FMDB *Transactions on Sustainable Health Science Letters*

Machine Learning-Based Cardiovascular Disease Detection and Intelligent Healthcare Workflow Automation

Oluebube F. Odoemena[1, *], Gabriel A. Elufidodo[2]

[1,2,] Department of Computer Science, University of Nigeria, Nsukka, Nigeria.
oluebube.ikwuagwu.pg91407@unn.edu.ng[1], gabriel.elufidodo@unn.edu.ng[2]

Abstract: Cardiovascular diseases (CVDs) are a major cause of death worldwide, emphasizing the importance of better early detection methods. This study introduces a machine learning model to detect CVDs by analyzing vital sign information. The model, which was trained using a cardiovascular dataset, utilizes three algorithms: Support Vector Machine (SVM), Naïve Bayes, and Decision Tree. It assesses the health status of patients by predicting vital sign values, allowing healthcare providers to receive timely alerts. Following the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, a proof-of-concept web application was created with object-oriented design principles and implemented using Python. This application predicts the likelihood of CVD and streamlines the scheduling of doctor appointments, promoting quick medical intervention. Our findings reveal that the Decision Tree classifier performed the best in accurately identifying patients who are at risk based on abnormalities in vital signs. This method can potentially enhance early detection of CVDs and improve the timing of medical care.

Keywords: Cardiovascular Diseases; Machine Learning; Early Detection; Vital Signs; Support Vector Machine; Naive Bayes; Decision Tree; Cross-Industry Standard Process for Data Mining (CRISP-DM); Web Application; Healthcare Automation.

Received on: 11/01/2024, **Revised on:** 06/03/2024, **Accepted on:** 04/04/2024, **Published on:** 01/06/2024

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSHSL

DOI: https://doi.org/10.69888/FTSHSL.2024.000174

Cite as: O. F. Odoemena, and G. A. Elufidodo, "Machine Learning-Based Cardiovascular Disease Detection and Intelligent Healthcare Workflow Automation," *FMDB Transactions on Sustainable Health Science Letters*., vol. 2, no. 2, pp. 81–93, 2024.

1. Introduction

Cardiovascular diseases (CVDs) stand as a formidable global health challenge, exerting a considerable toll on public health by contributing significantly to morbidity and mortality across diverse populations. The World Health Organization (WHO) highlights that cardiovascular diseases claim a staggering 17.9 million lives annually [1], cementing their status as the foremost cause of death on a global scale. Early disease detection and precise diagnosis are imperative, as timely interventions can effectively mitigate associated risks, bolstering treatment outcomes. In recent times, the convergence of medical insights and computational prowess has paved the way for transformative advancements in healthcare. Machine learning (ML), as an aspect of artificial intelligence (AI), has emerged as a potent ally in addressing intricate medical challenges, particularly the early detection and prediction of cardiovascular diseases [2]. The confluence of data-driven methodologies with clinical expertise has the potential to usher in a paradigm shift in how healthcare practitioners approach risk assessment, diagnosis, and patient care. The timely and accurate detection of cardiovascular diseases (CVDs) is critical for effective treatment and better patient outcomes [16]. Recent progress in medical knowledge, especially in machine learning (ML) and artificial intelligence (AI), has

*Corresponding author.

paved the way for transformative innovations in healthcare [17]. This research represents a significant leap forward, focusing on developing an advanced web-based interface driven by machine learning technology [18].

This research centers on creating a cutting-edge web-based platform designed to automatically assign patients to appropriate doctors, facilitating early-stage detection of cardiovascular diseases [19]. By harnessing the power of machine learning, this interface acts as a dynamic and intelligent tool, streamlining the process of patient allocation and disease identification. This innovative approach aims to revolutionize traditional risk assessment, diagnosis, and patient care methods within the healthcare sector [20]. By integrating medical expertise with data-driven methodologies, we aim to redefine how healthcare professionals approach cardiovascular disease management [21]. Through this research, we aspire to create a user-friendly, accessible tool tailored for medical practitioners and healthcare workers [22]. This tool enhances the efficiency of patient assignment and contributes significantly to the overall improvement of cardiovascular disease management, ultimately leading to better healthcare outcomes for patients globally [23].

2. Literature Review

2.1. Cardiovascular Diseases

Integrating machine learning (ML) techniques into cardiovascular disease (CVD) detection is anchored in a robust theoretical framework that amalgamates data-driven methodologies with predictive modeling [24]. As a subset of artificial intelligence (AI), ML offers a paradigm shift in healthcare by capitalizing on its ability to uncover intricate patterns and latent relationships within voluminous and complex datasets. In CVD detection, ML algorithms present an avenue to unearth subtle precursors of disease onset, enabling timely intervention and improved patient outcomes [25].

Supervised learning algorithms are central to the theoretical foundation, underpinning the crux of CVD prediction. These algorithms encompass classification and regression models, each serving as an instrumental tool in discerning patterns and making informed predictions. Classification algorithms, such as logistic regression, operate on labeled data to quantify the relationship between input variables and the likelihood of a binary outcome, such as the presence of a cardiovascular condition [3]. On the other hand, Support Vector Machines (SVMs) navigate the intricacies of classification by leveraging a decision boundary to segregate data points into discrete categories, making them particularly adept at identifying complex patterns within multidimensional data [4].

Ensemble methods, exemplified by Random Forest, further enrich the theoretical landscape. These methods amalgamate multiple base models to enhance predictive accuracy. In the context of CVD detection, a Random Forest generates an ensemble of decision trees and amalgamates their outputs, leading to predictions of heightened robustness [26]. Additionally, Neural Networks, inspired by the human brain's neural structure, bring forth the domain of deep learning and hierarchical feature extraction. The unparalleled capacity of Neural Networks to model intricate nonlinear relationships within data is particularly advantageous in deciphering complex medical domains, including cardiovascular disease detection [27].

2.2. Prediction Models

Prediction models are used to analyze and assess the performance, future outcomes, or trends [5]. A sustainable forecasting model that improves crop assessment adoption and helps identify how to grow crops while managing diseases was presented herein. This model aims to improve the timeliness, effectiveness, and foresight to control crop diseases while reducing yield loss, environmental impact, and economic cost. The data set used in this approach was a combination of pathogen, host, and environment, and they were used under different assumptions. This design supported datasets, feasibility testing, evaluation of different models, and forecast metrics. Its framework integrates multivariate spatial assumptions and threshold-based infection, which improves previous approaches. It covers climate covariates, dynamic disease parameters, spatial dependence, and assumptions on disease climate. Southern Alberta, Canada, was the area of study, and it is the major area in western Canada that produces lots of agricultural produce with a growing period of about 123 days (May-August).

The Alberta Climate Information Service provided satellite measurements of the area. This report covered major disease risk variables like liquid water on the canopy surface and temperature between 1961 and 2016. A site-specific model was evaluated to predict the disease outbreak and implemented using a validated R code [28]. The hhh4 (a class of spatiotemporal models used for analyzing subnational case count data implemented in the R package) spatiotemporal-endemic model was also implemented to compare the spatial dependence assumptions with the CLR (Common Language Runtime) site-specific model predictions. At the end of the study, the integrated framework offered a feasible way to combine diverse datasets and models of different assumptions to explain uncertainty and variability in crop disease [29]. The hhh4 and CLR models predict infection time correctly, but the CLR model predicts the timing of the infection a week earlier than it occurred. In order to cover a larger

area, an expanded evaluation of the model will require large heterogeneous assumptions and a seasonal airborne surveillance program. Predictive models have the potential to control health events [30].

In order to know the probability of stratifying patients having a certain outcome, a prediction model was proposed using regression technique and the s methodology of prediction modeling. The model aims to help identify which patients have a better chance of recovering if a particular event happened to them and which treatment method is more adaptable for them [31]. This model was built to consider events like the probability of having a tumor and the chance of surviving a side effect or cancer. Linear regression and binary regression models were used while considering the metrics available. A time-to-event analysis of when this event would occur was also carried out. It was difficult for the model to perform well beyond what it was trained for [32]. This led to an understanding that a flexible algorithm should be used, which makes it easy to follow the data points in the training set, which will reduce errors. Since the model is a general model to predict several events, the number of features was reduced to improve its generalization and increase its interpretability. This model was tested to evaluate its performance using a separate dataset from its training set. It showed generalizability and faster computation time [33].

Liu et al. [6] predicted total crop output using a Wavelet neural network (WNN). This paper aimed to predict agricultural production in China from 2002 to 2010. Wavelet neural networks, which constituted neural network theory, and wavelet analysis theory were used to build the algorithm. The neural network has good fault tolerance capability and self-learning function, and the wavelet transform has a good time-frequency localization property, and both combined have a powerful advantage in the algorithm. A comparison test was conducted between the traditional back propagation (BP) network prediction model and the model trained using a gradient-descent algorithm. The results showed that this model had better prediction accuracy and a faster convergence rate. This model showed superior to the traditional BP neural network model.

Siontis et al. [7] Crop yield prediction review analyzed and investigated five hundred sixty-seven studies carefully. It was found that several crop yield prediction models are based on features like soil type, rainfall, and temperature. This study aims to know the features and algorithms used in crop yield prediction studies. A review protocol was defined where research questions were set and used to select relevant studies from the database. Once the studies were selected using specific filters, they were accessed using a quality criterion and set of exclusions. All the relevant studies used in this review were synthesized to ensure they respond to the research questions. It was noticed that models with many features did not provide the best performance, but those with fewer features did. This study showed that the algorithm most applied to these models is Artificial Neural Networks based on machine learning projects. It analyzed an additional 50 papers in deep learning literature, and it showed that the most widely used deep learning algorithm is Convolutional Neural Networks (CNN), and other commonly used algorithms are Long-Short Term Memory (LSTM) and Deep Neural Networks (DNN). At the same time, commonly used models are neural networks, random forests, gradient-boosting trees, and linear regression. This work showed a need for further research on crop yield prediction as no specific conclusion has been drawn as to what model is best for crop yield.

In precision agriculture, crop yield prediction is a challenging part. Many prediction models have been proposed and validated; every data set depends on different factors. It could be the use of fertilizer, climate, soil, seed variety, and weather [8] and the design of an integrated climatic assessment indicator (ICAI) for wheat production [34]. This study covered the Jiangsu Province in China and aimed to predict wheat yield variations. Several meteorological factors affect crops in that region's climatic suitability. Datasets produced from meteorological assimilations combined with observation data were used to construct the indicator. While modeling the ICAI algorithm, Support Vector (SVM) and Random Forest (RF) were compared to find which machine learning algorithm performs better in an independent test set [35]. In order to detect the climatic yield, Monte Carlo simulations were carried out to determine a reasonable division for the Kolmogorov-Smirnov (KS) test. Three values were generated using the indicator, including normal and yield increment yield loss. These values' spatial and temporal prediction accuracy increased from 67.8% to 100% during the Central, Northern, and Southern Jiangsu tests. A more advanced indicator must be produced to cover a larger region [36].

Prediction models in agriculture can estimate the actual crop yield, but a better performance in the prediction is still needed. This shows that prediction in agriculture is not easy; it is a task with several complicated steps [9] to predict crop yield using machine learning. This work covered several large farms in Western Australia, and the crop yield of these farms was checked using diverse machine-learning datasets. Datasets such as soil test results, apparent electrical conductivity, Moderate Resolution Imaging Spectroradiometer (MODIS), and rainfall were used to predict the crop yield. Rather than access just one dataset like other predictive models, this study aimed to use all the datasets from 11000 to 17000-hectare canola, barley, and wheat crops from 2013, 2014, and 2015.

About 10m of the grid was processed for yield data, and for each observation point, the time and associated predictor variables in space were collated. Then, a 100m spatial resolution was processed for yield data using a modeling yield [37]. Three random forest models were created based on mid-season, pre-sowing, and late-season conditions to predict the crop yield of the three crops using the same dataset [38]. The results showed that the models accurately predicted the crop yield relatively. This was

partly because the datasets were available for every season, improving predictions [39]. Future work should elaborate on this yield monitor data by exploring integrating more data sets into the models, increasing the possibility of the yield forecast to guide management decisions, and focusing on predicting finer spatial resolution within fields [40].

2.3. Machine Learning Algorithms

The two fields share much regarding the relationship between data mining and machine learning. Machine learning is a field tasked with predicting from a known data property. Data mining, on the other hand, is a field tasked with discovering the properties of the data. Data mining is used interchangeably with knowledge discovery, which uses machine learning but with different needs and purposes [41]. In contrast, machine learning uses models such as unsupervised learning to find unknown phenomena in datasets. The types of machine learning models will be explained later in this section. The relationship between optimization and machine learning is centered on how machine learning models use optimization functions to minimize errors in their prediction [42].

After developing any machine learning with the least loss, the next challenge is how the model generalizes any other unseen data. The application of machine learning is becoming common, and people are accepting machine learning to make choices, from music on YouTube to partners on dating apps to healthcare centers and practitioners on healthcare databases. The ability of machine learning to function better on unseen data will keep machine learning models making these choices for the foreseeable future. Generalization has become an area of machine learning borrowed from mathematics to do this; just like data mining, the difference between machine learning and statistics is not an application but rather a goal. Machine learning aims to find ways to generalize data with the minimum loss, while statistics aims to find inferences in the dataset [10].

The development of machine learning was centered on making computer systems intelligent; because of this, machine learning has a large application area. The application of machine learning is almost in every sector, as far as the sector produces data to train ML models. Machine learning (ML) applications range from healthcare, agriculture, medicine diagnosis and discovery, military, banking, marketing, internet security, etc. Research has found that machine learning has proved its application in solving complex healthcare images using computer vision techniques. For example, machine learning algorithms can identify malignant cells in cancer imagery. Another useful technique is Natural Language Processing (NLP), which processes large medical records [11]. Machine learning is learned in four different ways: supervised learning, unsupervised learning, reinforcement learning, and semi-supervised learning. The nature or approach machine learning uses to learn depends on the nature of the data given and the kind of problem the machine learning is set to solve. In the following sections, we will explain the first two of the four approaches in machine learning, provide examples, explain their application, and explain how they work.

2.4. Supervised Learning

The supervised approach of machine learning learns by example; the data with which this type of machine learning is developed comes with input objects and the desired output [12]. The various algorithms generate a function that maps inputs to desired outputs. One standard formation of the supervised learning task is the classification problem: the learner is required to learn (to approximate the behavior of) a function that maps a vector into one of several classes by looking at the input and output examples of the function. The model then makes predictions on the unseen dataset. The model is expected to perform the task reasonably, with learning bias (inductive bias) in mind, because data is believed to be changing with time [13].

Supervised learning works in the way whereby given a set of N as the training data $(x_1, y_1), \ldots (x_N, y_N)$ whereby x_i is the feature vector of the i^{th} and y_i is its label, a learning algorithm seeks a function$g: X \to Y$, where X is the input space, and Y is the output space. The function g is an element of some space of possible function$G$, Usually called the hypothesis space. It is sometimes convenient to represent g using a scoring function $f: X \times Y \to \mathbb{R}$ such that g is defined as returning the y value that gives the highest score: $g(x) = arg_y \max f(x, y)$. Let F denote the space of scoring functions. Machine learning algorithms have been applied in so many fields. One algorithm solves many problems, but hyperparameter tuning makes an algorithm useful for different problems [43].

Hyperparameter tuning is used to tune a machine learning algorithm's hyperparameter towards generalizing the problem more correctly without falling for overfitting [14]. Therefore, the hyperparameter sets values of the supervised learning (or unsupervised learning algorithm) before training the model. Hyperparameter tuning is as important as clean data to machine learning or feature extraction. After training an algorithm, the way the algorithm reacts to new data is called generalization. The need for an algorithm to understand new data that it was not trained on or tested makes the algorithm useful. Learning models are probabilistic models, where g takes the form of a conditional probability model $g(x) = P(y|x)$, or f takes the form of a joint probability model where $f(x, y) = P(x, y)$. Machine learning models such as the Naïve Bayes algorithm use the joint probability model, and models like the Logistic regression use the conditional probability model [44]. The most important thing

to consider when developing supervised learning is the dataset, whereby the nature of the data differs from one dataset to another. El-Kady et al. [15] explain that there are two approaches to choosing f or g, namely empirical risk minimization and structural risk minimization. The empirical risk minimization looks for the function that best fits the dataset, while the structural risk minimization includes a penalty function that controls the bias/variance tradeoff.

There are many supervised learning algorithms, each with strengths and weaknesses, but no one can perform every supervised learning task [45]. There are factors to consider when choosing a supervised learning model. The factors include bias-variance tradeoff, a property of a supervised learning algorithm where the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters [46]. Function complexity and data size: The amount of data is very important when training a machine learning algorithm. Many machine learning algorithms can learn from a dataset with a small size because of the complexity of the learning function [47]. At the same time, some machine learning algorithms need a large data size because of the complexity of the interaction among many different input features and behave differently in different parts of the input space. The issue of dataset size is important, but the data quality is also very important [48]. As many machine learning experts say, "Your model is as good as your data."

However, the data size mostly when training a deep learning algorithm is also very important, as simple machine learning algorithms can learn on a simple and small dataset [49]. The dimensionality of the data: the dimension of the dataset is also important to consider when choosing a supervised learning algorithm. The dimensionality of a dataset here refers to the size of data or the features available in input x data [50]. Data with large dimensions make the learning problem very difficult, and even if the true function relied on a few of the dataset's features to learn, large dimensionality can cause the model to have high variance because the model will be confused about which features are important [51]. To solve this problem, some data mining methods are used to select the most important features, but selecting the best feature is not the only solution. Dimensionality reduction methods are also used to simplify the dataset [52].

These are the most important things to watch when choosing a supervised learning model; however, there are other important things, such as the noise in the data due to human error when creating the dataset in the first place. Other issues are heterogeneity of the output data and redundancy in the dataset [53]. The noise in the dataset is very important as the most recent controversial discussions in machine learning have been raised because of noise or sometimes prejudices in the dataset the model created. In 2016, The Guardian published a report of the first beauty contest judged by a machine learning model, which was found to be biased [54]. The model was found not to like people with darker skin color. Another important report by Forbes titled Why Artificial Intelligence is set up to fail LGBTQ people [55]. These and many other examples are concluded to be a result of human prejudice and or because of the nature of machine learning, whereby much of what happens in the hidden layers is considered a Black box [56]. The machine learning algorithms, especially the deep learning algorithms, are Black boxes in a way where the activities in the hidden layers of the algorithms are unknown to some extent, which raises an entire field of research on machine learning explain-ability [57]. The field of machine learning explain-ability is the process of understanding how machine learning algorithms work and how they make their choices, or it is the complete opposite of the Blackbox concept [58]. There are two classes of supervised learning, namely, the regression algorithm and the classification algorithm.

2.5. Unsupervised Learning

Four approaches are used to make unsupervised learning: clustering, anomaly detection, latent variable models, and neural networks (neural networks can be supervised, unsupervised, and reinforcement learning). Clustering is a model under unsupervised learning that groups data points into groups that look alike or have similar properties. Cluster analysis or clustering is the main activity in exploratory data analysis, usually used in data mining for information retrieval, pattern recognition, and image analysis [59]. However, machine learning is used to make models that predict which data points are similar and thus belong to the same group. The notion of the cluster in the clustering unsupervised is a very loose term. Different algorithms are defined under this class as every algorithm has its way of making predictions, but most importantly, the group data is in classes of data points [60]. Anomaly or outlier detection is a form of unsupervised learning used to detect noise in a dataset. The application of this class of unsupervised machine learning is very wide, as it is used in finding rare items, events, or observations that raise suspicions in a dataset [61].

A latent variant is a form of unsupervised learning that assumes the responses on the indicators, or manifest variables, are the results of an individual's position on the latent variable and that the manifest variables have nothing in common after controlling for the latent variable. As mentioned earlier, artificial neural networks (ANNs) can be both unsupervised and supervised learning; however, the model simulates the human brain when making decisions or understanding the world.

3. Methods

3.1. Data Collection and Preprocessing

This section outlines the methodology employed in developing an efficient machine learning model for the early detection of cardiovascular diseases (CVD) and creating a web-based application for real-time predictions and interactions with healthcare professionals. The methodology includes data collection, preprocessing, model selection, evaluation, and web-based interface development. The research adopted the Cross-Industry Standard Process for Data Mining (CRISP-DM) for model development and Object-Oriented Development Design (OODM) for the web-based interface. CRISP-DM, a widely recognized cyclic framework, guides the analytical aspects of this study. It encompasses six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The initiation involves defining project success criteria and a well-defined problem scope. This guides subsequent steps, starting with data collection. The dataset for this research was sourced from Kaggle, an online data repository. It contains 319,000 records in a comma-separated values (CSV) format. The data includes past physical and medical history of patients with and without cardiovascular disease. This extensive dataset provides a comprehensive basis for developing predictive models.

Data preprocessing is crucial for refining raw data into a coherent dataset. This process includes data cleaning, normalization, and feature extraction. Data cleaning involves handling missing values, removing duplicates, and correcting errors. Ensuring data quality at this stage is essential for the model's reliability. Normalization standardizes data ranges to improve the model's performance, ensuring all features contribute equally to the analysis and avoiding biases due to varying scales. Feature extraction involves selecting relevant information from the dataset, where features are categorized into attributes and targets, aligning with the research objectives. Dimensionality reduction techniques enhance robustness, reducing the dataset's complexity without losing critical information. The data preparation phase consumes a substantial portion of the project timeline, significantly impacting the project's success. Properly processed data lays the foundation for effective model training and prediction.

3.2. Model Development

The model development phase involves selecting appropriate machine learning algorithms, designing a testing strategy, building executable models, and evaluating them to choose the optimal one. Four machine learning algorithms were selected for this study: Support Vector Machine (SVM), Naive Bayes classifier, Random Forest, and Logistic Regression. These algorithms were trained using the prepared training dataset. SVM is effective in high-dimensional spaces, making it suitable for CVD prediction. The Naive Bayes classifier is simple, efficient, and useful for predicting CVD likelihood based on demographic and clinical data. Random Forest, an ensemble method, improves prediction accuracy by combining multiple decision trees. Logistic regression is useful for binary classification problems like predicting the presence or absence of CVD. The models were iteratively evaluated using performance metrics such as precision, recall, and F1-score to ensure they were fit for use. The highest accuracy model was selected for system prediction.

3.3. Web-Based Application Development

The development of the web-based application followed the Object-Oriented Design Methodology (OODM), which includes phases such as requirements gathering, system design, implementation, and testing. This approach ensures the creation an intuitive, navigable, and interactive interface. Requirements gathering involves understanding user needs and defining the application's functionality. System design establishes the application's architecture, ensuring scalability and ease of use. Implementation involves coding and developing the application based on the design specifications. Testing ensures the application works as intended and is free of bugs. The web-based application facilitates real-time predictions and interactions with healthcare professionals and individuals concerned about their cardiovascular health. The system classifies patients' health status and alerts medical experts based on the predicted vital signs values.

4. Results

The primary objective of our research was to construct a predictive model capable of identifying individuals at an increased risk of developing cardiovascular diseases (CVD) at an early stage using machine learning. We utilized a dataset from Kaggle, consisting of over 319,000 records with 18 features, including physical metrics, lifestyle choices, medical history, and demographic details. This dataset, devoid of missing values, provided a solid foundation for developing our machine-learning models.

4.1. Dataset Description

The dataset's numerical features revealed intriguing trends. The average BMI of 28.3 suggested a prevalence of overweight and obesity. Physical and mental health scores averaged 3.37 and 3.89 out of 30, respectively, with potential skew towards lower scores. The average sleep duration was 7.09 hours, with most individuals falling between 6 and 8 hours, although extreme values were present.

4.2. Model Development and Testing

In the model development phase, the dataset was partitioned into training (80%) and testing (20%) subsets to evaluate the models' performance on unseen data. We employed four machine learning algorithms: Support Vector Machine (SVM), Logistic Regression, Naive Bayes, and Random Forest Classifier. Table 1 summarizes the accuracy scores of these algorithms.

Table 1: Summary of Algorithm's Accuracy

Algorithm	Accuracy Score (%)
Support Vector Machine	91.72
Logistic Regression	91.64
Naive Bayes	86.56
Random Forest Classifier	91.43

The Support Vector Machine (SVM) achieved the highest accuracy at 91.72%, followed closely by Logistic Regression at 91.64% and Random Forest at 91.43%. Naive Bayes had a lower accuracy of 86.56% (Figure 1).

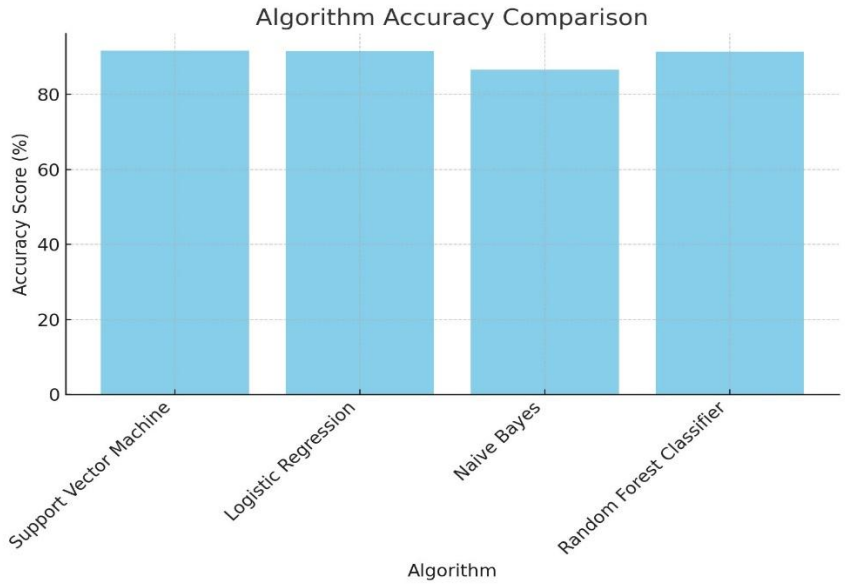

Figure 1: Algorithm Accuracy Comparison

Further analysis using precision, recall, and F1-score provided deeper insights into the models' performance across different classes. Table 2 summarizes these metrics for each algorithm.

Table 2: Summary of Algorithm's Classification Report

Algorithm	Target	Precision	Recall	F1-Score
Support Vector Machine	0	0.92	1.00	0.96
	1	1.00	0.00	0.00
Logistic Regression	0	0.92	0.99	0.96
	1	0.48	0.11	0.18
Naive Bayes	0	0.95	0.90	0.93
	1	0.29	0.44	0.35
Random Forest Classifier	0	0.92	0.99	0.95
	1	0.43	0.11	0.17

SVM displayed strong performance for class 0 but struggled with class 1, resulting in an F1-score of 0.00 for class 1. Logistic Regression and Random Forest showed similar trends, performing well for class 0 but poorly for class 1 (Figure 2).

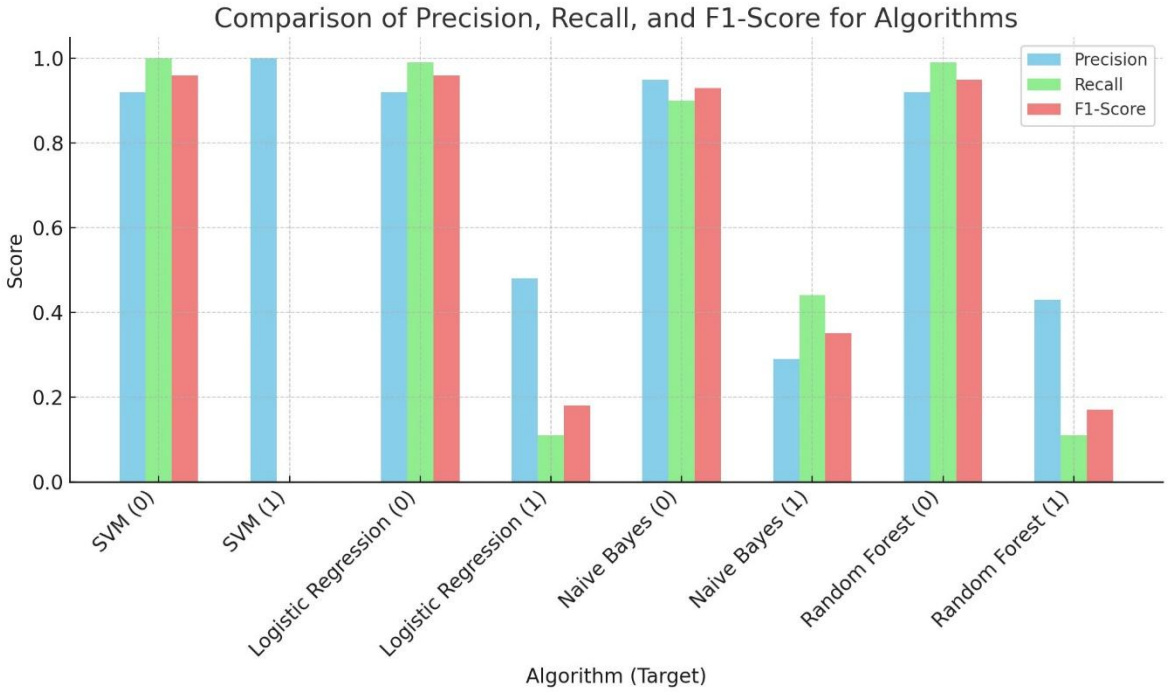

Figure 2: Comparison of Precision, Recall, and F1-Score for Algorithms

Naive Bayes exhibited a more balanced performance across both classes but with lower overall precision and recall for class 1.

5. Discussion

The study aimed to develop a predictive model for early detection of cardiovascular diseases (CVD) using machine learning techniques. Machine learning has become an invaluable tool in healthcare because it can process large datasets, detect complex patterns, and make predictions that can aid in early diagnosis. Our models, including Support Vector Machine (SVM), Logistic Regression, and Random Forest, demonstrated promising results in classifying individuals at risk of CVD. However, while the models exhibited high accuracy, challenges emerged regarding their ability to handle the class imbalance within the dataset, specifically in predicting the minority class (class 1), representing individuals with CVD. The SVM model achieved the highest accuracy, indicating its strong performance distinguishing between at-risk individuals and those not at risk. However, the model exhibited a significant shortcoming in its ability to recall instances of class 1. Although it accurately identified individuals from the majority class (class 0, those without CVD), it often failed to classify individuals in the minority class correctly. This issue points to a bias in the SVM model towards the majority class, which, in medical applications, is a critical limitation. Failing to identify individuals at risk of CVD can have severe consequences, potentially leading to delayed or missed diagnoses.

Logistic regression, although slightly less accurate than SVM, faced similar challenges. Logistic regression is often preferred in healthcare settings for its interpretability; however, like the SVM model, it struggled with the minority class. The imbalance in the dataset caused Logistic Regression to overfit the majority class, reducing its ability to identify the minority class. This imbalance meant that individuals with CVD were frequently misclassified as not at risk, diminishing the model's overall utility for early detection. Random Forest, an ensemble learning method, also achieved high accuracy in predicting CVD risk, comparable to Logistic Regression. While Random Forest is typically robust against overfitting and performs well in classification tasks, it still encountered difficulties with class imbalance. Like the other models, Random Forest exhibited a bias towards class 0 and struggled to recall instances of class 1 accurately. This limitation highlights a recurring challenge in medical datasets, where class imbalance can significantly impact the effectiveness of machine learning models, particularly when detecting critical, yet less frequent, conditions like CVD.

The problem of class imbalance is a common issue in healthcare datasets, where most individuals are often healthy, while a smaller proportion suffer from specific conditions. This imbalance can lead to models that perform well in the majority class but fail to identify high-risk individuals in the minority class. In our study, the minority class represented individuals with CVD,

and the models' failure to accurately classify this group underscores the need for specialized techniques to address class imbalance. Simply relying on accuracy as a metric can be misleading in such cases, as high accuracy may be achieved by correctly classifying the majority class while ignoring the minority class. Several techniques can be employed to improve the models' ability to detect individuals in the minority class. One of the most effective methods for handling class imbalance is oversampling, which involves generating synthetic data for the minority class. The Synthetic Minority Over-sampling Technique (SMOTE) is commonly used for this purpose, creating synthetic instances by interpolating between existing minority class examples. This method helps balance the dataset, allowing the model to learn better decision boundaries for the minority class and improve recall for class 1.

Undersampling, another technique, reduces the number of instances in the majority class to match the size of the minority class. While this can help address class imbalance, it also risks discarding valuable information from the majority class, which may negatively impact the model's overall performance. Advanced machine learning algorithms, such as ensemble methods and cost-sensitive learning, offer alternative solutions to address class imbalance. Ensemble methods, like XGBoost or AdaBoost, combine multiple models to improve classification accuracy, while cost-sensitive learning assigns higher penalties for misclassifying the minority class. These techniques encourage the model to focus on correctly identifying instances from the minority class. Despite the challenges posed by class imbalance, the richness of the dataset used in this study provided a solid foundation for the machine learning models. The dataset included a wide range of features, such as physical health metrics (e.g., blood pressure, cholesterol levels), lifestyle factors (e.g., smoking habits, exercise frequency), and demographic details (e.g., age, gender). These features enabled the models to consider various factors when predicting CVD risk, increasing the likelihood of identifying at-risk individuals.

Nevertheless, while the dataset was rich in features, further refinement of the models is necessary to improve their performance, particularly for the minority class. One promising approach is feature engineering, which involves creating new features or modifying existing ones to capture the underlying patterns in the data better. For example, interactions between certain variables, such as age and blood pressure, could provide more insightful information about an individual's CVD risk than either variable considered independently. The models' predictive power could be enhanced by carefully crafting new features or selecting the most relevant ones. In addition to feature engineering, incorporating additional relevant features from external sources, such as genetic information or comprehensive medical histories, could further improve the models' accuracy. Including such data could give the models a more complete view of each individual's health status, helping them better identify those at risk for CVD. Additionally, synthetic data generation techniques, such as Generative Adversarial Networks (GANs), could be explored in future research to generate synthetic samples for the minority class. These techniques have shown promise in other fields for creating realistic data and could prove valuable for addressing class imbalance in medical datasets.

Another critical consideration for adopting machine learning models in healthcare is the interpretability and explainability of these models. Healthcare professionals and end-users must trust the models' predictions and understand the factors influencing those predictions. Suppose a model predicts that an individual is at high risk of CVD. In that case, healthcare practitioners need to know which factors such as high blood pressure, smoking, or family history—contributed most to that prediction. This transparency allows practitioners to make informed decisions about treatment and intervention. Explainability techniques, such as SHapley Additive exPlanations (SHAP) values, provide a powerful method for interpreting machine learning models. SHAP values assign importance scores to each feature, explaining how each contributed to the final prediction. By using SHAP values, healthcare practitioners can gain clear insights into why a model made a particular prediction, increasing trust in the model's outputs and ensuring that predictions are based on medically relevant factors.

Explaining model predictions is crucial for building trust with healthcare providers and ensuring that models do not rely on irrelevant or biased features, which could lead to incorrect or unfair predictions. For example, if a model consistently assigns high risk to individuals based on irrelevant demographic features rather than their health status, this could lead to discriminatory outcomes. Ensuring transparency in machine learning models can help mitigate these risks and promote the ethical and effective use of machine learning in healthcare. In light of these findings, future work in this area should focus on several key areas. First, addressing class imbalance remains a top priority. Techniques such as oversampling, undersampling, and advanced algorithms should be further explored to improve the models' ability to detect high-risk individuals in the minority class. Additionally, incorporating new features from external data sources, refining feature engineering approaches, and exploring synthetic data generation methods will be essential for improving the predictive power of machine learning models for CVD risk.

Moreover, enhancing model transparency and interpretability should be a key focus of future research. Techniques such as SHAP values can provide healthcare practitioners with clear insights into the models' decision-making processes, fostering trust and ensuring that machine learning models are used responsibly and effectively in medical practice. While machine learning models such as SVM, Logistic Regression, and Random Forest have shown promise in predicting CVD risk, class imbalance, and model transparency must be addressed to improve their utility in healthcare settings. By employing advanced

techniques to address class imbalance, refining feature selection, and enhancing model interpretability, machine learning can play a pivotal role in the early detection and prevention of cardiovascular diseases, ultimately saving lives.

6. Conclusion

This study successfully developed a machine learning-based model for the early detection of cardiovascular diseases, achieving high accuracy with Support Vector Machine, Logistic Regression, and Random Forest classifiers. The research utilized a robust dataset from Kaggle, encompassing various physical, lifestyle, and demographic features, to train and test the models. Despite the high accuracy, the models faced challenges in accurately predicting the minority class, underscoring the need for future improvements in handling class imbalance. The findings highlight the potential of machine learning in transforming cardiovascular disease management by enabling early detection and timely intervention. The proposed system can significantly enhance patient outcomes and contribute to the broader shift towards preventive healthcare by integrating predictive analytics into healthcare workflows. Future work will address the identified limitations, enhance model interpretability, incorporate real-time data, expand the dataset, and develop a user-friendly interface for practical application. These advancements will improve the model's predictive capabilities and ensure its seamless integration into real-world healthcare settings, ultimately contributing to the early detection and prevention of cardiovascular diseases on a larger scale.

Acknowledgment: We thank our mentors for their guidance, our development team for their hard work, and our beta testers for their valuable feedback. We are also grateful to our families and friends for their support.

Data Availability Statement: Data supporting this study are available upon request and subject to privacy and ethical guidelines.

Funding Statement: The support and funding provided by the Laboratory of the Department of Computer Science, University of Nigeria, Nsukka.

Conflicts of Interest Statement: The authors declare no conflicts of interest.

Ethics and Consent Statement: This study was ethically approved, and informed consent was obtained from all participants.

References

1. T. J. Yeo, Y.-T. L. Wang, and T. T. Low, "Have a heart during the COVID-19 crisis: Making the case for cardiac rehabilitation in the face of an ongoing pandemic," Eur. J. Prev. Cardiol., vol. 27, no. 9, pp. 903–905, 2020.
2. M. S. Ibrahim and S. Saber, "Machine Learning and Predictive Analytics: Advancing Disease Prevention in Healthcare," Machine Learning and Predictive Analytics, vol. 7, no.1, pp. 53–71, 2023.
3. D.W. Hosmer, S. Lemeshow, R.X. Sturdivant. Applied logistic regression. John Wiley & Sons.vol.398., no.8, pp. 1-12, 2013.
4. C. Cortes, & V. Vapnik, Support-vector networks. Machine learning, vol. 20, no.3, pp. 273-297.1995.
5. F. J. W. M. Dankers, A. Traverso, L. Wee, and S. M. J. van Kuijk, "Prediction modeling methodology," in Fundamentals of Clinical Data Science, Cham: Springer International Publishing, pp. 101–120. 2019.
6. X. Liu, L. Faes, M.J Calvert, A.K Denniston, & T.J Ford. Machine learning in cardiology. Heart, vol.107, no.21, pp. 1750-1758, 2021.
7. K. C. Siontis, P. A. Noseworthy, Z. I. Attia, and P. A. Friedman, "Artificial intelligence-enhanced electrocardiography in cardiovascular disease management," Nat. Rev. Cardiol., vol. 18, no. 7, pp. 465–478, 2021.
8. S. Wang, H. Xu, Q. Zou, and M. Deng, "Identification of potential cardiovascular disease-associated long non-coding RNAs and their mechanisms of action," Frontiers in Genetics, vol. 12, no.1, pp. 1-15, 2021.
9. M. Abdar, M. A. Rezaei, and A. Mehdizadeh, "Hybrid machine learning models for the prediction of heart disease," Biocybernetics and Biomedical Engineering, vol. 40, no. 2, pp. 535–548, 2020.
10. Z. I. Attia et al., "Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram," Nat. Med., vol. 25, no. 1, pp. 70–74, 2019.
11. J.A Damen, L. Hooft, E. Schuit, T.P Debray, G.S Collins, I. Tzoulaki, K.G Moons Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ, Vol. 353, no. 5, p.1416, 2020.
12. Z. Vinicius, R. David, S. Adam, B. Victor, L. Yujia, B. Igor, Deep reinforcement learning with relational inductive biases. iclr (pp. 1-18). London, UK: DeepMind, 2019
13. L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," Neurocomputing, vol. 415, no.11, pp. 295–316, 2020.

14. S. R. Sain and V. N. Vapnik, “The nature of statistical learning theory,” Technometrics, vol. 38, no. 4, p. 409, 1996.
15. A. M. El-Kady, M. M. Abbassy, H. H. Ali, and M. F. Ali, “Advancing Diabetic Foot Ulcer Detection Based on Resnet And Gan Integration,” Journal of Theoretical and Applied Information Technology, vol. 102, no. 6, pp. 2258–2268, 2024.
16. A. Salah and N. Hussein, “Recognize Facial Emotion Using Landmark Technique in Deep Learning,” in 2023 International Conference on Engineering, Science and Advanced Technology (ICESAT 2023), Iraq, pp. 198–203, 2023.
17. A. Veena and S. Gowrishankar, “An automated pre-term prediction system using EHG signal with the aid of deep learning technique,” Multimed. Tools Appl., vol. 83, no. 2, pp. 4093–4113, 2024.
18. A. Veena and S. Gowrishankar, “Context based healthcare informatics system to detect gallstones using deep learning methods,” International Journal of Advanced Technology and Engineering Exploration, vol. 9, no. 96, pp. 1661–1677, 2022.
19. C. Elayaraja, J. Rahila, P. Velavan, S. S. Rajest, T. Shynu, and M. M. Rahman, “Depth sensing in AI on exploring the nuances of decision maps for explainability,” in Advances in Computational Intelligence and Robotics, IGI Global, USA, pp. 217–238, 2024.
20. C. S. Kumar, B. S. Kumar, G. Gnanaguru, V. Jayalakshmi, S. S. Rajest, and B. Senapati, “Augmenting chronic kidney disease diagnosis with support vector machines for improved classifier accuracy,” in Advances in Medical Technologies and Clinical Practice, IGI Global, USA, pp. 336–352, 2024.
21. D. R. Bhuva and S. Kumar, “A novel continuous authentication method using biometrics for IOT devices,” Internet of Things, vol. 24, no. 11, p. 100927, 2023.
22. G. Gnanaguru, S. S. Priscila, M. Sakthivanitha, S. Radhakrishnan, S. S. Rajest, and S. Singh, “Thorough analysis of deep learning methods for diagnosis of COVID-19 CT images,” in Advances in Medical Technologies and Clinical Practice, IGI Global, USA, pp. 46–65, 2024.
23. H. O. Rasul, B. K. Aziz, D. D. Ghafour, and A. Kivrak, “Discovery of potential mTOR inhibitors from Cichorium intybus to find new candidate drugs targeting the pathological protein related to breast cancer: an integrated computational approach,” Mol. Divers., vol. 27, no. 3, pp. 1141–1162, 2023.
24. H. O. Rasul, B. K. Aziz, D. D. Ghafour, and A. Kivrak, “In silico molecular docking and dynamic simulation of eugenol compounds against breast cancer,” J. Mol. Model., vol. 28, no. 1, p. 17, 2022.
25. H. O. Rasul, B. K. Aziz, D. D. Ghafour, and A. Kivrak, “Screening the possible anti-cancer constituents of Hibiscus rosa-sinensis flower to address mammalian target of rapamycin: An in silico molecular docking, HYDE scoring, dynamic studies, and pharmacokinetic prediction,” Mol. Divers., vol. 27, no. 5, pp. 2273–2296, 2023.
26. I. Khalifa, H. Abd Al-Galil, and M. M. Abbassy, “Mobile hospitalization,” International Journal of Computer Applications, vol. 80, no. 13, pp. 18–23, 2013.
27. I. Khalifa, H. Abd Al-Galil, and M. M. Abbassy, “Mobile hospitalization for Kidney Transplantation,” International Journal of Computer Applications, vol. 92, no. 6, pp. 25–29, 2014.
28. K. Singh, L. M. M. Visuwasam, G. Rajasekaran, R. Regin, S. S. Rajest, and T. Shynu, “Innovations in skeleton-based movement recognition bridging AI and human kinetics,” in Explainable AI Applications for Human Behavior Analysis, IGI Global, USA, pp. 125–141, 2024.
29. M. A. Abdelazim, M. M. Nasr, and W. M. Ead, “A survey on classification analysis for cancer genomics: Limitations and novel opportunity in the era of cancer classification and Target Therapies,” Annals of Tropical Medicine and Public Health, vol. 23, no. 24, p. 7, 2020.
30. M. Awais, A. Bhuva, D. Bhuva, S. Fatima, and T. Sadiq, “Optimized DEC: An effective cough detection framework using optimal weighted Features-aided deep Ensemble classifier for COVID-19,” Biomed. Signal Process. Control, p. 105026, 2023.
31. M. G. Hariharan, S. Saranya, P. Velavan, E. S. Soji, S. S. Rajest, and L. Thammareddi, “Utilization of artificial intelligence algorithms for advanced cancer detection in the healthcare domain,” in Advances in Medical Technologies and Clinical Practice, IGI Global, USA, pp. 287–302, 2024.
32. M. Gandhi, C. Satheesh, E. S. Soji, M. Saranya, S. S. Rajest, and S. K. Kothuru, “Image recognition and extraction on computerized vision for sign language decoding,” in Explainable AI Applications for Human Behavior Analysis, IGI Global, USA, pp. 157–173, 2024.
33. M. Kuppam, M. Godbole, T. R. Bammidi, S. S. Rajest, and R. Regin, “Exploring innovative metrics to benchmark and ensure robustness in AI systems,” in Explainable AI Applications for Human Behavior Analysis, IGI Global, USA, pp. 1–17, 2024.
34. M. M. Abbassy and A. A. Mohamed, “Mobile Expert System to Detect Liver Disease Kind,” International Journal of Computer Applications, vol. 14, no. 5, pp. 320–324, 2016.
35. M. M. Abbassy, “Opinion mining for Arabic customer feedback using machine learning,” Journal of Advanced Research in Dynamical and Control Systems, vol. 12, no. SP3, pp. 209–217, 2020.

36. M. M. Abbassy, “The human brain signal detection of Health Information System IN EDSAC: A novel cipher text attribute based encryption with EDSAC distributed storage access control,” Journal of Advanced Research in Dynamical and Control Systems, vol. 12, no. SP7, pp. 858–868, 2020.
37. N. J. Hussein and F. Hu, “An alternative method to discover concealed weapon detection using critical fusion image of color image and infrared image,” in 2016 1st IEEE International Conference on Computer Communication and the Internet (ICCCI 2016), Wuhan, China, pp. 378–383, 2016.
38. N. J. Hussein and S. K. Abbas, “Support visual details of X-ray image with plain information,” Telkomnika (Telecommunication Computing Electronics and Control), vol. 19, no. 6, pp. 1975–1981, 2021.
39. N. J. Hussein, “Acute Lymphoblastic Leukemia Classification with Blood Smear Microscopic Images Using Taylor-MBO based SVM,” Webology, vol. 18, pp. 357–366, 2021.
40. N. J. Hussein, F. Hu, H. Hu, and A. T. Rahem, “IR and Multi Scale Retinex image enhancement for concealed weapon detection,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 1, no. 2, pp. 399–405, 2016.
41. N. J. Hussein, H. A. Abdulameer, and R. H. Al-Taie, “Deep Learning and Histogram Gradient Algorithm to Detect Visual Information Based on Artificial Intelligent,” in ACM International Conference Proceeding Series, pp. 577–581, 2024.
42. N. J. Hussein, S. R. Saeed, and A. S. Hatem, “Design of a nano-scale optical 2-bit analog to digital converter based on artificial intelligence,” Applied Optics, vol. 63, no. 19, pp. 5045–5052, 2024.
43. N. J. Jasim Hussein, “Robust Iris Recognition Framework Using Computer Vision Algorithms,” in 2020 4th International Conference on Smart Grid and Smart Cities (ICSGSC 2020), Osaka, Japan, pp. 101–108, 2020.
44. N. J. Jasim Hussein, F. Hu, and F. He, “Multisensor of thermal and visual images to detect concealed weapon using harmony search image fusion approach,” Pattern Recognition Letters, vol. 94, no.12, pp. 219–227, 2017.
45. P. Paramasivan, S. S. Rajest, K. Chinnusamy, R. Regin, and F. J. John Joseph, Eds., Advancements in Clinical Medicine, IGI Global, USA, 2024.
46. P. Paramasivan, S. S. Rajest, K. Chinnusamy, R. Regin, and F. J. John Joseph, Eds., Cross-Industry AI Applications, IGI Global, USA, 2024.
47. R. A. Sadek, D. M. Abd-alazeem, and M. M. Abbassy, “A new energy-efficient multi-hop routing protocol for heterogeneous wireless sensor networks,” International Journal of Advanced Computer Science and Applications, vol. 12, no. 11, p.14, 2021.
48. R. Boina, “Assessing the Increasing Rate of Parkinson’s Disease in the US and its Prevention Techniques”,” International Journal of Biotechnology Research and Development, vol. 3, no. 1, pp. 1–18, 2022.
49. R. C. A. Komperla, K. S. Pokkuluri, V. K. Nomula, G. U. Gowri, S. S. Rajest, and J. Rahila, “Revolutionizing biometrics with AI-enhanced X-ray and MRI analysis,” in Advances in Medical Technologies and Clinical Practice, IGI Global, USA, pp. 1–16, 2024.
50. R. Regin, A. A. Khanna, V. Krishnan, M. Gupta, S. R. Bose, and S. S. Rajest, “Information design and unifying approach for secured data sharing using attribute-based access control mechanisms,” in Recent Developments in Machine and Human Intelligence, IGI Global, USA, pp. 256–276, 2023.
51. R. Regin, S. S. Rajest, M. Bhattacharya, D. Datta, and S. S. Priscila, “Development of predictive model of diabetic using supervised machine learning classification algorithm of ensemble voting,” Int. J. Bioinform. Res. Appl., vol. 19, no. 3, pp. 151–169, 2023.
52. S. Padmaja, S. Mishra, A. Mishra, J. J. V. Tembra, P. Paramasivan, and S. S. Rajest, “Insights into AI systems for recognizing human emotions, actions, and gestures,” in Advances in Computational Intelligence and Robotics, IGI Global, USA, pp. 389–410, 2024.
53. S. S. Rajest, B. Singh, A. J. Obaid, R. Regin, and K. Chinnusamy, “Recent developments in machine and human intelligence,” Advances in Computational Intelligence and Robotics, IGI Global, USA, 2023.
54. S. S. Rajest, B. Singh, A. J. Obaid, R. Regin, and K. Chinnusamy, “Advances in artificial and human intelligence in the modern era,” Advances in Computational Intelligence and Robotics, IGI Global, USA, 2023.
55. S. Sengupta, D. Datta, S. S. Rajest, P. Paramasivan, T. Shynu, and R. Regin, “Development of rough-TOPSIS algorithm as hybrid MCDM and its implementation to predict diabetes,” Int. J. Bioinform. Res. Appl., vol. 19, no. 4, pp. 252–279, 2023.
56. S. Singhal, A. Kotagiri, L. S. Samayamantri, and S. S. Rajest, “Interpretable machine learning models for human action and emotion deciphering,” in Advances in Computer and Electrical Engineering, IGI Global, USA, pp. 449–468, 2024.
57. U. Srilakshmi, I. Sandhya, J. Manikandan, A. H. Bindu, R. Regin, and S. S. Rajest, “Design and implementation of energy-efficient protocols for underwater wireless sensor networks,” in Advances in Computer and Electrical Engineering, IGI Global, USA, pp. 417–430, 2024.
58. V. Chunduri, S. A. Hannan, G. M. Devi, V. K. Nomula, V. Tripathi, and S. S. Rajest, “Deep convolutional neural networks for lung segmentation for diffuse interstitial lung disease on HRCT and volumetric CT,” in Advances in Computational Intelligence and Robotics, IGI Global, USA, pp. 335–350, 2024.

59. V. Chunduri, S. S. Rajest, V. R. Allugunti, D. Verma, R. Aarthi, and D. K. Sharma, “Deep neural network for brain tumour segmentation using guaranteed time slots (GTS) algorithm,” in Advances in Computer and Electrical Engineering, IGI Global, USA, pp. 371–390, 2024.
60. W. M. Ead and M. M. Abbassy, “IoT based on plant diseases detection and classification,” in 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Tamil Nadu, India, 2021.
61. Y. J. K. Nukhailawi and N. J. Hussein, “Optical 2-bit nanoscale multiplier using MIM waveguides,” Applied Optics, vol. 63, no. 3, pp. 714–720, 2024.