Proceedings of the International Conference on Industrial Engineering and Operations Management
Harare, Zimbabwe, December 8-10, 2020

Predicting Congenital Heart Diseases Using Machine Learning

John Tinashe Meda and Tawanda Mushiri
Department of Mechanical Engineering
University of Zimbabwe
Harare, Mt Pleasant, 630 Churchill Avenue, Zimbabwe
jtmeda@gmail.com, **tmushiri@eng.uz.ac.zw**

Abstract

The researcher was working on the idea that machine learning, a branch of artificial intelligence has proved more accurate than humans in predicting heart diseases. Congenital Heart Diseases (CHDs) are responsible for a greater percentage of infant mortality worldwide, hence the need to detect and predict the diseases at birth for the administration of proper and on time treatment. The target of the project was to come up with an algorithm that can be linked to a device(s) for inputs or that can receive inputs entered manually and give a prediction of potential CHDs. This has been previously addressed as the prediction of heart disease in general but there has not been enough focus on Congenital Heart Diseases hence for this study. The researcher made use of the K-Nearest Neighbours (K-NN), Logistic Regression and Support Vector Machine (SVM) to develop three prediction algorithms in Python Programming. The prediction accuracies of the three in algorithms were compared and the one with the highest score was selected. According to the selected SVM algorithm the researcher saw 73% accuracy in prediction. One recommendation is that more specific data on CHD should be collected to create better datasets.

Keywords
Congenital Heart Disease (CHD), K-Nearest Neighbours, Prediction, Linear Regression, Support Vector Machine (SVM)

1. Introduction

A Congenital Heart Disease (CHD), also referred to as a Congenital Heart Defect, is a problem in the structure of the heart, already present at birth [1]. CHDs involve the heart's wall, valves, arteries, and veins close to the heart, and these defects have the potential to disrupt the normal blood flow through the heart – slowing it down, misdirecting it, or preventing it completely [2].

Some commonly known symptoms of heart defects in severe cases, those that can be fatal if untreated, include: *cyanosis* (blueish lips, skin and fingernails), *fatigue*, *rapid and short breath* especially during feeding and poor blood circulation seen by *swollen legs, abdomen and area surrounding the eyes*. The serious CHDs are usually prevalent during pregnancy, at birth or during the first few months of life and the less intense ones are usually discovered later in the child's life seen by symptoms such as *quickly getting short of breath during exercise*, *fainting during exercise*, *easily tiring* and *swollen hands, feet or ankles* [3]. A blowing, whooshing or rasping sound in heartbeat known as a heart murmur, pounding heart and weak pulse can also be symptoms of CHDs [4].

According to the Department of Health and Human Services in the United States, Centres for Disease Control and Prevention in 2017, heart disease was the leading cause of human death[5]. Currently, in young children and infants, congenital heart disease contributes a large 30% to 50% to the mortality resulting from birth defects. Although mortality resulting from CHD during childhood and infancy is reportedly decreasing, its prevalence among adults is increasing.

2. Artificial Intelligence in the Health Sector

Currently, the common artificial intelligence techniques used in the health sector include machine learning for structured data (for example, the classical support vector machine) and deep learning, neural network, and natural

© IEOM Society International

Proceedings of the International Conference on Industrial Engineering and Operations Management
Harare, Zimbabwe, December 8-10, 2020

language processing for use with unstructured data. Through this, the major disease areas using artificial intelligence are cancer, cardiology and neurology [6]. Artificial intelligence software can be used for cardiovascular and respiratory monitors in intensive care and high dependence units by the interpretation of vital signs. Natural language processors in virtual assistants can also be used to communicate medical information with patients after a hospital visit and schedule follow ups. This system has been found to increase cooperation and reliability of the follow up [7]*).*

Current Heart Disease AI Applications
This section mentions some of the company's in the world that have actively taken up machine learning in disease prediction.

- Medical Imaging

The training of machine learning algorithms to improve the accuracy of patient scans to better detect the disease. Some existing developers of the algorithms include: Zebra Medical Vision and Analytics for Life [8].

- Risk Prediction

The use of machine learning to predict the risk of cardiovascular diseases and its related impacts. Kensci Platform and Heartflow are examples of currently existent firms. Heartflow, leverages deep learning to create a personalized 3D model of the potential patient's arteries, in detecting and diagnosing coronary artery diseases [8].

- ECG Monitoring

This is the use of deep learning to assist in automating the process of atrial fibrillation (AFib), the most common abnormal heart rhythm disease. An example is Cardiologs, which reads ECG recording from any digital device such as a smart watch or Holter Monitor.

Research Gap
A look at studies previously done on predicting the presence of heart disease using machine learning shows that most research has been done towards prediction of heart disease in general and coronary heart disease. This project will focus on bridging the research gap as it will focus on the prediction of congenital heart disease specifically at birth and a few weeks after.

3. Justification

- Reduction in the World's Death Toll Resulting from Heart Diseases:
  Most CHD patient fatalities occur when the disease is discovered at a later stage. Early detection and/or prediction will result in a reduction in the number of these deaths.

- More equipped medical doctors providing better quality health care:
  As the saying goes, "a work man is only as good as his tools" [9] and the proverb forewarned is forearmed, likewise, a well-informed, well equipped, doctor with the ability to predict heart disease before it actually occurs is in a better position to cure or prevent it.

- Saving of Resources, namely time and money:
  In 2014, Healthcare Business and Technology approximated the cost of a single open-heart surgery to US$324 000, which made it the world's seventh most expensive medical procedure [10]. An article from Medical News Today says that basic open-heart surgery, such as coronary artery bypass, requires a hospital stay of about 7-10

Proceedings of the International Conference on Industrial Engineering and Operations Management
Harare, Zimbabwe, December 8-10, 2020

days with at least 1 day in the Intensive Care Unit after the operation, and the procedure itself takes about 3-6 hours [11].

- Improved standard of living of a country:
  A mere look at the National Budget Speech for Zimbabwe by Mthuli Ncube shows that there is a proposed budget of US$694.5 million to be allocated to the Ministry of Health and Childcare *(Ncube, 2018).* Taking into considering the country's struggling economy, predicting heart diseases means better prevention of these diseases and therefore more free finances for the government to be allocated to other departments or improvement within the health sector itself.

4. Aim

The main goal of this study is to develop a machine learning algorithm that can be integrated with heart diagnostic devices in the smart Health 4.0 environment to predict the occurrence of several congenital heart diseases from currently existing data.

5. Objectives

- The selected prediction algorithm must give an accuracy of at least 70%.
- The designed system must have a user-friendly interface that can be understood by any person with basic literacy
- The system must also be secure system, by keeping user information private, limited access to the relevant people only

6. Machine Learning Algorithms

Three algorithms will be used in this study and their accuracy scores will be compared to choose the one which will be used to develop the heart prediction system. Some of the different algorithms that can be applied in machine learning for heart disease prediction are:

Decision Tree
A supervised learning algorithm commonly used I classification problems working effortlessly with continuous and categories attributes based on most significant *predictors* [12]. Firstly, the algorithm calculates the entropy of every attribute:

Equation 6-1 : Entropy S of attributes in decision tree algorithm

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

The dataset is then split with the aid of the variables/predictors with minimum entropy or maximum information gain: Gain (S, A):

Equation 6-2 : Decision Tree maximum information gain

$$Gain(S, A) = Entropy(S) - \sum_{v=values(A)} \frac{|S_i|}{|S|} \times Entropy(S)$$

Support Vehicle Machine SVM Algorithm
Defined as a finite-dimensional vector space consisting of a dimension for every attribute of an object, this algorithm has a very useful classification accuracy [14]. The algorithm classifies by finding the hyperplane that'll

©IEOM Society International

Proceedings of the International Conference on Industrial Engineering and Operations Management
Harare, Zimbabwe, December 8-10, 2020

maximise the margin between the two classes as shown in Figure 6.1. The vectors (case) defining the hyperplane are the support vectors.

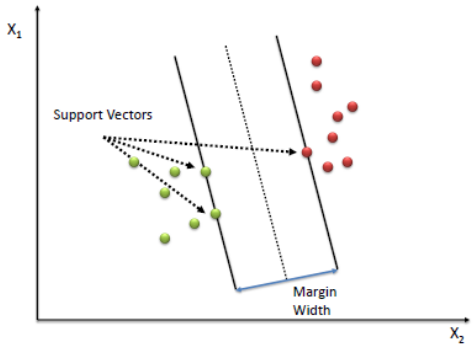

Figure 6.1: The Support Vector Machine Algorithm Illustration

Linear Regression

An algorithm based on supervised learning that works by performing a regression task mostly used to bring out the relationship between variables and forecasting It models a target prediction value based on independent variables. Regression models vary based on the relationship between dependent and independent variables [15]. The algorithm performs tasks to predict a dependent variable value (y) based on a given independent variable (x) and finds out a linear relationship between the input and output, x and y respectively, with the hypothesis function shown in Equation 6-3:.

Equation 6-3 : Hypothesis Function for Linear Regression

$$y = \theta_1 + \theta_2 x$$

When training the model:

- x: input training data
- y: labels to data (supervised learning)

When training the model:

Model fits best line to predict value of y for a given value of x

- θ_1: intercept
- θ_2: coefficient of x

And, with θ_1 and θ_2 we get the line of best fit so that when the model/algorithm is applied it will predict a value of y for the input value of x.

K-Nearest Neighbours (KNN)

K-NN is a supervised learning classification algorithm. K-NN algorithm predicts the class label of a new input; K-NN utilizes the similarity of new input to its input's samples in the training set. If the new input is same the samples in the training set. A case is classified based on its surrounding neighbours, with the case being assigned to the class

©IEOM Society International

Proceedings of the International Conference on Industrial Engineering and Operations Management
Harare, Zimbabwe, December 8-10, 2020

most common amongst its K nearest neighbours measured by a distance function. If K = 2, then the case is simply assigned to the class of its nearest two neighbours. The distance to be calculated between the value under investigation can be calculated using four methods, these are shown Equation 6-4, Equation 6-5, Equation 6-6 and Equation 6-7 that is the Euclidean, Minkowsky, Correlation and Chi-Square respectively:

- The Euclidean Distance Between Neighbours:

Equation 6-4 : Euclidean distance between KNN algorithm neighbours

$$Euclidean\ Distance = \sum_{i=1}^{k} (X_i - Y_i)^2$$

Where k is the number of neighbours being used, X and Y are the horizontal and vertical displacements, respectively.

Minkowsky Distance Between Neighbours

Equation 6-5 : Minkowsky distance between KNN algorithm neighbours

$$Minkowsky(A,B) = \left(\sum_{i=1}^{m} |x_i - y_i|^r \right)^{\frac{1}{r}}$$

Correlation Distance Between Neighbours

Equation 6-6 : Correlation distance between KNN algorithm neighbours

$$Correlation\ Distance\ (A,B) = \frac{\sum_{i=1}^{m} (x_i - \mu_i)(y_i - \mu_i)}{\sqrt{\sum_{i=1}^{m} (x_i - \mu_i)^2 \sum_{i=1}^{m} (y_i - \mu_i)^2}}$$

Chi Square Distance Between Neighbours

Equation 6-7 : Chi Square distance between KNN neighbours

$$Chi\ Square\ \ Distance\ (A,B) = \sum_{i=1}^{m} \frac{1}{sum_i} \left(\frac{x_i}{size_Q} - \frac{y_i}{size_I} \right)$$

7. Dataset Analysis

The original dataset used in conventional heart disease prediction had a 303-patient sample size and 76 features. In this study, however, the sample was reduced to 297 patients with 9 features i.e. monitored parameters, namely: *sex, blood pressure, resting ECG, maximum heart rate, birth weight, weight at 4 weeks, weigh difference at birth and 4 weeks, blood oxygen saturation and body temperature*.

Figure 7.1 is a summary of how the data is processed from the main dataset through the algorithm predictions all the way to the prediction accuracy scores. The filtered and edited dataset is the one that is fed into the Python program, which consists of the 297 patients with 9 features named above. The input data is obtained from an online form. The parents and child's doctor will fill out an online form as they build information to add to the dataset. Firstly, just after birth, an account on the hospital database created using MySQL with PHP is created. The doctor then goes on to the child vitals page where he or she enter the nine cardiac related parameters being monitored for prediction. The

©IEOM Society International

Proceedings of the International Conference on Industrial Engineering and Operations Management
Harare, Zimbabwe, December 8-10, 2020

parents' contact details are also entered on a separate. When all in this information has been captured, the child vitals database with values for the monitored parameters is exported to dataset. This dataset feeds parameters into the machine learning algorithm that performs the heart disease prediction.

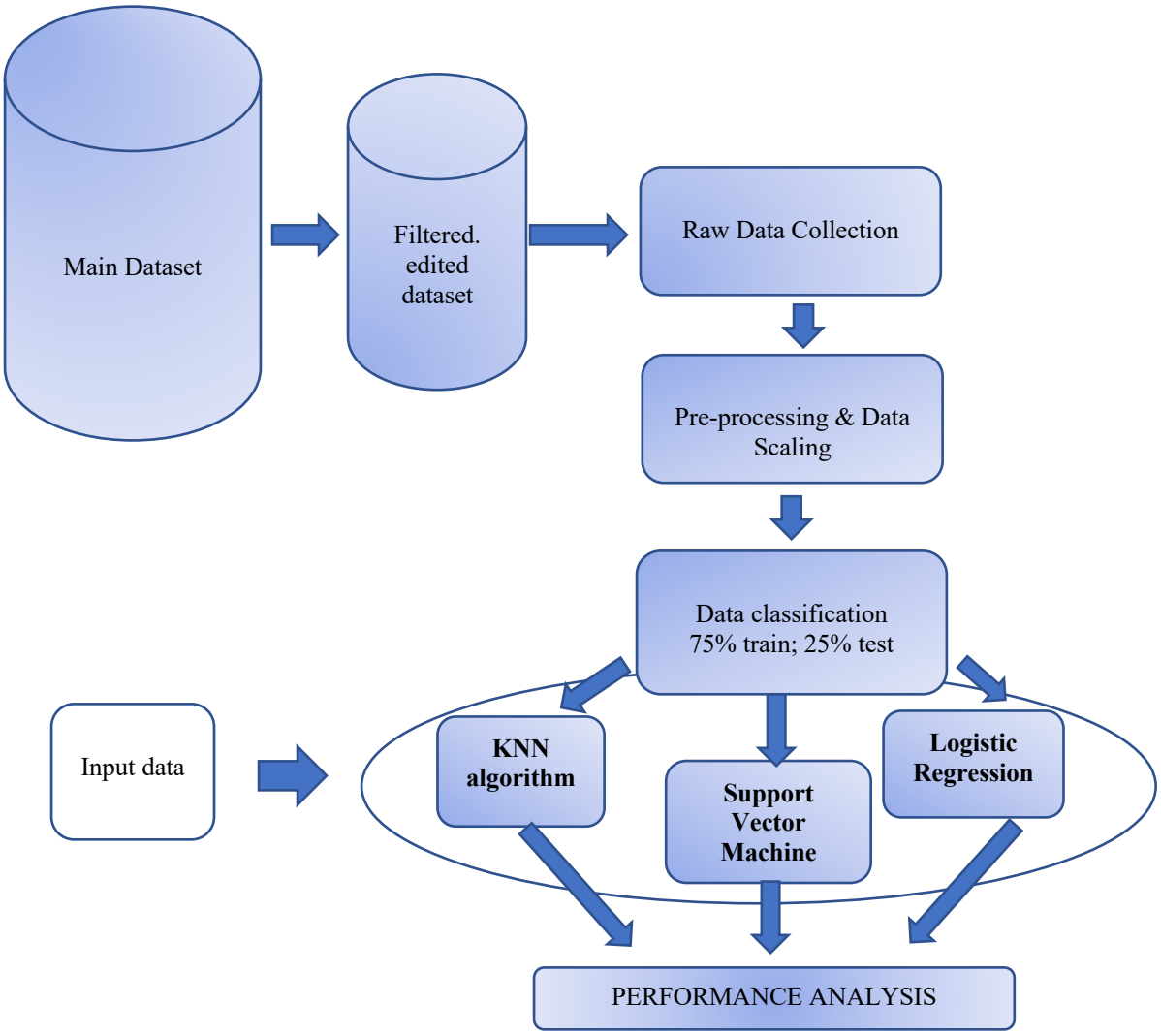

Figure 7.1: Machine Learning Prediction Process Overview

8. The Data

Number of CHD diagnosed babies in the dataset

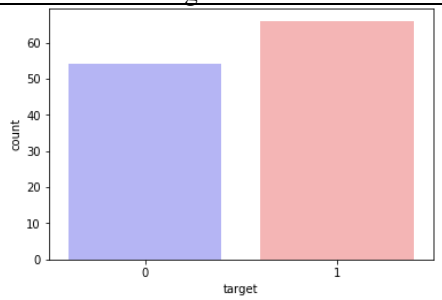

Figure 8.1: Number of babies flagged for heart disease

The sex distribution in the dataset

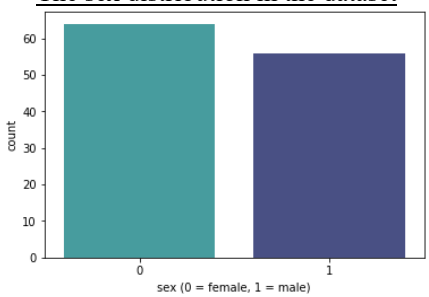

Figure 8.2: The number of male and female babies

©IEOM Society International

Proceedings of the International Conference on Industrial Engineering and Operations Management
Harare, Zimbabwe, December 8-10, 2020

In the dataset used, as illustrated in Figure 8.1 above, we see that of the 120 babies, 66 were diagnosed with a congenital heart disease and 54 were not.

In the dataset, as shown in Figure 8.2, there are 66 female babies and 54 male babies.

9. Selection of Algorithm

Three algorithms out of the many compatible ones stated above were selected to be tested for use in this study. The algorithm with the highest accuracy score was then selected as the one that will be a part of the prediction model – Support Vector Machine Algorithm. Shown below are the test results for the three algorithms (KNN, Linear Regression & SVM) that were tested with the dataset:

a) Using KNN Algorithm

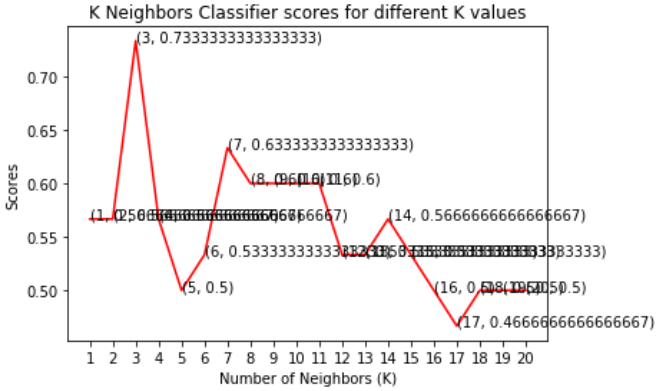

Figure 9.1: Plot of K values showing prediction accuracy of each

From the confusion matrix of the algorithm: $\begin{bmatrix} 6 & 5 \\ 3 & 16 \end{bmatrix}$ the accuracy of the system in predicting CHD is given by:

Equation 9-1 : Accuracy of ML algorithm used in K-Nearest Neighbours

$$Accuracy = \frac{Correct\ Predictions}{Correct + Incorrection\ Predictions} \times 100\%$$

$$Accuracy = \frac{6+16}{3+16+5+6} \times 100 = 73.333\%$$

73.333% accuracy as shown in Figure 9.1 is achieved when the number of neighbours (k) is 3. The figure shows the different accuracy scores achieved by the algorithm at different values of k neighbours. The lowest accuracy was seen at k = 17 which gave a score of 46%. The KNN algorithm overall did well with the data with the optimum number of K neighbours selected.

©IEOM Society International

Proceedings of the International Conference on Industrial Engineering and Operations Management
Harare, Zimbabwe, December 8-10, 2020

b) Using Logistic Regression Algorithm

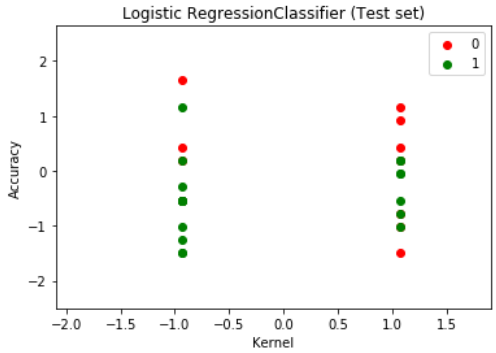

Figure 9.2: Prediction Output using Logistic Regression Algorithm

From the confusion matrix of the algorithm: $\begin{bmatrix} 6 & 5 \\ 7 & 12 \end{bmatrix}$ the accuracy of the system in predicting CHD is given by:

Equation 9-2 : Accuracy of ML algorithm used in Logistic Regression

$$Accuracy = \frac{Correct\ Predictions}{Correct + Incorrection\ Predictions} \times 100\%$$

$$Accuracy = \frac{6+12}{7+12+5+6} \times 100 = 60.0\%$$

Here we see that the accuracy of the prediction is 60%. However, Figure 9.2 was originally supposed to be a scatter graph with points well spread out over the axes but it did not produce such a result with our data showing some form incompatibility and the logistic regression algorithm could not therefore be used in the development of the congenital heart disease prediction system. The green dots symbolize correct predictions and the red dots incorrect ones, but the algorithm did not do well with the data as the points are too few.

c) Using Support Vector Machine Algorithm

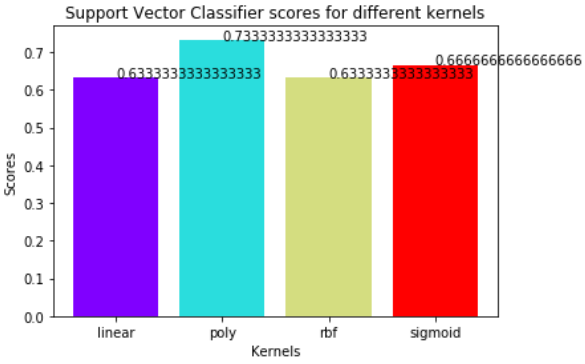

Figure 9.3: Prediction accuracy scores using the Support Vector Machine algorithm

The highest accuracy as shown above in Figure 9.3 under poly kernel is 73.33% which is consistent with the score achieved by the K-NN classification. Figure 9.3 also shows prediction accuracy scores obtained within the Support Vector Machine algorithm using three other **different kernels**: linear, RBF and sigmoid, which had accuracies of 63%, 63% and 66% respectively. This shows that prediction done by these other three was of lower accuracy that that of the poly kernel and gave values that are below the specification target of 70% and could therefore not be considered to be taken up for use in the heart disease prediction system.

©IEOM Society International

Proceedings of the International Conference on Industrial Engineering and Operations Management
Harare, Zimbabwe, December 8-10, 2020

10. Model CHD prediction system

Figure 10.1 shows a system that was proposed to help in the prediction of heart disease using machine learning. After the algorithm performance analysis shown in Section 5: Dataset Analysis the dataset is fed into the selected algorithm (Support Vector Machine), which is the first step in the prediction system model shown in Figure 10.1. At the doctor monitors the 9 heart parameters stated earlier using the instruments shown at the bottom of the model diagram. User accounts are created, and the parent(s) details are also captured to add the needed layer of security. The cloud server serves the purpose of storing data and making accessible from anywhere in the world. This supports the Health 4.0 revolution.

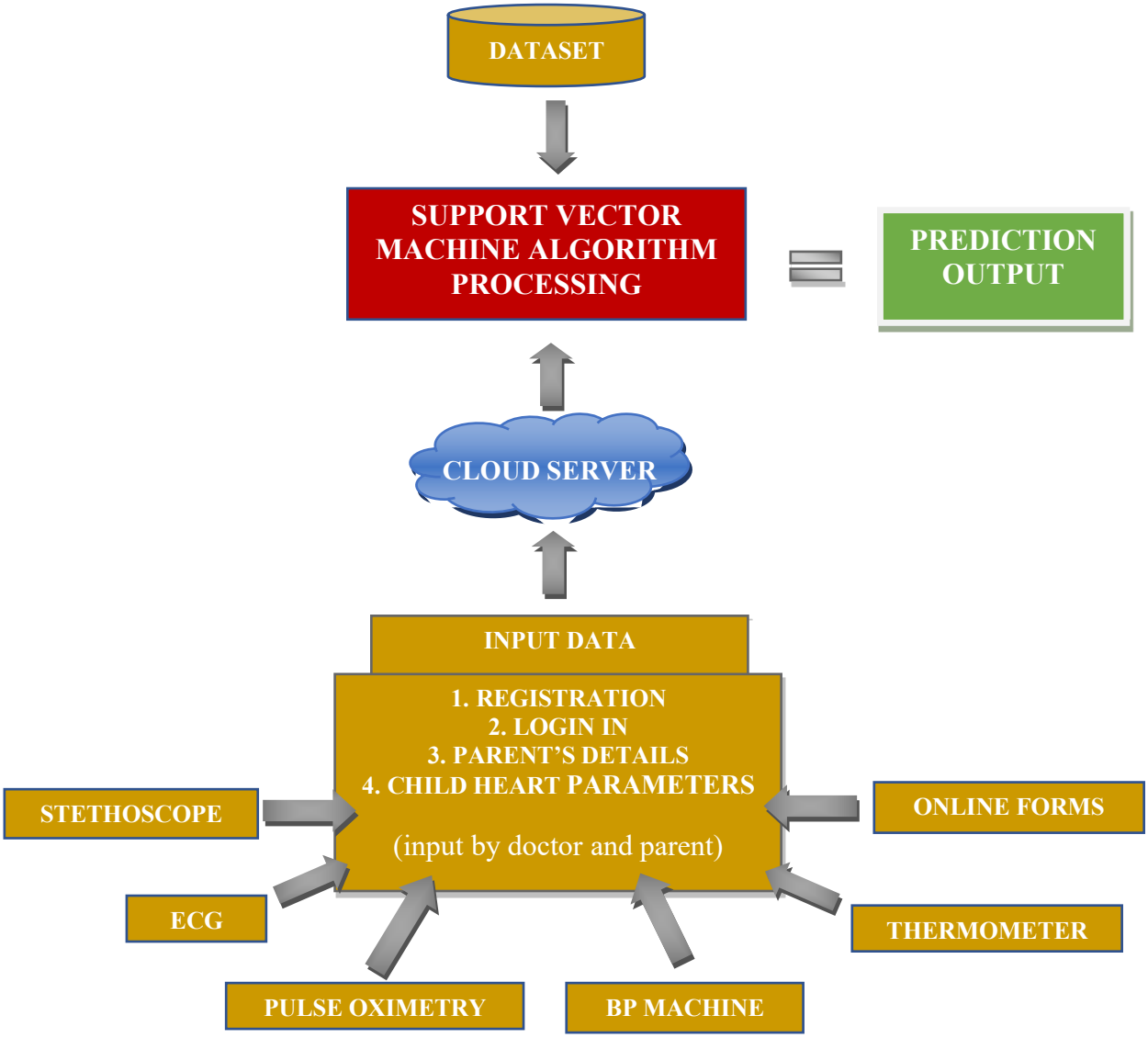

Figure 10.1: Proposed Congenital Heart Disease Prediction System

© IEOM Society International

Proceedings of the International Conference on Industrial Engineering and Operations Management
Harare, Zimbabwe, December 8-10, 2020

11. Conclusion

The main purpose of this project was to develop a system that will allow for the prediction of congenital heart diseases at birth and some weeks after birth that uses machine learning to improve the accuracy of prediction. Three algorithms (K-Nearest Neighbours, Logistic Regression and Support Vector Machine) were tested with the dataset to select the one with the highest accuracy. The support vector machine algorithm was taken up and the disease prediction system was developed. The proposed system achieved an accuracy of 73% which met our objective target of 70%. In the proposed system the database is password protected making information stored secure and the front-end interface developed in PHP and HTML is simple and easy to use.

12. Recommendations

There are several limitations that were encountered in this study that when addressed in future studies would make for more efficient, reliable, and convenient congenital heart disease prediction systems. Firstly, more data directly focussing on congenital heart disease should be collected to build specific datasets that will be used to train the machine learning algorithms. Also, in future studies, before coming to a concrete conclusion more machine learning algorithms should be tested with the dataset to pave way for more accurate CHD predictions. As the data volumes grow computers with greater processing power will also be required to perform the predictions. In the light of the full implementation of Health 4.0 to keep up with the times, portable heart disease input device(s) that can continually monitor the parameters sending the collected results to the cloud server remotely should be developed.

13. Acknowledgements

I would like to acknowledge my co-author and supervisor Dr. Tawanda Mushiri for his support and guidance through this study and the Department of Mechanical Engineering of the University of Zimbabwe for creating the study platform for me.

Special mention also goes to the Royal Academy of Engineering and HEP SSA for availing a platform for presenting and publishing our research papers.

14. References

[1] A. Dodge-Khatami, Transl. Pediatr. 5 (2016) 109–111.
[2] J.G. Harold, Circulation 130 (2014).
[3] R. Sun, M. Liu, L. Lu, Y. Zheng, P. Zhang, Cell Biochem. Biophys. 72 (2015) 857–60.
[4] M.A. Mari, M.M. Cascudo, J.C. Alchieri, Brazilian J. Cardiovasc. Surg. 31 (2016) 31–37.
[5] S.L. Murphy, J. Xu, K.D. Kochanek, E. Arias, NCHS Data Brief (2018) 1–8.
[6] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Stroke Vasc. Neurol. 2 (2017) 230–243.
[7] N.L. Narayana, N.V.S. Raju, K. Chaithanya, P.R. Reddy, 13 (2013).
[8] K. Sennaar, Emerj 05 (2019) 1–3.
[9] E. Awah, Living Beyond...: How to Sweep-Away the Obstacles on Your Path to Success, Booktango, 2013.
[10] N. Kijima, K. Tanaka, F. Marumo, Acta Cryst B39 (1983) 557.
[11] M. Zubrzycki, A. Liebold, C. Skrabal, H. Reinelt, M. Ziegler, E. Perdas, M. Zubrzycka, J. Pain Res. 11 (2018) 1599–1611.
[12] V. V. Ramalingam, A. Dandapath, M. Karthik Raja, Int. J. Eng. Technol. 7 (2018) 684.
[13] (n.d.).
[14] M. Aljanabi, M. Qutqut, M. Hijjawi, Int. J. Eng. Technol. 7 (2018) 5373–5379.
[15] T. Ahmad, L.H. Lund, P. Rao, R. Ghosh, P. Warier, B. Vaccaro, U. Dahlström, C.M. O'Connor, G. Michael Felker, N.R. Desai, J. Am. Heart Assoc. 7 (2018).

15. Biographies

John Tinashe Meda is a Continuous Improvement Consultant at Kaizen Institute South Africa and holds a Bachelor of Science Honours Degree in Mechanical Engineering obtained at the University of Zimbabwe.