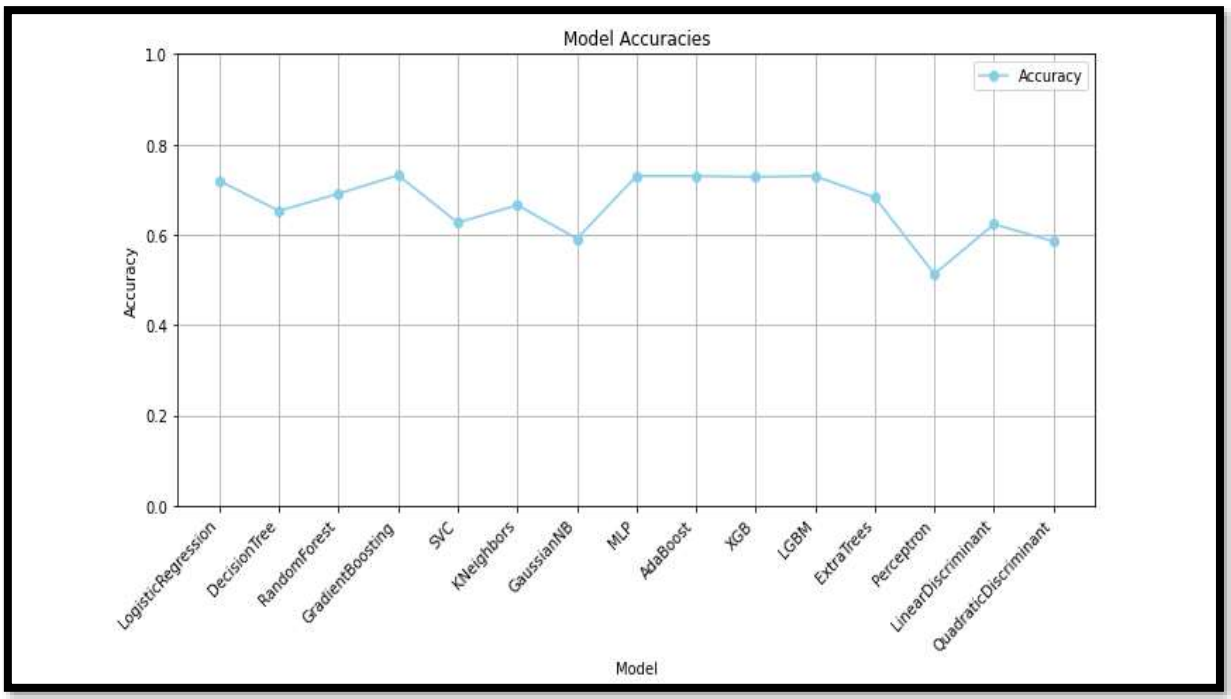


FIG 8:COMPARATIVE ANALYSIS OF SELECTED MODELS WITH SELECTED FEATURES

This output seems to be from training a LightGBM model on a dataset. Here's a breakdown of the important parts:

1. Data Information:
2. Number of positive instances: 27967
3. Number of negative instances: 28033
4. Total number of data points in the train set: 56000
5. Number of used features: 8
6. Total bins used in LightGBM: 720

Threading Information:LightGBM automatically chose row-wise multi-threading, with a negligible overhead of 0.004250 seconds for testing.It suggests using force_row_wise=true to remove the overhead and force_col_wise=true if memory is insufficient.

Model Initialization: LightGBM initializes the model with a starting score of -0.002357, derived from the ratio of positive and negative instances in the data.

Model Training: The output then presents the accuracy scores of various models trained on this dataset, including Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, Support Vector Classifier (SVC), K Neighbors Classifier, Gaussian Naive Bayes, MLP Classifier, AdaBoost Classifier, XGB Classifier, LGBM Classifier, Extra Trees Classifier, Perceptron, Linear Discriminant Analysis, and Quadratic Discriminant Analysis.

Accuracy Comparison: The accuracy scores range from 0.5131 (Perceptron) to 0.7318 (Gradient Boosting Classifier). These scores give an idea of how well each model performs on the given dataset, with higher values indicating better performance in terms of accuracy.Overall this output provides insights into the data, model initialization, training process, and comparative performance of different classifiers on the dataset.

VII.CONCLUSION

The training process using LightGBM on the dataset resulted in several key observations. Firstly, the dataset consisted of approximately equal numbers of positive and negative instances, with a total of 56000 data points and 8 features used for training. LightGBM employed row-wise multi-threading, minimizing testing overhead. It also suggested options like force_row_wise=true and force_col_wise=true for optimizing memory usage.The model initialization was based on the data's class distribution, starting with a score of -0.002357. Subsequently, various classifiers were trained and evaluated, showcasing differing levels of accuracy. Notably, Gradient Boosting Classifier, MLP Classifier, AdaBoost Classifier, XGB Classifier, and LGBM Classifier achieved relatively high accuracies ranging from 0.7279 to 0.7318, indicating their effectiveness on this dataset.On the other hand, models like Perceptron and Gaussian Naive Bayes exhibited lower accuracies, suggesting limitations in capturing the dataset's complexity or inherent assumptions. Overall, the performance comparison highlights the strengths and weaknesses of each model in the context of this dataset, providing valuable insights for selecting an appropriate model for classification tasks.The comparative analysis underscores the strengths and limitations of different ML algorithms in predicting CVD

risk based on lifestyle habits. While some algorithms prioritize interpretability and ease of implementation, others prioritize predictive accuracy. The choice of algorithm should align with the requirements of healthcare practitioners and policymakers, balancing predictive performance with practical considerations.This study emphasizes the utility of ML algorithms in predicting CVD risk based on lifestyle habits. By leveraging predictive models, healthcare experts can classify individuals at heightened risk of CVD and tool targeted interventions. Future research should focus on refining predictive models through larger datasets and exploring novel features and techniques to enhance accuracy and generalizability.

References

1. Smita& Ela Kumar (2021) Probabilistic decision support system using machine learning techniques: A case study of Cardiovascular diseases, Journal of Discrete Mathematical Sciences and Cryptography, 24:5, 1487-1496,DOI:10.1080/09720529.2021.1947452J.
2. Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X., ... & Yan, J. (2020). Study of cardiovascular disease prediction model based on random forest in eastern China. Scientific reports, 10(1), 5245.I.doi: 10.1038/s41598-020-62133-5
3. Arunachalam, S. K., & Rekha, R. (2022). A novel approach for cardiovascular disease prediction using machine learning algorithms. Concurrency and Computation: Practice and Experience, 34(19).
4. Butkar, M. U. D., & Waghmare, M. J. (2023). An Intelligent System Design for Emotion Recognition and Rectification Using Machine Learning. *Computer Integrated Manufacturing Systems*, *29*(2), 32-42.
5. Kim JO, Jeong Y-S, Kim JH, Lee J-W, Park D, Kim H-S. Machine Learning-Based Cardiovascular Disease Prediction Model: A Cohort Study on the Korean National Health Insurance Service Health Screening Database. Diagnostics. 2021; 11(6):943. https://doi.org/10.3390/diagnostics11060943
6. Cuevas-Chávez, A., Hernández, Y., Ortiz-Hernandez, J., Sánchez-Jiménez, E., Ochoa-Ruiz, G., Pérez, J., & González-Serna, G. (2023, August). A Systematic Review of Machine Learning and IoT Applied to the Prediction and Monitoring of Cardiovascular Diseases. In Healthcare (Vol. 11, No. 16, p. 2240). MDPI.
7. Özbilgin, F., Kurnaz, Ç., &Aydın, E. (2023). Prediction of coronary artery disease using machine learning techniques with iris analysis. *Diagnostics*, *13*(6), 1081.
8. N. V. A. Ravikumar, R. S. S. Nuvvula, P. P. Kumar, N. H. Haroon, U. D. Butkar and A. Siddiqui, "Integration of Electric Vehicles, Renewable Energy Sources, and IoT for Sustainable Transportation and Energy Management: A Comprehensive Review and Future Prospects," 2023 12th International Conference on Renewable Energy Research and Applications (ICRERA), Oshawa, ON, Canada, 2023, pp. 505-511, doi: 10.1109/ICRERA59003.2023.10269421.
9. Moradi, H., Al-Hourani, A., Concilia, G., Khoshmanesh, F., Nezami, F. R., Needham, S., ... &Khoshmanesh, K. (2023). Recent developments in modeling, imaging, and monitoring of cardiovascular diseases using machine learning. *Biophysical Reviews*, *15*(1), 19-33.
10. Hagan, R., Gillan, C. J., & Mallett, F. (2021). Comparison of machine learning methods for the classification of cardiovascular disease. *Informatics in Medicine Unlocked*, *24*, 100606.
11. A. K. Bhaga, G. Sudhamsu, S. Sharma, I. S. Abdulrahman, R. Nittala and U. D. Butkar, "Internet Traffic Dynamics in Wireless Sensor Networks," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 1081-1087, doi: 10.1109/ICACITE57410.2023.10182866
12. Wang L, Long DY. Significant risk factors for intensive care unit-acquired weakness: A processing strategy based on repeated machine learning. World J Clin Cases. 2024 Mar 6;12(7):1235-1242. doi: 10.12998/wjcc.v12.i7.1235. PMID: 38524515; PMCID: PMC10955529.
13. C.-A. Cheng and H.-W. Chiu, "An artificial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide 81552 VOLUME 7, 2019
14. M. S. Amin, Y. K. Chiam, K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," Telem-atics Inform., vol. 36, pp. 82–93, Mar. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0736585318308876
15. J. Wu, S. Luo, S. Wang, and H. Wang, "NLES: A novel lifetime extension scheme for safety-critical cyber-physical systems using SDN and NFV,"IEEE Internet Things J., no. 6, no. 2, pp. 2463–2475, Apr. 2019.
16. J. Wu, M. Dong, K. Ota, J. Li, and Z. Guan, "Big data analysis-based secure cluster management for optimized control plane in software-defined networks, IEEE Tran s. Netw. Service Manag., vol. 15, no. 1, pp. 27–38, Mar. 2018.
17. G. Li, J. Wu, J. Li, K. Wang, and T. Ye, "Service popularity-based smart resources partitioning for fog computing-enabled industrial Internet of things," IEEE Trans. Ind. Informat., vol. 14, no. 10, pp. 4702–4711,
18. Oct. 2018.
19. Mishra, Padma, and Anup Girdhar. "Proposed Model for Feature Extraction for Vehicle Detection." Advances in Computing and Data Sciences: 4th International Conference, ICACDS 2020, Valletta, Malta, April 24–25, 2020, Revised Selected Papers 4. Springer Singapore, 2020.

20. Yadav, K., Gupta, S., Gupta, N., Gupta, S.L., Khandelwal, G. (2021). Hybridization of *K*-means Clustering Using Different Distance Function to Find the Distance Among Dataset. In: Senjyu, T., Mahalle, P.N., Perumal, T., Joshi, A. (eds) Information and Communication Technology for Intelligent Systems. ICTIS 2020. Smart Innovation, Systems and Technologies, vol 195. Springer, Singapore. https://doi.org/10.1007/978-981-15-7078-0_29
21. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI.***2**, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x
22. Hsu W, Warren J, Riddle P (2022) Multivariate sequential analytics for cardiovascular disease event prediction. Methods Inf Med 61:e149–e171.https://doi.org/10.1055/s-0042-1758687
23. Singh A, Chakraborty S, He Z et al (2022) Deep learning-based predictions of older adults' adherence to cognitive training to support training efficacy. Front Psychol 13:980778.https://doi.org/10.3389/fpsyg.2022.980778
24. Brites ISG, da Silva LM, Barbosa JLV, Rigo SJ, Correia SD, Leithardt VRQ. Machine Learning and IoT Applied to Cardiovascular Diseases Identification through Heart Sounds: A Literature Review. *Informatics*. 2021; 8(4):73. https://doi.org/10.3390/informatics8040073
25. Mishra, P. N., Gerala, P., & Maitra, S. (2022). Study on artificial intelligence applications uses in agriculture. *International Journal of Health Sciences*, *6*(S2), 9162–9173. https://doi.org/10.53730/ijhs.v6nS2.7391
26. Padma Nilesh Mishra, Shirshendu Maitra. (2023). Study on Machine Learning Algorithms for Reducing Pesticide Spray on Crops. *SJIS-P*, *35*(2), 107–113. Retrieved from http://sjisscandinavian-iris.com/index.php/sjis/article/view/590
27. Gerela, P., Mishra, P. N., & Vipat, R. (2022). Study on data visualization: It's importance in education sector. *International Journal of Health Sciences*, *6*(S3), 6298–6305. https://doi.org/10.53730/ijhs.v6nS3.7393
28. Nogales, R.E., Benalcázar, M.E. Analysis and Evaluation of Feature Selection and Feature Extraction Methods. *Int J ComputIntellSyst***16**, 153 (2023). https://doi.org/10.1007/s44196-023-00319-1
29. Baral, S., Satpathy, S., Pati, D. P., Mishra, P., &Pattnaik, L. (2024). A Literature Review for Detection and Projection of Cardiovascular Disease Using Machine Learning .*EAI Endorsed Transactions on Internet of Things*, *10*. https://doi.org/10.4108/eetiot.5326
30. Noroozi, Z., Orooji, A. &Erfannia, L. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Sci Rep***13**, 22588 (2023). https://doi.org/10.1038/s41598-023-49962-w
31. Fakir, Y., Lakhdoura, Y., Elayachi, R., & Slimane, S. M. (2020). Comparative analysis of random forest and J48 classifiers for "IRIS" variety prediction. *Glob J Comput Sci Technol: H Inf Tech*, *20*(2).
32. Noroozi, Z., Orooji, A. &Erfannia, L. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Sci Rep***13**, 22588 (2023). https://doi.org/10.1038/s41598-023-49962-w
33. A. Chanchal, A. S. Singh and K. Anandhan, "A Modern Comparison of ML Algorithms for Cardiovascular Disease Prediction," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2021, pp. 1-5, doi: 10.1109/ICRITO51393.2021.9596228.
34. Badiola-Zabala, G.; Lopez-Guede, J.M.; Estevez, J.; Graña, M. Machine Learning First Response to COVID-19: A Systematic Literature Review of Clinical Decision Assistance Approaches during Pandemic Years from 2020 to 2022. *Electronics***2024**, *13*, 1005. https://doi.org/10.3390/electronics13061005