Machine Learning-Driven Cardiovascular Risk Assessment: Synthesizing Contemporary Methods, Barriers, and Emerging Horizons

Bashir Ssimbwa[1]; Azizi Wasike[1,2]; Jamir Ssebadduka[3,4]

[1]Kampala International University
[2]Muni University
[3]International University of East Africa
[4]Islamic University in Uganda

Abstract: This review explores the application of machine learning (ML) models in cardiovascular disease (CVD) prediction, emphasizing their potential to revolutionize risk assessment and healthcare delivery. By examining traditional models such as Logistic Regression, Support Vector Machines, and Random Forest, alongside advanced approaches like XGBoost, the study highlights their strengths, limitations, and performance in various healthcare contexts. The review underscores the growing role of hybrid and explainable ML architectures, as well as the integration of deep learning techniques, in enhancing accuracy, scalability, and clinical trust. Despite notable advancements, challenges such as data quality, imbalance, and ethical considerations in underrepresented regions persist, underscoring the need for collaborative efforts among stakeholders to ensure equitable and efficient implementation. By addressing these gaps and leveraging robust models like XGBoost, ML has the potential to significantly reduce the global burden of CVDs and drive transformative change in precision medicine and patient-centred care.

I. INTRODUCTION

➤ *Background*

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, accounting for approximately 17.9 million deaths annually, which represents nearly one-third of all global deaths[1]. Notably, over 75% of these fatalities occur in low- and middle-income countries, underscoring a significant health disparity[2]

The global burden of CVDs has been escalating over the past decades. In 2000, around 14 million individuals succumbed to cardiovascular conditions; by 2019, this number had risen to nearly 18 million[3]. This increase is attributed to factors such as population growth, aging demographics, and the prevalence of risk factors like hypertension, obesity, and diabetes. Countries with the highest number of CVD-related deaths include China, India, Russia, the United States, and Indonesia[4].

Examining the continental landscape, Eastern Europe exhibits the highest age-standardized CVD mortality rates, reaching 432.3 per 100,000 individuals as of 2022[4]. In contrast, regions like the High-income Asia Pacific report significantly lower rates, approximately 73.6 per 100,000. These disparities highlight the varying impact of CVDs across continents, influenced by factors such as healthcare infrastructure, socioeconomic conditions, and lifestyle behaviours[5].

In Africa, the burden of CVDs is substantial and growing. The continent is home to over 1 billion people and contributes significantly to the global CVD burden. In 2013, an estimated 1 million deaths were attributed to cardiovascular diseases in Africa. Hypertension is identified as the primary driver of CVDs in the region, with the African continent having the highest prevalence of hypertension globally. Factors such as urbanization, aging populations, social stress, and limited access to healthcare services exacerbate this issue.

Focusing on Uganda, CVDs are a leading health concern. In 2019, the country reported 28,149 deaths due to cardiovascular diseases. This places Uganda in the bottom 40% of countries concerning age-standardized mortality from CVDs. The nation faces challenges such as high rates of air pollution and elevated blood pressure among females compared to global averages. Despite these challenges, Uganda has implemented seven out of eight key CVD-related policies and has a national action plan for cardiovascular diseases in place. However, the absence of an operational unit within the Ministry of Health dedicated to non-communicable diseases indicates areas needing further development.

➢ *Objective of the Review*

The objective of this review is to evaluate the effectiveness of machine learning models in predicting cardiovascular diseases

➢ *Machine Learning Overview*

Machine Learning (ML), a subset of artificial intelligence (AI), enables computers to learn from data and make predictions without explicit programming. By recognizing patterns, ML algorithms assist in classification and decision-making across various fields, including natural language processing, computer vision, and robotics. ML is broadly categorized into supervised learning, unsupervised learning, and reinforcement learning, each serving distinct purposes. Its success depends on the quality and quantity of training data. ML is transforming industries such as healthcare, finance, transportation, and e-commerce, automating processes and enhancing decision-making.

However, ML has challenges. Issues such as ensuring the quality of data, making models interpretable, and addressing ethical concerns remain pressing. These challenges underscore the need for adaptation, especially in resource-limited areas like Uganda. Despite obstacles, ML can drive innovation by enabling efficient data analysis, improving predictions, and supporting informed decision-making, delivering transformative benefits to communities and industries.

II. MACHINE LEARNING ARCHITECRURES OF CVD PREDICTION

Machine learning (ML) architectures process data to enable automation and data-driven insights. These range from Traditional ML models like decision trees and linear regression, valued for efficiency in low-data environments, to Deep Learning architectures, which use neural networks for high-dimensional tasks but require significant computational power. Hybrid ML architectures combine traditional and deep learning models, enhancing scalability for complex applications. Explainable AI (XAI) architectures improve transparency, crucial in sensitive fields like healthcare and finance. Multi-modal Learning architectures integrate diverse data types, such as text and images, to enhance accuracy in tasks like NLP and autonomous systems. As ML advances, balancing efficiency, interpretability, and computational demands remains a key challenge.

A. Traditional Machine Learning Architectures

Traditional machine learning (ML) Architectures have played a significant role in predicting cardiovascular diseases (CVD) by utilizing structured medical datasets. This section delves into the commonly used ML approaches, their applications in CVD prediction, notable studies that have implemented these methods, and the strengths and limitations of each architecture.

Machine learning techniques such as Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, Naive Bayes, k-Nearest Neighbors, and Gradient Boosting Machines have been widely applied to this field, yielding varying degrees of success depending on the context. By reviewing recent advancements, this discussion sheds light on how these traditional ML architectures have been used, their practical benefits, and the challenges they face in accurately predicting and managing cardiovascular disease.

➢ *Logistic Regression (LR)*

Logistic Regression (LR) is a widely used machine learning model in healthcare, particularly for binary classification tasks like determining whether a patient is at risk for cardiovascular disease (CVD). A notable example is a study by[6], which employed LR to estimate 10-year CVD risk based on routine physical examination data.[19] The model is favoured in clinical settings for its simplicity and high interpretability, allowing healthcare professionals to clearly understand how individual risk factors contribute to predictions. However, its reliance on linear relationships between predictors and outcomes can limit its ability to capture more complex interactions. Additionally, it may struggle when applied to high-dimensional datasets or scenarios involving non-linear patterns.

➢ *Decision Trees (DT)*

Decision Trees (DTs) are widely appreciated for their intuitive visualization and straightforward decision-making process, making them especially useful in clinical settings. For example, a study in South Korea successfully utilized DTs to classify cardiovascular disease (CVD) risk factors, showcasing their ability to pinpoint critical predictors[7]. These models are not only easy to interpret but also capable of capturing non-linear relationships between variables and outcomes. However, they come with challenges, such as a tendency to overfit, particularly with smaller datasets, and sensitivity to minor data changes, which can drastically alter the tree structure.[7]

➢ *Random Forests (RF)*

Random Forest (RF) is a powerful ensemble learning method that improves disease prediction by analysing complex medical data. It enhances classification accuracy by combining multiple decision trees, reducing overfitting, and ensuring reliable outputs. RF has been highly effective in diagnosing conditions like cardiovascular diseases and diabetes, achieving up to **91% accuracy** in genomic

analysis and healthcare diagnostics[8]. Additionally, it highlights key disease predictors, aiding medical research and decision-making[9]. However, RF demands high computational power, making it resource-intensive for large datasets. Its ensemble structure also reduces interpretability, posing challenges in clinical applications where transparency is essential. While simpler models may sometimes be preferable for clarity, RF remains a top choice for disease prediction due to its accuracy and robustness.[9]

➢ *Support Vector Machines (SVM)*

Support Vector Machines (SVMs) are widely used in disease prediction due to their ability to classify medical data by identifying the optimal hyperplane that separates different health conditions. Their strength lies in handling high-dimensional datasets, making them particularly useful for analysing genetic and metabolic markers. Additionally, SVMs reduce overfitting risks by maximizing the margin between classes, ensuring stable predictions even with limited data. Their adaptability through kernel functions also allows them to model complex, non-linear relationships in disease diagnostics. However, SVMs require careful hyperparameter tuning, which can be time-consuming and affects model performance[10]. Their high computational demands limit scalability for large medical datasets, making them less suitable for real-time applications. Additionally, their black-box nature makes interpretation challenging, reducing transparency in clinical decision-making[10]. While SVMs remain a strong choice for disease prediction, neural networks may offer better interpretability and scalability for specific healthcare needs.

➢ *Naïve Bayes (NB)*

Naïve Bayes is a probabilistic machine learning algorithm that classifies data based on conditional independence, making it effective for disease prediction. It has been successfully applied to heart disease, diabetes, chronic kidney disease, and cancer, achieving 85% accuracy in heart disease prediction and 95% in intelligent health risk assessment [11]. Its efficiency, ease of implementation, and ability to handle large datasets make it a valuable tool in medical diagnostics [12]. Additionally, it can sometimes outperform more complex models, proving its reliability in real-world healthcare applications. However, its assumption that features are independent can reduce accuracy when dealing with correlated medical variables. It also faces the zero-probability problem, which requires techniques like Laplace smoothing to improve prediction reliability. Despite these limitations, Naïve Bayes remains a fast, scalable, and effective algorithm for disease classification.

➢ *K-Nearest Neighbours (K-NN)*

K-NN is a non-parametric algorithm that classifies data based on proximity to labelled points, making it useful in disease prediction. In heart disease detection, it evaluates factors like age, gender, and medical history, outperforming traditional methods It also aids breast cancer diagnosis by classifying tumors as benign or malignant based on attributes like texture and perimeter. k-Fold Cross Validation (k-FCV) helps optimize k for better accuracy [12]. Beyond diagnosis, K-NN supports personalized health supplement recommendations by matching disease profiles with functional foods

K-NN's versatility enables applications across medical fields. Its non-parametric nature suits real-world datasets but makes it computationally expensive, requiring distance calculations for each input. Sensitivity to noise and irrelevant features can affect predictions, and performance depends on choosing the right k-value and distance metric. Despite these challenges, careful preprocessing and parameter tuning enhance its effectiveness in disease prediction.

➢ *Gradient Boosting Machines (GBM)*

XGBoost is a powerful machine learning algorithm that enhances prediction accuracy by combining multiple weak models iteratively. In heart disease prediction, XGBoost outperformed other models by analysing key indicators such as fasting blood sugar and cholesterol [13]. Similarly, in Parkinson's disease, it provided superior accuracy over traditional diagnostic methods, effectively tracking disease progression. The integration of boosting techniques like AdaBoost and Light GBM has further improved performance, with some models achieving up to 99.75% accuracy [11]. XGBoost is valued for its high accuracy, flexibility, and ability to capture complex patterns in medical data[11]. It can process diverse datasets, including clinical records and genetic data, making it versatile for different diseases. However, its performance depends on data quality, limiting effectiveness in cases of missing or biased data. Additionally, its complexity makes interpretation difficult, posing challenges for clinical integration. XGBoost is also computationally intensive, requiring significant resources. Despite these challenges, its predictive power makes it a valuable tool in disease forecasting, provided efforts are made to improve interpretability and data handling.

B. Deep Learning Architectures

Deep learning (DL) architectures have emerged as powerful tools for the prediction and diagnosis of cardiovascular diseases (CVDs), leveraging their ability to process complex, high-dimensional, and multi-modal data. This section reviews the use of various deep learning models in CVD prediction, highlighting their applications, strengths, weaknesses, and contributions in the health sector at large.

➢ *Convolutional Neural Networks (CNNs)*

Convolutional Neural Networks (CNNs) are deep learning models designed for processing complex image data, making them highly effective in disease prediction. They have been applied in diagnosing conditions like Parkinson's and malaria by analysing MRI scans and cell images, enabling early and accurate detection [14] CNNs also support home-based disease detection, allowing individuals to monitor their health remotely. Their ability to extract spatial features makes them ideal for medical imaging, with models like VGG16 and ResNet50 achieving accuracy rates of up to 97% in Parkinson's diagnosis[14]. Advanced techniques such as dropout and batch normalization further improve precision and reliability.

Despite their effectiveness, CNNs face challenges in healthcare. They require large datasets and high computational power, limiting accessibility. Privacy concerns and data quality issues hinder adoption, while mathematical validation is crucial for reliability. Overcoming these challenges with ethical use and data security will enhance their impact in disease prediction.

➢ *Recurrent Neural Networks (RNNs)*

Recurrent Neural Networks (RNNs) are deep learning models designed to process sequential data, making them ideal for analyzing time-series information in healthcare. They are used to predict disease progression by analyzing patient history and genetic profiles, improving early detection and risk assessments [15]. Their ability to capture temporal dependencies allows for more precise prognostication compared to traditional methods, enhancing diagnostic accuracy and accessibility[15]. RNNs also adapt well across different healthcare applications, making them valuable for medical data analysis.

Despite their benefits, RNNs face challenges. They require significant computational resources for training and deployment and are sensitive to data quality and privacy issues, which can limit broader adoption. Traditional RNNs also struggle with the vanishing gradient problem, affecting their ability to learn long-term dependencies. Addressing these issues through hybrid models, such as combining RNNs with other neural networks, could improve their efficiency and predictive performance in healthcare.

➢ *Deep Belief Learning Architectures*

Deep Belief Networks (DBNs) are deep learning models designed for handling high-dimensional data, making them effective in disease prediction. They have been used to optimize feature extraction and classification in healthcare, improving accuracy and efficiency. For instance, an Optimized DBN integrated with CNN-LSTM and UFDSOM achieved 98.23% accuracy in disease classification. DBNs have also been applied in Alzheimer's detection using DNA methylation and gene expression data, effectively addressing the High Dimension Low Sample Size issue and outperforming traditional models

DBNs offer high accuracy, robust feature extraction, and effective integration with other models like CNNs and LSTMs. However, their training process is complex due to multiple learning variables, often requiring optimization techniques like the Ant Lion Optimization algorithm. They are also sensitive to noise in medical data, necessitating advanced learning algorithms for reliability. Overcoming these challenges through improved training methods and hybrid models could further enhance DBNs' role in disease prediction.

➢ *Autoencoders*

Autoencoders are unsupervised learning models that extract efficient data representations, making them valuable in medical data analysis and disease prediction. They have been used to enhance feature extraction, improve genetic risk assessment, and predict disease associations. The Autoencoder-based Broad Learning model combines denoising autoencoders with incremental learning, achieving 98.50% accuracy in dynamic healthcare settings [16]. Genetic autoencoders help predict disease-associated single nucleotide polymorphism, improving risk scores for conditions like multiple sclerosis. The End-to-End Supervised Autoencoder model surpasses traditional methods in predicting co-existing diseases using diagnostic codes, while the SARMDA model integrates adversarial regularization with graph neural networks to predict microbe-disease associations effectively. Autoencoders excel in feature extraction, noise reduction, and high prediction accuracy, particularly in complex datasets [16]. However, they require significant computational resources, especially in genetic analysis, and may struggle to capture global patterns without advanced regularization. Addressing these challenges through optimized learning algorithms will further enhance their role in disease prediction.

➢ *Transformers*

Transformers are deep learning models known for processing complex data and identifying patterns, making them highly effective in disease prediction. They have been adapted for diagnosis, prognosis, and clinical outcome forecasting. DiagnoAI, for instance, utilizes a pre-trained transformer to predict diseases from symptom descriptions, excelling in multilingual settings[17]. TransformEHR predicts disease outcomes from electronic health records, aiding in early detection of conditions like pancreatic cancer. OncoBERT leverages unstructured EHR notes for cancer outcome prediction, improving both accuracy and interpretability. DeepProg focuses on disease progression analysis using images and auxiliary data, while transformers have also been explored for immunotherapy prognosis, enhancing early disease detection.

Transformers offer high accuracy, adaptability across different diseases, and improved interpretability, as seen with OncoBERT's transparency in decision-making. However, they require large datasets, which can be a limitation for rare diseases, and their training demands significant computational resources, restricting accessibility in lower-resource settings. Overcoming these challenges could further expand their role in clinical applications.

➢ *Hybrid Deep Learning Architectures*

Hybrid deep learning architectures combine multiple neural network models to enhance disease prediction accuracy and efficiency. By leveraging the strengths of different models, these architectures tackle complex medical data challenges. For instance, a hybrid model integrating genetic algorithms, stacked autoencoders, and Softmax classifiers outperformed traditional methods in early diabetes detection, leading to a web-based diagnosis tool [18]. Similarly, a CNN-LSTM model achieved high accuracy in chronic disease prediction, with 98% for cancer and 99% for kidney disease[18]. Another approach used a Coati Optimization Algorithm within an ANN framework, improving cardiovascular disease prediction, while a Bi-GRU-LSTM model effectively handled small, complex datasets for diabetes-related heart disease monitoring

Hybrid models excel in accuracy, versatility, and advanced feature extraction, making them highly adaptable across different diseases [19],[18]. However, their complexity makes training and optimization challenging and they require large, high-quality datasets, which may not always be available. Simplifying these models and improving efficiency could enhance their accessibility for wider healthcare applications.

C. Hybrid Architectures

Hybrid Machine Learning architectures integrate multiple computational models to enhance performance, scalability, and robustness. These architectures address challenges like adversarial attacks, distributed processing, and model efficiency by leveraging specialized layers and hybrid topologies. They improve neural network resilience in tasks like image classification, enhance emotion recognition in speech, and optimize distributed learning by reducing computational costs in wireless networks as seen below.

➢ *Convolutional Neural Networks (CNN) with Recurrent Neural Networks (RNNs)*

Hybrid CNN-RNN architectures have proven effective in disease prediction by combining CNNs for spatial feature extraction with RNNs for capturing temporal patterns, enhancing diagnostic accuracy. These models have been applied across various medical domains, including anaemia detection, where a CNN-RNN model achieved 90.27% accuracy, outperforming standalone CNNs and RNNs[20]. Similarly, a CNN-LSTM model demonstrated superior accuracy in predicting diseases from Electronic Health Records compared to SVMs and standalone deep learning models. In chronic disease detection, a CNN-LSTM hybrid achieved 98%-99% accuracy for kidney disease [18] while a CNN-RNN model utilizing transfer learning excelled in COVID-19 detection via X-ray analysis. These models offer enhanced feature extraction, improved accuracy across multiple datasets [20], and versatility in handling different medical data types. However, they pose challenges such as increased computational complexity, requiring more resources and longer training times, and dependence on large, high-quality datasets. Future research could refine these models by integrating attention mechanisms to improve efficiency while maintaining high predictive performance.

➢ *Support Vector Machines (SVM) with Neural Networks*

Hybrid SVM-NN models enhance disease prediction by leveraging both approaches. A hybrid ANN-SVM model improved diabetes classification to 89.7% accuracy using non-linear kernel SVMs. In medical imaging, SVMs acted as primary classifiers while ANNs refined results, boosting accuracy. The Neural Support Vector Machine outperformed standard SVMs and multi-layer perceptrons in regression and dimensionality reduction tasks. These models offer high accuracy effective dimensionality, and adaptability via hyperparameter tuning. However, they demand high computational resources and require extensive hyperparameter optimization. Despite challenges, SVM-NN hybrids advance disease prediction if computational constraints are managed.

➢ *Autoencoders with Decision Trees*

Autoencoder-based hybrid models enhance disease prediction by combining feature extraction with interpretability. Autoencoder Trees, introduced by Ozan İrsoy and Ethem Alpaydın, use decision trees for encoding and decoding, leveraging soft decision trees with multivariate splits for hierarchical data representation [20]. They achieve strong reconstruction accuracy on datasets like handwritten digits and news data [20]. Encoder Forest, proposed by Ji Feng and Zhi-Hua Zhou, integrated tree ensembles with autoencoders, offering lower reconstruction error and faster training than deep neural networks. Its robustness makes it applicable to disease prediction. Hybrid frameworks, such as those explored by P. Epsiba et al., apply these models to medical prognosis, proving their utility in healthcare. While these architectures improve interpretability and efficiency, they risk overfitting and performance variability based on dataset characteristics. Nonetheless, they provide a promising direction for disease prediction and complex data analysis.

➢ *CNNs with Transformers*

Hybrid CNN-Transformer models enhance medical image analysis by combining CNNs' local feature extraction with Transformers' global context modelling, improving disease prediction accuracy. Li Wang et al. developed a CNN-Transformer model for N-staging and survival prediction in non-small cell lung cancer (NSCLC), outperforming standalone CNNs and Transformers [21]. ReFormer, introduced by, employs frequency-based bridging to refine feature fusion in medical image segmentation. These models improve predictive accuracy, optimize feature fusion, and adapt across imaging tasks. However, their complexity and high computational costs pose challenges. Feature misalignment can also hinder effective fusion. Despite these limitations, refining integration strategies can enhance efficiency, making CNN-Transformer hybrids vital for medical diagnostics

D. Explainable Arttificial Intelligence Architectures

Explainable AI (XAI) architectures enhance transparency by making AI decisions interpretable, using techniques like LIME. They improve trust in applications like pathology and cyber security but face challenges balancing accuracy and interpretability as seen below.

➢ *Explainable Boosted Trees*

Explainable Boosted Trees architectures balance predictive accuracy and interpretability, crucial for disease prediction in healthcare. By integrating techniques like LIME[22], EBTs provide insights into decision-making processes, improving trust and clinical decision-making in areas like chronic kidney disease and ocular disease detection. Their strengths include enhanced interpretability[22] ,supporting better communication in healthcare, and high accuracy. However, challenges exist in implementation complexity and limited generalizability across conditions.

➢ *Decision Trees with Rule Based Explanations*

Decision trees with rule-based explanations are essential in explainable AI (XAI), particularly in disease prediction, by providing transparent, interpretable decision-making processes. For example, a decision tree model for chronic kidney disease diagnosis uses Gini Importance and SHAP values for interpretation [23]. The SRules method enhances the interpretability of complex models by creating concise, understandable rules, and integrated systems like decision trees with random forests use SHAP and LIME for medical predictions, increasing trust. These models improve transparency and decision-making and offer personalization for better predictions However, there are trade-offs between interpretability and accuracy, and creating surrogate models can be resource-intensive.

➢ *Explainable Neural Networks*

Explainable Neural Networks (XNNs) aim to improve transparency in AI-driven medical diagnostics by integrating explainability techniques into neural network architectures. These models maintain high diagnostic accuracy while providing interpretable outputs, crucial for clinical adoption. For instance, Thakur's research enhances deep learning models with XAI techniques for medical diagnosis and combine large language models with explainability algorithms for better disease predictions. Jere's study uses ensemble learning and SHAP for heart disease prediction, achieving high performance [24], while Mathew et al. integrate SHAP and LIME in decision tree models for medicine recommendations. XNNs offer high transparency [24] and robust diagnostic performance empowering clinicians with understandable explanations. However, these models can be complex to implement and may risk overfitting with limited data along with scalability challenges. Despite these, XNNs provide crucial insights into medical AI systems, promoting trust and understanding.

➢ *Bayesian Networks*

Bayesian Networks (BNs) are a key architecture in explainable AI (XAI), particularly for disease prediction. BNs utilize probabilistic graphical models to capture dependencies and causal relationships, making them ideal for medical diagnostics, where uncertainty is common. Researchers like Jose A. Quesada and colleagues have extensively reviewed BNs' use in medical diagnostics over the past 40 years [25]. They are widely used in disease diagnosis and prognosis, with metrics like sensitivity and specificity to assess performance. BNs' strengths include clear explainability, integration of expert knowledge, and effective uncertainty handling, crucial for medical applications. However, they face challenges in managing complexity as the number of variables grows and their unstructured application limits effectiveness. Addressing these challenges, researchers aim to create structured frameworks for better interpretability and communication in BNs

E. Multi-Modal Learning Architectures

Multimodal learning architectures combine data from different modalities (e.g., text, images, audio) to improve predictions. These models capture both inter- and intra-

modality relationships, enhancing decision-making [26]. Integration can occur via deep or early fusion, but challenges include data heterogeneity and complexity.

➢ *Deep Neural Networks in Multimodal Learning*

DNNs in multimodal learning have revolutionized disease prediction by integrating various data types like omics, imaging, and structured data to improve model accuracy. For example, Stochastic Training with Regularization and Gradient Descent by Ohnuki predicts drug-disease relationships using multi-omics data, while [16] ABL model enhances feature extraction in medical data with a Denoising Auto Encoder [16]. Other models like IDCGN and MFDL use advanced graph-based techniques and multi-omics data to enhance prediction accuracy [27]. These models are strong in integrating diverse data, improving accuracy, and handling noisy data[16]. However, challenges include high computational demands, model complexity, and reliance on high-quality multimodal data. Future work should focus on optimizing these models to balance performance and complexity in diverse healthcare applications.

➢ *Graph Neural Networks (GNN)*

Graph Neural Networks (GNNs) play a crucial role in multimodal learning for disease prediction by modelling complex relationships across various data types, such as clinical, demographic, and imaging data. Models like IDCGN and STRGNN use multi-omics data and advanced graph techniques to enhance prediction accuracy [27],[28]. GNNs are widely used in applications like heart disease prediction and veterinary epidemiology, where they integrate data from diverse sources to improve predictive performance. Their strengths include the ability to capture intricate relationships and scale with large datasets, but they face challenges such as data noise and high computational demands, [27] Future research may focus on refining GNNs for better performance and interpretability.

III. MACHINE LEARNING MODELS

Machine learning (ML) models are computational systems that analyse data to identify patterns and make predictions without explicit programming. In cardiovascular disease (CVD) prediction, ML enhances early detection, risk assessment, and personalized treatment by uncovering hidden patterns in complex datasets. Key models include Random Forest, which reduces overfitting for reliable predictions, and Support Vector Machines (SVM), known for handling non-linear data with high accuracy [29],[30] Logistic Regression which offers interpretability, K-Nearest Neighbors (KNN) that provides strong classification, and XGBoost excels in processing large datasets [31],[32]Beyond prediction, ML aids biomarker discovery and improves clinical workflows, though challenges like data quality, interpretability, and ethical concerns must be addressed for effective healthcare integration[33], [34]

➢ *Support Vector Machine (SVM):*

Support Vector Machines are supervised learning models used for classification and regression, excelling in

high-dimensional data analysis through kernel functions. In cardiovascular disease (CVD) prediction, SVMs have shown high accuracy across studies. In India, they outperformed models like XGBoost [35], while in China, SVMs with radial basis functions excelled in coronary heart disease classification. Bangladesh researchers found polynomial SVMs more accurate than linear ones. Comparative studies consistently highlight SVM's strong predictive power.

Despite their accuracy and robustness, SVMs face challenges such as computational intensity, overfitting risks, and limited interpretability. Optimizing hyperparameters and improving model explainability could enhance their clinical applicability.

➢ *XGBoost (eXtreme-Gradient Boost Algorithm):*
XGBoost is an ensemble learning algorithm based on decision trees, designed for speed and efficiency in handling large datasets [36],[37]. Widely used in cardiovascular disease (CVD) prediction, it has demonstrated high accuracy in identifying heart attack and stroke risks. In the UK, researchers developed a mixed-effects XGBoost model to predict 30-day mortality post-cardiac surgery, improving predictions by accounting for hospital-level variations [36]. Other studies found XGBoost outperformed models like Random Forest and Logistic Regression in stroke prediction, highlighting its precision in classifying at-risk individuals[38].

XGBoost excels in processing complex medical data with superior predictive accuracy, making it more effective than models like KNN and Naïve Bayes[37]. However, it struggles with correlated data without modifications and requires careful hyperparameter tuning, which can be computationally demanding [36]. Optimizing its implementation and integrating clinical insights can enhance its practical use in healthcare.

➢ *K-Nearest Neighbours (KNN) Classifier:*
K-Nearest Neighbors (KNN) is a supervised learning algorithm used in cardiovascular disease (CVD) prediction to classify individuals based on their risk of heart-related conditions. It works by identifying the 'k' closest data points to an input and predicting outcomes based on the majority class. In a study using Kaggle data, KNN was compared with Logistic Regression and Random Forest, where it performed well but was outmatched by Random Forest in accuracy[39]. In Indonesia, researchers applied KNN with Principal Component Analysis (PCA) for feature selection, achieving 90.16% accuracy in heart attack risk classification, demonstrating its clinical potential [39]. Another study highlighted KNN's role in reducing misdiagnosis and improving patient classification.

KNN is valued for its simplicity and effectiveness in high-dimensional data when paired with proper feature selection [40],[39]. However, it is computationally intensive for large datasets and sensitive to irrelevant features, which can impact performance if not properly filtered[40],[39]. While KNN is useful in CVD prediction, models like

Random Forest and SVM may offer better accuracy depending on the dataset and clinical application.

➢ *Random Forest Classifier:*
Random Forest Classifier is an ensemble machine learning model used for cardiovascular disease (CVD) prediction that analyses multiple risk factors. It constructs numerous decision trees and combines their outputs to enhance accuracy and reduce overfitting. Studies in India and other countries have utilized this model to integrate clinical, demographic, and genetic data for improved CVD risk assessment [41]. The goal is to enhance sensitivity and specificity, enabling early diagnosis and personalized treatment [42]

Random Forest is known for its high accuracy, outperforming models like Logistic Regression and SVM[43]. It also provides insights into feature importance, helping identify key risk factors. However, its complexity makes interpretation difficult, and its performance depends on data quality [42]. Despite these challenges, optimizing the model and integrating additional predictive factors could further improve its effectiveness in clinical settings.

➢ *Logistic Regression Classifier:*
Logistic Regression (LR) is a statistical machine learning model used for binary classification, making it suitable for predicting the likelihood of CVD. It estimates the probability of a condition based on predictor variables, offering a simple yet effective approach to early diagnosis. Studies in India and other countries have employed LR to improve CVD risk assessment. For example, [44] used regularized LR on the UCI dataset, achieving 90% accuracy, while Miao [45] enhanced LR with K-means++ and PCA for better data quality. Azis [46] applied 5-fold cross-validation, reporting accuracy between 80% and 88.29%, demonstrating the model's robustness. Additionally, Mahajan found LR to be one of the most reliable models, achieving nearly 90% accuracy.

LR is favoured for its simplicity, ease of interpretation, and strong predictive capabilities, making it valuable in medical diagnostics[46],[44]. It can also integrate with methods like PCA to improve performance [45]. However, its reliance on linear relationships may limit accuracy in complex datasets, and it is sensitive to outliers. Despite these drawbacks, LR remains a trusted tool in CVD prediction, with the potential for further enhancement through hybrid modelling approaches.

IV. DATA SETS

A dataset is a structured collection of data used to train machine learning models. In CVD prediction, key datasets include NIH's(National Institute of Health) multi-omics data for genetic risk assessment [47], UCI(University of California Irvine machine learning repository) Heart Disease for ML benchmarks[48] , IEEE Dataport(Institute of Electrical and Electronics Engineers) for ensemble models [49], UK(United Kingdom) Biobank for deep learning [50], and Kaggle for ML experimentation [51]. While these

datasets enhance prediction, challenges like data quality and model interpretability require further validation[51],[48]

➢ *NIH's Dataset*

The NIH dataset, particularly from NHANES(National Health and Nutrition Examination Survey), has been instrumental in training ML models for CVD prediction by providing extensive patient data, including demographics, medical history, and lifestyle factors[52]. It has also been used alongside CAC (collection of coronary artery calcium)scans in AI-CVD studies to predict CVD events and mortality through AI-driven screening[43]. Researchers like Surya et al. applied neural networks and gradient boosting for early diagnosis, while M. Naghavi *et al.* [43]. leveraged AI for enhanced risk assessments [52], [43]. Strengths of the dataset include its comprehensive variables and large sample size, which improve model reliability [53]. However, challenges such as limited population diversity and intensive data preprocessing require further validation for broader clinical application [54].

➢ *UCI Heart Disease dataset*

The UCI Heart Disease dataset is widely used in ML research for CVD prediction, serving as a benchmark for testing various models. It includes diverse patient attributes, enabling the development of both traditional and advanced algorithms for early diagnosis. Studies have applied centralized and federated ML approaches, with SVM achieving 83.3% accuracy in a centralized setting and 73.8% in federated learning, enhancing privacy [53]. Regularized logistic regression demonstrated 90% accuracy, highlighting its predictive power[44] , while CNN-based deep learning models outperformed traditional methods in efficiency and accuracy [55]. Researchers such as Rodriguez & Nafea explored model interpretability, Singh validated logistic regression, and Mansoor optimized deep learning techniques. The dataset's strengths include versatility across ML applications and support for privacy-preserving methods [56],[44]. However, its limited size and class imbalance may hinder model generalization [56]. Addressing these challenges could enhance its real-world applicability.

➢ *The UK Biobank*

The UK Biobank dataset, with data from over 500,000 participants, is vital for ML-based CVD prediction[57]. Researchers have used ML models to analyze genetic, lifestyle, and clinical factors. Nine models, including LSVM(Lagrangian Support Vector Machine), RF(Random Forest), and NN(Neural Network), utilized cardiac imaging and baseline characteristics. ML tools like REGENIE and VariantSpark fine-mapped coronary artery disease loci while survivalFM improved risk prediction by modeling factor interactions. Strengths include comprehensive data and advanced ML techniques, enhancing sensitivity, Challenges include complexity and computational demands, requiring solutions for better clinical integration.

➢ *IEEE Dataport*

The IEEE Dataport dataset has been instrumental in improving ML models for CVD prediction, leveraging demographic and biometric data to identify high-risk individuals. Researchers have applied models like support vector machines, ensemble learning, and neural networks, achieving AUC values between 0.85 and 0.92. Logistic regression, random forest, and K-nearest neighbors enhanced classification accuracy, while AI-integrated simulations reached 97% accuracy via a web-based tool. Strengths include diverse data, high model accuracy, and user-friendly applications. However, demographic biases and dataset size limitations remain challenging. Expanding the dataset and addressing bias could further refine predictive models for clinical use.

➢ *Kaggle Dataset*

The Kaggle dataset, comprising 70,000 clinical records, is widely used in ML-based CVD prediction. Researchers have applied various models to improve accuracy and reliability. Ensemble learning combining Decision Tree, k-NN, and Naïve Bayes achieved over 99% accuracy [54], while deep neural networks outperformed traditional ML methods. Feature selection using PCA with XGBoost enhanced model performance . Random Forest demonstrated robustness with 88.04% accuracy and SMOTE(Synthetic Minority Over-sampling Technique) addressed class imbalance. Strengths include a large dataset and support for diverse ML techniques [54] However, data imbalance and limited generalizability remain challenges [54] , validation is needed to improve real-world applicability.

V. TRAINING METHODS

Model training optimizes algorithms to recognize patterns and make accurate predictions. For cardiovascular disease (CVD) prediction, five key models—SVM, Logistic Regression, Random Forest, XGBoost, and KNN—rely on specialized training methods. SVM requires parameter tuning and feature scaling[58], while Logistic Regression benefits from feature selection and regularization. Random Forest improves accuracy through hyperparameter optimization and ensemble learning[58]. XGBoost enhances performance via gradient boosting, regularization, and pruning[58]. KNN depends on distance metric selection and feature scaling[58]. Computational efficiency varies, with ensemble models like Random Forest and XGBoost requiring more resources than Logistic Regression and KNN [58] .Applying optimal training methods ensures effective CVD prediction.

➢ *Parameter Tuning and Feature Scaling*

Hyperparameter tuning optimizes model settings to enhance performance, while feature scaling standardizes input data for balanced influence. In training SVM models for CVD prediction, GridSearchCV systematically explored parameter combinations, improving accuracy to 81.96%[59]. Pandu et al, also applied GridSearchCV and RandomizedSearchCV, though with a broader focus on ML models. Kernel selection proved crucial, as Hoque et al. found polynomial SVM outperformed linear SVM. Feature scaling, essential for distance-based calculations, was emphasized in preprocessing strategies. While these

techniques enhance precision, they can be computationally demanding and risk overfitting if misapplied.

➢ *Feature Selection and Regularization*

Feature selection identifies the most relevant variables, while regularization prevents overfitting by penalizing large coefficients. In training logistic regression models for CVD prediction, Redhaei et al. applied Genetic Algorithm (GA), Particle Swarm Optimization , and Arithmetic Optimization Algorithm , achieving 98.59% accuracy[60]. Behki & Pal integrated PSO to refine feature selection, enhancing performance and reducing complexity. Regularization techniques improved model generalization, with Singh et al. achieving 90% accuracy in CVD prediction. Despite improving accuracy, these methods require careful parameter tuning and can reduce interpretability, affecting clinical adoption.

➢ *Ensemble Learning and Hyperparameter Optimization*

Ensemble learning, which combines multiple models for improved accuracy, and hyperparameter optimization, which fine-tunes parameters for optimal performance, have been effectively used in training Random Forest models for CVD prediction. Goad & Deore applied ensemble-based techniques, significantly boosting prediction accuracy[61]. Chithra et al. integrated a CNN-based Random Forest model, achieving 94.8% accuracy. Bayesian optimization further enhanced model robustness on datasets like Cleveland and Framingham [24]. Despite improved accuracy and feature importance identification, challenges like computational cost, data dependency, and limited interpretability remain, necessitating explainable AI techniques such as SHAP for clinical adoption [24]

➢ *Boosting, Regularization and Pruning Methods*

Boosting, a technique that enhances weak learners iteratively, is key to improving XGBoost's accuracy in cardiovascular disease (CVD) prediction. Studies by Zhou and Harshit et al. show that XGBoost outperforms KNN and Naïve Bayes in accuracy and F1 scores, making it highly effective for medical data. On the other hand Regularization, which penalizes model complexity to prevent overfitting, ensures XGBoost remains generalizable to unseen data, crucial in clinical applications. The use of L1 (Lasso) and L2 (Ridge) regularization further enhances interpretability. Pruning, which removes unimportant branches in decision trees, refines model simplicity and boosts reliability in clinical settings. Despite these strengths, hyperparameter tuning and high computational demands pose challenges, requiring optimized implementation to balance overfitting and underfitting for improved CVD prediction.

➢ *Distance Metric Selection and Feature Scaling Method*

Distance metric selection, which determines how similarity is measured between data points, and feature scaling, which ensures uniform contribution of all features, are critical in optimizing KNN for cardiovascular disease prediction. Prasetyo et al. found that using Cosine distance with 7 neighbors achieved 90.761% accuracy, outperforming Euclidean distance[179]. Feature scaling and selection

further enhanced performance, with Arif et al. demonstrating that Principal Component Analysis (PCA) improved accuracy from 83.6% to 90.16%[180], while Islam et al.used Information Gain and Correlation Heatmaps for dimensionality reduction. These methods boost KNN's predictive power, but challenges remain, as performance depends on careful distance metric selection, dataset size, and feature distribution.

VI. EVALUATION OF MACHINE LEARNING MODELS

Model evaluation is the process of assessing a machine learning model's performance to ensure its reliability in real-world applications, particularly in cardiovascular disease (CVD) prediction. Key metrics include accuracy, AUC-ROC, sensitivity, specificity, and confusion matrix analysis. Accuracy measures correct classifications, with XGBoost and ensemble methods achieving high scores[62], [63]. AUC-ROC assesses class separation, where XGBoost and LightGBM excel [63], [64] Sensitivity gauges CVD case detection, improved by Random Forest with SMOTE[65]. Specificity, reducing false positives, is strong in logistic regression and decision trees[66]. Confusion matrix analysis details classification outcomes, widely used in SVM and logistic regression [62]. Addressing class imbalances with SMOTE further improves model performance[65],[64]

➢ *Accuracy, Precision, Recall and F1 Score*

XGBoost consistently delivers the highest performance in CVD prediction, reaching up to 98.5% accuracy, 99.1% precision, 98.3% recall, and a 98.7% F1 score, thanks to its powerful ensemble learning techniques [67], [63]. However, its computational demands can be a drawback [58]. Random Forest follows closely with an accuracy of 89%, offering strong precision and recall due to its robustness in handling complex data [68]. Logistic Regression, while simpler, achieves 81–85% accuracy and is particularly effective at identifying true-positive cases [63]. SVM performs moderately, with 74% accuracy, but its sensitivity to parameter tuning affects consistency. KNN struggles the most, reaching only 69% accuracy, largely due to its reliance on proper feature scaling. While XGBoost leads in predictive power, models like Random Forest and Logistic Regression provide a balance between accuracy and computational efficiency, making them viable options for real-world applications.

➢ *AUC-ROC*

Random Forest stands out as the top performer in CVD prediction, achieving an AUC-ROC of 0.89, especially when combined with ensemble techniques like voting and stacking[49]. It also maintains high accuracy at 88.04%, surpassing XGBoost, SVM, and KNN [69]. The use of SMOTE further strengthens its ability to handle class imbalances [65]. XGBoost follows closely with an AUC of 0.79 but is generally outperformed by Random Forest in predictive accuracy [49]. SVM shows promise but does not exceed Random Forest in AUC-ROC performance, though resampling techniques like SMOTE improve its [68], [69], KNN struggles in comparison, being highly sensitive to

dataset size and feature selection, which limits its performance [68][189], Logistic Regression, often used as a baseline model, remains less effective than ensemble methods, even with LASSO regularization [65], [68]. While Random Forest dominates in AUC-ROC performance, the best model choice depends on dataset characteristics and specific clinical needs.

➤ *Sensitivity*

Random Forest and XGBoost show the highest sensitivity in CVD prediction, excelling in identifying true positives[65], [192] Random Forest performs best with SMOTE for class imbalance [65], while XGBoost benefits from hyperparameter tuning[68]. Logistic Regression, though simpler, achieves good sensitivity with LASSO. SVM improves with SMOTE but lags behind ensemble models [65], [68].KNN struggles, as performance depends on parameters and dataset quality[68]. Model choice depends on data and interpretability needs.

➤ *Confusion Matrix*

Random Forest consistently outperforms other models in CVD prediction, achieving high accuracy (88.04%) and excelling in sensitivity and specificity, especially when combined with SMOTE[65], [69] XGBoost is a strong contender, offering high accuracy with proper hyperparameter tuning but requiring more computational power[68], [69]. Logistic Regression, valued for its simplicity, serves as a reliable baseline but lags in accuracy and sensitivity [65],[68]. SVM improves with resampling techniques but remains less effective than ensemble models. KNN struggles with accuracy and is highly dependent on parameter tuning [68]. While Random Forest leads overall, model choice depends on data constraints, computational resources, and interpretability needs.

From the evaluation, XGBoost is a top choice for CVD prediction due to its high accuracy, efficiency, and ability to handle complex data. Its ensemble learning captures non-linear relationships better than models like Logistic Regression, SVM, and KNN, achieving ROC AUC scores up to 0.9973[38]. It manages correlated data like hospital-level effects [36] and integrates feature selection methods such as Grey Wolf Optimization to reduce overfitting. Unlike CNNs, which struggle with sparse tabular data, XGBoost excels in medical datasets. While its computational demands are higher, its predictive power makes it ideal for CVD risk assessment in clinical settings.

VII. CHALLENGES AND OPPORTUNITIES IN ML BASED CVD PREDICTION

➤ *Model Interpretability*

Interpretability challenges hinder ML adoption for CVD prediction in Uganda. Complex models lack transparency, making clinical trust difficult [70] .Uganda's diverse healthcare data adds complexity, requiring advanced feature selection[71] Explainable AI techniques, like rule extraction and attention mechanisms, improve transparency[72] . Better interpretability fosters trust, adoption, and personalized care, essential for AI-driven CVD prediction.

➤ *Data Quality and Imbalance*

This includes incomplete or skewed datasets bias models toward majority classes, reducing sensitivity for high-risk patients [73],[74]Traditional models struggle with imbalanced data, affecting accuracy and recall [75]. Resampling techniques like SMOTE and ensemble methods like boosting improve predictions [76], [75] Advanced solutions, including weighted Random Forests and GANs, further mitigate imbalance [73]. Fine-tuning XGBoost and standardized preprocessing enhance accuracy and generalizability [74],[77]. Adapting these strategies to Uganda's healthcare system is key to improving ML-driven CVD prediction and reducing mortality.

➤ *Computational and Infrastructure Limitation*

ML-based CVD prediction in developing countries like Uganda faces challenges in data availability, computational resources, and system integration. Limited access to quality datasets affects model training, requiring better data collection and integration [78] .Infrastructure constraints hinder the deployment of complex models, though cloud-based solutions can provide scalable resources. Advanced models like deep learning demand sophisticated implementation, which may be difficult in resource-limited settings, emphasizing the need for simpler, efficient alternatives[79]. Developing user-friendly interfaces can improve adoption in clinical workflows. Strategic partnerships and tailored ML approaches are essential to overcoming these barriers and enhancing CVD prediction in Uganda.

➤ *Generalization Across Diverse population*

ML-based CVD prediction faces challenges in generalizing across diverse populations, especially in Uganda and other developing countries, due to variations in data quality, risk factors, and inherent biases. Limited access to standardized datasets affects model accuracy when applied to different regions[77]. Population-specific factors, such as genetic markers and chronic inflammation in HIV-positive individuals, must be integrated to improve predictions[80]. Biases in sampling can lead to disparities, requiring techniques like resampling to enhance fairness[81]. Adaptable models using data optimization and ensemble learning perform better across diverse datasets [78]. Overcoming these challenges demands not just technical solutions but also a deeper understanding of local health profiles and socio-economic influences.

➤ *Opportunities for Advancing ML in CVD Prediction*

Advancing machine learning (ML) in cardiovascular disease (CVD) prediction in Uganda and developing countries presents a major opportunity to improve early detection and healthcare outcomes. ML models like Random Forest, Gradient Boosting, and Neural Networks have achieved up to 99% accuracy, enhancing risk assessment and timely interventions [82],[78], Ensemble learning further boosts accuracy by capturing complex patient data patterns [80]. Leveraging diverse datasets, including those

from the UCI repository, refines model performance [83], while feature engineering and hyperparameter tuning enhance predictive accuracy[78]. Integrating ML into cloud-connected healthcare systems ensures accessibility for clinicians, enabling better risk assessment and personalized treatments[84]. However, challenges like data privacy, healthcare integration, and region-specific risk factors—such as HIV-related CVD risks in Sub-Saharan Africa—must be addressed for effective implementation[80].

VIII.   FUTURE DIRECTIONS

➤ *Improving Data Quality*

Improving data quality is crucial for advancing machine learning ML in CVD prediction in Uganda and other developing countries. Hybrid feature selection methods, such as the CVD Prediction Framework , enhance data reliability by filtering irrelevant features, improving model performance[85].Data cleaning and standardization, as seen in Uganda's weather data processing, can be applied to CVD datasets to eliminate inconsistencies and improve accuracy [86], [87]. Health policies promoting risk reduction strategies—such as tobacco control and physical activity—can also lead to better data collection and quality [88]. Data optimization frameworks, including standardization and oversampling techniques, enhance model generalizability across populations [77]. However, challenges like limited data collection points and resource constraints persist, requiring collaboration between policymakers, healthcare providers, and technologists to ensure sustainable improvements in predictive accuracy.

➤ *Advancing Model Interpretability and Trust*

This involves balancing accuracy through use of ensemble models like XGBoost, when integrated with user-friendly tools like Streamlit, improve clinical usability[79]. The Randelistic algorithm, combining multiple classifiers, enhances accuracy and reduces false negatives [89]. AI (XAI) methods, such as SHAP and LIME, address the "black box" issue, fostering trust among clinicians [90], [91]. Exploring alternative algorithms and optimizing features can boost generalization and reduce bias, ensuring broader applicability. Future efforts should focus on expanding datasets, clinical trials, and cost-effective AI integration to enhance real-world adoption.

➤ *Deploying Scalable and Resource-Efficient Solutions*

Addressing high computational demands in ML-based CVD prediction in Uganda and developing countries requires scalable, cost-effective solutions. Leveraging existing healthcare infrastructure, such as HIV/AIDS care systems, can support cardiovascular care by training specialists and establishing regional centres. Mobile-based solutions like CVD Magic enable early risk detection without expensive lab tests, improving accessibility. Innovative approaches, including DNA computing, reduce computational load through parallel processing, while optimized deep learning models enhance prediction accuracy with minimal resources. Grid computing consolidates available resources to boost efficiency in financially constrained settings. Overcoming challenges in

technical expertise, funding, and long-term sustainability is crucial for successful implementation.

IX.   CONCLUSION

This review underscores the potential of machine learning in cardiovascular disease prediction, highlighting both its advancements and challenges in diverse healthcare settings. Among various models, XGBoost, a powerful gradient boosting algorithm, emerges as the most effective due to its superior accuracy, robustness, and ability to handle imbalanced datasets[17], [68]. Its efficiency in feature selection, scalability, and adaptability makes it well-suited for real-world clinical applications, particularly in resource-limited settings like Uganda. While challenges such as data quality, interpretability, and computational constraints persist, leveraging XGBoost with optimized preprocessing and explainability techniques presents a promising path toward more reliable and impactful CVD prediction models.

REFERENCES

[1].   R. Kakkar and R. T. Lee, "Cardiovascular diseases," Drug Discovery Today: Disease Models. Accessed: Feb. 05, 2025. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases?utm_source=chatgpt.com#tab=tab_1

[2].   J. J. Joseph et al., "Comprehensive Management of Cardiovascular Risk Factors for Adults with Type 2 Diabetes: A Scientific Statement from the American Heart Association," Circulation, vol. 145, no. 9, pp. 722–759, Mar. 2022, doi: 10.1161/CIR.0000000000001040/SUPPL_FILE/SUPPLEMENTAL.

[3].   G. A. Roth et al., "Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study," J. Am. Coll. Cardiol., vol. 76, no. 25, pp. 2982–3021, Dec. 2020, doi: 10.1016/J.JACC.2020.11.010.

[4].   S. S. Virani et al., "Heart Disease and Stroke Statistics—2021 Update," Circulation, vol. 143, no. 8, pp. E254–E743, Feb. 2021, doi: 10.1161/CIR.0000000000000950.

[5].   G. A. Mensah, G. A. Roth, and V. Fuster, "The Global Burden of Cardiovascular Diseases and Risk Factors: 2020 and Beyond," J. Am. Coll. Cardiol., vol. 74, no.

20, pp. 2529–2532, Nov. 2019, doi: 10.1016/J.JACC.2019.10.009.

[6]. J. Wu et al., "Lifestyle behaviors and risk of cardiovascular disease and prognosis among individuals with cardiovascular disease: a systematic review and meta-analysis of 71 prospective cohort studies," Int. J. Behav. Nutr. Phys. Act., vol. 21, no. 1, pp. 1–17, Dec. 2024, doi: 10.1186/S12966-024-01586-7/FIGURES/5.

[7]. H. C. Kim et al., "2018 Korean Society of Hypertension guidelines for the management of hypertension: Part I-epidemiology of hypertension," Clin. Hypertens., vol. 25, no. 1, pp. 1–6, Aug. 2019, doi: 10.1186/S40885-019-0121-0/TABLES/3.

[8]. Y. Anistyasari, S. C. Hidayati, R. Harimurti, and Ekohariadi, "A Random Forest Algorithm for Predicting Computer Programming Skill Associated with Learning Styles," 2023 6th Int. Conf. Vocat. Educ. Electr. Eng. Integr. Scalable Digit. Connect. Intell. Syst. Green Technol. Educ. Sustain. Community Dev. ICVEE 2023 - Proceeding, pp. 162–166, Oct. 2023, doi: 10.1109/ICVEE59738.2023.10348199.

[9]. K. I. Taher, A. M. Abdulazeez, and D. A. Zebari, "Data Mining Classification Algorithms for Analyzing Soil Data," Asian J. Res. Comput. Sci., pp. 17–28, May 2021, doi: 10.9734/AJRCOS/2021/V8I230196.

[10]. R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," Information, vol. 15, no. 4, Apr. 2024, doi: 10.3390/INFO15040235.

[11]. I. Mohammed Hassoon, "Boosting Learning Algorithms for Chronic Diseases Prediction: A Review," Iraqi J. Comput. Informatics, vol. 50, no. 2, pp. 22–30, Oct. 2024, doi: 10.25195/IJCI.V50I2.506.

[12]. S. Chatakondu and K. Zhai, "An Analysis of the k-Nearest Neighbor Classifier to Predict Benign and Malignant Breast Cancer Tumors," J. Student Res., vol. 12, no. 4, Nov. 2023, doi: 10.47611/JSRHS.V12I4.5577.

[13]. Y. Wu, "Heart Disease Prediction Using Gradient Boosting Decision Trees," Proc. 1st Int. Conf. Eng. Manag. Inf. Technol. Intell., pp. 527–535, Jan. 2024, doi: 10.5220/0012958300004508.

[14]. B. R. Wankar, N. V. Kshirsagar, A. V. Jadhav, S. R. Bawane, and S. M. Koshti, "Innovative Deep Learning Approach for Parkinson's Disease Prediction: Leveraging Convolutional Neural Networks for Early Detection," EAI Endorsed Trans. Pervasive Heal. Technol., vol. 10, May 2024, doi: 10.4108/EETPHT.10.6190.

[15]. N. . Shangaranarayanee, K. Hareesh Kumar, T. Jagadeesh, and S. Karthik, "Disease Prognosis Using Artificial Intelligence Neural Networks," J. Artif. Intell. Capsul. Networks, vol. 6, no. 1, pp. 1–14, Mar. 2024, doi: 10.36548/JAICN.2024.1.001.

[16]. H. Byeon et al., "Deep Neural Network Model for Enhancing Disease Prediction using Auto Encoder based Broad Learning.," SLAS Technol., vol. 29, no.

3, pp. 100145–100145, May 2024, doi: 10.1016/J.SLAST.2024.100145.

[17]. N. R. Barman, K. Sharma, and R. Hazra, "A transformer-based approach to automate disease prediction from patient descriptions," 2023 IEEE 7th Conf. Inf. Commun. Technol. CICT 2023, pp. 1–5, Dec. 2023, doi: 10.1109/CICT59886.2023.10455356.

[18]. S. K. Binu and C. Shanthi, "Clinical Insight: Comparative Analysis of Deep Learning Models for Disease Prediction across Multifaceted Datasets," 3rd IEEE Int. Conf. Distrib. Comput. Electr. Circuits Electron. ICDCECE 2024, Apr. 2024, doi: 10.1109/ICDCECE60827.2024.10548473.

[19]. M. A. Bülbül, "A novel hybrid deep learning model for early stage diabetes risk prediction," J. Supercomput., vol. 80, no. 13, pp. 19462–19484, May 2024, doi: 10.1007/S11227-024-06211-9.

[20]. G. Airlangga, "A Hybrid CNN-RNN Model for Enhanced Anemia Diagnosis: A Comparative Study of Machine Learning and Deep Learning Techniques," Indones. J. Artif. Intell. Data Min., vol. 7, no. 2, p. 366, May 2024, doi: 10.24014/IJAIDM.V7I2.29898.

[21]. L. Wang, C. Zhang, and J. Li, "A Hybrid CNN-Transformer Model for Predicting N Staging and Survival in Non-Small Cell Lung Cancer Patients Based on CT-Scan," Tomography, vol. 10, no. 10, pp. 1676–1693, Oct. 2024, doi: 10.3390/TOMOGRAPHY10100123.

[22]. M. S. Arif, A. U. Rehman, and D. Asif, "Explainable Machine Learning Model for Chronic Kidney Disease Prediction," Algorithms, vol. 17, no. 10, pp. 443–443, Oct. 2024, doi: 10.3390/A17100443.

[23]. V. N. Manju, N. Aparna, and K. Krishna Sowjanya, "Decision Tree-Based Explainable AI for Diagnosis of Chronic Kidney Disease," Proc. 5th Int. Conf. Inven. Res. Comput. Appl. ICIRCA 2023, pp. 947–952, Aug. 2023, doi: 10.1109/ICIRCA57980.2023.10220774.

[24]. I. D. Mienye and N. Jere, "Optimized Ensemble Learning Approach with Explainable AI for Improved Heart Disease Prediction," Information, vol. 15, no. 7, pp. 394–394, Jul. 2024, doi: 10.3390/INFO15070394.

[25]. C. S. Muñoz-Valencia, J. A. Quesada, D. Orozco-Beltran, and X. Barber, "Bayesian networks for disease diagnosis: What are they, who has used them and how? (Preprint)," Apr. 2023, doi: 10.2196/PREPRINTS.48570.

[26]. "A Framework for Multi-modal Learning: Jointly Modeling Inter- & Intra-Modality Dependencies," May 2024, doi: 10.48550/ARXIV.2405.17613.

[27]. S. Li and R. Zhang, "A novel interactive deep cascade spectral graph convolutional network with multi-relational graphs for disease prediction.," Neural Networks, vol. 175, pp. 106285–106285, Apr. 2024, doi: 10.1016/J.NEUNET.2024.106285.

[28]. Y. Ohnuki, M. Akiyama, and Y. Sakakibara, "Deep learning of multimodal networks with topological regularization for drug repositioning," J.

Cheminform., vol. 16, no. 1, Aug. 2024, doi: 10.1186/S13321-024-00897-Y.

[29]. W. Wang, "Heart disease prediction using machine learning models," Theor. Nat. Sci., vol. 51, no. 1, pp. 9–17, Oct. 2024, doi: 10.54254/2753-8818/51/2024CH0144.

[30]. S. Yadav, A. Yadav, D. S. Srivastava, and P. Rai, "Heart Disease Prediction Using Machine Learning," Indian Sci. J. Res. Eng. Manag., vol. 08, no. 008, pp. 1–4, Sep. 2024, doi: 10.55041/IJSREM37304.

[31]. L. S, "Predicting Heart Disease through Machine Learning Methods," Int. J. Innov. Sci. Res. Technol., pp. 829–842, Sep. 2024, doi: 10.38124/IJISRT/IJISRT24SEP382.

[32]. J. Hassan, "Heart Disease Prediction Using Machine Learning Algorithms," J. Innov. Comput. Emerg. Technol., vol. 4, no. 2, Oct. 2024, doi: 10.56536/JICET.V4I2.145.

[33]. Z. Ahmed, "Abstract 4113047: Discovering novel biomarkers and predicting cardiovascular disease using AI/ML techniques for precision medicine," Circulation, vol. 150, no. Suppl_1, Nov. 2024, doi: 10.1161/CIRC.150.SUPPL_1.4113047.

[34]. "Advancements In Heart Disease Prediction: A Machine Learning Approach For Early Detection And Risk Assessment," Oct. 2024, doi: 10.48550/ARXIV.2410.14738.

[35]. S. Tomar, D. Dembla, and Y. Chaba, "Analysis and Enhancement of Prediction of Cardiovascular Disease Diagnosis using Machine Learning Models SVM, SGD, and XGBoost," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 4, pp. 469–479, 2024, doi: 10.14569/IJACSA.2024.0150449.

[36]. T. Dong et al., "Enhancing Cardiovascular Risk Prediction: Development of an Advanced Xgboost Model with Hospital-Level Random Effects," Bioengineering, vol. 11, no. 10, pp. 1039–1039, Oct. 2024, doi: 10.3390/BIOENGINEERING11101039.

[37]. Y. Zhou, "Prediction and Analysis of Cardiovascular Diseases Based on XGBoost," 2024 IEEE 2nd Int. Conf. Image Process. Comput. Appl. ICIPCA 2024, pp. 1877–1881, Jun. 2024, doi: 10.1109/ICIPCA61593.2024.10709218.

[38]. D. Bhardwaz, Y. Singla, R. Kaur, and L. Sharma, "Advancing Cardiovascular Health Risk Assessment: A Comprehensive Study on Heart Attack Prediction Using XGBoost Models and Cross-Validation," 2024 Int. Conf. Adv. Comput. Res. Sci. Eng. Technol. ACROSET 2024, pp. 1–5, Sep. 2024, doi: 10.1109/ACROSET62108.2024.10743413.

[39]. S. N. N. Arif, A. M. Siregar, S. Faisal, and A. R. Juwita, "Klasifikasi Penyakit Serangan Jantung Menggunakan Metode Machine Learning K-Nearest Neighbors (KNN) dan Support Vector Machine (SVM)," J. media Inform. Budidarma, vol. 8, no. 3, p. 1617, Jul. 2024, doi: 10.30865/MIB.V8I3.7844.

[40]. C. Yan, Y. Xing, S. Liu, E. Gao, and J. Wang, "Machine Learning Models for Cardiovascular Disease Prediction: A Comparative Study," Jun. 2024, doi: 10.1101/2024.05.27.596092.

[41]. Suman Kumar Swarnkar, "Random Forest-Based Prediction Models for Assessing Cardiovascular Disease Risk: Integrating Clinical and Genetic Factors," J. Electr. Syst., vol. 20, no. 3s, pp. 663–672, Apr. 2024, doi: 10.52783/JES.1352.

[42]. V. R. Burugadda, V. Dutt, Mamta, and N. Vyas, "Personalized Cardiovascular Disease Risk Prediction Using Random Forest: An Optimized Approach," Proc. - 2023 IEEE World Conf. Appl. Intell. Comput. AIC 2023, pp. 226–232, Jul. 2023, doi: 10.1109/AIC57670.2023.10263915.

[43]. M. Naghavi et al., "Abstract 4144083: AI-CVD: Artificial Intelligence-Enabled Opportunistic Screening of Coronary Artery Calcium Computed Tomography Scans for Predicting CVD Events and All-Cause Mortality: The Multi-Ethnic Study of Atherosclerosis (MESA)," Circulation, vol. 150, no. Suppl_1, Nov. 2024, doi: 10.1161/CIRC.150.SUPPL_1.4144083.

[44]. S. P. Singh, A. Singh, and M. Kumari, "Beyond Traditional Methods: Utilizing Regularized Logistic Regression for Accurate Heart Disease Forecasting," Proc. - 2024 Int. Conf. Emerg. Innov. Adv. Comput. INNOCOMP 2024, pp. 247–251, May 2024, doi: 10.1109/INNOCOMP63224.2024.00048.

[45]. H. Miao, "Logistic regression for cardiovascular diseases prediction by integrating PCA and K-means ++," Theor. Nat. Sci., vol. 38, no. 1, pp. 126–132, Jun. 2024, doi: 10.54254/2753-8818/38/20240569.

[46]. H. Azis, "Assessing the Performance of Logistic Regression in Heart Disease Detection through 5-Fold Cross-Validation," Int. J. Artif. Intell. Med. Issues, vol. 2, no. 1, pp. 1–11, May 2024, doi: 10.56705/IJAIMI.V2I1.137.

[47]. W. DeGroat et al., "Multimodal AI/ML for discovering novel biomarkers and predicting disease using multi-omics profiles of patients with cardiovascular diseases," Dent. Sci. reports, vol. 14, no. 1, p. 26503, Nov. 2024, doi: 10.1038/S41598-024-78553-6.

[48]. Raza Naeem, "Machine and deep learning techniques for cardiovascular disease detection," J. Innov. Comput. Emerg. Technol., vol. 4, no. 2, Oct. 2024, doi: 10.56536/JICET.V4I2.131.

[49]. N. Ghaniaviyanto Ramadhan, Adiwijaya, W. Maharani, and A. Akbar Gozali, "Prediction of Cardiovascular Disease (CVD) in the Upcoming Year Using Tree-Based Ensemble Model," 2024 12th Int. Conf. Inf. Commun. Technol. ICoICT 2024, pp. 210–216, Aug. 2024, doi: 10.1109/ICOICT61617.2024.10698310.

[50]. R. Bhuvaneswari, S. Karthigeyan, and T. Sabhanayagam, "CVNET - Cardiovascular Disease Detection using Deep Convolutional Neural Network," 2024 Asia Pacific Conf. Innov. Technol. APCIT 2024, pp. 1–6, Jul. 2024, doi: 10.1109/APCIT62007.2024.10673606.

[51]. N. V. DURGA SAI SIVA VARA PRASAD RAJU and P. N. DEVI, "AI-Assisted Medical Imaging and Heart Disease Diagnosis: A Deep Learning Approach for Automated Analysis and Enhanced Prediction

Using Ensemble Classifiers," J. Artif. Intell. Gen. Sci. ISSN3006-4023, vol. 6, no. 1, pp. 210–229, Oct. 2024, doi: 10.60087/JAIGS.V6I1.242.

[52]. B. Surya, B. Venkatesh, S Vijayalakshmi, A. H. Narayanan, and R. Syed, "AI for Early CVD Diagnosis and Personalized Care," Int. Res. J. Med. Surg., vol. 01, no. 02, pp. 09–18, Jan. 2024, doi: 10.47857/IRJMEDS.2024.V01I02.008.

[53]. S. Dalal et al., "Application of Machine Learning for Cardiovascular Disease Risk Prediction," Comput. Intell. Neurosci., vol. 2023, no. 1, pp. 1–12, Mar. 2023, doi: 10.1155/2023/9418666.

[54]. I. Ilham, "Enhancing Cardiovascular Disease Prediction Accuracy through an Ensemble Machine Learning Approach," Int. J. Artif. Intell. Med. Issues, vol. 2, no. 2, pp. 95–103, Nov. 2024, doi: 10.56705/IJAIMI.V2I2.157.

[55]. C. Mansoor, S. K. Chettri, and H. Naleer, "Advancing Heart Disease Prediction through Synergistic Integration of Machine Learning and Deep Learning Techniques," Proc. - 3rd Int. Conf. Adv. Comput. Commun. Appl. Informatics, ACCAI 2024, May 2024, doi: 10.1109/ACCAI61061.2024.10602447.

[56]. "Centralized and Federated Heart Disease Classification Models Using UCI Dataset and their Shapley-value Based Interpretability," Aug. 2024, doi: 10.48550/ARXIV.2408.06183.

[57]. C. Sudlow et al., "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," PLoS Med., vol. 12, no. 3, Mar. 2015, doi: 10.1371/JOURNAL.PMED.1001779.

[58]. R. G. - et al., "Review on Heart Disease Prediction using Machine Learning Approaches," Int. J. Multidiscip. Res., vol. 6, no. 5, Oct. 2024, doi: 10.36948/IJFMR.2024.V06I05.29475.

[59]. D. P. Mishra, H. K. Gupta, G. Saajith, and R. Bag, "Optimizing Heart Disease Prediction Model with GridsearchCV for Hyperparameter Tuning," 2024 1st Int. Conf. Cogn. Green Ubiquitous Comput. IC-CGU 2024, pp. 1–6, Mar. 2024, doi: 10.1109/IC-CGU58078.2024.10530772.

[60]. A. Al Redhaei, W. A. Fares, and M. A. Al Betar, "Utilizing Optimization Techniques in Feature Selection for Effective Cardiovascular Disease Prediction," 2023 24th Int. Arab Conf. Inf. Technol. ACIT 2023, pp. 1–6, Dec. 2023, doi: 10.1109/ACIT58888.2023.10453772.

[61]. P. M. Goad and P. J. Deore, "Predicting and Analyzing Cardiovascular Disease through Ensemble Learning Approaches," Int. Res. J. Multidiscip. Technovation, vol. 6, no. 5, pp. 153–163, Sep. 2024, doi: 10.54392/IRJMT24510.

[62]. H. Liu, Y. Tian, and D. Yu, "Prediction of cardiovascular and cerebrovascular diseases based on machine learning models," Appl. Comput. Eng., vol. 46, no. 1, pp. 35–44, Mar. 2024, doi: 10.54254/2755-2721/46/20241068.

[63]. Nishat Anjum et al., "Improving Cardiovascular Disease Prediction through Comparative Analysis of Machine Learning Models," J. Comput. Sci. Technol. Stud., vol. 6, no. 2, pp. 62–70, Apr. 2024, doi: 10.32996/JCSTS.2024.6.2.7.

[64]. Y. Dou, J. Liu, W. Meng, and Y. Zhang, "Comparative analysis of supervised learning algorithms for prediction of cardiovascular diseases.," Technol. Heal. Care, vol. 32, pp. 241–251, Apr. 2024, doi: 10.3233/THC-248021.

[65]. U. Hasanah, A. M. Soleh, and K. Sadik, "Effect of Random Under sampling, Oversampling, and SMOTE on the Performance of Cardiovascular Disease Prediction Models," J. Mat. Stat. dan Komputasi, vol. 21, no. 1, pp. 88–102, Sep. 2024, doi: 10.20956/J.V21I1.35552.

[66]. M. Chandrika, R. Kiran, P. Vaidya, and R. Kodnad, "Evaluation of Machine Learning Models for Cardiovascular Risk Assessment," 2nd Int. Conf. Intell. Cyber Phys. Syst. Internet Things, ICoICI 2024 - Proc., pp. 993–996, Aug. 2024, doi: 10.1109/ICOICI62503.2024.10696637.

[67]. M. N. Sailaja, N. Ramakrishnaiah, and K. Swaroopa, "Cardiovascular Wellness ThroughTechnology: A Closer Look at Machine Learning Techniques," 2nd IEEE Int. Conf. Adv. Inf. Technol. ICAIT 2024 - Proc., pp. 1–8, Jul. 2024, doi: 10.1109/ICAIT61638.2024.10690421.

[68]. "Comparative Study of Machine Learning Algorithms in Detecting Cardiovascular Diseases, Dayana," May 2024, doi: 10.48550/ARXIV.2405.17059.

[69]. K. Niharika, U. B. Sofi, B. F. Ahmed, M. Arun, J. Ravi, and V. G. Krishnan, "Cardiovascular Disease Prediction Through Machine Learning Algorithms," Proc. 3rd Int. Conf. Appl. Artif. Intell. Comput. ICAAIC 2024, pp. 583–588, Jun. 2024, doi: 10.1109/ICAAIC60222.2024.10574997.

[70]. R. K. P. Tripathi and S. Tiwari, "Unravelling the Enigma of Machine Learning Model Interpretability in Enhancing Disease Prediction," Adv. Syst. Anal. Softw. Eng. high Perform. Comput. B. Ser., pp. 125–153, Dec. 2023, doi: 10.4018/978-1-6684-8531-6.CH007.

[71]. A. H. Elmi, A. Abdullahi, and M. A. Barre, "A machine learning approach to cardiovascular disease prediction with advanced feature selection," Indones. J. Electr. Eng. Comput. Sci., vol. 33, no. 2, pp. 1030–1041, Feb. 2024, doi: 10.11591/IJEECS.V33.I2.PP1030-1041.

[72]. C. Skouteli, N. Prenzas, A. Kakas, and C. S. Pattichis, "Explainable AI Modeling in the Prediction of Cardiovascular Disease Risk.," Stud. Health Technol. Inform., vol. 316, pp. 978–982, Aug. 2024, doi: 10.3233/SHTI240574.

[73]. "Mitigating Learning Bias in Healthcare Datasets," Kumbhalwar, Jul. 2024, doi: 10.31979/ETD.98WG-H7W3.

[74]. A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, "Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases.," Diagnostics, vol. 14, no. 2, Jan. 2024, doi: 10.3390/DIAGNOSTICS14020144.

[75]. S. Naganjaneyulu, G. Akanksha, S. Shaheeda, and M. Sadhak, "HMLF_CDD_SSBM: A Hybrid Machine Learning Framework for Cardiovascular Disease Diagnosis Prediction Using the SMOTE Stacking Method," Lect. Notes Networks Syst., vol. 537 LNNS, pp. 571–585, Jan. 2023, doi: 10.1007/978-981-99-3010-4_47.

[76]. K. V. Tompra, G. Papageorgiou, and C. Tjortjis, "Strategic Machine Learning Optimization for Cardiovascular Disease Prediction and High-Risk Patient Identification," Algorithms, vol. 17, no. 5, Apr. 2024, doi: 10.3390/A17050178.

[77]. P. C. Kanth, S. Vijayalakshmi, and T. S. Palathara, "Machine Learning Model Enabled with Data Optimisation for Prediction of Coronary Heart Disease," pp. 1–5, Mar. 2024, doi: 10.1109/TQCEBT59414.2024.10545194.

[78]. A. Kachhawa and J. Hitt, "An Intelligent System for Early Prediction of Cardiovascular Disease using Machine Learning," J. Student Res., vol. 11, no. 3, Mar. 2023, doi: 10.47611/JSRHS.V11I3.2989.

[79]. M. A. Sufian et al., "Enhancing Clinical Validation for Early Cardiovascular Disease Prediction through Simulation, AI, and Web Technology," Diagnostics, vol. 14, no. 12, pp. 1308–1308, Jun. 2024, doi: 10.3390/DIAGNOSTICS14121308.

[80]. V. O. Eguavoen, F. I. Amadin, and E. Nwelih, "Cardiovascular Disease Risk Prediction For People Living With Hiv Using Ensemble Deep Neural Network," Int. Conf. Sci. Eng. Bus. Driv. Sustain. Dev. Goals, SEB4SDG 2024, pp. 1–9, Apr. 2024, doi: 10.1109/SEB4SDG60871.2024.10629982.

[81]. F. Li, J. Zhao, P. Wu, H. H. Ong, W. Wei, and J. F. Peterson, "Abstract 15589: Evaluating Methods to Mitigate the Bias for Machine Learning-Based Cardiovascular Risk Model," Circulation, vol. 146, no. Suppl_1, Nov. 2022, doi: 10.1161/CIRC.146.SUPPL_1.15589.

[82]. C. J. Ejiyi et al., "Enhanced Cardiovascular Disease Prediction Modelling using Machine Learning Techniques: A Focus on CardioVitalnet.," Netw. Comput. Neural Syst., pp. 1–33, Apr. 2024, doi: 10.1080/0954898X.2024.2343341.

[83]. M. S. Hossain, M. A. Talukder, and M. Z. Mahmud, "Advancements in Cardiovascular Disease Detection: Leveraging Data Mining and Machine Learning," bioRxiv, Mar. 2024, doi: 10.1101/2024.03.09.584222.

[84]. A. Petreska, "Cardiovascular disease prediction with machine learning techniques," J. Cardiol. Curr. Res., vol. 17, no. 2, pp. 41–51, Apr. 2024, doi: 10.15406/JCCR.2024.17.00603.

[85]. T. Shilpa and anal paul, "CVDPF: A Hybrid Feature Selection Method with Data-Driven Approach for Cardiovascular Disease Prediction Framework using Machine Learning," Dec. 2022, doi: 10.21203/RS.3.RS-2323170/V1.

[86]. A. G. Tumusiime, O. S. Eyobu, I. Mugume, and T. J. Oyana, "A weather features dataset for prediction of short-term rainfall quantities in Uganda," Data Br., vol. 50, Sep. 2023, doi: 10.1016/J.DIB.2023.109613.

[87]. "Cleaned Weather Dataset for Uganda," gahwera, Jan. 2023, doi: 10.7910/DVN/PQLYHP.

[88]. Alexandros Argyriadis, Chrisi Vlachou, Ioannis Andriopoulos, and Ioannis Dimitrakopoulos, "Health promotion strategies for cardiovascular disease in the community of Uganda: Policies, challenges, and pathways to a healthier future," Int. J. Sci. Res. Arch., vol. 10, no. 1, pp. 360–365, Sep. 2023, doi: 10.30574/IJSRA.2023.10.1.0759.

[89]. M. Ananthi, A. S. Narayanan, T. P. Dhiraj Prasad, and R. Jai Vignesh, "Cardiovascular Disease Prediction Using Randelistic Algorithm," Proc. Int. Conf. Circuit Power Comput. Technol. ICCPCT 2023, pp. 20–25, Aug. 2023, doi: 10.1109/ICCPCT58313.2023.10245957.

[90]. M. E. Shirley, N. H. Kasujja, and G. Marvin, "Shapley Additive Explanations (SHAP) for Cardiovascular Diseases Prediction," 2nd Int. Conf. Sustain. Comput. Smart Syst. ICSCSS 2024 - Proc., pp. 1429–1437, Jul. 2024, doi: 10.1109/ICSCSS60660.2024.10625027.

[91]. M. Madhavilatha and M. G. T. P. Kumari, "Interpretable Artificial Intelligence in Cardiovascular Health: An In-depth Analysis of Heart Disease Data," Indian Sci. J. Res. Eng. Manag., vol. 08, no. 02, pp. 1–13, Feb. 2024, doi: 10.55041/IJSREM28549.

AUTHORS

Mr. Ssimbwa Bashir: Ssimbwa Bashir received his Bachelor's degree in Computer Science from Kampala International University in November, 2024. His research interests include Data Engineering, Machine Learning Engineering and Data Science & Analytics.

Mr. Wasike Azizi ; Mr. wasike Azizi is a computer scientist, researcher, and lecturer with extensive expertise in Natural Language Processing, Machine Translation, Automata Theory, and Rule-Based Machine Translation. He is currently a PhD candidate in Computer Science (2022–

2025) at Kampala International University, where he is conducting research in advanced computational linguistics and AI-driven language technologies. He holds a Master of Science in Computer Science (2018–2020) and a Bachelor of Science in Computer Science (2009–2012) from the same institution. Additionally, he earned a Postgraduate Diploma in Education (2021–2022), equipping him with skills in teaching ICT and Physics. With a strong background in academia and research, Mr. Wasike Azizi is dedicated to advancing AI-driven innovations in language processing and computational modelling while contributing to education and technology development.

Mr. Ssebadduka Jamir is a computer scientist, researcher, and lecturer specializing in advanced computing and artificial intelligence. He is currently a PhD candidate in Computer Science (expected completion: November 2025) at Kampala International University, where he is conducting research in emerging technologies and computational methodologies.

He holds a Master's Degree in Computer Science, further strengthening his expertise in data-driven systems and software development. In addition to his academic pursuits, Mr. Ssebadduka Jamir is a lecturer at the International University of East Africa (IUEA) and the Islamic University in Uganda (IUIU), where he is committed to shaping the next generation of computer scientists through teaching and research.

His work focuses on innovative AI applications, system optimization, and technological advancements, contributing significantly to both academia and industry.