ISSN : 2350-0743

www.ijramr.com

International Journal of Recent Advances in Multidisciplinary Research

Vol. 10, Issue 10, pp.9084-9089, October, 2023

RESEARCH ARTICLE

MACHINE LEARNING-BASED PREDICTIVE MODELS FOR CARDIOVASCULAR RISK ASSESSMENT IN DATA ANALYSIS, MODEL DEVELOPMENT, AND CLINICAL IMPLICATIONS

Dharshika Singarathnam, *SwathiGanesan, Sangita Pokhrel and Nalinda Somasiri

Department of Computer Science, York St John University, United Kingdom

ARTICLE INFO

Article History:

Received 25th July, 2023
Received in revised form
28th August, 2023
Accepted 29th September, 2023
Published online 30th October, 2023

Key Words:

Cardiovascular Diseases, Machine Learning, CVD risk level, Random Forest Classifier, Risk Assessment.

ABSTRACT

Cardiovascular diseases (CVDs) remain a leading global cause of morbidity and mortality. Timely identification of individuals at risk is paramount for effective interventions and prevention. This study endeavors to develop machine learning approaches for predicting the initial cardiovascular risk level analyzing the dataset encompassing patient demographics, medical history, lifestyle factors, and clinical indicators. Patient characteristics, including age, gender, diabetes or hypertension presence, smoking status, and physical activity level, along with medical indicators such as blood pressure, cholesterol, and glucose levels, are considered. Diverse machine learning algorithms—logistic regression, decision tree classifier, random forests, linear SVC, naive bayes, and neural network—are employed to train and optimize predictive models. Evaluation metrics (accuracy, precision, recall, F1 score, and AUC-ROC) assess model performance. Accurate risk prediction models hold significance in aiding healthcare decisions, optimizing resource allocation, and enhancing patient outcomes. Identifying high-risk individuals early enables preventive strategies and personalized interventions, reducing the CVD burden. Study objectives encompass dataset preprocessing, exploratory analysis, feature selection and engineering, model training and optimization, and performance evaluation. Findings contribute to cardiovascular risk prediction, presenting a robust model for accurate risk assessment and improved patient outcomes.

INTRODUCTION

Cardiovascular diseases (CVDs) posed a significant global health burden, responsible for numerous deaths and disabilities worldwide (Roth *et al*., 2017). According to the World Health Organization (WHO), CVDs stood as the leading cause of mortality globally, with an estimated 17.9 million deaths in 2019 (World Health Organization, 2023). These encompassed conditions such as coronary artery disease, stroke, heart failure, and peripheral artery disease, necessitating early identification for effective management and prevention strategies (Wang *et al*., 2011). The research problem addressed in this study was the imperative to develop machine learning approaches for accurately predicting the initial cardiovascular risk level based on relevant patient characteristics and medical indicators. The objective was to leverage machine learning techniques to enhance early identification and risk stratification, enabling targeted interventions and personalized healthcare strategies. Accurate prediction models for the initial cardiovascular risk level provided healthcare professionals with valuable tools for risk assessment, optimizing resource allocation and delivering personalized care (Greenland *et al*., 2010). These models had implications for clinical practice, public health planning, and economic considerations by enabling early identification, targeted interventions, and cost savings associated with managing advanced stages of CVDs (Vallejo-Torres *et al*., 2014).

The study aimed to construct machine learning models using patient characteristics and medical indicators to forecast the initial cardiovascular risk (D'Agostino *et al*., 2008). Specific patient characteristics and medical indicators, including demographic information, medical history, lifestyle factors, and clinical indicators, were considered (Lloyd-Jones *et al*., 2010). Various machine learning algorithms, such as Logistic regression, Decision tree, Random forests, Linear SVC, Naïve bayes, and neural network, were employed, and the models' performance was evaluated using relevant metrics (D'Agostino *et al*., 2008). The study's objectives included pre-processing, feature selection, model training, evaluation, and providing insights for the potential use of developed models in clinical practice (D'Agostino *et al*., 2008). The ultimate goal was to contribute to the realm of cardiovascular risk prediction and improve preventive approaches for enhanced patient outcomes (Krittanawong *et al*., 2020).

LITERATURE REVIEW

Cardiovascular diseases (CVDs) represent a significant global health burden, necessitating timely identification and precise assessment of individual cardiovascular risk to prevent the onset of these conditions. This risk assessment involves evaluating various factors such as age, gender, blood pressure, cholesterol levels, smoking status, diabetes, and family history of CVDs.

Accurate cardiovascular risk assessment offers several pivotal advantages. Firstly, it enables risk stratification, categorizing individuals into low, intermediate, or high-risk groups, thereby facilitating tailored interventions and treatment plans (D'Agostino *et al*., 2008). (Geminiganesan *et al*., 2021)in the recent genomic advancements reveal the molecular basis of steroid-resistant nephrotic syndrome, with ANLN mutation identified as a causative factor in focal segmental glomerulosclerosis, underscoring the importance of genetic testing to guide treatment decisions.Secondly, it supports early implementation of preventive measures and lifestyle modifications, including promoting a healthy diet, regular physical activity, smoking cessation, and management of modifiable risk factors like hypertension and hyperlipidemia (Gaziano *et al*., 2009). (Thirdly, accurate risk assessment allows for efficient allocation of healthcare resources by identifying high-risk individuals who stand to benefit most from intensive interventions, thereby optimizing resource utilization and improving cost-effectiveness in managing CVDs (Conen *et al*., 2011). Lastly, it fosters shared decision-making between healthcare providers and patients, empowering individuals to actively participate in discussions about potential interventions, treatment options, and medication adherence, ultimately enhancing overall cardiovascular care (Sridhar *et al*., 2012). In recent years, deep learning has excelled in accurate sentiment analysis(Ganesan *et al*, 2023).The integration of big data technology, particularly sentiment analysis of using convolutional neural network showed promise with a 96.12% accuracy rate (Pokhrel *et al*, 2022). In essence, cardiovascular risk assessment plays a crucial role in identifying at-risk individuals, implementing preventive measures, optimizing resource allocation, and promoting patient engagement (D'Agostino *et al*., 2008; Gaziano *et al*., 2009; Conen *et al*., 2011; Sridhar *et al*., 2012)

Challenges in Initial Cardiovascular Risk Prediction: Predicting an individual's initial cardiovascular risk level is accompanied by several challenges that need to be addressed to improve accuracy and effectiveness. These challenges include:

- Complex Interactions: Cardiovascular diseases (CVDs) are influenced by a multitude of risk factors, and their interactions can be intricate. Traditional risk prediction models often consider risk factors individually, assuming linear relationships. However, this approach may not capture the complex interplay between risk factors and their synergistic effects on cardiovascular risk. Incorporating and understanding these complex interactions is essential for enhancing risk prediction accuracy (Kannel & Wilson, 1995; Ridker& Cook, 2013).
- Individual Variability: Each individual's response to risk factors and disease progression may vary, leading to heterogeneity in cardiovascular risk. Traditional risk prediction models often provide average risk estimates for a population, but they may not accurately account for individual variations. Incorporating personalized factors such as genetics, lifestyle, and biomarkers can help improve the precision of risk prediction models (Thanassoulis& Vasan, 2010; Kathiresan & Srivastava, 2012).
- Temporal Dynamics: The progression of cardiovascular risk can be dynamic, and risk factors may change over time. Traditional risk prediction models often provide static risk estimates based on baseline measurements, neglecting the dynamic nature of risk factors and their effects. Incorporating temporal dynamics and longitudinal data into risk prediction models can enhance their accuracy and provide more reliable risk estimates (Berry *et al*., 2007; D'Agostino *et al*., 2001).
- Incorporating Novel Risk Factors: Traditional risk prediction models may not account for emerging or novel risk factors that have been identified through ongoing research. For instance, biomarkers, genetic variants, and imaging techniques have shown promise in improving risk prediction. Incorporating these novel risk factors into predictive models can enhance their accuracy and better capture an individual's cardiovascular risk (Wang *et al*., 2006).

Addressing these challenges in initial cardiovascular risk prediction is crucial for developing more accurate and personalized risk assessment models. By accounting for complex interactions, individual variability, temporal dynamics, and incorporating novel risk factors, future risk prediction models can provide improved risk estimates and enable more effective preventive interventions.

Roles of Machine Learning in Cardiovascular Risk Assessment

Machine learning techniques have emerged as powerful tools in cardiovascular risk assessment, offering improved accuracy and personalized predictions. Machine learning algorithms can effectively analyze large and complex datasets, capture intricate patterns, and identify relevant risk factors. Here are some key aspects highlighting the role of machine learning in cardiovascular risk assessment:

- Risk Prediction Models: Machine learning algorithms can be utilized to develop robust risk prediction models by leveraging various data sources such as electronic health records, genetic profiles, lifestyle data, and medical imaging. These models can incorporate a wide range of risk factors and their complex interactions to provide more accurate and individualized risk estimates (Musunuru & Kathiresan, 2010; Khera *et al*., 2018).

- Feature Selection and Risk Factor Identification: Machine learning algorithms can automatically identify relevant risk factors and features from high-dimensional datasets. By applying feature selection techniques, machine learning models can identify the most informative variables for risk prediction, reducing noise and improving model performance (Doğan & Yıldız, 2015; Rizk & Sabbagh, 2017).

- Risk Stratification and Personalized Medicine: Machine learning algorithms can stratify individuals into different risk categories based on their unique characteristics and risk profiles. This enables personalized medicine by tailoring preventive strategies and interventions to individuals at higher risk, optimizing healthcare resource allocation, and improving patient outcomes (Rumsfeld *et al*., 2016; Li *et al*., 2020).

- Early Detection and Prevention: Machine learning algorithms can help identify early signs and patterns indicative of cardiovascular risk, enabling timely interventions and preventive measures. Through the analysis of an extensive array of data encompassing genetic markers, clinical measurements, and lifestyle elements, machine learning models can play a pivotal role in pinpointing individuals with elevated risk levels, even prior to the manifestation of evident symptoms (Zheleva *et al*., 2017).

RESEARCH METHODOLOGY

Research Design: This study utilized a comparative analysis research design to evaluate various Machine Learning models for predicting initial cardiovascular risk. Employing algorithms like logistic regression, support vector machines, random forests, and neural networks, the research aimed to identify the most accurate model by systematically comparing their performance on a standardized dataset. The approach allowed for a comprehensive assessment of strengths and limitations, contributing insights into different algorithms' effectiveness in addressing the research objective. To ensure validity, cross-validation was employed, mitigating overfitting risks and providing a robust evaluation of model performance. This research design, driven by the goal of informing researchers about the most effective model for cardiovascular risk prediction, contributes to the advancement of the field and the development of accurate predictive models.

METHODOLOGY

The data collection process for this research involved sourcing the "Cardiovascular Disease Dataset" from Kaggle, a comprehensive compilation of clinical and lifestyle variables. Ethical considerations were prioritized through anonymization and de-identification to safeguard individuals' privacy. Various data sources, including health records and surveys, contributed to a dataset capturing diverse cardiovascular factors. Ethical guidelines and privacy protection measures were strictly adhered to, with data access limited to authorized personnel. The subsequent data pre-processing phase ensured the dataset's cleanliness and suitability for analysis. Techniques such as handling missing values, managing outliers, addressing noise, and transforming data were employed. These steps aimed to enhance the robustness of machine learning models. Feature engineering techniques, including creating interaction terms and generating polynomial features, were applied to improve the model's accuracy in predicting cardiovascular risk. In model selection and training, diverse algorithms, such as logistic regression, decision trees, random forests, support vector machines, and neural networks, were chosen to explore various modelling approaches. The data was split into training and test datasets, adhering to standard practices to evaluate model performance. A validation set further allowed for fine-tuning and optimizing models, ensuring reliability in real-world scenarios. This comprehensive methodology positions the research to contribute valuable insights to cardiovascular risk prediction.

RESEARCH FINDING

Data Study: The Cardiovascular Disease dataset, encompassing 70,000 patient records with 11 features and a target variable, serves as a comprehensive resource for cardiovascular research. Categorized into three types—Objective, Examination, and Subjective Features—the dataset captures diverse information. Objective Features, grounded in standardized measurements, include age, height, weight, blood pressure, cholesterol, and glucose, providing vital physiological insights. Examination Features, derived from medical tests and evaluations, offer valuable data on heart health, including ECG and stress test results. Subjective Features, gleaned from patient self-reports, shed light on lifestyle factors such as smoking, alcohol consumption, and family history, enhancing our understanding of cardiovascular risk. Table 1 succinctly outlines each feature typeand its significance in this dataset.

Data Preprocessing: In the pursuit of advancing cardiovascular research and healthcare applications, meticulous data pre-processing serves as a cornerstone for the successful implementation of machine learning models and analytical insights. This comprehensive approach, involves addressing missing values, removing duplicate rows, and strategically handling outliers in the cardiovascular dataset. Missing values were meticulously dealt with using a specialized function, allowing for informed decision-making regarding imputation or data cleansing techniques. Duplicate rows were systematically eliminated to ensure dataset integrity, and outliers, identified in variables such as 'height,' 'weight,' 'hi_bp,' and 'low_bp,' were rigorously removed to enhance model performance and data reliability. Notably, records presenting physiological inconsistencies, where diastolic pressure surpassed systolic pressure, were also excluded. Furthermore, a crucial transformation converted the 'age' feature from days to years, enhancing interpretability and aligning the dataset with the cardiovascular research context. The decision to retain all features through a thoughtful analysis of the cardiovascular disease dataset underscored their inherent relevance, contributing to a comprehensive understanding of factors influencing disease prediction. Additionally, the application of 5-fold cross-validation provided a robust evaluation metric, striking a practical balance between assessment thoroughness and computational efficiency. This method ensures reliable insights into the model's predictive capabilities while mitigating over fitting risks. Finally, the dataset was judiciously split into training, validation, and test sets to facilitate effective model training, fine-tuning, and unbiased evaluation on unseen data. This comprehensive data split strategy is instrumental in ensuring the model's generalization to new, unseen samples, ultimately contributing to the depth and reliability of our cardiovascular research findings.

Model Development and Results: The research employed a diverse set of machine learning models, including Logistic Regression, Decision Tree Classifier, Random Forest, Naive Bayes, Linear Support Vector Classifier (Linear SVC), and Neural Network, to predict cardiovascular outcomes and address research inquiries. Cross-validation techniques were integrated for robust and unbiased assessments, involving data partitioning into subsets to mitigate biases and overfitting. Model performance was thoroughly analysed, utilizing visual aids such as confusion matrices, ROC curves, and Precision-Recall Curves (PRC). These visualizations provided insights into the models' abilities to accurately classify outcomes. The training, validation, and test accuracies for each model were presented in Table 2, with a visual comparison shown in Figure 1. Comprehensive classification metrics, including precision, recall, and F1-score for each model, were detailed in Table 3. The results showcase varying performances of different machine learning models in predicting cardiovascular outcomes. The Logistic Regression model exhibits commendable accuracy rates of 69.17%, 70.34%, and 70.90% on the Training, Validation, and Test sets, respectively.

Table 1. Data Features

Feature	Type	Description
Age	Objective	Patient's age in days (integer)
Height	Objective	Patient's height in centimetres (integer)
Weight	Objective	Patient's weightin kilograms (float)
Gender	Objective	Gender of the patient(categorical code)
Systolic blood pressure	Examination	Systolic blood pressure at the time of examination
Diastolic blood pressure	Examination	Diastolic blood pressure at the time of examination
Cholesterol	Examination	Cholesterol level categorized as: 1: Normal 2: Above normal 3: Well abovenormal
Glucose	Examination	Glucose level categorized as: 1: Normal 2: Above normal 3: Well above normal
Smoking	Subjective	Binary variable indicating if the patient smokes
Alcohol intake	Subjective	Binary variable indicatingif the patient consumes alcohol
Physical activity	Subjective	Binary variable indicating if the patient engages in physical activity
Presence or absence of cardiovascular disease	Target Variable	Binary variable indicating the presence or absence of cardiovascular disease

Table 2. Model Accuracy

Model	Training Accuracy	Validation Accuracy	Test Accuracy
Logistic Regression	0.6917	0.7034	0.7090
Decision Tree	0.6304	0.6249	0.6376
Random Forest	0.7205	0.7190	0.7264
Linear SVC	0.5909	0.6380	0.6277
Naive Bayes	0.5786	0.5736	0.5837
Neural Network	0.5350	0.5318	0.5389

Table 3. Comprehensive Classification Metrics for Different Models

Model	Precision (Class 0)	Recall (Class 0)	F1-Score (Class 0)	Precision (Class 1)	Recall (Class 1)
Logistic Regression	0.6811	0.7754	0.7252	0.7450	0.6438
Decision Tree	0.6320	0.6423	0.6371	0.6433	0.6330
Random Forest	0.7117	0.7523	0.7314	0.7426	0.7009
Linear SVC	0.5789	0.9110	0.7079	0.8002	0.3498
Naive Bayes	0.5471	0.9265	0.6879	0.7744	0.2474
Neural Network	0.7343	0.1079	0.1881	0.5235	0.9617

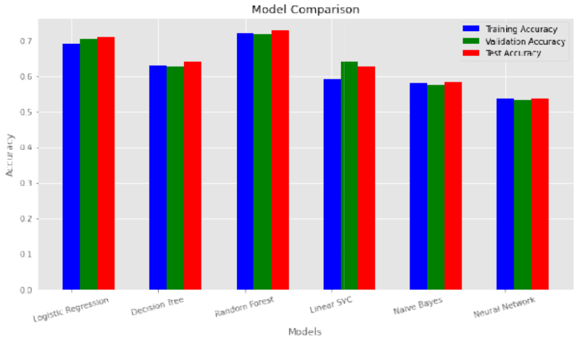

Figure 1. Accuracy Comparison of Different Classification Models

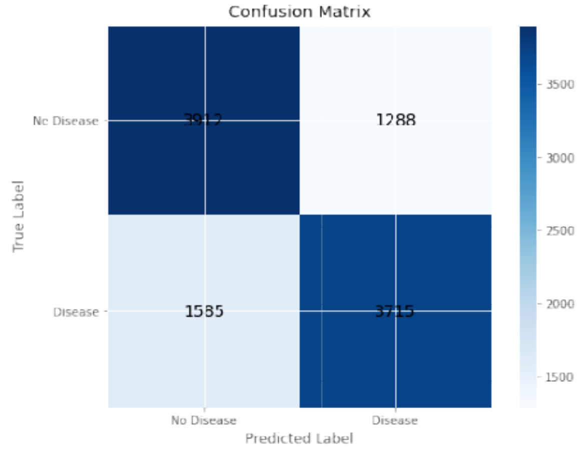

Figure 2. Confusion Matrix of Random Forest

Its balanced precision, recall, and F1-score values for both classes indicate robust performance, avoiding biases toward one class. In contrast, the Decision Tree model, while achieving moderate accuracy rates of 63.04%, 62.49%, and 63.76%, shows lower precision, recall, and F1-score compared to Logistic Regression, suggesting potential overfitting or limited generalization. The Random Forest model emerges as a standout performer with higher accuracy rates of 72.05%, 71.90%, and 72.64%. Notably, it demonstrates remarkable precision, recall, and F1-score values for both classes, showcasing effectiveness in discriminating between instances of class 0 and class 1, crucial in medical diagnostics. Conversely, the Linear SVC model exhibits relatively lower accuracy rates of 59.09%, 63.80%, and 62.77%. Its subpar precision, recall, and F1-score for class 0 indicate a struggle to identify instances of this class, possibly due to the linear nature of the SVC.

The Naive Bayes model demonstrates lower accuracy rates of 57.86%, 57.36%, and 58.37%, with particularly poor recall and F1-score for class 1. Naive Bayes assumptions might not align well with intricate feature interdependencies in medical datasets. The Neural Network model shows the lowest accuracy rates of 53.50%, 53.18%, and 53.89%. Challenges in identifying class 0 and its relatively better recall and F1-score for class 1 suggest room for improvement, potentially through optimizing architecture and hyper parameters.

Among the evaluated algorithms, the Random Forest stands out as the best-performing model due to its ensemble approach, reducing the risk of overfitting and improving generalization. Its ability to capture non-linear relationships is particularly valuable in medical scenarios with complex interactions between health indicators. The confusion matrix in Figure 2 provides a detailed breakdown of the Random Forest model's performance, highlighting true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). This offers insights into the model's classification accuracy and potential errors. The ROC curve (Figure 3) illustrates the Random Forest model's discrimination ability by showcasing the trade-off between true positive rate (sensitivity) and false positive rate. A higher area under the ROC curve (AUC-ROC) indicates superior discrimination between classes. This visualization provides a clear understanding of how well the model distinguishes between positive and negative instances.

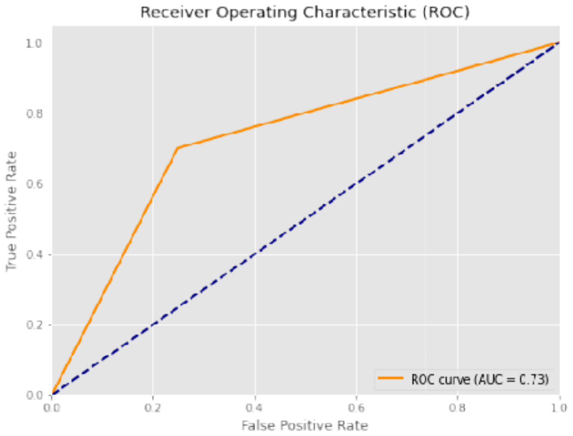

Figure 3. ROC curve of Random Forest

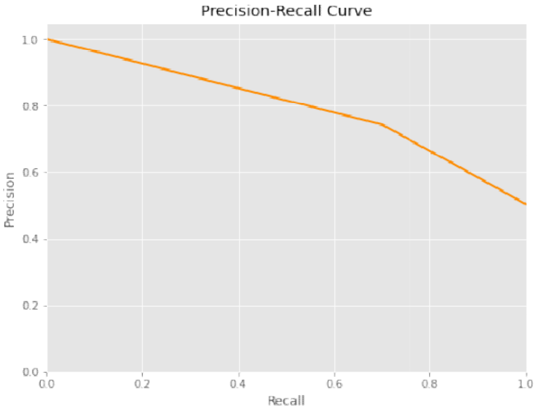

Figure 4. Precision-recall curve of Random Forest

The Precision-Recall curve (P-R curve) in Figure 4 emphasizes the balance between precision and recall. The AUC-PR quantifies the model's ability to maintain high precision while achieving high recall. Analysing the Random Forest's P-R curve provides insights into its precision-recall trade-off and performance across different threshold levels. These visualizations collectively offer a holistic view of the Random Forest model's classification prowess, showcasing its ability to navigate the delicate balance between sensitivity, specificity, precision, and recall. The ROC curve and P-R curve serve as valuable tools for assessing and comparing the performance of different classification models, aiding in the selection of the most suitable algorithm for cardiovascular risk prediction. The Neural Network model, while having the potential to capture intricate patterns, demonstrates the lowest performance. Enhancements could involve optimizing the model architecture, experimenting with layers, nodes, and activation functions, applying regularization techniques, and fine-tuning hyper parameters for efficient convergence during training.

CONCLUSION

The research aimed to assess the effectiveness of various machine learning models in predicting initial cardiovascular risk levels and to compare their performance through training, validation, and test accuracies. This investigation provided insights into the potential utility of these models for accurate risk assessment and intervention planning. The results of the experimentation shed light on the predictive capabilities of different machine learning algorithms for initial cardiovascular risk level prediction. Notably, the Random Forest algorithm demonstrated the highest test accuracy of 0.7264, suggesting its proficiency in capturing intricate data relationships and enhancing prediction accuracy. Similarly, the Logistic Regression model exhibited competitive performance with a test accuracy of 0.7090, showcasing its simplicity and potential as a practical choice for risk assessment purposes. The Decision Tree model, despite achieving a relatively lower accuracy, contributed valuable insights into the underlying data structure. Conversely, the Linear Support Vector Classifier (Linear SVC), Naive Bayes, and Neural Network models displayed comparatively lower accuracies, indicating room for improvement through possible refinement or feature engineering. Importantly, beyond predictive accuracy, the clinical relevance of these models remains a key consideration. Their interpretability, ability to identify crucial features, and potential for guiding medical decisions are pivotal aspects when applying these predictions within healthcare scenarios. The study underscores the promise of machine learning models in predicting initial cardiovascular risk levels. The Random Forest and Logistic Regression models emerge as notable candidates for accurate risk assessment, yet further exploration and optimization are essential to ascertain their viability in clinical contexts. As the landscape of machine learning advances, the integration of these models into comprehensive cardiovascular risk assessment frameworks holds potential to enhance patient outcomes and facilitate well-informed medical interventions.

Recommendations

The study's findings, evaluating machine learning models for cardiovascular risk prediction, yield actionable recommendations for future research and practical implementation. Notably, the Random Forest algorithm consistently outperforms others, emphasizing its suitability for accurate risk assessment. However, the interpretability of models like Logistic Regression and Linear SVC shouldn't be overlooked, allowing clinicians insight into risk predictions. Decision Trees, despite slightly lower accuracy, offer simplicity and interpretability, with the potential for mitigating over fitting through techniques like pruning. Addressing data quality and feature engineering is pivotal for overall model improvement. Continuous validation, updates, and consideration of Neural Network refinement are crucial, acknowledging their untapped potential. Ensemble strategies, building on the success of Random Forest, further enhance predictive accuracy. The study underscores the importance of model selection, interpretability, and continuous validation, offering a roadmap for developing precise tools aiding early risk assessment and personalized interventions for improved cardiovascular health outcomes.

REFERENCES

Berry, J. D., Dyer, A., Cai, X., Garside, D. B., Ning, H., Thomas, A. and Lloyd-Jones, D. M. (2007). Lifetime risks of cardiovascular disease. New England Journal of Medicine, 356(25), 2542-2550.

Conen, D., Ridker, P. M., Mora, S., Buring, J. E. and Glynn, R. J. (2011). Blood pressure and risk of developing type 2 diabetes mellitus: the Women's Health Study. European heart journal, 32(23), 2937-2942.

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M. and Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation, 117(6), 743-753.

Doğan, U. Ç. and Yıldız, O. T. (2015). Feature selection and classification methods for multi-dimensional patient data. Journal of Medical Systems, 39(8), 1-11.

Ganesan, S, Somasiri, N, &Colombage, C. (2023). Deep Learning Approaches for Accurate Sentiment Analysis of Online Consumer. doi: 10.1109/ICCCI56745.2023.10128231

Gaziano, T. A., Young, C. R., Fitzmaurice, G., Atwood, S., Gaziano, J. M. and Laboratory Medicine Division, BWH. (2009). Laboratory-based versus non-laboratory-based method for assessment of cardiovascular disease risk: the NHANES I Follow-up Study cohort. The Lancet, 371(9616), 923-931.

Greenland, P., Alpert, J. S., Beller, G. A., *et al.* (2010). 2010 ACCF/AHA guideline for assessment of cardiovascular risk in asymptomatic adults: A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Journal of the American College of Cardiology, 56(25), e50-e103.

Geminiganesan, S., Ganesan, S., Jayaraj, J., Barathi, G., Muthu Kumar, S. and Samy, N.K.(2021). A Puffy Child – A Rare Case of Steroid Resistant Nephrotic Syndrome with ANLN Mutation. 32(3), 385-391.

Kannel, W. B. and Wilson, P. W. (1995). Risk factors that attenuate the female coronary disease advantage. Archives of internal medicine, 155(20), 57-61.

Kathiresan, S. and Srivastava, D. (2012). Genetics of human cardiovascular disease. Cell, 148(6), 1242-1257.

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H. and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics, 50(9), 1219-1224.

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M. and Kitai, T. (2020). Artificial intelligence in precision cardiovascular medicine. Journal of the American College of Cardiology, 75(23), 2867-2879.

Li, Y., Armstrong, P. W., Lam, C. S. and Lee, T. H. (2020). Machine learning in heart failure: ready for prime time. Current Cardiology Reports, 22(1), 1-11.

Lloyd-Jones, D. M., Leip, E. P., Larson, M. G., *et al.* (2010). Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. Circulation, 121(4), 506-514.

Musunuru, K. and Kathiresan, S. (2010). Algorithms for enhancing the utility of DNA sequence data in clinical settings. Nature Reviews Genetics, 11(5), 267-278.

Pokhrel, A.S., Somasiri, N., Jeyavadhana, C.R and Ganesan, S. (2022). Web Scraping Technology using TF-IDF to Enhance the Big Data Quality on Sentiment Analysis. [Online] ray.yorksj.ac.uk. Available at: https://ray.yorksj.ac.uk/id/eprint/7202/.

Ridker, P. M. and Cook, N. R. (2013). Statins: new American guidelines for prevention of cardiovascular disease. The Lancet, 382(9907), 1762-1765.

Rizk, M. and Sabbagh, R. (2017). Machine learning for feature selection in heart disease diagnosis: A systematic review. Expert Systems with Applications, 83, 205-217.

Roth, G. A., Johnson, C., Abajobir, A., *et al.* (2017). Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. J Am Coll Cardiol, 70(1), 1-25.

Rumsfeld, J. S., Alexander, K. P., Goff Jr, D. C., Graham, M. M., Ho, P. M., Masoudi, F. A. and Peterson, E. (2016). Cardiovascular health: the importance of measuring patient-reported health status: a scientific statement from the American Heart Association. Circulation, 127(22), 2233-2249.

Sridhar, G. R., Sanjana, R. and Bhanu, C. (2012). A study on cardiac risk assessment and lipid profile in ischemic heart disease patients. Indian journal of clinical biochemistry, 27(1), 28-32.

Sridhar AR, Young JM, Marwick TH. (2012). Will screening individuals at high risk of cardiovascular events deliver substantial population benefits? BMC Medicine, 10, 12.

Thanassoulis, G. and Vasan, R. S. (2010). Genetic cardiovascular risk prediction: will we get there?. Circulation, 122(22), 2323-2334.

Vallejo-Torres, L., García-Lorenzo, B., Serrano-Aguilar, P., *et al.* (2014). Estimating a cost-effectiveness threshold for the Spanish NHS. Health Economics, 23(6), 730-747.

Wang, T. J., Gona, P., Larson, M. G., Tofler, G. H., Levy, D., Newton-Cheh, C. and Benjamin, E. J. (2006). Multiple biomarkers for the prediction of first major cardiovascular events and death. New England Journal of Medicine, 355(25), 2631-2639.

Wang, Y., Tuomilehto, J., Jousilahti, P., *et al.* (2011). Lifestyle factors in relation to heart failure among Finnish men and women. Circ Heart Fail, 4(5), 607-612.

World Health Organization. (2023). Cardiovascular Diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Zheleva, B., Turner, A. J. and Aggarwal, C. (2017). Machine learning for early detection of heart failure hospitalization. Annals of Translational Medicine, 5(10), 211.

****