Heart Disease Prediction Using Machine Learning

Sreejit Ramakrishnan
Assistant Professor
https://orcid.org/0009-0004-6154-2724

ABSTRACT

Machine Learning (ML), which is one of the most prominent applications of Artificial Intelligence, is doing wonders in the research field of study. In this paper machine learning is used in detecting if a person has a heart disease or not. A lot of people suffer from cardiovascular diseases (CVDs), which even cost people their lives all around the world. Machine learning can be used to detect whether a person is suffering from a cardiovascular disease by considering certain attributes like chest pain, cholesterol level, age of the person and some other attributes. Classification algorithms based on supervised learning which is a type of machine learning can make diagnoses of cardiovascular diseases easy.

The diagnosis and prognosis of cardiovascular disease are crucial medical tasks to ensure correct classification, which helps cardiologists provide proper treatment to the patient. Machine learning applications in the medical niche have increased as they can recognize patterns from data. Using machine learning to classify cardiovascular disease occurrence can help diagnosticians reduce misdiagnosis. This research develops a model that can correctly predict cardiovascular diseases to reduce the fatality caused by cardiovascular diseases. This paper proposes a method of k-modes clustering with Huang starting that can improve classification accuracy. Models such as random forest (RF), decision tree classifier (DT), multilayer perceptron (MP), and XGBoost (XGB) are used.

Keywords—heart disease; machine learning; k-modes; classification; multilayer perceptron; model evaluation

Introduction:

Globally, cardiovascular disease (CVDs) is the primary cause of morbidity and mortality, accounting for more than 70% of all fatalities. According to the 2017 Global Burden of Disease research, cardiovascular disease is responsible for about 43% of all fatalities . Common risk factors for heart disease in high-income nations include lousy diet, cigarette use, excessive sugar consumption, and obesity or excess body fat . However, low- and middle-income nations also see a rise in chronic illness prevalence. Between 2010 and 2015, the global economic burden of cardiovascular diseases was expected to reach roughly USD 3.7 trillion (Mozaffarian et al., 2015; Maiga et al., 2019).

In addition, technologies such as electrocardiograms and CT scans, critical for diagnosing coronary heart disease, are sometimes too costly and impractical for consumers. The reason mentioned above alone has resulted in the deaths of 17 million people. Twenty-five to thirty percent of firms' annual medical expenses were attributable to employees with cardiovascular disease. Therefore, early detection of heart disease is essential to lessen its physical and monetary cost to people and institutions. According to the WHO estimate, the overall number of deaths from CVDs would rise to 23.6 million by 2030, with heart disease and stroke being the leading causes. To save lives and decrease the cost burden on society, it is vital to apply data mining and machine learning methods to anticipate the chance of having heart disease.

Heart disease, specifically cardiovascular disease (CVDs), is a leading cause of morbidity and mortality worldwide, accounting for over 70% of all global deaths. According to the Global Burden of Disease Study 2017, CVD accounts for more than 43% of all deaths. Common risk factors associated with heart disease include unhealthy food, tobacco, excessive sugar, and overweight or extra body fat, often found in high-income countries. However, low- and middle-income countries are also seeing an increase in the prevalence of chronic diseases. The economic burden of CVDs worldwide has been estimated to be approximately USD 3.7 trillion between 2010 and 2015.

Furthermore, devices such as electrocardiograms and CT scans, essential for detecting coronary heart disease, are often too expensive and infeasible for many low- and middle-income countries. Therefore, early determination of heart disease is crucial to decrease its physical and financial burden on individuals and organizations. According to a WHO report, by 2030, the total number of deaths due to CVDs will increase to 23.6 million, mainly from heart disease and stroke. Therefore, it is crucial to use data mining and machine learning techniques to predict the likelihood of developing heart disease in order to save lives and reduce the economic burden on society.

In the medical field, a vast amount of data is generated daily using data mining techniques, and we can find hidden patterns that can be used for clinical diagnosis. Therefore, data mining plays a vital role in the medical field, which can be proved by the work conducted in the past few decades. Many factors, such as diabetes, high blood pressure, high cholesterol, and abnormal pulse rate, need to be considered when predicting heart disease. Often, the medical data available need to be completed, affecting the results in predicting heart disease.

Machine learning plays a crucial role in the medical field. Using machine learning, we can diagnose, detect, and predict various diseases. Recently, there has been a growing interest in using data mining and machine learning techniques to predict the likelihood of developing certain diseases. The already-existing work contains applications of data mining techniques for predicting the disease. Although some studies have attempted to predict the future risk of the progression of the disease, they have yet to find accurate results. The main goal of this paper is to accurately predict the possibility of heart disease in the human body.

In this research, we aim to investigate the effectiveness of various machine learning algorithms in predicting heart disease. To achieve this goal, we employed a variety of techniques, including random forest, decision tree classifier, multilayer perceptron, and XGBoost, to build predictive models. In order to improve the convergence of the models, we applied k-modes clustering to preprocess the dataset and scale it. The dataset used in this study is publicly available on Kaggle. All the computation, preprocessing, and visualization were conducted on Google Colab using Python. Previous studies have reported accuracy rates of up to 94% using machine learning techniques for heart disease prediction. However, these studies have often used small sample sizes, and the results may not be generalizable to larger populations. Our study aims to address this limitation by using a larger and more diverse dataset, which is expected to increase the generalizability of the results.

2. Literature Survey

In recent years, the healthcare industry has seen a significant advancement in the field of data mining and machine learning. These techniques have been widely adopted and have demonstrated efficacy in various healthcare applications, particularly in the field of medical cardiology. The rapid accumulation of medical data has presented researchers with an unprecedented opportunity to develop and test new algorithms in this field. Heart disease remains a leading cause of mortality in developing nations, and identifying risk factors and early signs of the disease has become an important area of research. The utilization of data mining and machine learning techniques in this field can potentially aid in the early detection and prevention of heart disease.

The purpose of the study described by Narain et al. (2016) is to create an innovative machine-learning-based cardiovascular disease (CVD) prediction system in order to increase the precision of the widely used Framingham risk score (FRS). With the help of data from 689 individuals who had symptoms of CVD and a validation dataset from the Framingham research, the proposed system—which uses a quantum neural network to learn and recognize patterns of CVD—was experimentally validated and compared with the FRS. The suggested system's accuracy in forecasting CVD risk was determined to be 98.57%, which is much greater than the FRS's accuracy of 19.22% and other existing techniques. According to the study's findings, the suggested approach could be a useful tool for doctors in forecasting CVD risk, assisting in the creation of better treatment plans, and facilitating early diagnosis.

In a study conducted by Shah et al. (2020), the authors aimed to develop a model for predicting cardiovascular disease using machine learning techniques. The data used for this purpose were obtained from the Cleveland heart disease dataset, which consisted of 303 instances and 17 attributes, and were sourced from the UCI machine learning repository. The authors employed a variety of supervised classification methods, including naïve Bayes, decision tree, random forest, and k-nearest neighbor (KKN). The results of the study indicated that the KKN model exhibited the highest level of accuracy, at 90.8%. The study highlights the potential utility of machine learning techniques in predicting cardiovascular disease, and emphasizes the importance of selecting appropriate models and techniques to achieve optimal results.

In a study by Drod et al. (2022), the objective was to use machine learning (ML) techniques to identify the most significant risk variables for cardiovascular disease (CVD) in patients with metabolic-associated fatty liver disease (MAFLD). Blood biochemical analysis and subclinical atherosclerosis assessment were performed on 191 MAFLD patients. A model to identify those with the highest risk of CVD was built using ML approaches, such as multiple logistic regression classifier, univariate feature ranking, and principal component analysis (PCA). According to the study, hypercholesterolemia, plaque scores, and duration of diabetes were the most crucial clinical characteristics. The ML technique performed well, correctly identifying 40/47 (85.11%) high-risk patients and 114/144 (79.17%) low-risk patients with an AUC of 0.87. According to the study's findings, an ML method is useful for detecting MAFLD patients with widespread CVD based on simple patient criteria.

In a study published by Alotalibi (2019) , the author aimed to investigate the utility of machine learning (ML) techniques for predicting heart failure disease. The study utilized a dataset from the Cleveland Clinic Foundation, and implemented various ML algorithms, such as decision tree, logistic regression, random forest, naïve Bayes, and support vector machine (SVM), to develop prediction models. A 10-fold cross-validation approach was employed during the model development process. The results naïvendicated 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 provides insight into the potential of ML techniques as an effective tool for predicting heart failure disease and highlights the decision tree algorithm as a potential option for future research.

Through a comparison of multiple algorithms, Hasan and Bao (2020) carried out a study with the main objective of identifying the most efficient feature selection approach for anticipating cardiovascular illness. The three well-known feature selection methods (filter, wrapper, and embedding) were first taken into account, and then a feature subset was recovered from these three algorithms using a Boolean process-based common "True" condition. This technique involved retrieving feature subsets in two stages. A number of models, including random forest, support vector classifier, k-nearest neighbors, naïve Bayes, and XGBoost, were taken into account in order to justify the comparative accuracy and identify the best predictive analytics. As a standard for comparison with all features, the artificial neural network (ANN) was used. The findings demonstrated that the most accurate prediction results for cardiovascular illness were provided by the XGBoost classifier coupled with the wrapper technique. XGBoost delivered an accuracy of 73.74%, followed by SVC with 73.18% and ANN with 73.20%.

The primary drawback of the prior research is its limited dataset, resulting in a high risk of overfitting. The models developed may not be appropriate for large datasets. In contrast, we utilized a cardiovascular disease dataset consisting of 70,000 patients and 11 features, thereby reducing the chance of overfitting. The table presents a concise review of cardiovascular disease prediction studies performed on large datasets, further reinforcing the effectiveness of using a substantial dataset.

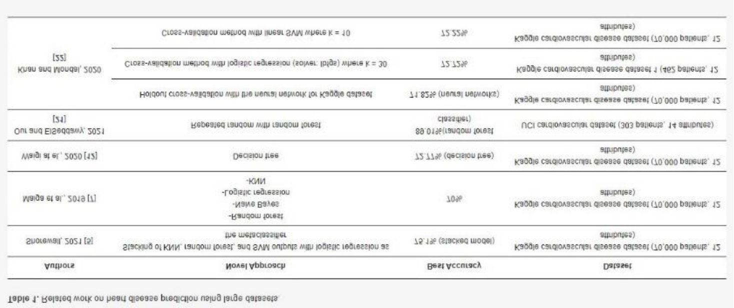


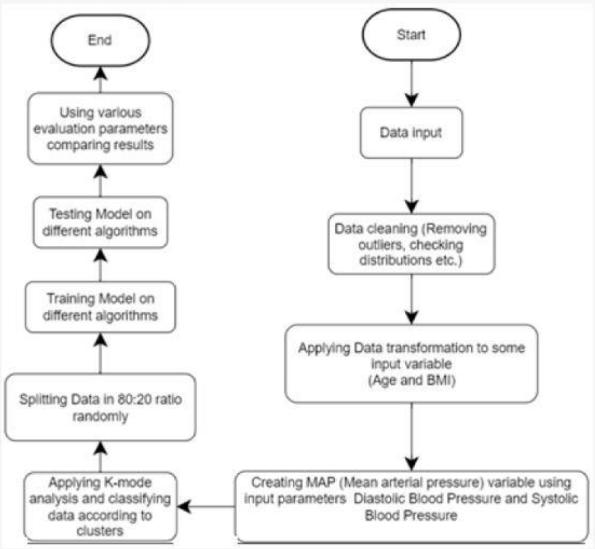
Table 1. Related work on heart disease prediction using large datasets.

Authors	Novel Approach	Best Accuracy	Dataset
Shorewall, 2021 [5]	Stacking of KNN, random forest, and SVM outputs with logistic regression as the metaclassifier	75.1% (stacked model)	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)
Maiga et al., 2019 [7]	-Random forest -Naive Bayes -Logistic regression -KNN	70%	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)
Waigi et al., 2020 [12]	Decision tree	72.77% (decision tree)	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)
Our and Elseddawy, 2021 [21]	Repeated random with random forest	89.01%(random forest classifier)	UCI cardiovascular dataset (303 patients, 14 attributes)
Khan and Mondal, 2020 [22]	Holdout cross-validation with the neural network for Kaggle dataset	71.82% (neural networks)	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)
	Cross-validation method with logistic regression (solver: lbfgs) where k = 30	72.72%	Kaggle cardiovascular disease dataset 1 (462 patients, 12 attributes)
	Cross-validation method with linear SVM where k = 10	72.22%	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)

3. **Methodology**

This study aims to predict the probability of heart disease through computerized heart disease prediction, which can be beneficial for medical professionals and patients. To achieve this objective, we employed various machine learning algorithms on a dataset and present the results in this study report. To enhance the methodology, we plan to clean the data, eliminate irrelevant information, and incorporate additional features such as MAP and BMI. Next, we will separate the dataset based on gender and implement k-modes clustering. Finally, we will train the model with the processed data.

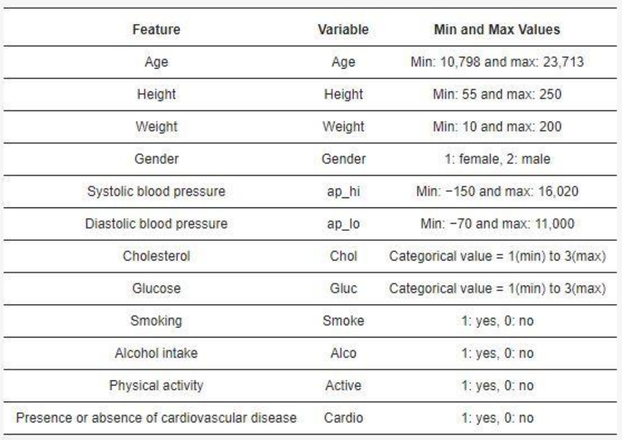
Figure 1. Flow diagram of Model.


Start
Data input
Data cleaning (Removing outliers, checking distributions etc.)
Applying Data transformation to some input variable (Age and BMI)
Creating MAP (Mean arterial pressure) variable using input parameters Diastolic Blood Pressure and Systolic Blood Pressure
Applying K-mode analysis and classifying data according to clusters
Splitting Data in 80:20 ratio randomly
Training Model on different algorithms
Testing Model on different algorithms
Using various evaluation parameters comparing results
End

3.1. Data Source

The dataset utilized in this study, as described, comprises 70,000 patient records with 12 distinct features, as listed in Table 2. These features include age, gender, systolic blood pressure, and diastolic blood pressure. The target class, “cardio,” indicates whether a patient has cardiovascular disease (represented as 1) or is healthy (represented as 0).

Table 2. Datasets attributes.



Feature	Variable	Min and Max Values
Age	Age	Min: 10,798 and max: 23,713
Height	Height	Min: 55 and max: 250
Weight	Weight	Min: 10 and max: 200
Gender	Gender	1: female, 2: male
Systolic blood pressure	ap_hi	Min: −150 and max: 16,020
Diastolic blood pressure	ap_lo	Min: −70 and max: 11,000
Cholesterol	Chol	Categorical value = 1(min) to 3(max)
Glucose	Gluc	Categorical value = 1(min) to 3(max)
Smoking	Smoke	1: yes, 0: no
Alcohol intake	Alco	1: yes, 0: no
Physical activity	Active	1: yes, 0: no
Presence or absence of cardiovascular disease	Cardio	1: yes, 0: no

3.2. Removing Outliers:

The presence of outliers in the dataset is evident. These outliers may have been the result of errors in data entry. The removal of these outliers has the potential to improve the performance of our predictive model. In order to address this issue, we removed all instances of ap_hi, ap_lo, weight, and height that fell outside of the range of 2.5% to 97.5%. This process of identifying and eliminating outliers was performed manually. As a result of this data cleaning process, the number of rows was reduced from 70,000 to 57,155.

3.3. Feature Selection and Reduction

We propose the use of binning as a method for converting continuous input, such as age, into categorical input in order to improve the performance and interpretability of classification algorithms. By categorizing continuous input into distinct groups or bins, the algorithm is able to make distinctions between different classes of data based on specific values of the input variables. For instance, if the input variable is “Age Group” and the possible values are “Young”, “Middle-aged”, and “Elderly”, a classification algorithm can use this information to separate the data into different classes or categories based on the age group of the individuals in the dataset

3.4. Clustering

Clustering is a machine learning technique where a group of instances is grouped based on similarity measures. One common algorithm used for clustering is the k-means algorithm, but it is not effective when working with categorical data. To overcome this limitation, the k-modes algorithm was developed. The k-modes algorithm, introduced by Huang in 1997, is similar to the k-means algorithm but utilizes dissimilarity measures for categorical data and replaces the means of the clusters with modes. This allows the algorithm to work effectively with categorical data.

3.5. Correlation Table

Further, a correlation table is prepared to determine the correlation between different categories. Mean arterial pressure (MAP_Class), cholesterol, and age were highly correlated factors. Intra-feature dependency can also be looked upon with the help of this matrix.

3.6. Modeling

A training dataset (80%) and a testing dataset (20%) are created from the dataset. A model is trained using the training dataset, and its performance is assessed using the testing dataset. Different classifiers, such as decision tree classifier, random forest classifier, multilayer perceptron, and XGBoost, are applied to the clustered dataset to assess their performance. The performance of each classifier is then evaluated using accuracy, precision, recall, and F-measure scores.

3.6.1. Decision Tree Classifier:

Decision trees are treelike structures that are used to manage large datasets. They are often depicted as flowcharts, with outer branches representing the results and inner nodes representing the properties of the dataset. Decision trees are popular because they are efficient, reliable, and easy to understand.

3.6.2. Random Forest:

The random forest [13] algorithm belongs to a category of supervised classification technique that consists of multiple decision trees working together as a group. The class with the most votes become the prediction made by our model.

3.6.3. Multilayer Perceptron:

The multilayer perceptron (MLP) is a type of artificial neural network that consists of multiple layers. Single perceptron can only solve linear problems, but MLP is better suited for nonlinear examples. MLP is used to tackle complex issues. A feedforward neural network with many layers is an example of an MLP

3.6.4. XGBoost

XGBoost is a version of gradient boosted decision trees. This algorithm involves creating decision trees in a sequential manner. All the independent variables are allocated weights, which are subsequently used to produce predictions by the decision tree. If the tree makes a wrong prediction, the importance of the relevant variables is increased and used in the next decision tree. The output of each of these classifiers/predictors is then merged to produce a more robust and accurate model. In a study by, the XGBoost model achieved 73% accuracy with the parameters ‘learning_rate’: 0.1, ‘max_depth’: 4, ‘n_estimators’: 100, ‘cross-validation’: 10 folds including 49,000 training and 21,000 testing data instances on 70,000 CVD dataset.

4. Results:

The dataset consisted of 70,000 rows and 12 attributes, but after cleaning and preprocessing, it was reduced to approximately 59,000 rows and 11 attributes. Since all attributes were categorical, outliers were removed to improve the model efficiency. The algorithms used in this study were random forest, decision tree, multilayer perception, and XGBoost classifier. This study used several measures of performance, namely, precision, recall, accuracy, F1 score, and area under the ROC curve. The dataset was split into two parts: 80% of the data used to train model and 20% used to test the model. We employed an automated approach for hyperparameter tuning by utilizing the GridSearchCV method. GridSearchCV takes in an estimator, a set of hyperparameters to be searched over, and a scoring method, and returns the best set of hyperparameters that maximizes the scoring method. This method is implemented in the scikit-learn library, and it uses k-fold cross-validation to evaluate the performance of different sets of hyperparameters.

Various machine learning classifiers, such as MLP, RF, decision tree, and XGBoost, were applied on the cardiovascular disease dataset to identify the presence of cardiovascular disease after hyperparameter tuning. The results indicate that the multilayer perceptron (MLP) algorithm obtained the highest cross-validation accuracy of 87.28%, along with high recall, precision, F1 score, and AUC scores of 84.85, 88.70, 86.71, and 0.95, respectively. All classifiers had an accuracy above 86.5%. The random forest algorithm’s accuracy was increased by 0.5% from 86.48% to 86.90% through hyperparameter tuning with GridSearchCV. Similarly, the accuracy of the XGBoost algorithm increased by 0.6% from 86.4% to 87.02% through hyperparameter tuning.

5. Conclusions:

The primary objective of this study was to classify heart disease using different models and a real-world dataset. The k-modes clustering algorithm was applied to a dataset of patients with heart disease to predict the presence of the disease. The dataset was preprocessed by converting the age attribute to years and dividing it into bins of 5-year intervals, as well as dividing the diastolic and systolic blood pressure data into bins of 10 intervals. The dataset was also split on the basis of gender to take into account the unique characteristics and progression of heart disease in men and women.

References:

1. Estes, C.; Anstee, Q.M.; Arias-Loste, M.T.; Bantel, H.; Bellentani, S.; Caballeria, J.; Colombo, M.; Craxi, A.; Crespo, J.; Day, C.P.; et al. Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016–2030. *J. Hepatol.* **2018**, *69*, 896–904.

2. Drożdż, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. *Cardiovasc. Diabetol.* **2022**, *21*, 240.

3. Murthy, H.S.N.; Meenakshi, M. Dimensionality reduction using neuro-genetic approach for early prediction of coronary heart disease. In Proceedings of the International Conference on Circuits, Communication, Control and Computing, Bangalore, India, 21–22 November 2014; pp. 329–332.

4. Benjamin, E.J.; Muntner, P.; Alonso, A.; Bittencourt, M.S.; Callaway, C.W.; Carson, A.P.; Chamberlain, A.M.; Chang, A.R.; Cheng, S.; Das, S.R.; et al. Heart disease and stroke statistics—2019 update: A report from the American heart association. *Circulation* **2019**, *139*, e56–e528.

5. Shorewala, V. Early detection of coronary heart disease using ensemble techniques. *Inform. Med. Unlocked* **2021**, *26*, 100655.

6. Mozaffarian, D.; Benjamin, E.J.; Go, A.S.; Arnett, D.K.; Blaha, M.J.; Cushman, M.; de Ferranti, S.; Després, J.-P.; Fullerton, H.J.; Howard, V.J.; et al. Heart disease and stroke statistics—2015 update: A report from the American Heart Association. *Circulation* **2015**, *131*, e29–e322.

7. Maiga, J.; Hungilo, G.G.; Pranowo. Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data. In Proceedings of the 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, 24–25 October 2019; pp. 45–48.

8. Li, J.; Loerbroks, A.; Bosma, H.; Angerer, P. Work stress and cardiovascular disease: A life course perspective. *J. Occup. Health* **2016**, *58*, 216–219.

9. Purushottam; Saxena, K.; Sharma, R. Efficient Heart Disease Prediction System. *Procedia Comput. Sci.* **2016**, *85*, 962–969.

10. Soni, J.; Ansari, U.; Sharma, D.; Soni, S. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *Int. J. Comput. Appl.* **2011**, *17*, 43–48.

11. Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access* **2019**, *7*, 81542–81554.

12. 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**, *7*, 1638–1645.

13. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.

14. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.

15. Gietzelt, M.; Wolf, K.-H.; Marschollek, M.; Haux, R. Performance comparison of accelerometer calibration algorithms based on 3D-ellipsoid fitting methods. *Comput. Methods Programs Biomed.* **2013**, *111*, 62–71.