IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 27, Issue 3, Ser. 5 (May. – June. 2025), PP 20-31
www.iosrjournals.org

Predicting Heart Disease with Body Composition Using a Hybrid Machine Learning Approach

Awal Mohammed; Daniel Kwadwo Nterful; Issah Bawah Muhammed
National Health Insurance Authority (NHIA)- Bekwai-Ghana
Directorate of ICT Services- Takoradi Technical University-Takoradi-Ghana
School of Peri-Operative and Critical Care Nursing –Korle Bu-Accra-Ghana

Abstract: *Heart diseases represent a significant global health concern, characterized by impaired heart function. Unfortunately, it's predicted that fatalities resulting from heart-related illnesses may escalate dramatically, reaching an astounding 24.2 million by the year 2030.Accurate prognosis and identification of cardiac diseases play a pivotal role in facilitating timely prevention, detection, and therapy. However, existing medical equipment such as electrocardiograms and computerized tomography scans employed for detecting heart disorders often pose difficulties owing to prohibitive costs and operational constraints, making access challenging for many individuals. By harnessing these capabilities, machine learning models can potentially provide more accurate predictions of heart disease risk based on body composition data. The use of machine learning algorithms in healthcare has already shown encouraging results in various applications, including disease diagnosis, treatment planning, and patient outcome prediction. In light of these developments, the study aims to create a hybrid machine learning model that leverages the strengths of multiple algorithms to predict heart disease risk based on body composition data to reduce death rate. This paper proposed six Machine Learning algorithms using body composition dataset, the algorithms are Decision tree model (DTM), XGBOOST, LIGHTGBM, Support Vector Machine (SVM), KNN and Hybrid model. The experimental result indicates that the HHP model outperformed others in precision, recall, F-score, with accuracy of 90.2 %*

Keywords: Machine Learning, body composition, classification, Heart Disease

Date of Submission: 20-06-2025 Date of Acceptance: 03-07-2025

I. INTRODUCTION

Heart disease remains one of the most significant health burdens globally, claiming over 17.9 million lives annually, according to the World Health Organization (WHO) [1]. The alarming rate of mortality underscores the pressing need for effective strategies to detect and prevent heart disease, particularly in its early stages. One promising approach to achieving this goal involves analyzing body composition, which has been linked to cardiovascular risk [2]. Traditionally, bioelectrical impedance analysis (BIA) has been a commonly used method for assessing body composition. However, BIA has several limitations, including low accuracy and practical challenges [3]. As a result, researchers have shifted their focus towards machine learning algorithms, which offer enhanced precision and the capacity to process complex data sets. By harnessing these capabilities, machine learning models can potentially provide more accurate predictions of heart disease risk based on body composition data. The use of machine learning algorithms in healthcare has already shown encouraging results in various applications, including disease diagnosis, treatment planning, and patient outcome prediction [4]. Within the context of body composition analysis, machine learning models can integrate diverse data points, such as anthropometric measurements, biochemical markers, and imaging results, to generate comprehensive profiles of individual risk factors. This integrated approach enables clinicians and health practitioners to make better-informed decisions regarding patient care and interventions. Moreover, advances in wearable technology and mobile devices have made it easier to collect and track body composition data remotely, providing an opportunity for continuous monitoring and personalized feedback [5]. When coupled with machine learning algorithms, these technologies empower individuals to take proactive steps toward reducing their risk of developing heart disease. In light of these developments, the study aims to create a hybrid machine learning model that leverages the strengths of multiple algorithms to predict heart disease risk based on body composition data. The study hypothesizes that the proposed model will demonstrate superior performance compared to traditional machine learning approaches. Ultimately, this paper puts forward the following contributions.

1. Develop a robust hybrid machine learning model that integrates diverse algorithms to effectively analyse body composition data and enhance the accuracy of predicting heart disease.

2. Investigate the most relevant and influential body composition variables for heart disease prediction, ensuring that the hybrid model considers and weighs these factors appropriately to improve diagnostic capabilities.
3. Evaluate the performance of the hybrid machine learning approach against existing traditional models and algorithms, assessing its effectiveness in providing more accurate and reliable predictions for heart disease based on body composition data. The subsequent sections of this paper are outlined as follows: The following section provides a concise description of the study area and the data source. The methodology employed is detailed in Section III, while Section IV encompasses the results and discussion. Lastly, the study's conclusion is presented in Section V.

II. Related Works

In recent years, the healthcare sector has witnessed considerable progress in data mining and machine learning technologies. These methods have gained widespread acceptance and shown promising results across diverse healthcare domains, especially in medical cardiology. The exponential growth of medical data has offered researchers a unique chance to create and evaluate novel algorithms in this field. Given that heart disease continues to be a primary cause of death in developing countries [6], identifying risk factors and early indicators has become a crucial area of investigation. By leveraging data mining and machine learning techniques, it may be possible to improve the early detection and prevention of heart disease. There has been significant advancement in the fields of data mining and machine learning within the healthcare sector in recent years. These methodologies have been widely adopted and have shown promising results in various healthcare applications, particularly in the realm of medical cardiology [3]. The rapid accumulation of medical data has presented researchers with a unique opportunity to develop and evaluate innovative algorithms in this area. Of note, heart disease remains a leading cause of mortality in developing countries, according to [6], making the identification of risk factors and early indicators a crucial area of investigation.

In a research work by [7] aimed to develop a cutting-edge machine learning based cardiovascular disease (CVD) prediction system to enhance the accuracy of the widely utilized Framingham risk score (FRS). By leveraging data from 689 individuals exhibiting CVD symptoms and a validation dataset from the Framingham study, the proposed system, which harnesses a quantum neural network to identify patterns of CVD, was experimentally validated and contrasted with the FRS. The novel approach demonstrated superior accuracy in predicting CVD risk, with a remarkable

98.57 % accuracy rate, vastly surpassing the FRS's 19.22 % and other existing methods. According to the study's results, the suggested approach has the potential to be a useful resource for doctors in predicting the risk of cardiovascular disease, creating more efficient treatment programs, and arriving at prompt diagnoses.

[8] conducted a research project aimed at developing a predictive model for cardiovascular disease using machine learning techniques. They used the Cleveland heart disease dataset, which consisted of 303 cases and 17 attributes, obtained from the UCI machine learning repository. The team applied several supervised classification methods, including naive Bayes, decision tree, random forest, and k-nearest neighbor (KNN). The study found that the KNN model achieved the highest accuracy rate, at 90.8 %. The research demonstrates the potential of machine learning techniques in predicting cardiovascular disease, while emphasizing the importance of selecting appropriate models and techniques to optimize results. In a recent study published in [9], the aim was to leverage machine learning (ML) techniques to identify the key risk factors for cardiovascular disease (CVD) in individuals with metabolic associated fatty liver disease (MAFLD). The study involved 191 MAFLD patients who underwent blood biochemical analysis and subclinical atherosclerosis assessment. Utilizing ML approaches, such as multiple logistic regression classifier, univariate feature ranking, and principal component analysis (PCA), the researchers built a model to identify those with the highest risk of CVD. The study found that hypercholesterolemia, plaque scores, and duration of diabetes were the most important clinical characteristics. The ML technique demonstrated excellent performance, accurately identifying 40/47 (85.11 %) highrisk patients and 114/144 (79.17 %) low-risk patients, with an AUC of 0.87. The study's findings suggest that an ML method is useful for detecting MAFLD patients with extensive CVD based on straightforward patient criteria.

In [10] The paper aimed to develop an optimization function based on Support Vector Machines (SVM) to predict cardiovascular disease. The Genetic Algorithm (GA) was employed to select the most relevant features for the SVM model.The GA efficiently performed feature selection, leading to accurate cardiovascular predictions using the SVM classifier, with an accuracy of 88.34 % in diagnosing cardiac illness using the selected attributes. The dataset used for the study was obtained from the Cleveland Heart Disease Database. [11] conducted a study to determine the most effective feature selection approach for predicting cardiovascular disease. They evaluated three popular feature selection methods (filter, wrapper, and embedding) and used a Boolean process based common" True condition to retrieve feature subsets in two stages. The study compared the accuracy of several models, including random forest, support vector classifier, k-nearest neighbors, naive Bayes, and XGBoost, to establish the best predictive analytics. The artificial neural network (ANN) was used as a baseline for comparison with all features. The results showed that the XGBoost classifier combined with the wrapper technique provided the most accurate prediction results for cardiovascular illness, with an accuracy of 73.74 %, followed by SVC with 73.18 % and ANN with 73.20 %. In a study conducted by [12], the author explored the effectiveness of machine learning (ML) techniques in predicting heart failure disease. The study utilized a dataset from the Cleveland Clinic Foundation and employed various ML algorithms, including decision tree, logistic regression, random forest, naive Bayes, and support vector machine (SVM), to develop prediction models. A 10-fold cross-validation approach was used during the model development process. The results showed that the decision tree algorithm achieved the highest accuracy in predicting heart disease, with a rate of 93.19 %, followed by the SVM algorithm at 92.30 %. This study demonstrates the potential of ML techniques as a valuable tool for predicting heart failure disease and highlights the decision tree algorithm as a promising option for further research. Previous studies have limitations, particularly in terms of dataset and differing model performance approaches. Given these advancements, the study aspires to develop a cutting-edge hybrid machine learning model that capitalizes on the advantages of numerous algorithms to estimate heart disease risk using body composition data [6]. The hybrid model combines the strengths of multiple models predict on the body composition dataset to produce improved result.

2.1. Techniques adopt in previous studies

To guarantee that the selected machine learning method was impartial and appropriate for the task, a thorough literature review was carried out. This involved examining and synthesizing existing research on machine learning techniques for predicting healthcare outcomes, with a focus on cardiovascular disease. The review encompassed a broad range of publications, including journal articles, conference proceedings, and book chapters with different techniques or methods used in these studies. Md [13], the study use methodology encompasses data preprocessing, feature selection, and the implementation of supervised classification methods. With a dataset comprising 13 features, the authors utilized the info-gain feature selection strategy to identify the most pertinent attributes for predicting heart disease. This led to the utilization of 10 attributes, including age, gender, height, weight, ap hi, ap lo, cholesterol, gluc, active, and cardio, which were identified as highly correlated through the feature selection approach. The authors utilized the mean imputation approach to address null values during data preprocessing, aiming to enhance the accuracy of the classification models. Subsequently, the study employed three supervised classification methods - K-Nearest Neighbors (KNN), Naive Bayes, and Random Forest - to predict the presence of early-stage cardiac disease in patients [14]. The effectiveness of these models was assessed through 10-fold cross-validation, with 70 % of the data allocated for training and 30 % for testing. The study's authors presented the outcomes of the classification models, which included the accuracy of the KNN algorithm. The KNN algorithm achieved an 87 % accuracy rate in predicting the existence of heart disease. The study's methodology encompassed data collection, dataset description, data pre-processing, feature engineering, and the application of machine learning algorithms. Utilizing the Cleveland UCI heart dataset, the study employed diverse data engineering strategies to enhance accuracy. The results of the study indicated that the best classification results were obtained when using the Random Forest (RF) classifier. The study achieved a maximum test accuracy of 95.4 % after optimizing the hyperparameters using the Random Search optimization method. The classification results using different classifiers (SVM, KNN, DT, and RF) were presented, with the RF classifier yielding the best results. Additionally, the study utilized several categorization performance metrics to assess the effectiveness of the techniques employed [10]

In another study by [7], the methodology used in the study involved employing various machine learning algorithms on a dataset to predict heart disease. The dataset, comprising 70,000 patient records with 12 distinct features, was preprocessed and cleaned. The authors applied k-modes clustering to preprocess the dataset and scale it, and the computation, preprocessing, and visualization were conducted using Python on Google Colab. The algorithms used in the study included random forest, decision tree, multilayer perception, and XGBoost classifier. The dataset was split into two parts: 80training the model and 20 % for testing the model. An automated approach for hyperparameters tuning was employed using the GridSearchCV method. The results of the study indicated that the dataset, after cleaning and preprocessing, was reduced to approximately 59,000 rows and 11 attributes. The study used several measures of performance, including precision, recall, accuracy, F1 score, and area under the ROC curve. The authors reported that the accuracy rates of up to 94 % have been achieved using machine learning techniques for heart disease prediction in previous studies. However, the authors aimed to address the limitation of small sample sizes by using a larger and more diverse dataset to increase the generalizability of the results.

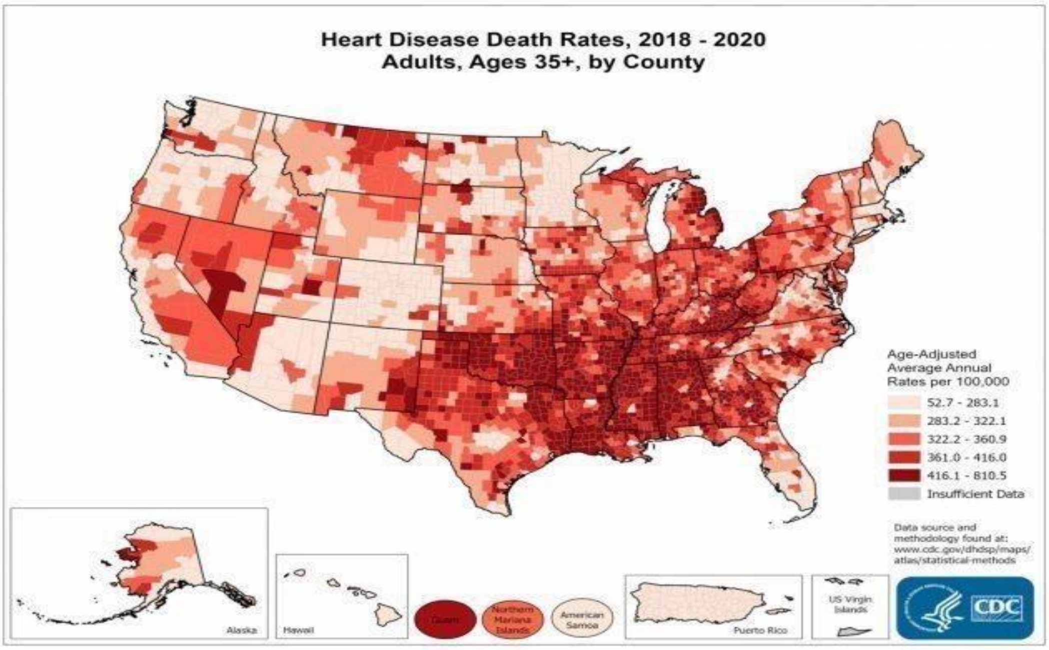
Heart Disease Death Rates, 2018 - 2020
Adults, Ages 35+, by County
Age-Adjusted Average Annual Rates per 100,000
52.7 - 283.1
283.2 - 322.1
322.2 - 360.9
361.0 - 416.0
416.1 - 810.5
Insufficient Data
Data source and methodology found at: www.cdc.gov/dhdsp/maps/atlas/statistical-methods
Alaska
Hawaii
Guam
Northern Mariana Islands
American Samoa
Puerto Rico
US Virgin Islands
CDC

Figure 1: Heart Disease Distribution Across Populations

Globally, the prevalence and impact of heart disease deaths from 2018 to 2020 varied across regions and populations. However, several overarching trends and statistics can provide insight into the general landscape of heart disease mortality during this period.

2.2. Types of Heart disease

1. Coronary artery disease: The narrowing or blockage of the coronary arteries, which supply blood to the heart muscle, leading to chest pain or a heart attack. Heart failure: When the heart is unable to pump enough blood to meet the body's needs, leading to fatigue, swelling, and shortness of breath.
2. Arrhythmias: Abnormal heart rhythms, which can be too fast, slow, or irregular, and can lead to palpitations, dizziness, or fainting.
Heart valve disease: Problems with the heart valves that control blood flow, leading to symptoms such as fatigue, swelling, and shortness of breath.
3.Cardiac arrest: A sudden stop of the heart's function, often caused by a heart attack or arrhythmia, which can lead to death if not treated promptly.
4.Congenital heart disease: Defects in the structure of the heart present at birth, which can range from minor to severe and may require surgical intervention.
5. Pulmonary hypertension: High blood pressure in the lungs, which can lead to shortness of breath, fatigue, and swelling.
6. Pericardial disease: Inflammation or fluid buildup around the heart, which can cause chest pain, fever, and difficulty breathing.
7. Endocarditis: An infection of the inner lining of the heart, which can damage heart valves and cause symptoms such as fever, chills, and joint pain.
8. Myocarditis: Inflammation of the heart muscle, which can lead to chest pain, fatigue, and shortness of breath.

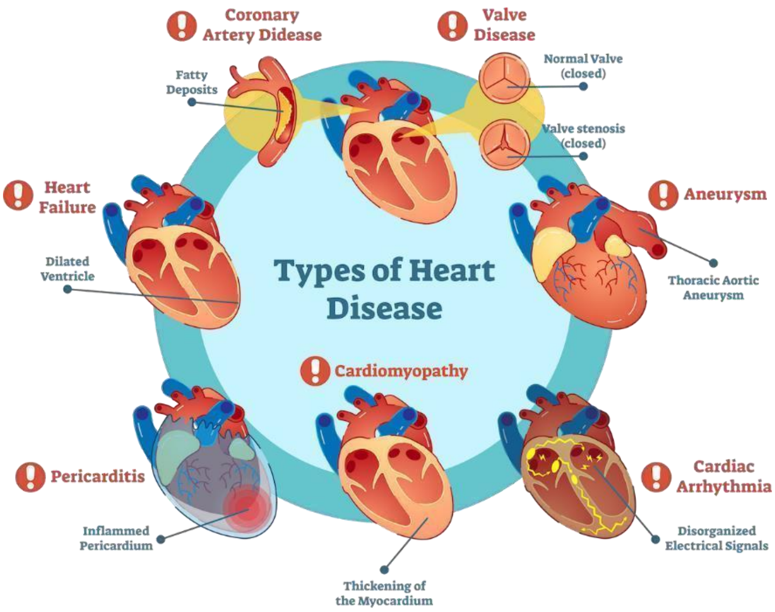

Figure 2: Types of Heart Diseases

Table Common Types of Heart Diseases, Their Descriptions, Factors, and Causes

Heart Disease (HD)	**Description**	**Factors/causes**
Coronary artery disease	Plaque accumulation inside the coronary arteries, the blood channels supplying the heart, results in coronary artery disease (CAD).	Elevated blood pressure, high cholesterol, cigarette smoking, Diabetes, Fatness, Absence of exercise Heart disease in the family history
Heart failure	Heart failure happens when the circulatory system is unable to pump enough blood to meet the body's requirements	The coronary artery elevated blood pressure, Diabetes heart valve issues, Heart rhythm issues, heart muscle illness, Congenital cardiac conditions
Arrhythmias	Abnormal heartbeats, or arrhythmias, can be excessively rapid, erratic, or both.	Heart muscle disease, Heart valve problems, Electrolyte imbalances, Medications Caffeine and nicotine
Heart Valve Problems	Heart valve issues can arise from damage or illness affecting the valves responsible for regulating blood flow through the heart	Aging, Rheumatic fever, Endocarditis (an infection of the heart valves) Congenital heart defects, Heart injury
Cardiomyopathy	Heart muscle diseases known as cardiomyopathies can weaken or expand the heart muscle.	High blood pressure Diabetes, Obesity Alcohol abuse, Viral infections, Genetic mutations
Pericardial Disease	The pericardial disease occurs when the membrane surrounding the heart becomes inflamed or infected	Infection, Autoimmune disorders, Radiation therapy Tumors, Trauma
Pulmonary Embolism	When a blood clot forms in the lungs and prevents blood from reaching the heart, it is known as a pulmonary embolism	Deep vein thrombosis Atrial fibrillation, Cancer, Injury, Surgery

2.3. Traditional methods

1. **Physical examination:** A doctor may listen to the patient’s heart sounds, check their pulse, and measure their blood pressure to detect any abnormalities.
2. **Electrocardiogram (ECG):** An ECG is a test that records the electrical activity of the heart. It can detect irregular heart rhythms, conduction problems, and other abnormalities.
3. **Chest X-ray**: A Chest X-ray can show the size and shape of the heart, as well as any fluid accumulation in the lungs, which can indicate heart failure.
4. **Blood tests Blood** tests can measure various markers, such as troponin, creatine kinase, and brain natriuretic peptide, which can indicate heart muscle damage or Strain.
5. Stress test: A stress test, also known as an exercise stress test, measures the heart’s ability to function during physical activity. It can help diagnose coronary artery disease, heart valve problems, and other conditions.
6. Echocardiogram: An echocardiogram is an ultrasound test that produces images of the heart. It can show the heart’s structure, function, and blood flow, and can help diagnose various heart conditions, such as heart valve problems and cardiomyopathy.
7. Cardiac catheterization: Cardiac catheterization is a test that involves inserting a thin tube (catheter) into a blood vessel in the arm or leg and guiding it to the heart. It can help diagnose coronary artery disease, heart

valve problems, and other conditions, and can also be used to perform procedures such as angioplasty and stenting.

8. Magnetic resonance imaging (MRI): MRI is a test that uses magnetic fields and radio waves to produce detailed images of the heart. It can help diagnose various heart conditions, such as cardiomyopathy, heart valve problems, and coronary artery Disease.

III. Methodology

This section provides an overview of the research methodology employed in this study. Specifically, it describes the data collection process, the preprocessing steps, and the statistical analysis methods used to investigate the relationship between body composition and heart disease. The database also contained information on patients' medical history, including their age, gender, smoking status, alcohol consumption, physical activity level, and medication use. Additionally, patients' serum lipid profiles, blood pressure, and fasting glucose levels were also recorded. Patients with a diagnosis of heart disease were identified based on their medical records, which included information on previous myocardial infarctions, coronary artery disease, heart failure, and arrhythmias. The aim of this study is to create a hybrid machine learning model that leverages the strength of multiple algorithms to predict heart disease risk based on body composition data.

3.1. Research Design Method

This study aims to predict the probability of heart disease through computerized heart disease prediction using a hybrid machine learning approach, which can be beneficial for medical professionals and patients. To achieve this objective, the study employed various machine learning algorithms on a dataset. To enhance the methodology, the plan to clean the data, eliminate irrelevant information, and incorporate additional features such as MAP and BMI. Next, the researcher separates the dataset based on gender and implement k-modes clustering.

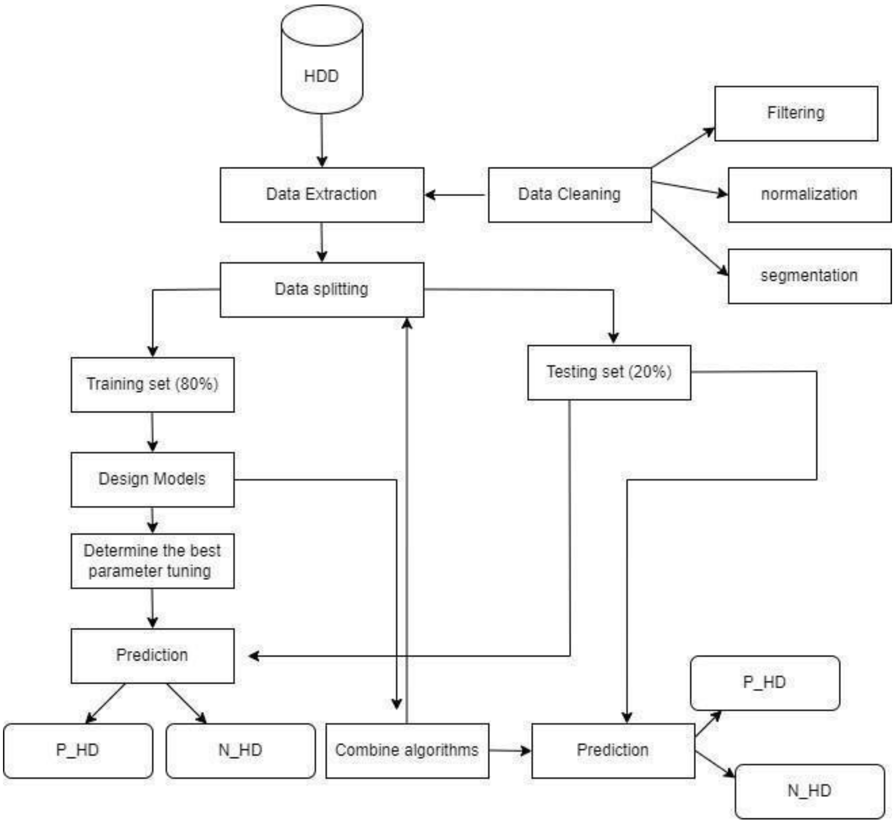

Figure 3: Methodology of the study

3.2. Technical Tools used

The study leveraged Python as the primary programming language for all data processing and manipulation tasks. To support these efforts, a variety of Python-based libraries were imported, and the Python environment was created using Anaconda. Additionally, Jupyter Notebook, a web-based IDE, allowed for the creation, execution, and visualization of Python scripts. The operating system used for the study was Windows, and the hardware consisted of a Quad-Core Processor machine with 6GB of installed memory, running on an x64 architecture. 1. NumPy provides mathematical functions for computing multidimensional array objects 2. Pandas Performs data structure operations, imports data, and analyzes data 3. Matplotlib Creates 2D plots 4. ScikitLearn Offers machine learning algorithms and evaluates model performance 5. Seaborn Builds upon the Matplotlib library for data visualization Source: Authors' construct, 2021 The above shows the libraries that were imported into python to aid in data processing, machine learning model implementation and data visualization. Scikit-learn contains a number machine learning model packages including that of Decision tree model (DTM) and XGBOOST, LIGHTGBM, Support Vector Machine (SVM), KNN and Hybrid model.

3.3. Data Preprocessing

In this research the dataset used was obtained from Kaggle.com, which is an open source community. Kaggle is the world's largest data science community with powerful tools and resources to help researcher to find their desire dataset for their various research. The dataset comprises 717 records with 12 distinct features as listed in Figure 3. Some key body composition features extracted from the raw dataset. Such as Gender, Age, Head Circumference, shoulder Width, Chest width, Belly, Waist, Hips, and Arm Length, ShoulderToWaist, WaistToKnee, Leg Length, and Total Height. Exploratory Data Analysis (EDA) and Preprocessing are crucial steps in the data science workflow. EDA involves examining and summarizing datasets to understand their structure, patterns, and relationships, while pre-processing focuses on cleaning, transforming, and preparing data for modeling or further analysis. These stages are vital in ensuring that the data is accurate, consistent, and ready for use in machine learning algorithms or other analytical techniques. Effective exploration and pre-processing can significantly impact the quality of insights gained from data analysis, making them essential skills for any data scientist or analyst. Python programming language will be used for the development of the models.

3.3.1. Two techniques for the feature's selection

i. Principal Component Analysis (PCA) is a linear dimensionality reduction technique that was used to transforms the original dataset into a new set of orthogonal features called principal components. The first few principal components often capture the most important features in the data.

ii. Linear Discriminant Analysis (LDA) LDA is a supervised feature extraction technique that was used to reduce the dimensionality of the data while preserving class reparability. It finds a lower-dimensional representation of the data that maximizes the separation between classes. LDA also can be describe as statistical method commonly used for dimensionality reduction and classification tasks. In the context of predicting heart disease, LDA can be employed to identify linear combinations of features that best discriminate between individuals with and without heart disease. To compute the mean vectors for each class (e.g., individuals with and without heart disease). To compute the mean vector uk for class k, we use the formula:

$$uk = \frac{1}{nk} \sum_{i=1}^{nk} x_{ki}$$

Where

- uk is the mean vector for class k.
- nk is the number of samples in class k.
- x_{ki} is the feature vector of the i^{th} sample in class k. To compute the covariance matrix for each class

$$Sk^{T} = \frac{1}{nk} \sum_{i=1}^{nk} (x_{ki} - uk)(x_{ki} - uk) - 1$$

Where

- nk Is the number of samples in class k
- Is the mean vector for class k.
- x_{ki} Is the feature vector of the i^{th} sample in class k.
- T denotes the transpose

Sum up the covariance matrices for all classes to get the within-class scatter matrix sw:

$$SW = \sum^{K} S_{k}$$

$k=1$ K is the total number of classes.

The covariance matrix for each class is computed by calculating the covariance matrix for each class individually and then summing them up to form the within-class scatter matrix. This matrix is a crucial component in the LDA algorithm for dimensionality reduction and classification.

3.4 **Partitioning of Data**

The data was split into two subsets using a random partitioning approach. The first subset, accounting for 80 % of the total data, was employed as training data, whereas the remaining 20 % served as test data [8]. A model was initially trained using the training data, and its performance was subsequently assessed using the test data. According to the researcher's preference, the data was divided in an 80:20 proportion, allocating 80 % for training and 20 % for validation purposes.

3.5 **Classification Analysis**

The medical field faces a significant challenge in tackling heart disease, which continues to grow in prevalence annually. To address this issue, numerous machine learning algorithms have been developed to predict the likelihood of heart disease. In particular, this study focused on six popular machine learning approaches, including Decision tree model (DTM) and XGBOOST, LIGHTGBM, Support Vector Machine (SVM), KNN and Hybrid model. These algorithms are well-suited for predicting binary dependent variables and have been widely applied in the healthcare sector. The machine learning models utilized in this research include: I. Decision tree model (DTM) Decision tree is a popular machine learning algorithm used for both classification and regression tasks. It works by recursively partitioning the data set into smaller subsets based on the values of the input features. Each internal node in the tree represents a feature selection and each leaf node represents a class label or predicted value. The decision tree algorithm uses a top-down approach, starting with the root node, and iteratively splits the data set into smaller subsets until a stopping criterion is met. Decision trees have several advantages, including ease of interpretation, handling missing values, and dealing with non-linear relationships.

3.5.1. **Hyperparameters used in Grid Search optimization method**

Table 3.2 below displays the hyperparameters selected for use in the GS optimization method, along with their definitions and default values, for various machine learning (ML) algorithms. These hyperparameters are crucial for enhancing the outcome of the ML models, and their tuning is essential for achieving optimal performance [15].

Table 2: ML Hyperparameters Used in Grid Search

Algorithm	Parameters	Definition	Default
Decision Tree	Max_depth	The tree's highest depth.	5
Random Forest	1. Max_depth 2.learning_rate	1. Maximum depth of the tree	6
		2. Rate of change of weights during training	0.1
Artificial Neural Networks(ANN)	1. num_leaves 2.learning_rate	1. Number of leaves in the decision tree 2.Rate of change of weights during training	50 0.06
Support Vector Machine(SVM)	C Gamma	Penalty parameter For misclassifications Kernel coefficient for 'rbf', 'poly', and 'sigmoid'	1
K-Nearest Neighbor	n-neighbors Weights Metric	In prediction, many neighbor functions are employed. Using a uniform measure to calculate the distance.	10 Uniform Minkowski
Hybrid model	linear_reg_alpha -> decision_tree_max_depth -> random_forest_n_estimators neural_networks_learning_rate	Regularization parameter for linear regression component Highest decision tree component depth The random forest component's tree count Rate at which weights change during neural network layer training	0.03 5 100 0.002

3.5.2. **Performance Metric measurement**

The assessment of the efficacy of numerous classification techniques necessitates the employment of appropriate evaluation metrics can be shown in table.

Table 3 Equations for Evaluating Classifiers' Performance

Metric Name	Equation
Recall	TP/(TP + FN)
Precison	TP (TP +FP)
F1 Score	2 * (Precision * Recall) / (Precision + Recall)
Accuracy	TP + TN) / (TP +FN +TN +FP)

IV. Experimental Evaluation

The experimental framework utilizes Anaconda version 6.5.4, a widely embraced Python programming language, and Jupyter Notebook for an immersive and dynamic development environment. Anaconda's versatility and utility are evident through its controlled ecosystem for each project, ensuring harmonious dependencies and streamlined progress. The installation on a Windows 11 desktop machine with a 1TB hard drive and 8GB memory provides a foundation for innovation and exploration. Python's programming prowess and rich library ecosystem make it the perfect language for the experiment's objectives. The vast repository of Python libraries, including pre built functions and modules, enhances the research's capabilities and infuses it with the collective wisdom of the programming community.

4.1. **Data obtained**

The dataset utilized for the study was retrieved from Kaggle.com, comprising a total of 717 instances with 13 features that were employed to train six machine learning algorithms for classification purposes. After data preparation, the dataset was divided into two subsets: a training set (80 %) and a testing set (20 %). The training set was used to train the machine learning models, whereas the testing set was used to evaluate their performance. This process allowed for the assessment of the models' accuracy and enabled the selection of the most effective model for heart disease risk prediction.

4.2 Comparison of proposed models

Table 4 Comparison of base models against the hybrid heart predictor model

Classifiers Names	Precision	Recall	F Score	No of samples	Accuracy %)
DT	89	87	88	102	88%
RF	82	78	80	102	80.5%
SVM	81	75	78	102	78.5%
KNN	81	75	78	102	78.5%
ANN	69	63	58	103	74.3 %
HHP	93	87	90	102	90.2%

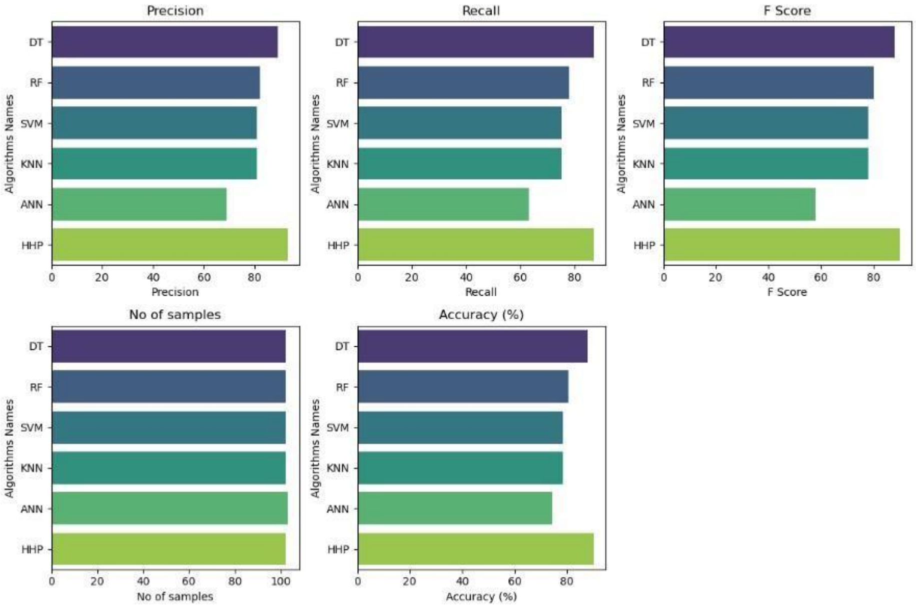

Figure 5: Comparison of proposed models

4.2.1 Comparison of Errors made by the algorithms using MSE

Table 5 Comparison of Errors made by the algorithms using MSE

Classifiers Names	Errors (%)
Decision Tree (DT)	Mean Squared Error (MSE): 0.43902439024390244
Random Forest (RF)	Mean Squared Error (MSE): 0.13658536585365855
Support Vector Machine (SVM)	Mean Squared Error (MSE): 0.28292682926829266
K-Nearest Neighbors (KNN)	Mean Squared Error (MSE) - KNN: 0.44390243902439025
Artificial Neural Networks (ANNs)	Mean Squared Error (MSE) - ANN: 0.2926829268292683
Hybrid heart predictor (HHP)	Mean Squared Error (MSE) - HHP Classifier: 0.0975609756097561

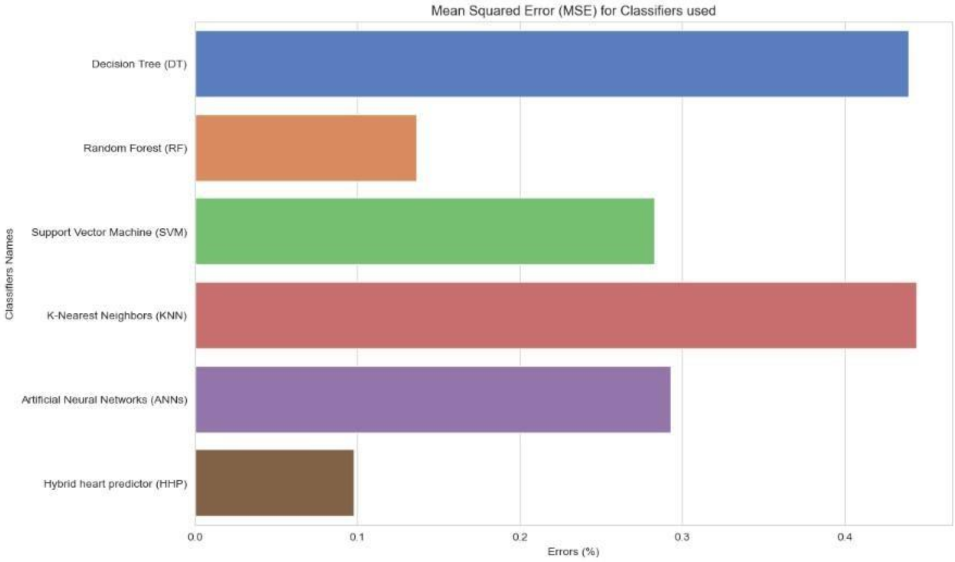

Figure 6 Comparison of errors made by the classifiers

V. Conclusion and Findings

This research work employs six supervised machine learning in predicting heart disease among individuals. The body composition dataset was obtained from Kaggle.com containing 717 instances with 13 features. The dataset was divided into training (80 %) and testing (20 %) sets for model training and evaluation. Six base algorithms, including Decision Tree, Random Forest,SVM, kNN, ANN, and HybridHeartPredictor (HHP) model, were used to diagnose various heart diseases. The performance of the base models was evaluated using confusion matrices to assess accuracy, precision, recall, and F1 score. The HHP model outperformed the other base models in terms of performance. Analysis of feature importance and detection rate per feature provided insights into the impact of different features on the HHP model's performance. A hybrid model was developed by combining predictions from the base models using a weighted average approach, with optimized weights determined through grid search. A comparative table displayed the performance metrics of both the base models and the HybridHeartPredictor (HHP) model. The HHP model outperformed others in precision, recall, F score, and accuracy. Specifically, for precision, the percentages were as follows: Decision Tree (DT) 89 %, Random Forest 82 %, SVM 81 %, KNN 81 %, ANN 69 %, and HHP 93 %. In terms of recall, DT achieved 87 %, Random Forest 78 %, SVM 75 %, KNN 75 %, ANN 63 %, and HHP 87 %. Similarly, for F Score, DT had 88 %, Random Forest 80 %, SVM 78 %, KNN 78 %, ANN 58 %, and HHP 90 %.

Lastly, in accuracy, DT achieved 88 %, Random Forest 80.5 %, SVM 78.5 %, KNN 78.5 %, ANN 74.3 %, and HHP excelled with 90.2 %. Furthermore, the Mean Squared Error (MSE) was used to evaluate the error rates of the models. The results showed that the HHP model made the least errors, with an MSE of 0.1 %, followed by the Random Forest model with an MSE of 0.13 %, the SVM model with an MSE of 0.3 %, the KNN model with an MSE of 44.3 %, the DT model with an MSE of 44 %, and the ANN model with an MSE of 0.3 %. Overall, the findings suggest that the HybridHeartPredictor model outperformed the other models in terms of predictive accuracy and minimizing errors, demonstrating its potential as a robust and reliable tool for heart disease diagnosis. The research made several recommendations. The study proposed that future research could investigate the inclusion of additional body composition features or health-related variables that could further enhance the predictive accuracy of the models. Longitudinal Data Analysis: Consider incorporating longitudinal data analysis to track changes in body composition over time and assess their impact on heart disease risk prediction. Interpretability of Models: Focus on enhancing the interpretability of machine learning models to provide insights into the factors influencing heart disease risk predictions based on body composition data. Validation on Diverse Populations: Validate the developed models on diverse populations to ensure their generalizability and effectiveness across different demographic groups. Integration of Genetic Data: Explore the integration of genetic data with body composition features to investigate the combined influence of genetic predisposition and lifestyle factors on heart disease risk. Real-time Prediction: Develop models capable of real-time prediction to enable timely interventions and personalized healthcare strategies for individuals at risk of heart disease. Clinical Implementation Studies: Conduct studies focusing on the practical implementation of machine learning models in clinical settings to evaluate their impact on patient outcomes and healthcare decision-making processes.

References

[1]. GHO, G. World Health Organization. Retrieved from World Health. World Health Organization. Retrieved from World Health 2016, 213, 259–265.

[2]. Kelly R, P. Y. J. Assessment of body composition and its relevance to cardiovascular disease. Journal of Investigative Medicine 2018, 66, 259–265.

[3]. Deurenberg, J.; Weststrate Measurements of body composition by bioelectrical impedance analysis. International Journal of Obesity 1999, 66, 259–265.

[4]. Miy, K.; R, H.; Liu, Y. Applications of machine learning in healthcare. Journal of Healthcare Engineering 2018, 18, 259–265.

[5]. Sarraf, M. Wearable sensors and machine learning for health monitoring systems. IEEE Reviews in Biomedical Engineering 2019, 20, 259–265.

[6]. Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. Eur. J. Mol. Clin. Med 2020, 20, 259–265.

[7]. Bhatt, C. M.; Patel, P.; Ghetia, T.; Mazzeo, P. L. Effective heart disease prediction using machine learning techniques. Algorithms 2023, 16, 88.

[8]. Shah, D.; Patel, S.; Bharti, S. K. Heart disease prediction using machine learning techniques. SN Computer Science 2020, 1, 345.

[9]. Shah, D.; Patel, S.; Bharti, S. K. Heart disease prediction using machine learning techniques. SN Computer Science 2020, 1, 345.

[10]. Kadhim, M. A.; Radhi, A. M. Heart disease classification using optimized Machine learning algorithms. Iraqi Journal For Computer Science and Mathematics 2023, 4, 31–42.

[11]. Garate-Escamila, A. K.; El Hassani, A. H.;´ Andr`es, E. Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked 2020, 19, 100330.

[12]. Jindal, H.; Agrawal, S.; Khera, R.; Jain, R.; Nagrath, P. In IOP conference series: materials science and engineering, 2021; 1022, 012072.

[13]. Pudjihartono, N.; Fadason, T.; KempaLiehr, A. W.; O'Sullivan, J. M. A review of feature selection methods for machine learning-based disease risk prediction. Frontiers in Bioinformatics 2022, 2, 927312.

[14]. Mohan, S.; Thirumalai, C.; Srivastava, G. Effective heart disease prediction using hybrid machine learning techniques. IEEE access 2019, 7, 81542–81554.

[15]. Bharti, R.; Khamparia, A.; Shabaz, M.; Dhiman, G.; Pande, S.; Singh, P., et al. Prediction of heart disease using a combination of machine learning and deep learning. Computational intelligence and neuroscience 2021, 2021.

[16]. Mandava, M. et al. MDensNet201-IDRSRNet: Efficient cardiovascular disease prediction system using hybrid deep learning. Biomedical Signal Processing and Control 2024, 93, 106147.

Author's Profile

Awal Mohammed holds an MSc. in Information Technology from KNUST, Ghana (2024), a B.Sc. (Hons) in Business Information Systems from the University of Education, Winneba (2013), and an HND in Computer Network Management from Koforidua Technical University (2009).

He is the Management Information Systems Manager at NHIA and a Technology Consultant, with professional certifications in networking and cybersecurity. His research interests span Information Systems, IT Project Management, Computer Networks and Security, E-learning, and Machine Learning.

Daniel Kwadwo Nterful is a Senior ICT Assistant at the Directorate of ICT Services, Takoradi Technical University. He received MSC in Information Technology from Kwame Nkrumah University of Science and Technology (KNUST) Kumasi. Bachelor of Technology in Information Technology (Application Management) Takoradi Technical University (TTU), Higher National Diploma from Ho Technical University (HTU).

He has mentored approximately 15 postgraduate students in their research work, including one PhD student.

His research interests include: AI, Machine learning, Deep learning, and Database management Systems.

Issah Bawah Muhammed is a seasoned information technology professional with a strong academic background and extensive teaching experience. He holds a Master of Science (MSc) in Information Technology from Kwame Nkrumah University of Science and Technology (KNUST). Bachelor of Science (BSc) in Information Technology, as well as an Advanced Diploma and Diploma in Information Systems, both awarded by the Institute for the Management of Information Systems (IMIS), United Kingdom.

Currently, Issah serves as an IT tutor at the School of Peri-Operative and Critical Care Nursing- Korle Bu.