Intelligent Control and Automation, 2024, 15, 9-27
https://www.scirp.org/journal/ica
ISSN Online: 2153-0661
ISSN Print: 2153-0653

A Machine Learning-Based Web Application for Heart Disease Prediction

Jesse Gabriel

Independent Researcher, Port Moresby, Papua New Guinea
Email: jessegabriel11@gmail.com

How to cite this paper: Gabriel, J. (2024) A Machine Learning-Based Web Application for Heart Disease Prediction. *Intelligent Control and Automation*, **15**, 9-27.
https://doi.org/10.4236/ica.2024.151002

Received: December 7, 2023
Accepted: February 15, 2024
Published: February 18, 2024

Abstract

This work leveraged predictive modeling techniques in machine learning (ML) to predict heart disease using a dataset sourced from the Center for Disease Control and Prevention in the US. The dataset was preprocessed and used to train five machine learning models: random forest, support vector machine, logistic regression, extreme gradient boosting and light gradient boosting. The goal was to use the best performing model to develop a web application capable of reliably predicting heart disease based on user-provided data. The extreme gradient boosting classifier provided the most reliable results with precision, recall and F1-score of 97%, 72%, and 83% respectively for Class 0 (no heart disease) and 21% (precision), 81% (recall) and 34% (F1-score) for Class 1 (heart disease). The model was further deployed as a web application.

Keywords

Heart Disease, US Center for Disease Control and Prevention, Machine Learning, Imbalanced Data, Web Application

1. Introduction

Recently, there have been major developments in computing; both hardware and software technologies including robust artificial intelligence (AI) systems and tools. AI systems use various algorithms, especially machine learning (ML), to learn from data and/or experience, make predictions and improve their prediction performance. This "ability" of AI has attracted wider adoption and application of AI technologies across multiple disciplines, organisations and industries including the health care industry. ML has been proven beneficial in giving an immeasurable platform in the medical field so that health care issues can be resolved effortlessly and expeditiously [1] [2] [3] [4]. Health care organisations

depend on computer-based technology in the gathering and storage of vast amounts of data in electronic formats. This has provided a good platform for automation [5], streamlining [6] and optimization [7] of workflows with ML and related technologies leading to improved efficiency, productivity and decision-making [8]. For instance, natural language processing (NLP) and ML (including deep learning) have been facilitating the development of intelligent chatbots like ChatGPT, digital health tools and applications that are significantly changing and redefining medical practice and service provision. Such technologies have established proof of concepts in some medical specialities such as radiology, psychiatry, pathology, and ophthalmology. The technologies can be further useful to assist early detection, prediction, diagnosis and management of different health conditions. In disease prediction, ML models are usually trained on a data set that contains the disease to be predicted along with all the recorded signs and symptoms associated with the disease. These signs and symptoms are termed features. Once the model has been trained and tested on how accurately it can predict the disease using the selected features, it can be deployed to make predictions on new data provided by users. Generally, the more the amount of data used to train the models, the better the prediction performance using new data. Researchers have explored the performance of different ML algorithms in predicting diseases. Recent publications in the literature indicate that the commonly used models for disease prediction include logistic regression (LR), random forest (RF), support vector machines (SVM), decision trees (DT) and gradient boosting (GB). Other models like the neural networks (deep learning), bayesian networks and naive bayes have been used by few. Apart from the conventional models, other researchers like [9] have also implemented hybrid models, comprising several models with or without novel techniques. Most of these studies involved predicting diabetes e.g. [10] [11] [12], depression, e.g. [13] [14] [15], hypertension, e.g. [16] [17], anxiety e.g. [18] [19], and heart disease, e.g. [10] [20] [21] [22]. Heart disease is a prevalent health issue globally and predicting the likelihood of heart disease is quite critical. The present work considered the above cited works and especially related works from [10] [20] [22] [23]. [10] used a two-stage machine learning ML models (logistic regression and Evimp functions) model to predict the co-occurrence of Diabetic mellitus (DM) and cardiovascular diseases (CVD). Their data involved 2000 participants who were older than 40 years old and who had to undergo specific requirements for the purpose of testing and data collection. The results of their work showed predictive accuracy of the ML model to detect co-occurrence of DM and CVD at 94.09%, sensitivity 93.5%, and specifcity 95.8%. [22] used five ML (LR, k-nearest neighbours—k-NN, Naïve Bayes, RF and Extreme Gradient Boosting) and two deep learning models (Multilayer perceptrons—MP and Convolutional neural networks—CNN) to predict eight most common chronic conditions including cardiovascular as defined by the Australian Government Department of Health (AGDoH) [24]. Their method involved a novel feature engineering technique

where the patient network was engineered from bipartite graphs. They concluded that the extreme gradient boosting (XGBoost) produced a highest accuracy of 95.05%. Similarly, [20] compared the performance of four ML models (SVM, RF, XGBoost, and k-NN) in predicting heart disease. They mentioned that their dataset was sourced from Cleveland Heart Disease Dataset. This preprocessed dataset has around 300 rows (instances) and 14 features and can be seen at the repository provided by [25]. Their results showed that the XGBoost along with SVM models exhibited the highest F1-score performance, reaching up to 88%. [21] used the same dataset [25] to train five machine learning models (naive bayes, artificial neural networks, SVM, RF and LR). Their result indicated an accuracy level of 97.53% accuracy from the SVM algorithm along with sensitivity and specificity of 97.50% and 94.94% respectively. They further extended their work in developing a cloud-based application where the patients can upload the physiological data for checking the status of their cardiac health.

The above cited works have many strengths, but there are few limitations. One is that the size of the dataset and the variety in the data is small. This can lead to models performing well during the development (training and testing) phase but they can perform poorly on new data. Another is the lack of clarity in the data preprocessing (e.g. handling of imbalanced datasets) and the subsequent selection of evaluation matrices. Usually datasets for such cases as disease prediction, fraud detection, etc., are imbalanced in nature where the number of positive cases are significantly less than that of negative cases. Appropriate methods to handle imbalanced data (e.g. resampling the minority class, class weighting, etc.) need to be implemented. In addition, the performance matrices selected should also take into account the performance of the model against each class as the overall performance such as accuracy may not provide clear insights on the models performance in predicting each class [26]. Finally, not many have attempted extending their work into creating software applications that can be practically used. In the cited works, only [25] have reported the development of a cloud-based application. Reliable software applications that are based on ML technology can provide many benefits. For instance, shortcomings particularly in developing countries (e.g. Papua New Guinea) include shortage of doctors and healthcare professionals, resulting in high strain on the limited workforce and resources [27] [28] [29]. This has seen health professionals and doctors deal with demanding situations to research signs and symptoms correctly and perceive illnesses at an early stage. Additionally, due to the isolation and lack of infrastructure in many rural communities, health care service delivery is a major challenge requiring significant attention and funding efforts [30] [31] [32] [33]. Tested and reliable open-source software applications can help in such situations.

Drawing from the above cited works and situations, the aim of this study was to train and test five of the commonly used ML models and to use the best performing model to create and deploy a web-based application that can reliably

predict heart disease using user-input data. The approach for this included data sourcing, data pre-processing, model training, evaluation of the results, and model deployment on the web. Each of these steps are described accordingly in the following sections and subsections. A brief discussion on the results is also provided along with notable strengths and limitations of this work, which can be improved with further work extended from this study.

2. Method

2.1. Data Source

The original dataset for this study was sourced from the Center for Disease Control and Prevention (CDC) [34] in the United States (U.S.). CDC collects the data through the Behavioral Risk Factor Surveillance System (BRFSS). BRFSS is a collaborative project between all of the states in the U.S. and participating U.S. territories and the CDC. BRFSS was initiated in 1984, with 15 states collecting surveillance data on risk behaviors through monthly telephone interviews. Over time, the number of states participating in the survey increased. BRFSS now collects data in all 50 states as well as the District of Columbia and participating U.S. territories. During 2020, all 50 states, the District of Columbia, Guam, and Puerto Rico collected BRFSS data. BRFSS completes over 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world. According to the CDC, heart disease is a leading cause of death for people of most races in the U.S. (African Americans, American Indians and Alaska Natives, and whites). About half of all Americans (47%) have at least 1 of 3 major risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicators include diabetes status, obesity (high BMI), not getting enough physical activity, or drinking too much alcohol. Identifying and preventing the factors that have the greatest impact on heart disease is very important in healthcare. In turn, developments in computing allow the application of machine learning methods to detect "patterns" in the data that can predict a patient's condition.

2.2. Data Preprocessing

The original dataset is in SAS Transport Format and contains 401,958 rows and 279 variables. Preprocessing of the data was done with Python 3 using pandas, numpy, sklearn and matplotlib modules. Preprocessing included handling of missing or NaN values, feature selection, encoding of categorical data, data standardization, data splitting into train and test sets, and weighting of the classes.

Missing values (*NaNs*): Through the data cleansing process, missing values (NaNs) or rows with incomplete entries were removed, leveraging the dataset's size to ensure data integrity.

Feature selection: Most features (columns) had significant portion of the values missing and had to be removed altogether. It was ensured that the selected features captured the key risk factors of heart disease. These two steps resulted in

the selection of 17 features (Table 1) and 319,795 instances. The selected features considered the person's Body Mass Indix (BMI) category, if the person had smoked at least 100 cigarettes (approx. 5 packets) in their lifetime, if the person has more than 14 drinks of alcohol (men) or more than 7 (women) in a week, if they stroke in the past, the number of days their physical health was not good over the last 30 days, the number of days their mental health was not good over the last 30 days, if the person has difficulty walking, the persons gender, their age category, race, if the person is a diabetic, if the person played any sports (running, biking, etc.) in the past month (physical activity), the persons general health, their average sleep hours in a day, if the person has asthma, if the person has kidney disease, and if the person has skin cancer.

Categorical data encoding: The majority of the gathered data comprises categorical variables, as depicted in Table 1. To facilitate ML model compatibility, label encoding was employed to convert non-numeric categories into corresponding numerical representations. For instance, binary categorical values such as "Yes" and "No" were transformed into values of 1 and 0 respectively.

Data standardization: Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance). The concept of standardization and the associated *z-score* formula ([35], p. 880) has been widely used in statistics and mathematics. By substituting the *arithmetic mean* [36] and the *standard deviation* [37] formulas into the *z-score* formula, Equation (1) can be obtained and used for data standardization.

Table 1. The first five rows of the data with the selected 17 features and the target variable (Heart Disease). Features 1 to 8 and in the top panel and 9 to 17 are in the bottom panel.

Heart Disease	BMI Category	Smoking	Alcohol Drinking	Stroke	Physical Health	Mental Health	Diff Walking
No	Underweight (BMI < 18.5)	Yes	No	No	3.0	30.0	No
No	Normal weight (18.5 ≤ BMI < 25.0)	No	No	Yes	0.0	0.0	No
No	Overweight (25.0 ≤ BMI < 30.0)	Yes	No	No	20.0	30.0	No
No	Normal weight (18.5 ≤ BMI < 25.0)	No	No	No	0.0	0.0	No
No	Normal weight (18.5 ≤ BMI < 25.0)	No	No	No	28.0	0.0	Yes

Sex	Age Category	Race	Diabetic	Physical Activity	Gen Health	Sleep Time	Asthma	Kidney Disease	Skin Cancer
Female	55 - 59	White	Yes	Yes	Very good	5.0	Yes	No	Yes
Female	80 or older	White	No	Yes	Very good	7.0	No	No	No
Male	65 - 69	White	Yes	Yes	Fair	8.0	Yes	No	No
Female	75 - 79	White	No	No	Good	6.0	No	No	Yes
Female	40 - 44	White	No	Yes	Very good	8.0	No	No	No

$$z = \frac{x - \frac{1}{N}\sum_{i=1}^{N} x_i}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{i=1}^{N} x_i\right)}} \tag{1}$$

where: z = standardized value for a given data point, x = original data point, N = total number of data points, $\frac{1}{N}\sum_{i=1}^{N} x_i$ = mean value of all data points, $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{i=1}^{N} x_i\right)}$ = standard deviation of all data points.

Class weighting: Class weighting is essential for imbalanced datasets to minimize model bias towards the majority class and poor prediction performance of the minority class. The dataset used in this study was imbalanced, where the total number of people who actually had heart disease was only 27,373 out of the total of the selected 319,795 (8.56% of the total). A good default method for determining weights is the *inverse-frequency* class weights given by Equation (2) [38]:

$$w_i = \frac{N}{K\sum_{n=1}^{N} t_{ni}} \tag{2}$$

where: N = total number of samples, K = number of classes, w_i = weight for the class i, and t_{ni} = indicator that the n^{th} sample belongs to the i^{th} class.

Note that $\sum_{n=1}^{N} t_{ni}$ is the total number of samples that belong to class i. For binary classification ($K = 2$), where the majority of the samples belong to class 0 (negative class) and minority of the samples belong to class 1 (positive class), Equation (2) can be simplified as Equation (3a).

$$\begin{aligned} w_{\text{positive}} &= \frac{N}{2 \times \text{number of positive samples}} \\ w_{\text{negative}} &= \frac{N}{2 \times \text{number of negative samples}} \end{aligned} \tag{3a}$$

The idea of class weighting is quite simple. For instance, in our case, the ratio of majority (negative) to minority (positive) class is 10.63:1; in order for the classifier to learn equally from both classes, the minority class should be accorded a weight of 10.63. Thus, in cases where the models (e.g. gradient boosting) require only the positive (minority) class weight, Equation (3b) can be used. These formulas are widely used by the machine learning community, e.g. [39] [40].

$$w_{\text{positive}} = \frac{\text{number of negative samples}}{\text{number of positive samples}} \tag{3b}$$

Data splitting: The final step involved splitting the data into training and testing sets. 80% of the data was used for training and 20% was used for testing the models.

2.3. Machine Learning Models and Training

Five (5) machine learning models were trained on the dataset: the random forest (RF), support vector machine (SVM), logistic regression, light gradient boosting (LightGBM) and extreme gradient boosting (XGBoost). These were implemented using the sklearn, lightgbm and xgboost modules in Python 3. The models were trained under two cases; Case 1 involved training the models prior to class weighting, and Case 2 involved training them after class weighting. In Case 1, the rationale was to observe how the models would perform when treating all classes equally. This would help in understanding the baseline performance and identifying any bias towards the majority class, especially in the presence of imbalanced data. In Case 2, training the models on the class-weighted dataset was crucial for addressing the imbalanced nature of the data. It was intended that by assigning higher weights to the minority class, the models would be better equipped to learn patterns in the minority class, potentially improving overall predictive performance and reducing bias towards the majority class. The comparison between Case 1 and Case 2 would allow for an evaluation of the impact of class weighting on model performance and its effectiveness in handling imbalanced datasets.

2.3.1. Random Forest (RF)

RF [41] is an ensemble algorithm that extends bootstrap aggregation (bagging) of decision trees for classification and regression problems. In bagging, multiple decision trees are created from different bootstrap samples, and predictions are averaged. Unlike typical decision trees, trees in the ensemble are unpruned, making them slightly overfit to the training data. Nevertheless, the forest prediction is the majority vote in classification. RF introduces randomness by selecting a subset of features at each split point, enhancing diversity among trees and reducing correlation in predictions. Key hyperparameters to tune include the number of randomly selected features at each split and the depth of decision trees, with a common heuristic of the square root of the total features for classification. The number of trees is increased until no further improvement is observed.

2.3.2. Support Vector Machine (SVM)

SVM is a versatile supervised learning algorithm applicable to classification and regression tasks [42]. In classification, common use cases of SVM include disease prediction, text and image classification and others. Key terms in SVM include *support vectors*, which are data points crucial for maximizing the margin or separation between classes; the *hyperplane*, a decision boundary in multidimensional space; *margin*, the distance between support vectors; and *kernels*, functions addressing non-linear data by mapping them into a higher-dimensional space. The optimization problem of finding the best hyperplane is crucial in SVM, and various kernel functions, like the RBF kernel, handle non-linear decision boundaries.

2.3.3. Logistic Regression (LR)

Logistic regression is widely used in classification problems, from predicting diseases to classifying text and images. Logistic regression computes the probability of an event occurrence, offering insights like the likelihood of a customer making a purchase or clicking on an advertisement link. It's particularly useful for binary classification tasks, providing a foundational approach for more complex models. Logistic regression encompasses three main types: binary logistic regression for two possible outcomes, multinomial logistic regression for three or more nominal categories, and ordinal logistic regression for three or more ordinal categories [43]. The algorithm uses the logistic or sigmoid function to map real-valued outputs into the range [0, 1], aiding in probability-based predictions.

2.3.4. Gradient Boosting: Extreme (XGBoost) and Light (LightGBM)

Gradient boosting utilizes an ensemble of weak learners (often decision trees) to enhance model performance in terms of efficiency, accuracy, and interpretability. Ensembles are constructed from decision tree models. Trees are added to the ensemble and fit to correct the prediction errors made by prior models. Extreme (XGBoost) and light (LightGBM) gradient boosting are two popular algorithms based on gradient-boosted decision tree (GBDT), each with its own strengths. XGBoost and LightGBM are favored in regression, classification, and ranking problems.

XGBoost: Developed by [44] in 2016 as part of the Distributed Machine Learning Community (DMLC), XGBoost is a scalable, distributed GBDT ML library that provides parallel tree boosting. While random forest uses bagging to build full decision trees in parallel from random bootstrap samples of the data set and the final prediction is the majority vote (classification), GBDTs iteratively train an ensemble of shallow decision trees, with each iteration using the error residuals of the previous model to fit the next model and the final prediction is a weighted sum of all of the tree predictions. XGBoost's boosted tree algorithms are built for model performance and computational speed.

LightGBM: Introduced by [45] in 2017, LightGBM focuses on improving training efficiency and scalability. Two key innovations, gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB), set LightGBM apart. GOSS optimizes the learning process by concentrating on instances with larger gradients, significantly reducing computational complexity. EFB, on the other hand, bundles sparse mutually exclusive features, such as one-hot encoded categorical variables, enhancing automatic feature selection. These innovations collectively accelerate training time by up to 20 times, making LightGBM a powerful GBDT implementation.

2.4. Performance Evaluation

Evaluation metrics are computed by the models. These are helpful when evaluating the performances of the trained ML models. Choosing the appropriate

evaulation can depend on factors including the nature of the problem, data distribution, domain-specific considerations and many others. *Accuracy*, *Precision*, *Recall* and *F1-score* are important evaluation matrices. [46] established the relationship between the precision and recall, which contributed to the development of appropriate formulas as provided by [47] and given in Equations (4)-(7).

In our case, we want to predict a class label, 1 or 0, corresponding to *heart disease* or *no heart disease*. This is called a deterministic classifier. To get a label prediction from our probabilistic classifiers, we need to choose a probability threshold t. The default is to predict label 1 (heart disease) if the predicted probability is larger than $t = 50$. All the following metrics implicitly use this default.

- *False negatives* (FN) and *false positives* (FP) are samples that were incorrectly classified.
- *True negatives* (TN) and *true positives* (TP) are samples that were correctly classified.
- *Accuracy* is the percentage of samples correctly classified [47]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

- *Precision* is the percentage of predicted positives that were correctly classified [47]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

- *Recall* is the percentage of actual positives that were correctly classified [47]:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

- *F*1*-Score* is a single metric that combines precision and recall into a harmonic mean. It provides a balance between precision and recall. F1-score is suitable for the imbalanced data, but it doesn't consider the entire precision-recall trade-off. It is calculated using Equation (7) [47]:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

3. Results and Discussion

The results from the models along with evaluation of their performances are presented in this section. The models were trained under two main cases; Case 1 (base case) involved training of the models without weighting the classes and Case 2 involved training of the models with the weighted classes. The results are summarised and presented accordingly.

Table 2 shows the performance of the models for Case 1 (unweighted classes) and **Table 3** shows the performance for Case 2 (weighted classes). The Accuracy represents the overall accuracy of each of the models while the Precision, Recall and F1-Score reflects the performance of each class (Class 0 and Class 1) from the models.

Table 2. Summary of performance metrics for unweighted classes (Case 1).

		CLASS 0			CLASS 1		
Algorithm	Accuracy	Precision	Recall	F1-Score	Precision	Recall	F1-Score
RF	91.33	91	100	95	75	1	2
SVM	91.26	91	100	95	0	0	0
LR	91.28	92	99	95	51	9	15
XGBoost	91.32	91	100	95	73	1	2
LightGBM	91.37	92	99	95	54	9	15

Table 3. Summary of performance metrics for weighted classes (Case 2).

		CLASS 0			CLASS 1		
Algorithm	Accuracy	Precision	Recall	F1-score	Precision	Recall	F1-score
RF	70.81	98	70	81	21	82	33
SVM	73.12	97	73	83	21	78	34
LG	74.25	97	74	84	22	76	34
XGBoost	72.51	97	72	83	21	81	34
LightGBM	74.66	97	74	84	22	78	35

While Accuracy can help in the evaluation of the model, it is not a helpful metric for dealing with the imbalanced dataset [26]. We can have over 99% accuracy by correctly predicting the majority class (Class 0) all the time. For Case 1, while the predictions for Class 0 were almost perfect, the model performed very poorly in predicting Class 1. Precision is crucial when the cost of false positives is high. In imbalanced datasets, where the minority class is of interest, precision helps ensure that the predicted positive instances are more likely to be true positives and not false positives. In our case, the cost of predicting a negative case of heart disease as positive (false positive) is less than predicting a positive case of heart disease as negative (false negative). In fact, tolerating a bit of false positives in the models can help people to seek early medical attention by providing alert of a likely positive case of heart disease. Recall is vital when it's important to avoid false negatives. In imbalanced datasets, recall ensures that the model identifies as many true positive (minority class) instances as possible. F1-Score is a balanced metric that considers both precision and recall. It is useful when achieving a trade-off between false positives and false negatives is essential. In our case, the focus was to correctly identify the positive cases of heart disease. Thus, our objective was to increase the Recall, without too much penalty to the Precision. **Figures 1-5** illustrate the results in confusion matrix for the models under the two cases (Case 1 & Case 2).

In the confusion matrix plots, it is seen that the predictions of Class 0 (majority class) in Case 1 were almost perfect (nearly 100% true negatives) while the predictions of the minority class (Class 1) were very poor (up to 99% false

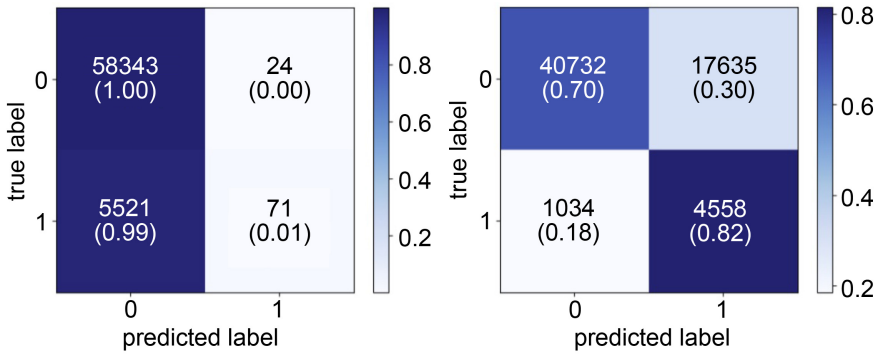

Figure 1. Confusion matrix of the RF classifier using the unweighted class (Case 1, left panel) and the weighted classes (Case 2, right panel).

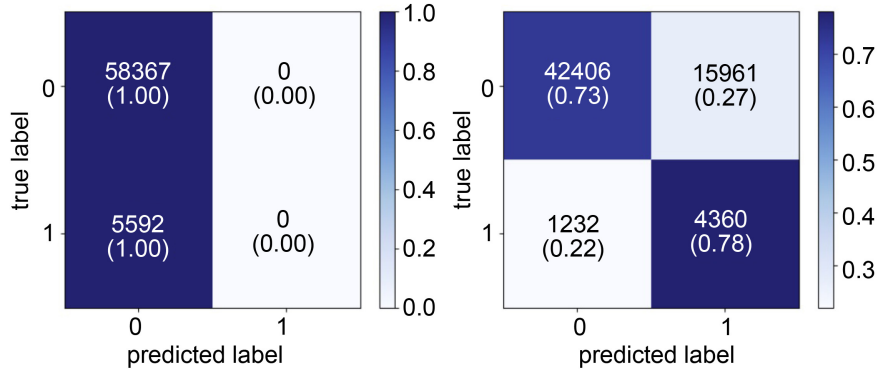

Figure 2. Confusion matrix of the SVM classifier using the unweighted class (Case 1, left panel) and the weighted classes (Case 2, right panel).

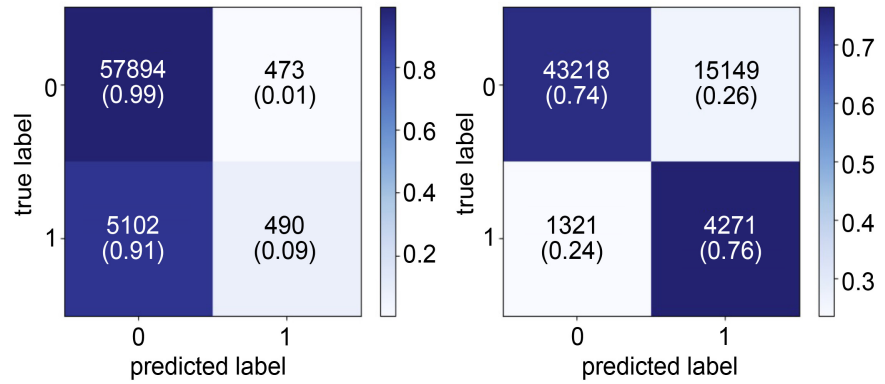

Figure 3. Confusion matrix of the logistic regression classifier using the unweighted class (Case 1, left panel) and the weighted classes (Case 2, right panel).

negatives). Class weighting improved the model performances by increasing the number of true positives significantly. For instance, class weighting in the RF classifier improved true positives from 1% to 82% (**Figure 1**) and class weighting in the XGBoost classifier improved true positives from 1% to 81% (**Figure 4**).

3.1. Model Selection

Looking at the class weighted case (Case 2), the results show no major differences

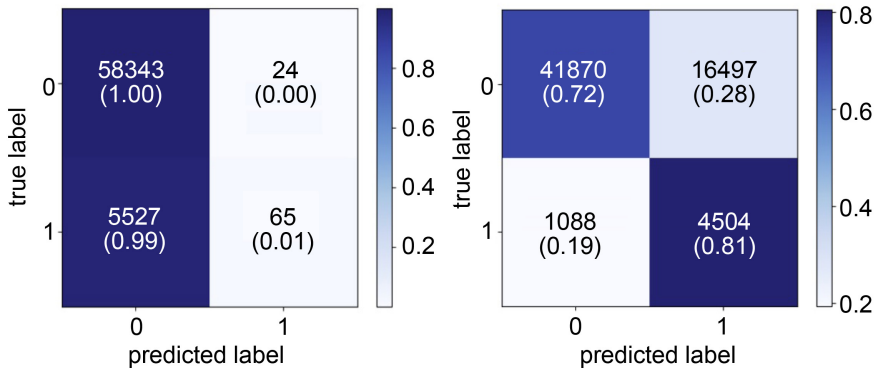

Figure 4. Confusion matrix of the XGBoost classifier using the unweighted class (Case 1, left panel) and the weighted classes (Case 2, right panel).

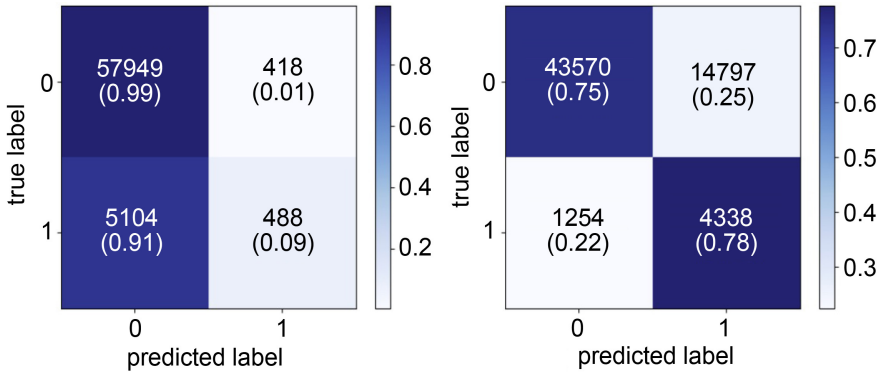

Figure 5. Confusion matrix of the LightGBM classifier using the unweighted class (Case 1, left panel) and the weighted classes (Case 2, right panel).

in the overall performance of the models in terms of the Accuracy, Precision, Recall and F1-Score in each of the two classes. We are more concerned with the correct identification of the positive cases of heart disease while minimising false positives. The ideal model would have a significantly high value of Recall along with high values in the others. The RF model produced the highest Recall of 82% followed by XGBoost at 81%; however, the XGBoost generally provides a better trade off in terms of Precision, F1-Score and Accuracy in both Class 0 and Class 1. Thus, the XGBoost model can provide the best results in predicting heart disease.

The plot in Figure 6 shows the feature importance in the selected model (trained XGBoost classifier) with the three commonly used measures (types) of feature importance: the *gain*, *cover* and *weight*. The *gain* measures the average improvement in (training set) loss brought by a feature. In other words, it tells us how much a feature helps to make accurate predictions in our training data. The DiffWalking (Difficulty Walking) feature contributed the most towards *gain*. The *cover* measures the number of data points a given feature affects by taking the average of the hessian values of all splits the feature is used in. The DiffWalking (Difficulty Walking) feature also contributed the most towards *cover*. The *weight* is the number of times a feature is used to split the data across

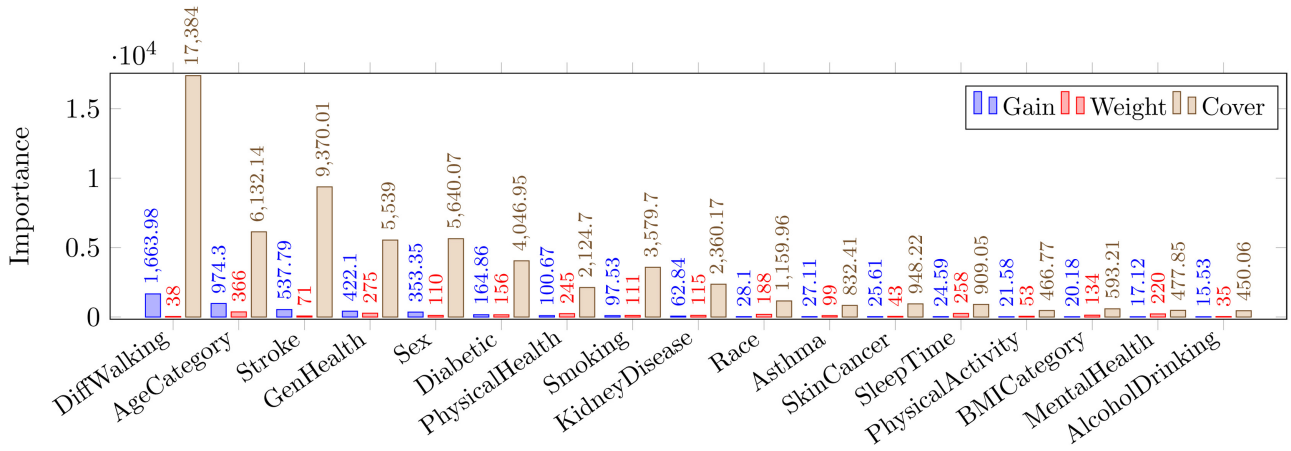

Figure 6. Feature importance in the XGBoost model in terms of the gain, weight and cover for the 17 features.

all trees. The AgeCategory feature contributed the most towards *weight*. In general, the plot shows that the XGBoost model considers difficulty in walking (DiffWalking) as a key indicator of a possible case of heart disease while others such as alcohol drinking (Alcohol Drinking) are considered the least important indicators of heart disease.

3.2. Strengths and Limitations

Some limitations and strengths can be noted in this work especially in light of the related works from [10] [20] [22] [23]. In terms of limitations, firstly, this work only focused on using the conventional ML models without any novel technique such as the feature engineering technique proposed by [22]. Secondly, nothing much was done with hyperparameter tuning in this work. Thus, model performance can be further improved through diligent hyperparameter tuning techniques such as using the *grid search* [48] method to determine optimal hyperparameters including the class weights and the probability threshold *t*. Finally, a combination of other performance evaluation matrices such as the area under the curve (AUC) may provide insights into model optimization possibilities.

In terms of strengths, firstly, this work dealt with a large dataset; thanks to the CDC [34]. The raw dataset contains 401,958 instances and 279 features and the variety in the samples is enormous (age, race, lifestyle, etc.) as compared to the datasets used by the above cited authors. Secondly, this work placed emphasis on the use of appropriate preprocessing techniques to handle imbalanced and showed that class weighting can significantly improve models performance. It was also emphasized that evaluation matrics such as Accuracy can not give clear indication of model performance. For instance, in this study, evaluation of the models under Case 1 (unweighted class) revealed high accuracy (over 90%), primarily driven by effective predictions of the majority class. However, this masked poor performance in predicting the minority class (heart disease cases). Upon introducing class weighting (Case 2), substantial improvements were ob-

served, especially in terms of true positives (Recall). This showed that better performance of identifying most true positives can be achieved at the expense of the overall accuracy of the model. Finally, as compared with most of the other cited and related work, the present study contributed an online web application which can be freely accessed by anyone. The use of this application is briefly described in the next subsection. With more rigorous study including model set-up and better or more data, efficient online tools and systems can be developed to optimize health care service provision.

3.3. Model Deployment: Online Web Application

The selected model (XGBoost classifier) was deployed online and can be accessed freely at https://drxgboost.streamlit.app/. By clicking the link, anyone can access and use the app. The use of the app is briefly described in the online user interface as indicated in Figure 7. In order to receive a prediction in terms of heart disease, the user will be required to provide all the 17 inputs (features) in a manner that best describe their health and living conditions and experiences. Once the inputs have been provided, the user can push the "predict" icon at the

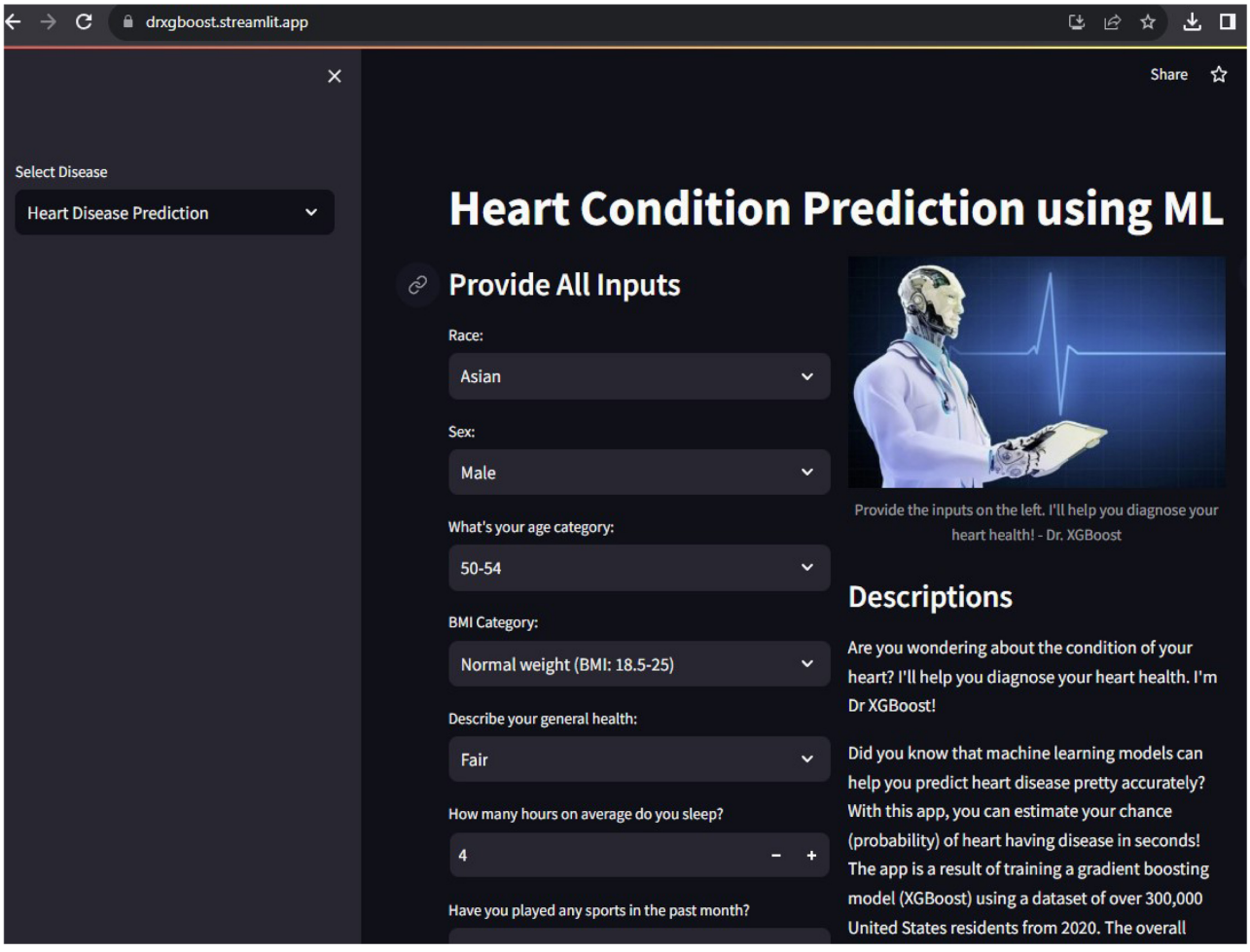

Figure 7. An image of the deployed online application based on the trained XGBoost ML model.

end and a result will be returned within seconds. The returned result will include a percentage probability (1% - 100%) of the user having a heart disease. For the sake of presentation of the results, the returned probabilities have been further classified into four groups, 0 - 25% as the "green zone", 25% - 50% as the "yellow zone", 50% - 75% as the "orange zone" and 75% - 100% as the "red zone". It has also been mentioned in the web application that results are not equivalent to a medical diagnosis as the model has less than perfect accuracy. However, the model can help in providing useful insights for early consultation of professional doctors.

4. Conclusion

Heart disease is one of the leading causes of deaths in the world. Early detection and treatment of potential heart disease can save many lives. The medical diagnosis of heart disease is a challenging process that requires accuracy and efficiency. Predictive modeling using machine learning is proving to be extremely valuable in providing insights and predicting heart disease. Out of the five ML models used in this study, the extreme gradient boosting model proved to have the best performance in predictive heart disease. Efficient data preprocessing, model setup including hyperparameter tuning techniques can lead to accurate predictions from the trained models. For instance, many problems (e.g. disease prediction, fraud detection, email spam detection, etc.) involve highly imbalanced class labels in the datasets and such datasets require diligent data preprocessing techniques including class weighting. The model trained in this study was deployed as a web application and it can be freely accessed. The results from the application can help in providing useful insights for early consultation of professional doctors. As technology continues to evolve, robust techniques in machine learning can be expected and this can boost AI-based applications in healthcare to assist health professionals and policymakers in making informed decisions to provide personalised patient care and improve healthcare services.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Chakraborty, C., Bhattacharya, M., Pal, S. and Lee, S. (2023) From Machine Learning to Deep Learning: An Advances of the Recent Data-Driven Paradigm Shift in Medicine and Healthcare. *Current Research in Biotechnology*, **7**, Article ID: 100164. https://doi.org/10.1016/j.crbiot.2023.100164

[2] Mbunge, E. and Batani, J. (2023) Application of Deep Learning and Machine Learning Models to Improve Healthcare in Sub-Saharan Africa: Emerging Opportunities, Trends, and Implications. *Telematics and Informatics Reports*, **11**, Article ID: 100097. https://doi.org/10.1016/j.teler.2023.100097

[3] Motwani, A., Shukla, P.K. and Pawar, M. (2022) Ubiquitous and Smart Healthcare

Monitoring Frameworks Based on Machine Learning: A Comprehensive Review. *Artificial Intelligence in Medicine*, **134**, Article ID: 102431. https://doi.org/10.1016/j.artmed.2022.102431

[4] Rasheed, K., Qayyum, A., Ghaly, M., *et al.* (2022) Explainable, Trustworthy, and Ethical Machine Learning for Healthcare: A Survey. *Computers in Biology and Medicine*, **149**, Article ID: 106043. https://doi.org/10.1016/j.compbiomed.2022.106043

[5] Liao, W., He, J., Luo, X., Wu, M., Shen, Y., Li, C. and Chen, N. (2022) Automatic Delineation of Gross Tumor Volume Based on Magnetic Resonance Imaging by Performing a Novel Semisupervised Learning Framework in Nasopharyngeal Carcinoma. *International Journal of Radiation Oncology Biology Physics*, **113**, 893-902. https://doi.org/10.1016/j.ijrobp.2022.03.031

[6] Pierre, K., Haneberg, A.G., Kwak, S., Peters, K.R., Hochhegger, B., Sananmuang, T., Tunlayadechanont, P., Tighe, P.J., Mancuso, A. and Forghani, R. (2023) Applications of Artificial Intelligence in the Radiology Roundtrip: Process Streamlining, Workflow Optimization, and Beyond. *Seminars in Roentgenology*, **58**, 158-169. https://doi.org/10.1053/j.ro.2023.02.003

[7] Zhai, K., Yousef, M.S., Mohammed, S., Al-Dewik, N.I. and Qoronfleh, M.W. (2023) Optimizing Clinical Workflow Using Precision Medicine and Advanced Data Analytics. *Processes*, **11**, Article No. 939. https://doi.org/10.3390/pr11030939

[8] Javaid, M., Haleem, A., Singh, R.P., Suman, R. and Rab, S. (2022) Significance of Machine Learning in Healthcare: Features, Pillars and Applications. *International Journal of Intelligent Networks*, **3**, 58-73. https://doi.org/10.1016/j.ijin.2022.05.002

[9] Behera, M.P., Sarangi, A., Mishra, D. and Sarangi, S.K. (2023) A Hybrid Machine Learning Algorithm for Heart and Liver Disease Prediction Using Modified Particle Swarm Optimization with Support Vector Machine. *Procedia Computer Science*, **218**, 818-827. https://doi.org/10.1016/j.procs.2023.01.062

[10] Abdalrada, A.S., Abawajy, J. and Al-Quraishi, T. (2022) Machine Learning Models for Prediction of Co-Occurrence of Diabetes and Cardiovascular Diseases: A Retrospective Cohort Study. *Journal of Diabetes & Metabolic Disorders*, **21**, 251-261. https://doi.org/10.1007/s40200-021-00968-z

[11] Chari, S., *et al.* (2023) Informing Clinical Assessment by Contextualizing Post-Hoc Explanations of Risk Prediction Models in Type-2 Diabetes. *Artificial Intelligence in Medicine*, **137**, Article ID: 102498. https://doi.org/10.1016/j.artmed.2023.102498

[12] Dworzynski, P., Aasbrenn, M., Rostgaard, K., Melbye, M., Gerds, T.A., Hjalgrim, H. and Pers, T.H. (2020) Nationwide Prediction of Type 2 Diabetes Comorbidities. *Scientific Reports*, **10**, Article No. 1776. https://doi.org/10.1038/s41598-020-58601-7

[13] Ojeme, B. and Mbogho, A. (2016) Selecting Learning Algorithms for Simultaneous Identification of Depression and Comorbid Disorders. *Procedia Computer Science*, **96**, 1294-1303. https://doi.org/10.1016/j.procs.2016.08.174

[14] Tennenhouse, L.G., Marrie, R.A., Bernstein, C.N., Lix, L.M. and CIHR Team in Defining the Burden and Managing the Effects of Psychiatric Comorbidity in Chronic Immunoinflammatory Disease (2020) Machine-Learning Models for Depression and Anxiety in Individuals with Immune-Mediated Inflammatory Disease. *Journal of Psychosomatic Research*, **134**, Article ID: 110126. https://doi.org/10.1016/j.jpsychores.2020.110126

[15] Wang, X., Eichhorn, J., Haq, I. and Baghal, A. (2021) Resting-State Brain Metabolic Fingerprinting Clusters (Biomarkers) and Predictive Models for Major Depression in Multiple Myeloma Patients. *PLOS ONE*, **16**, e0251026.

https://doi.org/10.1371/journal.pone.0251026

[16] Farran, B., Channanath, A.M., Behbehani, K. and Thanaraj, T.A. (2013) Predictive Models to Assess Risk of Type 2 Diabetes, Hypertension and Comorbidity: Machine-Learning Algorithms and Validation Using National Health Data from Kuwait—A Cohort Study. *BMJ Open*, **3**, e002457. https://doi.org/10.1136/bmjopen-2012-002457

[17] Nikolaou, V., *et al.* (2021) The Cardiovascular Phenotype of Chronic Obstructive Pulmonary Disease (COPD): Applying Machine Learning to the Prediction of Cardiovascular Comorbidities. *Respiratory Medicine*, **186**, Article ID: 106528. https://doi.org/10.1016/j.rmed.2021.106528

[18] Glauser, T., *et al.* (2020) Identifying Epilepsy Psychiatric Comorbidities with Machine Learning. *Acta Neurologica Scandinavica*, **141**, 388-396. https://doi.org/10.1111/ane.13216

[19] Linden, T., De Jong, J., Lu, C., Kiri, V., Haeffs, K. and Fröhlich, H. (2021) An Explainable Multimodal Neural Network Architecture for Predicting Epilepsy Comorbidities Based on Administrative Claims Data. *Frontiers in Artificial Intelligence*, **4**, Article ID: 610197. https://doi.org/10.3389/frai.2021.610197

[20] Asih, P.S., Azhar, Y., Wicaksono, G.W. and Akbi, D.R. (2023) Interpretable Machine Learning Model for Heart Disease Prediction. *Procedia Computer Science*, **227**, 439-445. https://doi.org/10.1016/j.procs.2023.10.544

[21] Nashif, S., Raihan, Md.R., Islam, Md.R. and Imam, M.H. (2018) Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System. *World Journal of Engineering and Technology*, **6**, 854-873. https://doi.org/10.4236/wjet.2018.64057

[22] Uddin, S., Wang, S., Lu, H., Khan, A., Hajati, F. and Khushi, M. (2022) Comorbidity and Multimorbidity Prediction of Major Chronic Diseases Using Machine Learning and Network Analytics. *Expert Systems with Applications*, **205**, Article ID: 117761. https://doi.org/10.1016/j.eswa.2022.117761

[23] Yang, P., Qiu, H., Wang, L. and Zhou, L. (2022) Early Prediction of High-Cost Inpatients with Ischemic Heart Disease Using Network Analytics and Machine Learning. *Expert Systems with Applications*, **210**, Article ID: 118541. https://doi.org/10.1016/j.eswa.2022.118541

[24] Australian Government Department of Health (2020) Chronic Conditions in Australia. https://www.health.gov.au/topics/chronic-conditions/chronic-conditions-in-australia

[25] Janosi, A., Steinbrunn, W., Pfisterer, M. and Detrano, R. (1988) Heart Disease. UCI Machine Learning Repository.

[26] Mortaz, E. (2020) Imbalance Accuracy Metric for Model Selection in Multi-Class Imbalance Classification Problems. *Knowledge-Based Systems*, **210**, Article ID: 106490. https://doi.org/10.1016/j.knosys.2020.106490

[27] Bangdiwala, S.I., Fonn, S., Okoye, O., *et al.* (2010) Workforce Resources for Health in Developing Countries. *Public Health Reviews*, **32**, 296-318. https://doi.org/10.1007/BF03391604

[28] Lamuri, A., *et al.* (2023) Burnout Dimension Profiles among Healthcare Workers in Indonesia. *Heliyon*, **9**, e14519. https://doi.org/10.1016/j.heliyon.2023.e14519

[29] Moyo, E., *et al.* (2023) Burnout among Healthcare Workers during Public Health Emergencies in Sub-Saharan Africa: Contributing Factors, Effects, and Prevention Measures. *Human Factors in Healthcare*, **3**, Article ID: 100039. https://doi.org/10.1016/j.hfh.2023.100039

[30] Asante, A. and Hall, J. (2011) A Review of Health Leadership and Management Capacity in Papua New Guinea. Human Resources for Health Knowledge Hub, University of New South Wales, Sydney. https://sph.med.unsw.edu.au/sites/default/files/sphcm/Centres_and_Units/LM_PNG_Report.pdf

[31] Mitchell, M., Thomason, J., Donaldson, D. and Garner, P. (1991) The Cost of Rural Health Services in Papua New Guinea. *Papua and New Guinea Medical Journal*, **34**, 276-284.

[32] Wiltshire, C., Watson, A.H.A., Lokinap, D. and Currie, T. (2020) Papua New Guinea's Primary Health Care System: Views from the Front Line. ANU and UPNG.

[33] World Bank Group (2017) Health Financing System Assessment Papua New Guinea. World Bank Publications, Washington DC. https://documents1.worldbank.org/curated/en/906971515655591305/pdf/122589-wp-p154901-public-23994-png-health-financing-system-assessment-web.pdf

[34] Centers for Disease Control and Prevention (CDC) (2020) Behavioral Risk Factor Surveillance System. Data Collected through the Behavioral Risk Factor Surveillance System. https://www.cdc.gov/brfss/annual_data/annual_2020.html

[35] Kreyszig, E. (1979) Advanced Engineering Mathematics. 4th Edition, Wiley, Hoboken, 880.

[36] Weisstein, E.W. (n.d.) Arithmetic Mean. From MathWorld—A Wolfram Web Resource. https://mathworld.wolfram.com/ArithmeticMean.html

[37] Weisstein, E.W. (n.d.) Standard Deviation. From MathWorld—A Wolfram Web Resource. https://mathworld.wolfram.com/StandardDeviation.html

[38] MathWorks. (n.d.) Sequence Classification Using Inverse Frequency Class Weights. https://www.mathworks.com/help/deeplearning/ug/sequence-classification-using-inverse-frequency-class-weights.html

[39] Stack Overflow Community (2019) How to Calculate Unbalanced Weights for BCEWithLogitsLoss in Pytorch. Stack Overflow. https://stackoverflow.com/questions/57021620/how-to-calculate-unbalanced-weights-for-bcewithlogitsloss-in-pytorch

[40] Tantai, H. (2023, February) Use Weighted Loss Function to Solve Imbalanced Data Classification Problems. Medium. https://medium.com/@zergtant/use-weighted-loss-function-to-solve-imbalanced-data-classification-problems-749237f38b75

[41] Liu, Y., Wang, Y. and Zhang, J. (2012) New Machine Learning Algorithm: Random Forest. In: Liu, B., Ma, M. and Chang, J., Eds., *Information Computing and Applications*, Lecture Notes in Computer Science, Vol. 7473, Springer, Berlin, 246-252. https://doi.org/10.1007/978-3-642-34062-8_32

[42] Kecman, V. (2005) Support Vector Machines—An Introduction. In: Wang, L., Ed., *Support Vector Machines: Theory and Applications*, Studies in Fuzziness and Soft Computing, Vol. 177, Springer, Berlin, 1-47. https://doi.org/10.1007/10984697_1

[43] Starbuck, C. (2023) Logistic Regression. In: Starbuck, C., Ed., *The Fundamentals of People Analytics*, Springer, Cham, 223-238. https://doi.org/10.1007/978-3-031-28674-2_12

[44] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[45] Ke, G., *et al*. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Proceedings of the* 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017, 314s9-3157. https://dl.acm.org/doi/10.5555/3294996.3295074

[46] Buckland, M. and Gey, F. (1994) The Relationship between Recall and Precision. *Journal of the American Society for Information Science*, **45**, 12-19. https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L

[47] Yu, L. and Zhou, N. (2021) Survey of Imbalanced Data Methodologies.

[48] Ogunsanya, M., Isichei, J. and Desai, S. (2023) Grid Search Hyperparameter Tuning in Additive Manufacturing Processes. *Manufacturing Letters*, **35**, 1031-1042. https://doi.org/10.1016/j.mfglet.2023.08.056