ZAMBIA INFORMATION COMMUNICATION TECHNOLOGY (ICT) JOURNAL
Volume 6 (Issue 1) (2022) Pages 38-43

Assisted Artificial Intelligence Medical Diagnosis System for Heart Disease

Mweemba Maambo[1], Jackson Phiri[2], Monica. M. Kalumbilo[3] and Leena Jaganathan[4]
Computer Science Department
University of Zambia (UNZA)
Lusaka, Zambia
maambomweemba4@gmail.com[1], Jackson.phiri@cs.unza.zm[2], mkabemba@gmail.com[3], and leena.kumar@unza.zm[4]

***Abstract*— In recent years increase of new and effective medical field applications has critical part in research. Artificial Intelligence (AI) Systems has great influence in the growth of these effective medical field applications and tools. One of the major health problems in both established and developing countries is heart disease. Therefore, diagnosis to regulate the heart disease is very vital, so that appropriate actions can be taken. The Artificial Intelligence System uses input medical data collected from an existing dataset from Kaggle and applies this data on the artificial intelligence application developed that uses data mining algorithm and a basic model on Zambian patients to see if the model will predict correctly. From the dataset collected 80% was used as training data and 20% was used as testing data. The Bayesian data mining algorithm was used for predicting the risk level and probability of heart disease. The system uses medical parameters to predict heart disease in patients and these parameters are age, sex, blood pressure, blood sugar (mg/dl), cholesterol (mm/dl), heart rate, exercise-induced angina, resting electrocardiogram, oldpeak, ST-slope and chest pain type. The data set collected by the system went through pre-processing which later supervised learning techniques and prediction model was conducted. Results were produced. Based on the results with the prediction accuracy of 90.97%, our results are in the same range as generated by other algorithms like KNN, Random Forest and Decision Tree algorithm.**

Keywords—Artificial Intelligence, Heart Disease, Data Mining Algorithms, Bayesian, Supervised Learning Techniques, Prediction Model

I. INTRODUCTION

Heart disease is a major health problem in today's time, and it is a term that assigns to many medical conditions related to the heart. These medical conditions describe the abnormal health conditions that directly influence the heart and all its parts [1]. It claims over 17.5 million lives a year, making it one of the world's biggest killers and by 2030 it is expected to rise to 23 million. Research collated in the World Heart Federation (WHF)'s CVD World Monitor shows that in Zambia, heart disease is responsible for 10 percent of all deaths in those aged between 30 and 70 [2]. Therefore, diagnosis to regulate heart disease is very vital, so that appropriate actions can be taken.

Recently, with the resurgence of AI and availability of data, there has been a growing trend of using this technology in health. For example, it has been used for health services management, predictive medicine, patient data and diagnostics, and clinical decision-making.

These technologies have also been used to combat heart diseases. Other scholars, Mirzajani et al. [3], Kumar et al. [4], and Enriko et al. [5], have used different types of techniques such as Genetic Algorithm (GA) to construct computational intelligence methods for the diagnosis of heart disease and some classification algorithms like, j48 decision tree, Naive Bayes (NB), KNN and SMO were analysed and compared to predict heart disease. In Zambia, heart disease is a major problem, it had contributed to 10% of all deaths in those aged between 30 and 70 in the year 2017 and in year 2020 the death rate increased by 2% [2].

Therefore, the proposed technique of an assisted medical diagnosis system will come in handy which can assist the healthcare medical practitioners with the diagnosis of heart disease. The intelligence diagnosis system is used as an assisted system that uses heart disease medical dataset collected from Kaggle as training and testing data which aims at accomplishing the following objectives; To develop a machine learning model that will assist in predicting heart disease, and To predict heart disease using a machine learning classification model.

And the mentioned objectives aim at answering the following questions; How can we develop a model to predict heart disease using machine learning classification? and How can we develop a prototype based on the model in (i) to predict heart disease using machine learning classification?

The rest of the paper is arranged as follows. In section II, we give the related work, in section III, we give the proposed work, in section IV, we give the results and discussion, and lastly in section V, we give the conclusion and recommendation.

II. RELATED WORK

Diagnosis to determine the level/type of heart disease is very important, so that appropriate action can be taken. Intelligence systems can be used to provide diagnosis support. According to Wiharto, Kusnanto and Herianto [6], the use of clinical decision support systems can help physicians deliver improved clinical practice and can reduce the occurrence of faulty diagnoses. Intelligence diagnosis system development will need existing health data as training data, in this case, data specifically related to heart disease.

A prediction and diagnosis of heart disease was conducted by Mirzajani and Siamak [3], were WEKA a powerful data mining tool, was used to apply the data mining algorithms. Experimental results from the implementation of selected classification algorithms, j48 decision tree, Naive Bayes (NB), KNN and SMO on heart disease dataset were analyzed and

compared. The dataset used contained 209 records and 8 features that were collected from a hospital in Iran, under control of health ministry. Description of dataset features used is given in Figure 1. Data was from one resource so there was no need for integration operations. Also, all the features in all the 209 samples contained value and there was no missing value problem. After comparison, results showed that the best classification accuracy is 83.73% which was achieved by j48 decision tree, the second-best classification accuracy was achieved by KNN and SMO with 82.78% followed by NB with 81.82%. Although accuracy is the most common measure in classification performance, other important performance measures such as sensitivity, Specificity, F-Measure, precision, and ROC indicators were considered to evaluate and compare classification efficiency of four selected algorithms.

Enriko [5], also used KNN, Naive Bayes and Decision Tree algorithms to predict heart disease, but with a different dataset collected from California University, Irvine (UCI) was taken to do the analysis, using 10 out of 76 parameters in available and they used a database software called MongoDB. And the accuracy results were not too far differed from each other. In the research, the 8 parameters KNN gave the best result with 81.85% accuracy.

#	NAME	POSSIBLE VALUES
1	Age	NUMERIC
2	Chest_pain_type	ASYMPT, ATYP_ANGINA, NON_ANGINAL,TYP_ANGINA
3	rest_bpress	NUMERIC
4	blood_sugar	TRUE, FALSE
5	rest_electro	Normal,left_vent_hyper, st_t_wave_abnormality,
6	max_heart_rate	NUMERIC
7	exercice_angina	YES, NO
8	Disease	NEGATIVE, POSITIVE

Figure 1: Dataset Description

According to Kumar, Anand et al. [4], Genetic Algorithm (GA) technique was described to construct computational intelligence methods for the diagnosis of heart disease. The performance of the model is validated using a 3-fold cross validation approach. The input heart disease data set was taken from UCIML repository. The average accuracy obtained using 3- folds with initial rule 25 was 81.83%, with 50 initial rule 86.83% and with 75 initial rules.

Zagorecki, Orzechowski, and Hołownia [7], designed a web-based software system responsible for providing medical diagnosis. At the heart of the system lies a distributed, parallel system of multiple Bayesian Networks (BN) engines that are responsible for performing queries to individual BNs. Each of these engines is based on SMILE general purpose BN software (http://genie.sis.pitt.edu). To achieve scalability and reliability, the BN engines are stateless. During software implementation, and analysis results were based on 97,000+ diagnoses made during the first few weeks after deployment of the system. Most of the diagnoses were made for users aged 25-39 (52.1% female and 47.9% male), which explained the symptoms such as depression, tiredness, tension headaches, or anxiety disorders, all of which are characteristic of the modern lifestyle. A review of the symptoms and location which were selected are the head (6.8%), genitals (4.4%) and lower abdomen (3.9%). Although there was no hard evidence; this may suggest that the system is often used to self-diagnose problems related to sexual health and other conditions users may feel uncomfortable going to a doctor about. The anus, for example, was indicated in 2.2% of cases, which was found unusually high (frequency like that of chest pain, sleepiness, or tiredness). Further they investigated the most common diagnoses that were produced for the age groups 55-70 and 70+. The remaining leading diagnoses for patients from the older age groups were consistently age-related problems, such as osteoarthritis, joint or bone trauma, ischemic heart disease, and gallstones.

A proposed classification techniques performance which is compared with prevailing techniques of SMO (Sequential Minimal Optimization), Bayes Net and MLP (Multi-Layer Perception) was conducted by Repaka, Anjan Nikhil Ravikanti et al. [8]. And an effective outcome was exhibited by the proposed Navies Bayesian with greater performance of 89.77% in contrast to rest of the techniques.

S. Chiwamba, J. Phiri, P. Obed et al.[9], [10], use convolutional Neural Network machine learning algorithm in the automated identification and capturing of fall armyworm moths. The achieved train accuracy was 45-60%, cross entropy was 70-80% and validation accuracy was 34-50%. This was all achieved by using the inception V3 in google TensorFlow which is a pre trained neural network and quicker experiment platform for researchers.

III. PROPOSED WORK

The proposed work predicts heart disease by using the Bayesian classification algorithm. The aim of this study is to propose an assisted medical diagnosis system for heart disease predictions based on features logged by medical practitioners which effectively predict the probability if the patient suffers from heart disease. The medical practitioners enter the input values from the patient's health report. The data is fed into a model which predicts the probability of having heart disease. Figure 2 shows the model process involved and Figure 3 shows the proposed system architecture.

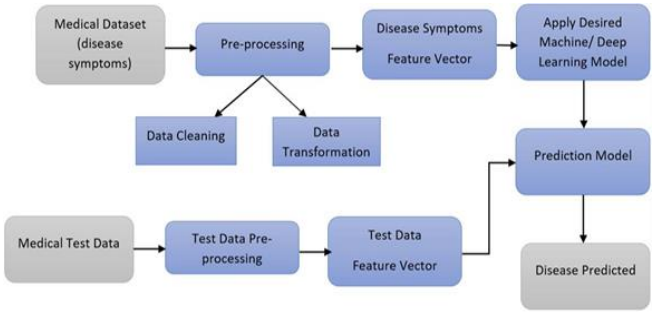


Figure 2: Model framework for Predicting Heart Disease

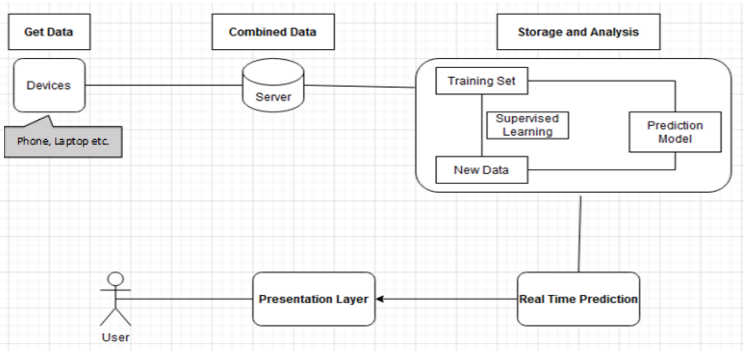


Figure 3: Proposed System Architecture

A. Data collection and pre-processing

The dataset used was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

1. Cleveland: 303 observations
2. Hungarian: 294 observations
3. Switzerland: 123 observations
4. Beach VA: 200 observations
5. Stalog (Heart) Data Set: 270 observations,

Making a total of 1190 observations. This dataset consists of a total of 11 features used. Therefore, we have used the already processed combined dataset accessible in the Kaggle website for our analysis [11]. The full description of the 11 features used in the proposed work is mentioned in Table 1 shown below.

Table 1: Features Description

No	Features Description	Distinct Values of Features
1	Age: age of the patient	Years
2	Sex: sex of the patient [M: Male, F: Female]	M, F
3	ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]	TA, ATA, NAP, ASY
4	RestingBP: resting blood pressure [mm Hg]	Values in mm/hg
5	Cholesterol: serum cholesterol [mm/dl]	Values in mm/dl
6	FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]	1, 0
7	RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]	Normal, ST, LVH
8	MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]	Numerical values between 60 & 202
9	ExerciseAngina: exercise-induced angina [Y: Yes, N: No]	Y, N
10	Oldpeak: oldpeak = ST [Numeric value measured in depression]	Numeric value measured in depression
11	ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]	Up, Flat, Down
12	HeartDisease: output class [1: heart disease, 0: Normal]	1, 0

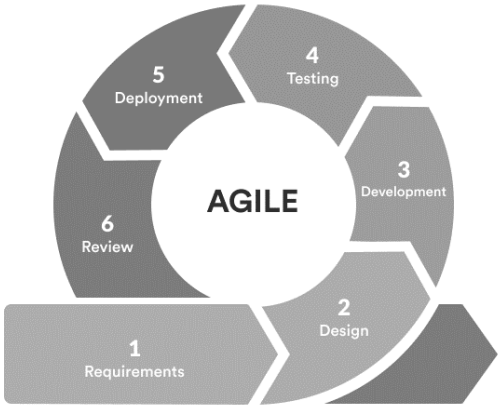
B. Bayesian based classification

The features mentioned in Table 1 are providing input to the Bayesian machine learning classifier. The input dataset is divided into 80% of the training dataset and 20% of the test dataset. Training dataset is the dataset which is used to train a model. Testing dataset is used to check the performance of the trained model. The performance is computed and analyzed based on different systems of measurement used such as precision, F1- scores, recall and accuracy.

Bayesian classification is used because it is a probabilistic approach to learning and inference based on a different view of what it means to learn from data, in which probability is used to represent uncertainty about the relationship being learnt [12].

$$P(m|k) = P(k|m) * P(m)/P(k) \quad (1)$$

Equation (1) states that the probability of m given k equals the probability of k given m times the probability of m, divided by the probability of k. In Equation (1), m is the theory to be tested and k is the proof connected with m.



C. System development methodology

Agile Systems Analysis and Development Life Cycle method was used for the development of the application. Agile methodologies seek to streamline the software development process and enable quick responses to changing needs without requiring a lot of reworks [13].

Figure 4: Agile Development Process

There are six phases involved in the agile development life cycle.

1. *Requirements gathering*: At this stage requirements are collected from the client on how they want the system to look like and how they want the system to perform. Requirement gathering for the system was very important for the project to move on to the next step.
2. *Design:* Once the requirements are collected, the step of designing takes over, which is mainly building the architecture of the project. This stage eliminates

improbable faults by establishing a norm and attempting to adhere to it.

3. *Development*: At this point, the software development process itself begins, while data recording continues in the background. Once the system is developed, the stage of implementation comes where the system goes through a trial study of unit testing to see if each unit of the system is functioning as it should be.
4. *System testing*: The system testing stage assessed the system for errors and bugs which are in the system.
5. *Deployment*: Once the system testing is done by both the developers and User the system is deployed for use.
6. *Review*: Once the system had passed through all the stages without any issues, the retrospective process started, where the developer verified whether the expected tasks time matches the actual one, also reasons for likely delay were found to avoid under- or overestimates in upcoming project. Last but not least, maintenance also happens here, where the system is periodically upgraded and maintained to keep up with changes.

D. Technologies

- Model Technologies: Jupyter notebook and Python
- Web Technologies: Python, Html, CSS, and JavaScript
- API Framework: Flask
- IDE: Visual Studio Code
- Management System: SQLite

IV. RESULTS AND DISCUSSION

The results obtained by Bayesian classification are presented in this section.

A. Data Analysis

Figure 5 shows the sample observations collected from the dataset.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	TA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	TA	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	TA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	TA	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	TA	150	195	0	Normal	122	N	0.0	Up	0

Figure 5:Sample Observations Collected

Notes: Age: person's age in years, Sex: person's sex (M: Male, F: Female), ChestPainType: (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic), Resting Blood Pressure: (measured in mm/hg), Serum Cholesterol (measured in mm/dl), Fasting Blood Sugar (1: if FastingBS > 120 mg/dl, 0: otherwise), Resting Electrocardiogram Results (Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria), Maximum Heart Rate (Numeric value between 60 and 202), Exercise-induced Angina (Y: Yes, N: No), Oldpeak = ST (Numeric value measured in depression), ST_Slope: the slope of the peak exercise ST segment (Up: upsloping, Flat: flat, Down: downsloping), HeartDisease: output class (1: heart disease, 0: Normal).

The distribution of categorical variables is shown in Figure 6 where number of patients is categorized by sex. 76.38% are male and 23.61% are female.

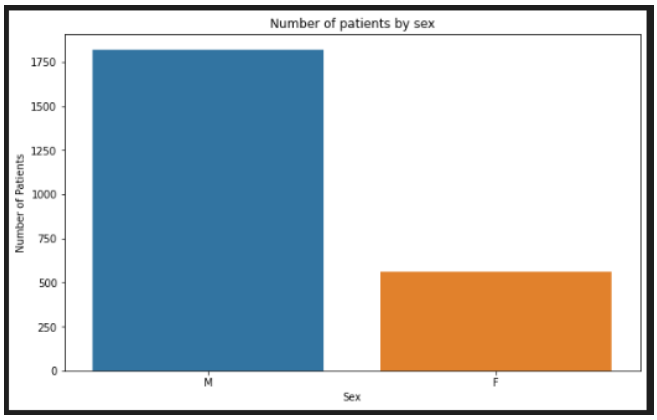


Figure 6: Number of patients by sex

88.87% of male patients have heart disease and 11.12% of female patients have heart disease. This is indicated in Figure 7.

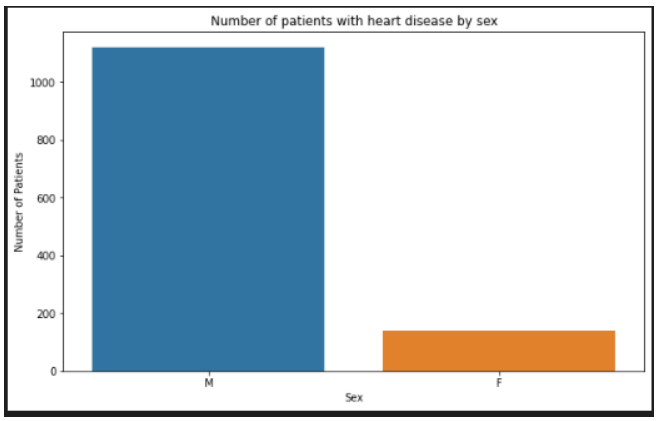


Figure 7: Number of patients with heart disease by sex

From the complete dataset used Figure 8 indicates that 52.85% of patients have heart disease and 47.14% do not have heart disease.

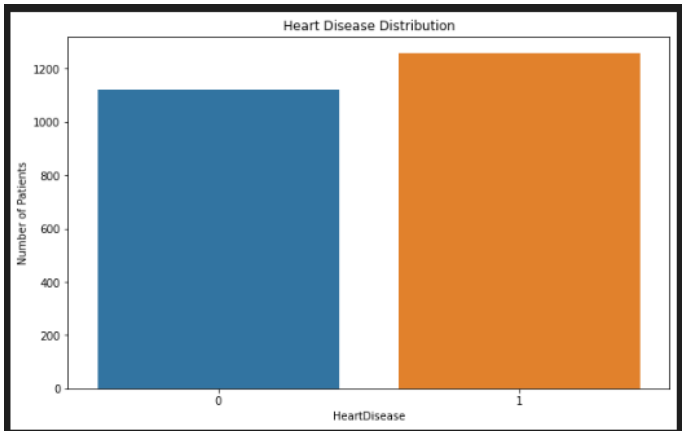


Figure 8: Heart Disease Distribution

B. Preprocessing

In the pre-processing stage the data is divided into two categories which is the training dataset with 80% of data and testing dataset with 20% of data.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope
1851	59	M	ASY	126	218	1	Normal	134	N	2.2	Flat
388	53	M	NAP	130	0	0	LVH	135	Y	1.0	Flat
2037	61	F	ASY	145	307	0	LVH	146	Y	1.0	Flat
230	37	F	ATA	130	173	0	ST	184	N	0.0	Up
1892	42	M	ASY	148	244	0	LVH	178	N	0.8	Up

Figure 9: Sample of the training dataset

C. Initial model training

CatBoostClassifier machine learning model is trained and tested with the 11 features for heart disease prediction with a learning rate of 0.018447 and the best test obtained was 91.28% with 3760 best iterations.

Evaluation is done on the prediction of test data with 98% accuracy. The precision, recall, f1-score, and support are shown int the confusion matrix in Figure 10.

	precision	recall	f1-score	support
0	0.98	0.98	0.98	226
1	0.98	0.98	0.98	250
accuracy			0.98	476
macro avg	0.98	0.98	0.98	476
weighted avg	0.98	0.98	0.98	476

Figure 10: Confusion matrix

D. Hyperparameter tuning

The training process itself is controlled by the data in hyperparameters. As it trains the model, the training product manages three types of data.:

- The training data is used to set up the model so that it can correctly anticipate fresh occurrences of similar data. The values in the training data, however, are never directly included in the model.
- The variables used by the selected machine learning technique to adapt to the data are the model's parameters.
- The controls over the training itself are known as the hyperparameters. Keep in mind that while hyperparameters are frequently stable during a job, parameters typically fluctuate during a training job.

In the course of the training 11 trails are conducted and trail 5 is the best with a value of 99%. The best Test value is 0.0849 and the best Iteration is 999.

Therefore, after training the final model with new parameters test data prediction accuracy is 99.07% and the training data prediction accuracy is 99.97%.

A. Web application simulation results

Heart diseases are the one of the top causes of death in Zambia. This can mainly be attributed to lack of software that can complement the efforts of medical practitioners in the prediction of heart disease.

To overcome the current diagnosis techniques for the diagnosis of heart disease, scholars Kaur et al. [1], Goma et al. [2], Mirzajani et al. [3], Kumar at al. [4], Enriko et al. [5], Wiharto et al. [6], Zagorecki et al. [7], and Repaka et al. [8], have used machine learning and deep learning approaches to construct a few AI-based forecasting systems. The suggested methods use a specific or limited dataset type or may not have a potential applicability in a clinical environment, despite the fact that the results provided are positive.

The purpose of the study was to design and develop a computer-assisted diagnosis system using artificial intelligence techniques (AI) which will assist lessening the death rate by providing decision support to medical practitioners allowing early diagnosis and treatment.

Below from Figure 11 to 14 analysis of the design and implementation for heart disease prediction is proposed. The system monitors the features logged and outputs the probability of heart disease of the features logged.

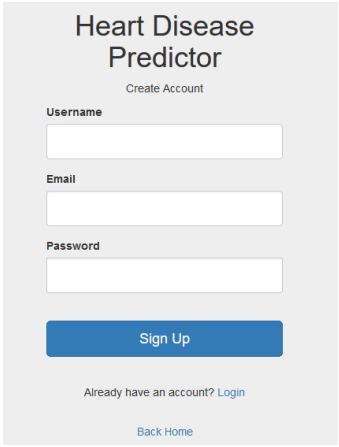


Figure 11: Medical practitioner must enter details to create an account

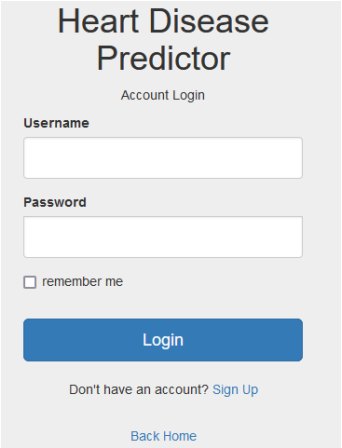


Figure 12: Medical practitioner must enter Username and Password to login

Figure 13: Medical practitioner must enter health details like blood pressure, heart rate, age etc., and click on 'run model' for the system to predict

Figure 14: Medical practitioner can see the probability of a patients having heart disease or not

V. CONCLUSION ANDS RECOMMENDATION

Development of a web application system to anticipate heart diseases precisely and effectively was required given the number of deaths in Zambia caused by heart ailments. The study's purpose was to propose an assisted medical diagnosis system for heart disease predictions based on features logged by medical practitioners. This study uses Bayesian model for predicting the probability of heart disease heart disease using a combined machine learning repository dataset from Kaggle website which consists of 11 features. The result of this study indicates the accuracy score of training data is 99.97% for prediction of heart disease using new parameters fed to the model and 99.07% accuracy with test data.

In future, a Zambian dataset with more features can be used to test, train, and create a model that can predict heart disease using different machine learning models. Also, a system model that can predict the exact type / level of heart disease can be developed.

ACKNOWLEDGMENT

The study was supported by the University of Zambia (UNZA). The dataset used for this study was collected from the Kaggle website. The author would like to thank the Computer Science Department (UNZA) for their input in the study.

REFERENCES

[1] B. Kaur and W. Singh, "Review on Heart Disease Prediction System using Data Mining Techniques," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 2, no. 10, pp. 3003–3008, 2014.

[2] F. Goma, W. Scholtz, O. Scarlatescu, G. Nel, and J. M. Fourie Zambia, "Zambia Country Report PASCAR and WHF Cardiovascular Diseases Scorecard project," *CardiovasCular Journal of afriCa* •, vol. 31, no. 4, 2020, doi: 10.5830/CVJA-2020-038.

[3] S. S. Mirzajani and siamak salimi, "Prediction and Diagnosis of Diabetes by Using Data Mining Techniques," *Avicenna Journal of Medical Biochemistry*, vol. 6, no. 1, pp. 3–7, 2018, doi: 10.15171/ajmb.2018.02.

[4] P. Siva Kumar, D. Anand, V. Uday Kumar, and D. Bhattacharyya, "A computational intelligence method for effective diagnosis of heart disease using genetic algorithm," *International Journal of Bio-Science and Bio-Technology*, vol. 8, no. 2, pp. 363–372, 2016, doi: 10.14257/ijbsbt.2016.8.2.34.

[5] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 8, no. 12, pp. 59–65, 2016.

[6] W. Wiharto, H. Kusnanto, and H. Herianto, "Intelligence system for diagnosis level of coronary heart disease with K-star algorithm," *Healthc Inform Res*, vol. 22, no. 1, pp. 30–38, 2016, doi: 10.4258/hir.2016.22.1.30.

[7] A. Zagorecki, P. Orzechowski, and K. Hołownia, "A system for automated general medical diagnosis using bayesian networks," *Stud Health Technol Inform*, vol. 192, no. 1–2, pp. 461–465, 2013, doi: 10.3233/978-1-61499-289-9-461.

[8] A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, "Design and implementing heart disease prediction using naives Bayesian," *Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019*, vol. 2019-April, no. April 2019, pp. 292–297, 2019, doi: 10.1109/icoei.2019.8862604.

[9] S. H. Chiwamba, J. Phiri, P. O. Y. Nkunika, M. Nyirenda, M. M. Kabemba, and P. H. Sohati, "Machine Learning Algorithms for automated Image Capture and Identification of Fall Armyworm (FAW) Moths," *Zambia ICT Journal*, vol. 3, no. 1, 2019, doi: 10.33260/zictjournal.v3i1.69.

[10] S. H. Chiwamba *et al.*, "An Application of Machine Learning Algorithms in Automated Identification and Capturing of Fall Armyworm (FAW) Moths in the Field Automatic Data Capturing At Satellite Depots Based On NFC Technology. View project An Application of Machine Learning Algorit," no. DECEMBER, pp. 119–124, 2018, [Online]. Available: https://www.researchgate.net/publication/331935302

[11] "Heart Failure Prediction Dataset | Kaggle." https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction (accessed Oct. 10, 2022).

[12] C. J. Du and D. W. Sun, "Object Classification Methods," *Computer Vision Technology for Food Quality Evaluation*, pp. 81–107, 2008, doi: 10.1016/B978-012373642-0.50007-7.

[13] I. Sommerville, *Software Engineering*, 9th ed. United States of America: Pearson Education, Inc., 2011. doi: 10.1016/B978-0-12-396961-3.00009-3.