

Comparison of Support Vector Machine Models in the Classification of Susceptibility to Schistosomiasis

D. OLANLOYE, R. OLASUNKANMI and E. ODUNTAN


Abstract— Schistosomiasis has become epidemic sending millions of people into untimely graves. A lot of contributing efforts in term of research have been made to eradicate or reduce the rate of this dangerous infection. In this research work, the concept of Machine Learning as one of the subdivision of Artificial Intelligence is being used to determine the level of susceptibility of Schistosomiasis. The research made a comparison of the various support vector machine models - Linear, Quadratic, Cubic, Fine Gaussian, Medium Gaussian and Coarse Gaussian model to determine the level of susceptibility to Schistosomiasis. The results obtained which include Confusion Matrix (CM), Receiver Operating Character (ROC) and Parallel Coordinate Plots (PCP) were interpreted in the form of accuracy, processing speed and execution time. It was finally concluded that Medium Gaussian is the best of all the six models.

Index Terms— Artificial Intelligence, Classification, Machine Learning, Schistosomiasis, SVM


I. INTRODUCTION

ONE OF the major problems in the Africa sub-region is the high mortality and morbidity rate in connection with the disease known as schistosomiasis. Schistosomiasis is a blood dwelling parasitic worm that represents a series of health problems in a tropical region. More than two hundred million people have been infected. It is one of the widely spread and prevalent parasitic diseases in the world today (Makolo and Akinyemi, 2016). Since the beginning of 20th century, this has been found endemic in several countries and its discontinuous geographical distribution


ODUNAYO OLANLOYE is with the Department of Computer and Information System Bowen University, Iwo, Nigeria, (e-mail: odun.olanloye@bowenuniversity.edu.ng).

 <https://orcid.org/0000-0002-3564-774X>

OLAWUMI OLASUNKANMI, is with the Department of Computer Science, Ladoko Akintola University of Technology, Ogbomoso, Nigeria, (e-mail: roolasunkanmi20@lautech.edu.ng).

 <https://orcid.org/0000-0002-8652-0626>

ESTHER ODUNTAN, is with the Department of Computer Science, The Federal Polytechnic, Ilaro Ogun State, Nigeria, (e-mail: esteemviac@gmail.com).

 <https://orcid.org/0000-0003-3522-0356>

Manuscript received November 28, 2019; accepted July 08, 2020.
DOI: [10.17694/bajece.651784](https://doi.org/10.17694/bajece.651784)

determine to a great extent by the occurrence of its intermediate hosts (freshwater snail) of Biomphalaria (Reynold and ArveLee, 2019). This call for strong measures, aimed at managing the scourge.

A lot of research work has been carried out in this area using various approaches or methods. Computer Scientists are putting in their best yet, enough has not been done to manage the scourge. This research is out to measure and to compare the performance of Support Vector Machine models at the various level of susceptibility of the scourge. The models will be implemented using MatLab Machine Learning (ML) tools.

A. Schistosomiasis

This is a typical type of disease also called bilharzia. It is caused by parasitic worm released by snail into a river. A victim is infected when in contact with contaminated water and the worm penetrated the skin of the victim. It enters into the body system where it continues to grow for several weeks and become an adult worm. The adult worm lives in the blood vessels where the female type continues to produce eggs. The egg when hatched releases free-swimming larva called miracidia into freshwater. The lava finds its way into a freshwater snail and the victim is infected when in contact with such water. The freshwater can also be contaminated when an already infected animal is in contact with freshwater through its urine or faeces. The worm when absorbed into the bloodstream find its level into the liver, intestine or other vital organs in the body system. Possible symptoms include muscular aches, itching skin, persistence cough, headache, stomach pain, joint pain etc. If not treated on time, the patient starts releasing blood in the urine or stew. This leads to retarded growth, especially in children. It can also cause bladder cancer as well as kidney or liver problem. Some other major complications include high blood pressure (hypertension), urinary problem and destruction of vital organs etc. It is a chronic communicable disease that can lead to death.

B. The Concept of Artificial Intelligence and Machine Learning

Between 1940 and 1950, Artificial Intelligence(AI) emerged as a separate and independent field of study. Though, earlier before this time, literature revealed that there has been in existence of different areas of research which studied different concept that form the basis of AI. Those areas were integrated to shape AI as one of the major and independent areas of study.

An attempt to create an “Artificial Intelligence” in a machine such that the machine too can think and solve real-life problems in a humanlike manner is what is being described as Artificial Intelligence. It is a field where systems are made to cope with a certain degree of uncertainty like the accuracy of unexpected events such as unpredictable changes in the world in which the system operates. AI can be defined as the simulation of human intelligence on a machine, to make the machine efficient to identify and use the

right piece of knowledge at a given step of solving a problem. Amint Konar (1999). One major subfield of AI is ML.

Learning could simply be defined as the process of acquiring knowledge or skill which could be applied in some application areas in futures. The process of making a machine to learn through codes or algorithm to acquire enough skill or knowledge to solve future problems is described as Machines Learning. So, the machine learns whenever it changes its structure, programme or data (based on its input or in response to external information) in such a manner that its expected future performance improves. (Kelvin, 2008).

C. Support Vector Machine

SVM is a typical example of ML algorithm. Support Vector Machine (SVM) was introduced by Boser, Guyon and Vapnik in COLT 92. It is a supervised learning algorithm used for classification and regression. It is referred to as generalized linear classifier. In other words, it is a classification and regression tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding overfitting of the data. (Vikramadiya, 2006).

Support vector machine (SVM) is a Machine Learning algorithm that learns by examples to assign labels to objects. For instances, an SVM can identify a non-fraudulent and fraudulent card (William, 2006). It maximizes a particular mathematical function for a given collection of data. Eventually, the algorithm presents an appropriate classification.

D. Related Works

Machine Learning Principles or tools have been used to carry out a series of research work in the area of disease classification or detection.

Stefanie et.al (2013) presented a comprehensive overview of schistosomiasis and explained the latest trend in the diagnosis and treatment of the disease most especially in children. It also revealed the number of people including young and old, recently treated with praziquantel. The article was a review and hence did not propose a specific methodology in detecting or diagnosing the disease.

Makolo and Akinyemi (2016) researched prediction of the risk of infection of schistosomiasis using machine learning approach. The methodology used was based on Neural Network with a single layer but with little variation in transformation function. The algorithm was implemented using JAVA Programming Language.

Guo Li et.al (2018) made use of three different machine learning models to diagnose patient with advance schistosomiasis residing in Hubei province. The data was collected from a previous study based on Hubei population sample including 4136 advanced schistosomiasis cases. Multivariate Logic Regression (MLR), Artificial Neural Network (ANN), Decision Tree (DT) were used for implementation. Hence, in term of sensitivity, the author established the fact that ANN outperformed the other two models.

Noura (2015) was able to apply Neural Networks in the diagnosis and treatment of heart diseases. The author specifically used backpropagation approach for the classification of two of groups – presence or absence of the diseases. Eventually, about 88% of the testing data set were classified successfully.

Bakpo and Kabari (2015) also presented a research work using Neural Network for diagnosis of skin diseases. Feedforward backpropagation was used to classify the absence or presence of the disease. When supplied with the relevant data, the system was able to record about 90% success.

Amosa et.al (2015) also developed an expert system for diagnosis and management of kidney disease making use of CLIPS expert system version 6.3. The system has about 76 rules and can detect

various types of renal disease.

Ramya and Radha (2016) also diagnosed chronic kidney disease using Machine Learning algorithm. The classification model was able to classify different stages of chronic kidney diseases. The result showed that Radial basic function is better than others with 85% level of accuracy.

Again, Support Vector Machine (SVM) as one of the major tools of Machine Learning has been applied in solving different medical problems.

Deepti and Sheetal (2013) used SVM and ANN for classification of heart disease in an attempt to assist the physicians to achieve a speedy diagnosis with accurate result. The diagnosis time is reduced with more accuracy in the result obtained. It was finally concluded that SVM performs better than ANN.

Prashasti and Disha (2016) predicted the spread of cardiovascular diseases using SVM and Bayesian classification. The research work predicted accuracy, and sensitivity using SVM and Bayesian classification. The research was able to predict whether a person has heart disease or not. Accuracy graph shows that SVM is better than Naive Bayes.

Shanshikant, Cheta and Ashak (2011) developed a heart disease diagnosis system using SVM. It is an expert system that can decide what type of heart disease a patient suffers for. In the research work, it was established that SVM with sequential minimum optimization is as good as Radial Basis function for diagnosis of heart disease.

E. Significance of the Research

This work analyses the predictive factors in the dataset to establish the spread of schistosomiasis amongst different age groups, across continents and maximally evaluates the six (6) SVM models in term of speed, accuracy and processing time. The result obtained will serve as an eye opener to the researchers working in the area of machine learning, especially those that are interested in using SVM models to carry out further research work in the area of identification, classification and prediction of schistosomiasis and other related diseases. The result obtained, will serve as a good platform for further research. work.

II. METHODOLOGY

A. Data

The research data was compiled across 4 age groups (5 - 24, 25 - 49, 50 - 74, above 74), sex (male and female), 3 different levels of exposure (2, 5, 10 for low, average and high). Following this trend, 3, 2, and 1 were assigned to age group 25 to 49, 50 to 74 and 1 to 4 respectively. 1 was also assigned to those that are above 74 years of age. This means that age groups 1 to 4 and those above 74 are in the same group simply because they are too young or too old to make attempt to swim in the river where they can easily be infected with schistosomiasis. In term of gender, males are perceived to swim in rivers than females and therefore stand a higher risk of contracting the disease and therefore assigned a higher risk value of 2 while female gender was assigned 1.

Considering the continent's level of development, Africa is at the highest risk of contracting the disease from the river and therefore assigned the highest risk factor of 4, followed by Asia, South America and others with risk factors of 3, 2 and 1 respectively. For the level of exposure, the values used are 2, 5 and 10 for low, average and high level of exposure.

B. Method

We tested various SVM models using MATLAB machine learning toolbox to classify the susceptibility of man to schistosomiasis infection. Five different predictors which include – age, location, sex, exposure and calculated score were used. The predictors were used as input in the classification learner. The scatter dot plot of the predictors.i.e. the plot of the calculated score against the other four predictors are shown in fig1. The six SVM models used are Linear, Quadratic, Cubic, Fine Gaussian, Medium Gaussian and Coarse Gaussian. We selected the six SVM models and used MATLAB to implement the classification model. Each algorithm was trained using the data set and the result of the training for each model was

evaluated using ROC curve, Confusion matrix and parallel coordinate plot.

III. RESULTS

Evaluation results obtained from each model using ROC curve, Confusion matrix and parallel coordinate plot are shown in fig 2-7. The performance of the models was also compared using accuracy, speed and training time as shown in table1. Table 1 further explains the model performance with further details depicted in fig 8-10.

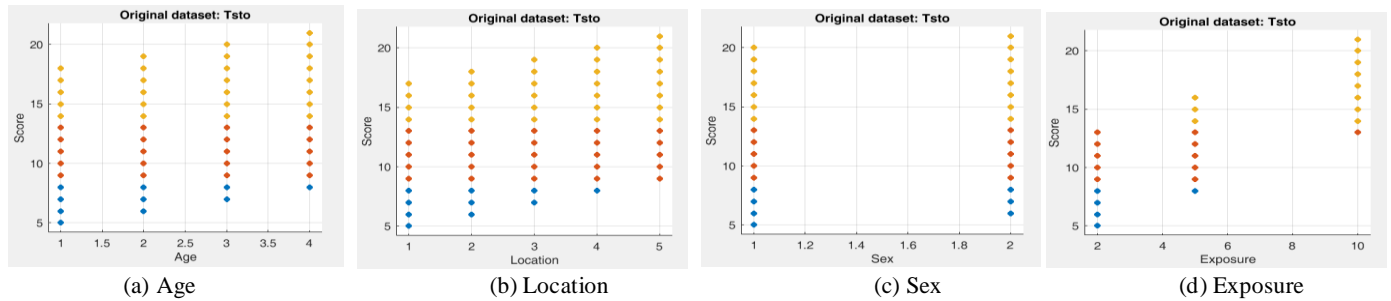


Fig.1. The Scatter Dot Plot of the Predictors

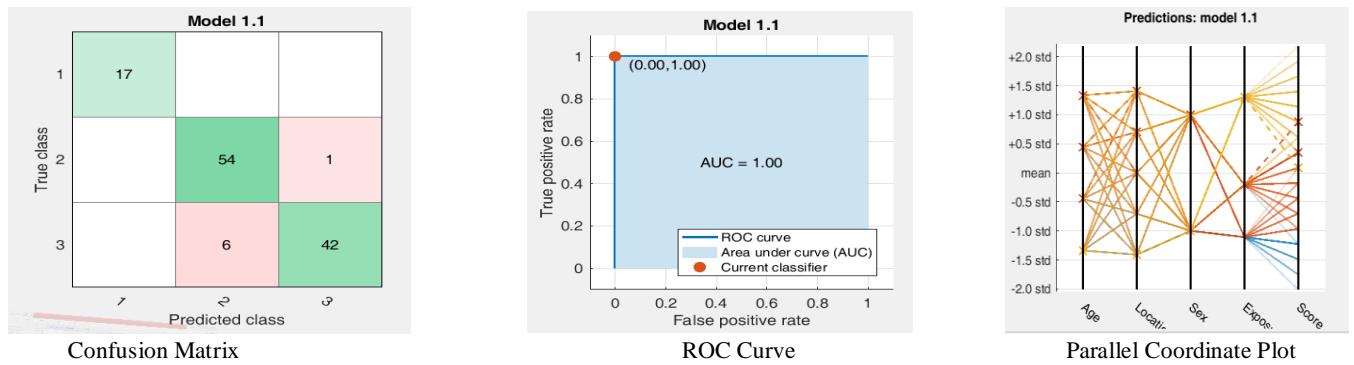


Fig.2. Confusion Matrix, ROC Curve and Parallel Coordinate Plot of the Linear Model

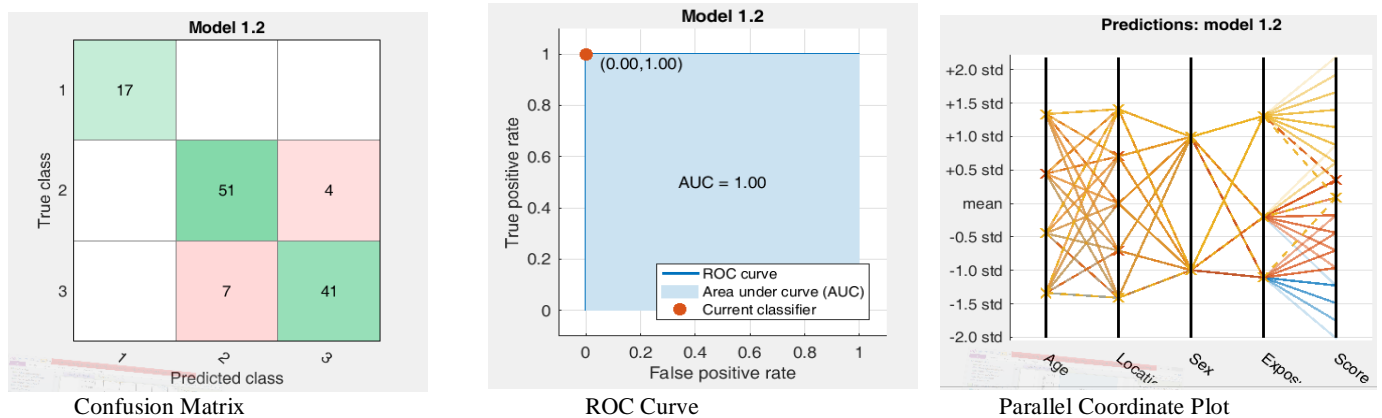


Fig.3. Confusion Matrix, ROC Curve and Parallel Coordinate Plot of the Quadratic Model

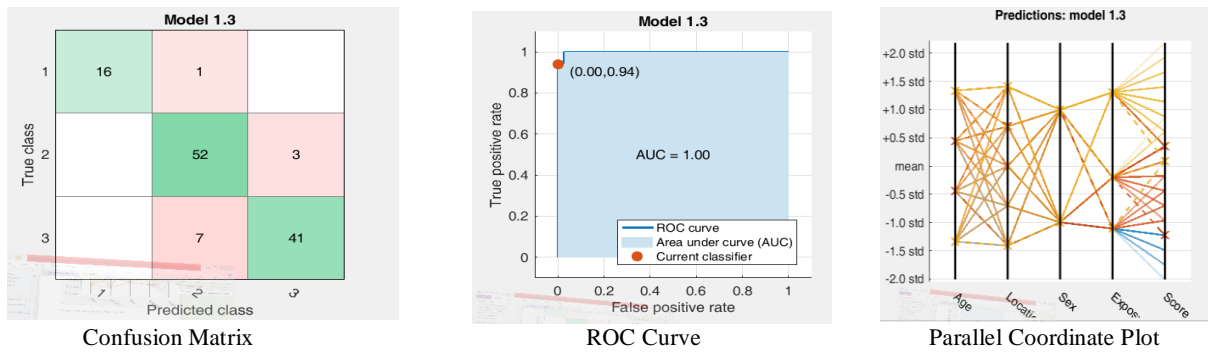


Fig.4. Confusion Matrix, ROC Curve and Parallel Coordinate Plot of the Cubic Model

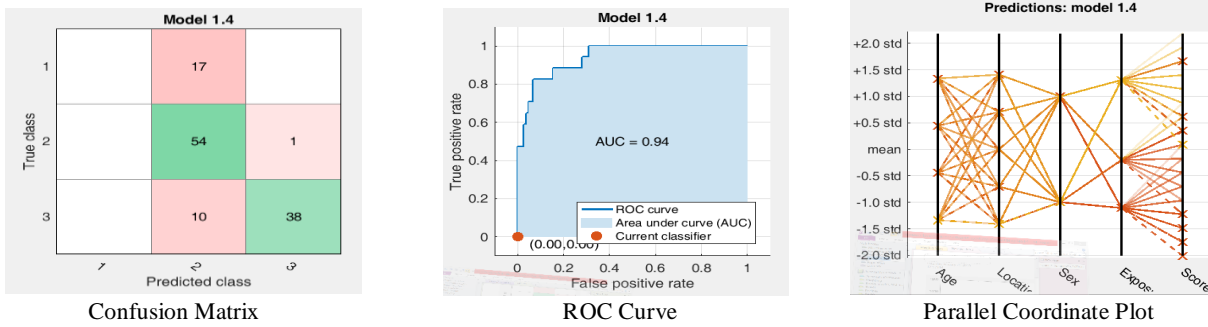


Fig.5. Confusion Matrix, ROC Curve and Parallel Coordinate Plot of the Fine Gaussian Model

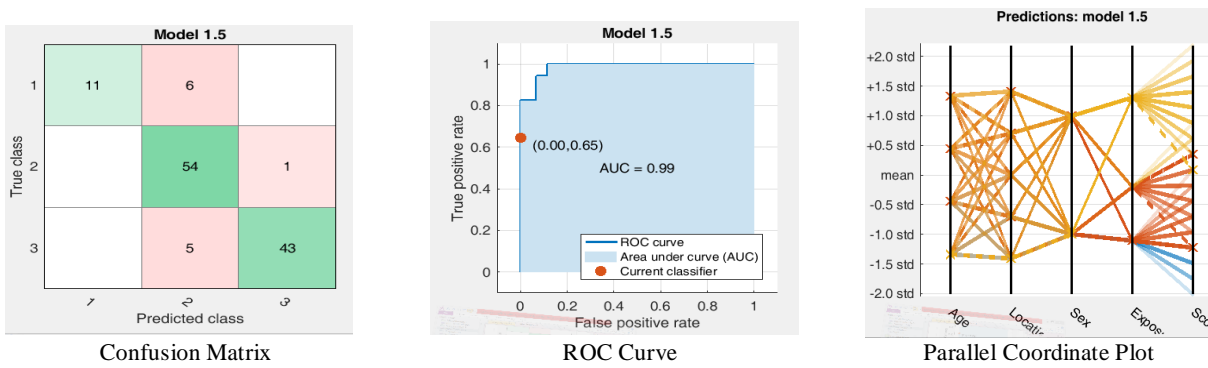


Fig.6. Confusion Matrix, ROC Curve and Parallel Coordinate Plot of the Medium Gaussian Model

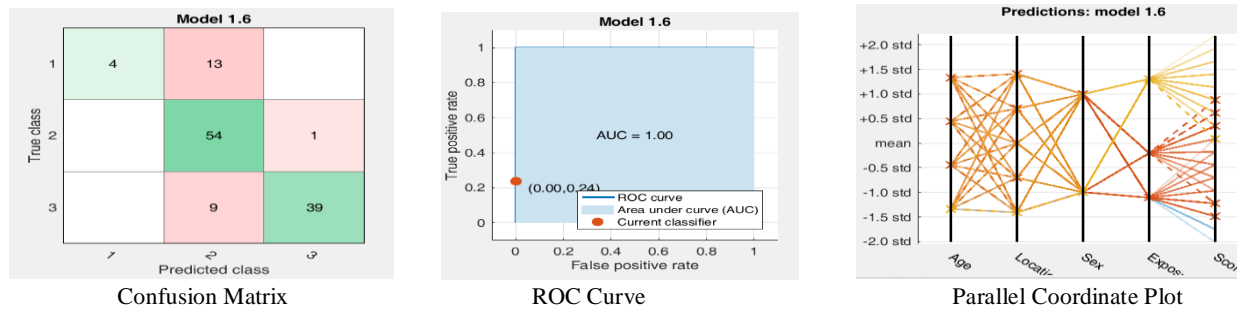


Fig.7. Confusion Matrix, ROC Curve and Parallel Coordinate Plot of the Coarse Gaussian Model

Table I. Model Performance (Accuracy, Prediction Speed and Training Time)

Model No	Model's Type	Accuracy (%)	Prediction Speed (Obs/sec)	Training Time (sec)
1.1	Linear	94.2 (1)	630 (5)	7.98 (5)
1.2	Quadratic	90.8 (2)	1400 (3)	1.34 (4)
1.3	Cubic	90.8 (2)	1700 (1)	1.06 (2)
1.4	Fine Gaussian	76.7 (5)	1600 (2)	1.12 (3)
1.5	Medium Gaussian	90.0 (3)	1700 (1)	1.03 (1)
1.6	Coarse Gaussian	80.8 (4)	1300 (4)	1.06 (2)

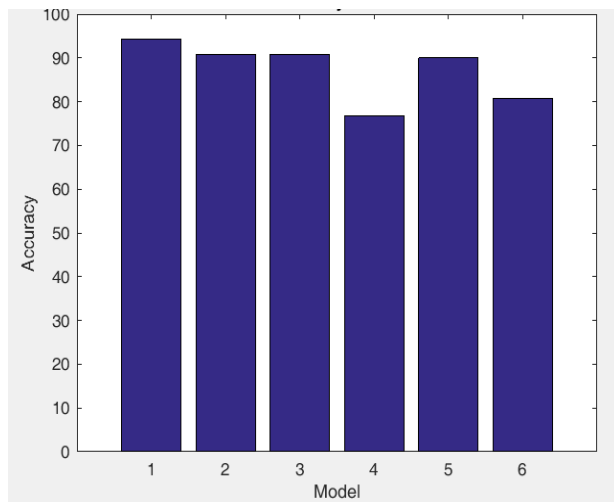


Fig.8. Plot of Accuracy for each Model

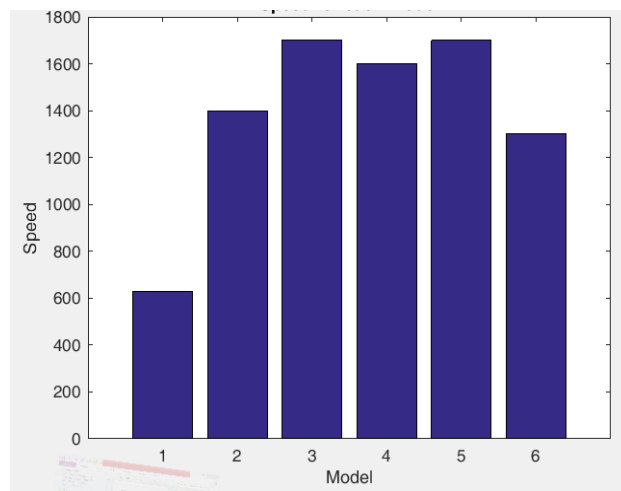


Fig.9. Speed for each Model

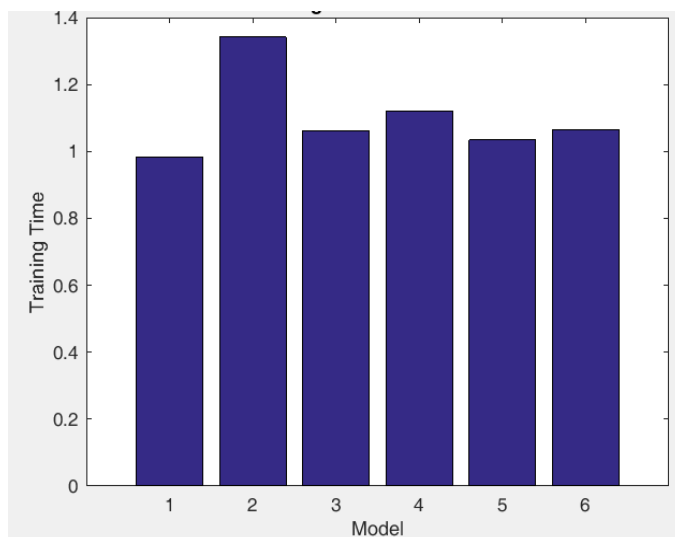


Fig.10. Training Time for each Model

IV. DISCUSSION OF RESULTS

In fig 1 (the scatter plot of the predictors) -yellow, brown and blue points indicates a high, average and low risk of schistosomiasis infection. In Fig 1a, age group 6-24 has the highest tendency of being infected with the disease followed by those between 25 – 49 since the two groups are the most active ones that can easily swim in rivers. The other two groups 1-4 and 50-74 are too young or too old to swim in rivers and hence, they are less susceptible to the infection. In fig 1b Africa (5) being an underdeveloped continent has the highest risk of contracting the disease compared to other continents. It is followed

by Asia (4), South America (3), Developed countries (2) and others (1). Fig 1c shows that males (1) stand at higher risk of being infected compared to female (2). There is a high tendency for males to swim in a river compared to female. The level of exposure is expressed by fig 1d. (5) is the highest level followed by medium (5) and the lowest level of exposure (2).

Fig 2-7 expressed the result obtained after the training with the behaviour of model 1.1, 1.2, 1.3, 1.4, 1.5, and 1.6 respectively. The plots are in the form of CM, ROC, and PCP and table 1 explains the behaviour of the models better. The table was expressed in term of Accuracy, Prediction Speed and Training Time for each of the models. It is clear from the ranking that none of the models is perfect in terms of the accuracy, speed and time. For instance, the Linear model came first in term of accuracy, while cubic and the Medium Gaussian outperforms others in prediction speed.

Medium Gaussian requires lower training Time. It can be observed that only Medium Gaussian performs best in term of Prediction Speed and Training Time but managed to come up at the third position under Accuracy. Therefore, it appears to be the best of all.

V. CONCLUSION

Schistosomiasis is presently one of the diseases that is spreading all over the world especially in the Africa sub-region. The disease has in recent time become endemic and destroying the lives of innocent people all over the world. A lot of research is going on to determine how to manage or eradicate or reduce this scourge.

In this research, efforts were made to look into the behaviour of support Vector Machine models in determining the likelihood of being infected. Six models were trained and the result obtained in term of CM, ROC and PCP were interpreted clearly in term of

accuracy, processing, speed and execution time. It was finally concluded that none of the models was perfect but Medium Gaussian appeared to be better than others.

REFERENCES

- [1] K. Amit (1999). Artificial Intelligence and Soft Computing Behavioural and Cognitive Modelling of the Human Brain CRIC Press, Boca ration London, New York, Washington, D. C
- [2] B. M. G. Amosa, O. K. Olalere, K. A. Kawonise, A. O. Fabiyi, and A. A. Fabiyi (2015). Expert System for Diagnosis and Management of Kidney Diseases. International Journal of Computer Trends and Technology (IJCTT) Volume 20 Number 3, December 2015
- [3] F. S. Bakpo and L. G Kabari (2005). Diagnosing Skin Diseases Using an Artificial Neural Network. Available at: <http://cdn.intechopen.com/pdfs-wm/14893.pdf> P. Saville, *Review of Radar Absorbing Materials*, Defence R&D Canada – Atlantic.
- [4] V. Deepti and S. Sheetal (2013). Classification of Heart Diseases using SVM and ANN. International Journal of Research in IJRCCCT Computer & Communication Technology Vol.2, Issue 9
- [5] L. Guo, Z. Xiaorong, L. Jianbiyi, H. Z. Yhanqichen, C. Yanyan, L. Jianhua, J. Hengbo, Y. Junsing, and N. Shaofa (2018) PLOS Neglected Tropical Diseases
- [6] P. M. Kelvin (2008). Machine Learning a Probabilistic Perspective. The MIT Press Cambridge, Massachusetts, London, England
- [7] U. A. Makolo and M. O. Akinyemi (2016) Predicting the Risk of Infection with SCHISTOSOMA HEAMATOBIUM using ML. International Journal of Computer Application (0975 – 8887) Volume 136, No 8 February 2016
- [8] A. Noura (2015). Heart Diseases Diagnoses using Artificial Neural Network, Network and Complex Systems, ISSN 2224-610X (Paper) ISSN 2225-0603 (Online), Vol.5, No.4, 2015
- [9] K. Prashasti and R. S. Disha (2016). Prediction of Cardiovascular Diseased using Support Vector Machine and Bayesian Classification. International Journal of Computer Application (0975 – 8887) Vol 156, No 2
- [10] H. Reynold and W. ArveLee (2019). Station of Schistomiasis Elimination in the Caribbean Region. Tropical Medical and Infection Disease.
- [11] S. Ramya and N. Radha (2016). Diagnosis of Chronic Kidney Disease using ML Algorithm. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 4, Issues 1,
- [12] G. Shashikant, P. Chetan, and G. Ashok (2011). Heart Disease Diagnosis using Support Vector. International Conference on Computer Science and Information Technology (ICCSIT'2011) Pattaya.
- [13] K. Stefanie, L. B. Soren, J. I. Katrin, K. Jennifer and U. Jurg (2013). Diagnosis and Treatment of Schistosomiasis in Children in the Era of Intensified Central Expert Reviews.
- [14] J. Vikramadiya (2006). Support Vector Machine. School of FFCS Washington State University, Pulman 99164.
- [15] S. N. William (2006). What is a Support Vector Machine PRIMEK: Computation Biology/Mature Biotechnology Volume 24, No 12 .

BIOGRAPHIES



OLANLOYE ODUNAYO received B.Sc. Mathematical Science (Computer Science option) from University of Agriculture, Abeokuta in 1994 and M.Sc. and Ph.D. in Computer Science from Nnamdi Azikiwe University Awka in 2002 and 2017 respectively. He also received Post Graduate Diploma in Education from University of Ibadan in 2004. He has authored and co-authored various books in Computer Science and has published quite a no of articles in reputable local and international journals. He is highly experienced in teaching Computer Science courses at various levels. He is currently a Senior Lecturer in the Department of Computer Science and Information Technology, Bowen University, Iwo, Osun State, His research interest include neural network, fuzzy logic, machine learning, deep learning.



OLASUNKANMI OLAWUMI bagged a B.Tech. Computer Science at Ladoke Akintola University of Technology, in 2018 and N.C.E. (Computer Science) from Emmanuel Alayande College of Education Oyo, in 2010. In 2019, She was a Program Assistant and Machine Learning Researcher at Data Science Nigeria. She is currently a Network Administrator and Machine Learning Researcher at Ladoke Akintola University of Technology. Her research interests include data mining, machine learning, deep learning and health informatics.



ODUNTAN ODUNAYO is a researcher and academia in Computer Science. She is blessed with PhD (Computer Science) from Ladoke Akintola University of Technology Ogbomoso; M.Sc. (Computer Science) from University of Lagos, Akoka; B.Sc.(Computer Science) from Ogun State University Ago -Iwoye and N.C.E. (Computer Science) from St. Andrews College of Education Oyo. A Principal Lecturer in the Department of Computer Science, The Federal Polytechnic, Ilaro Ogun State. She has published innovating scholar articles in reputable international and local journals including a book of readings. She is experienced in teaching computer science at all levels of education with the use of simple terms that enables students to understand difficult concepts in computing. She is a Full member of Computer Professionals of Nigeria (CPN), Nigeria Computer Society (p89908NCS), and a Professional Member of Nigerian Women in Information Technology (NIWIT). She happily married and blessed with children.