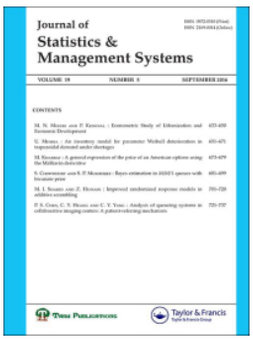

Taylor & Francis
Taylor & Francis Group

Journal of Statistics and Management Systems

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tsms20

Birth weight prediction of new born baby with application of machine learning techniques on features of mother

Zakir Hussain & Malaya Dutta Borah

To cite this article: Zakir Hussain & Malaya Dutta Borah (2020) Birth weight prediction of new born baby with application of machine learning techniques on features of mother, Journal of Statistics and Management Systems, 23:6, 1079-1091, DOI: 10.1080/09720510.2020.1814499

To link to this article: https://doi.org/10.1080/09720510.2020.1814499

Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 23 (2020), No. 6, pp. 1079–1091
DOI : 10.1080/09720510.2020.1814499

Taylor & Francis
Taylor & Francis Group

Birth weight prediction of new born baby with application of machine learning techniques on features of mother

Zakir Hussain *
Malaya Dutta Borah †
Department of Computer Science and Engineering
National Institute of Technology Silchar
Cachar
Silchar 788010
Assam
India

Abstract

The degree of malnutrition is very high in India. Early detection of the possibility for a child to be affected by malnutrition can combat the situation to some extent. Birth weight prediction of new born baby is necessary as parent and doctors can prepare themselves for precautionary and curative measures for the development of physical and mental health. In this study, birth weight prediction of new born baby has been carried out using two machine learning techniques called Gaussian Naïve Bayes and Random Forest. These two models have been trained and tested on a self-created dataset containing 445 instances with eighteen numbers of features of mother. The dataset contains a label with two classes: low-weight and normal-weight. We got 86% accuracy for Gaussian Naïve Bayes and 100% accuracy for Random Forest. Both the techniques have shown significant improvement compared to existing studies.

Subject Classification: *68Q32.*

Keywords: *Birth weight prediction, Gaussian naïve bayes, Random forest, Features of mother, Malnutrition, Under-nutrition, Nutritional status.*

1. Introduction

Birth weight can be used as an indicator for the nutritional status of new born babies. Most of the times, only under-nutrition is thought of as

*E-mail: zak08hussain@gmail.com (Corresponding Author)
†E-mail: malayaduttaborah@gmail.com

malnutrition, but actually malnutrition indicates both under-nutrition and over-nutrition [9]. There are a lot of reasons to develop malnutrition. Some of those include intake of inadequate diet or consumption of less calorie, imbalanced calorie intake during the period of suffering from disease, occurred complication due to illness like less absorption of nutrient and excessive loss of nutrient etc. [8]. Malnutrition has high impact on losing physical and mental potential and as a result it decreases the potential to handle stress [11]. Food is the main source of nourishment of the body as it supplies necessary chemicals needed for development and probability of suffering from malnutrition increases if there is a case of refusal to the food [16]. Since 80% of the brain development happens within two years from the birth, hence early identification of the susceptibility toward suffering from malnutrition is very much necessary. If the development in the early ages is disturbed, then the possibility of becoming the victim of the consequences of less development of brain increases. Some other consequences like excessive muscle loss, higher infection, higher complication, increased morbidity and mortality, impaired wound healing are also associated with malnutrition. As a result we will get children with less performance in almost all aspects. That means our country will have less efficient and less compatible human resource that will definitely hamper the economic development of the country [16].

As per the World Health Organization (WHO) report, India has a big number of children and adults that are the victims of malnutrition [13]. India has a very high degree of malnutrition. WHO and UNICEF review of 2018 suggested that the Sustainable Development Goals (SDG) of eliminating all types of malnutrition by 2030 was aspirational but that may not be achieved [12]. As a curative measure, early detection of the susceptibility and providing necessary treatment can be considered. To detect it early, the studies and experiments can be carried out on the mother who is carrying the baby in her womb [16]. We can find some distinct features on the mother who is supposed to deliver a child with a probability of becoming the victim of malnutrition. Those features can be considered for applying machine learning techniques to classify the children to identify nutritional status.

1.1 *Motivation*

From past few years it has been noticed that some unsocial, dangerous and inhuman activities like rape, accident, theft, easy influence by the extremist are increasing. Most of the abduction cases, if analysed well then it could be found that they are suffering from some mental disorders in some form. Most of the accidents are caused by drug addicts

or some other type of addicted people. This drug addiction or other types of addictions can be related to malnutrition. The thought for the theft also come from some kind of social factors or mental issues or upbringing and types of people they are surrounded by. This type of thoughts only comes into a mind residing in unhealthy body. People with low self-esteem are easily influenced by the extremist. There are many factors for this but malnutrition can also be considered as one of those. Most of the cases, it has been seen that they are affected by some mental or physical disorders. Malnutrition can be thought of as a cause of such disorders. It is always said that a healthy mind resides in a healthy body. So, for a healthy body there must have proper balance in the nutrients. The solution to the consequences of malnutrition in the early ages will reduce the rate of these dangerous activities significantly.

1.2 *Research Objectives*

Many theoretical researches are carried out in this problem. Till date many causes for this problem have been assumed and medical science is trying to solve the problems with their traditional approaches that are very time consuming and costly for the people underlying Below Poverty Line (BPL) category. It has been seen that use of improved techniques in this area is very less. Application of machine learning techniques can be used to address this problem and that will reduce the cost and time significantly. Machine learning techniques have the potential to predict the nutritional status in new born with high accuracy and doctors can take better decision for right treatment. In this research, machine learning techniques have been used to get prediction with high accuracy using some important features. Since this study is considering the prediction for the new born babies, so the features of mother will be the base.

The main objective of this research is to alert the parent and the doctors about the health status by predicting the weight of a new born baby. The aim of this research is to find the solution for the following research questions to achieve the above objectives:

1. Is there any effect of birth weight in the development of physical and mental health?
2. Is it possible to predict birth weight of new born baby by using features of mother?
3. Can machine learning techniques correctly predict the birth weight?

The first and the second questions are totally based on medical science only. To get the answers to these questions, two Medical and Health Officers have helped a lot. To answer the third question, data were collected from Geramari MPHC, Dhubri, Assam and Kazigaon SD, Kokrajhar, Assam and machine learning techniques were applied.

2. Literature Review

Literature review for this study has focused on the prediction of birth weight of new born babies. Some surveys have also been carried out for the babies under the age of five years to extract the analysis report. The study has covered a vast area related to prediction and classification considering different aspects. The study in [1] is about the prediction of birth weight of infant in maternal hypertensive and non-hypertensive condition. They have used WEKA for performance measurement of Naïve Bayes method [1]. The study in [2] is about a model based on the logical decision tree algorithm and decision tree algorithm. Their claim indicates that logical decision tree algorithm had the highest predictive capabilities in terms of recall [2]. Along with that they have mentioned that the model based on the decision tree algorithm with low pruning had the highest precision [2]. The study in [3] is about the assessment of maternal risk factors linked with the neonates having low birth weight. They have also shown a comparison between Random Forest and Logistic Regression. They had considered 600 volunteer pregnant women to carry out the experiment [3]. They have identified four top rank variables like age of pregnancy, BMI during third trimester of pregnancy, age of mother and BMI during first trimester of pregnancy [3]. Their claim shows that Random Forest outperformed Linear Regression. The study in [4] is about prediction and classification of low birth weight data in Indonesia. Binary Logistic Regression and Random Forest approach have been used in IDHS (Demographic and public Health Survey in Indonesia) data of 2012 [4]. They have carried out the experiment considering the features like place of residence, time zone, mother's education level, father's education level, wealth index, age of mother, job of the mother, and number of children [4]. Their claim indicates that Linear Regression showed good performance in prediction, but poor performance in classification but Random Forest had good performance for both prediction and classification [4]. The aim of the study in [5] is to compare logistic regression and data mining techniques in identifying promising predictor variables as well as to come up with a decision support system to help the physicians for making better decision

in case of low weight child birth [5]. They have used the methods like Logistic Regression, Support Vector Machine and Neural Network, Naïve Bayes, Decision Tree, Random Forest, and Data mining techniques [5]. Their experiment was carried out on data from Baystate Medical Centre, Springfield, Massachusetts of 1986. They used 189 instances with 11 attributes like ID-identification number, mother's age in years (AGE), the weight before pregnancy (LWT), number of doctor visits during the first trimester of pregnancy (FTV), race (RACE), lifestyle information such as smoking (SMOKE), a history of previous preterm delivery (PTL), the existence of uterine irritability (UI), and hypertension (HT) [5]. Their study had identified some highly influenced variables like- Mother's last weight before becoming pregnant, Mother's age, Number of doctor visits during the first trimester, parity to predict LBW [5]. The aim of the study in [6] is to extract useful information from health indicators of pregnant woman for early detection of potential low birth weight cases using machine learning techniques [16]. They have used Bayes minimum error rate on Indian healthcare data for construction of decision rules and they have used 18 attributes and got an accuracy of 96.77% [6]. The study in [7] is about development of Artificial Neural Network for predicting birth weight. They used Multi-layer concept topology considering the features like age, smoke, race, weight (lbs) before pregnancy, uterine irritability, number of doctor visits in 1st trimester, hypertension and claimed 100% accuracy. The study in [10] aims to analyse malnutrition on the basis of food intake, wealth index, age group, educational level, occupation and they have applied Decision tree, Artificial Neural network in a dataset of family health survey with 254 instances and 9 attributes [10]. They have carried out the experiment for children under the age of five and found 68.50% accuracy for ID3, 77.17% for Random Forest and 77.17% for Multilayer perceptron.

3. Methodology

3.1 *Proposed Architecture*

The architecture in Figure 1 shows the work flow starting from the data collection to the performance evaluation of machine learning techniques that are used in this study. The collected data have been arranged to create the dataset. After creating the dataset, it has been processed to select the required features of mother to apply in the machine learning techniques. After selecting appropriate features, the machine learning techniques

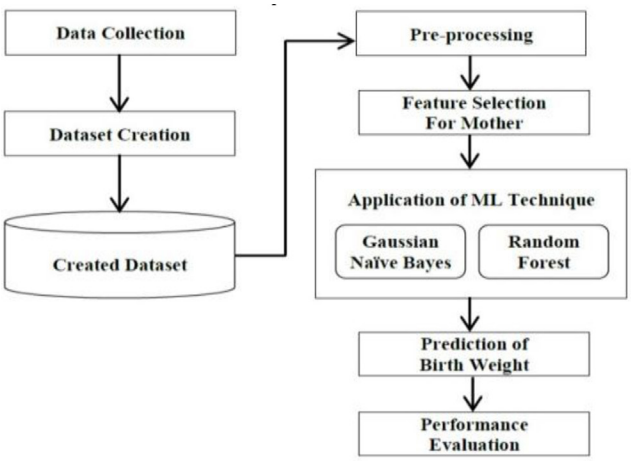

Figure 1

Block diagram of the work flow

have been applied on the dataset. The applied techniques have predicted and created the result and based on the result, the performance evaluation has been carried out for those methods using statistical tools.

3.2 *Data Collection and Dataset Creation*

The data required for this experiment has been collected from two government health centres. Since the study is about identifying birth weight of new born babies, so we have gathered data of mothers. During data collection, it has been seen that the health centres keep record of mother up to the born baby reaches 5 years of age. We have considered the data only up to the birth as our objective is to predict the birth weight of the new born baby. The necessary data from the registers maintained in the health centres were picked up and the dataset has been created with a name CBWDB.csv. The details of the created dataset can be found in Table 1.

3.3 *Pre-processing and Feature Selection*

During the collection of the data from the health centres, we noticed that, there are many columns that keep different types of information about mother and child. As all the information recorded in the registers were not suitable for applying the machine learning techniques, we have selected only eighteen numbers of features and one label for the

Table 1
Overview of created dataset

Name of the dataset:	CBWDB.csv	
No. of instances:	445	
No. of class labels:	02	L and N: L-Low weight, N-normal weight
No. of attributes:	18	
Details of attributes:	SEC	Socio-Economic Condition
	Age(years)	Age of mother at the time of pregnancy
	Height(cm)	Height of mother at the time of pregnancy
	BGroup	Blood Group of mother
	Parity	No. of previous pregnancy crossing the period of viability
	ANC	Antenatal Check
	Iwt(kg)	Initial weight of mother
	Fwt(kg)	Final weight of mother (Last ANC)
	IBP_sys	Initial systolic Blood Pressure
	IBP_dias	Initial diastolic Blood Pressure
	FBP_sys	Final systolic Blood Pressure (Last ANC)
	FBP_dias	Final diastolic Blood Pressure (last ANC)
	IHb(gm%)	Initial Haemoglobin level
	FHb(gm%)	Final Haemoglobin level (Last ANC)
	BS(RBS)	Blood Sugar (Random)
	TermPreterm	Term: 37 to 40 weeks, Preterm: <37 weeks
	Sex	Sex of new born baby
	BWt(kg)	Baby birth weight
Missing value:	There were some missing values. Those have been replaced by the average of the nearest values.	
Source:	Geramari MPHC, Dhubri, Assam, India and Kazigaon SD, Kokrajhar, Assam, India	
Creator:	Zakir Hussain, Research scholar, NIT Silchar, Assam, India	

classification. During creation of dataset, the blank spaces were filled up with the average of the values present in the nearby cells.

Eighteen features that were selected for the application of machine learning techniques after discussion with doctors are: SEC, Age(years), Height(cm), Bgroup, Parity, ANC, Iwt(kg), Fwt(kg), IBP_sys, IBP_dias, FBP_sys, FBP_dias, IHb(gm%), FHb(gm%), BS(RBS), TermPreterm, Sex, BWt(kg) [16]. Table 1 contains the details of these features. One more column with name LNH has been included in the dataset to record the birth weight label.

3.4 *Used Machine Learning Techniques*

There are lots of machine learning algorithms to address classification and prediction problems. In this study, two machine learning techniques called Gaussian Naïve Bayes and Random Forest have been implemented in our created dataset.

Gaussian Naïve Bayes: Naïve Bayes classifier is a probability based machine learning model whose crux is based on the Bayes theorem [15]. We can write:

$$P(c \mid F) = \frac{P(A \mid c)P(c)}{P(A)}$$

Here, c has been considered as class variable which is supposed to be predicted and A has been considered as Attributes or parameters or features [16]. In Bayes theorem the attributes or parameters or features or predictors are considered as independent and the presence on one doesn't affect the other and hence they are called naïve.

A can have many instances. Let us consider that A has n number of instances denoted by $a_1, a_2, ..., a_n$. So, substituting A by its instances, the above equation can be written as:

$$c = \arg\max_c P(c)\prod_{i=1}^{n} P(a_i \mid c)$$

For Gaussian Naïve Bayes classifier, the above equation changes to an equation containing Gaussian function and it looks like:

$$P(a_i \mid c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(\frac{(a_i - \mu_c)^2}{2\sigma_c^2}\right)$$

Gaussian Naïve Bayes classifier has been used as it works well for both binary (two class) classification and multi-class classification. It

works simply by calculating the mean and standard deviation. It does not overlook any features for estimating the probability.

Random Forest: Random forests are also called as random decision forests. Decision trees are combined to form random forest [17]. The constituent of random forests are the decision trees. Individual decision trees of random forest operate as ensemble. Each individual tree in the random forests denotes a predicted class [14]. The class level with highest vote becomes the prediction of the model.

Bootstrap aggregating or bagging is used for training the algorithm of random forests. For a given training set $T = t_1,\ldots,t_n$ with responses $R = r_1,\ldots,r_n$, selection of random sample as a replacement of training set is carried out by repetition of bagging (K times). Then the trees are fitted to these samples.

For $k = 1, \ldots, K$

- Replacement of sample, n examples for training from T, R call these T_k, R_k.
- Training of regression or classification tree X_k on T_k, R_k.

After training, new unseen samples t' can be predicted by averaging the predictions from all individual regression trees on t'.

$$\hat{X} = \frac{1}{K}\sum_{k=1}^{K} X_k(t')$$

Majority vote is considered in case of classification trees.

Along with this, estimation of uncertainty for prediction can be performed by the standard deviation of the predictions from all the individual regression trees on t'.

$$\sigma = \sqrt{\frac{\sum_{k=1}^{K}\left(X_k(t') - \hat{X}\right)^2}{K-1}}$$

The above is about the bagging algorithm. Random forests uses feature bagging, i.e. at each candidate split, a random subset of the features is used. This is done to get rid of the association of the trees in ordinary bootstrap sample.

Random forest has been used as it has the ability to increase the power of algorithm and most importantly it prevents over fitting. It has the ability to handle multiple input features without feature deletion.

Table 2
Performance of used Machine Learning techniques

ML Algorithm	Accuracy (%)		Precision		Recall		F1-Score	
	Training	Testing	L	N	L	N	L	N
Gaussian Naïve Bayes	86	86	0.64	0.91	0.64	0.91	0.61	0.91
Random Forests	100	100	1	1	1	1	1	1

4. Results and Discussions

The background study reveals about the limited works that are carried out in predicting the birth weight of new born babies. Most of the studies are about the prediction and analysis of nutritional status of the child under the age of five years. Here, a novel study of predicting whether a new born baby will be under-weight or normal using the features of mother has been carried out. This study has assumed that certain features of mother during the pregnancy (from conceive to birth) can give a platform to predict the birth weight of new born babies and both the machine learning algorithms have been implemented in our created dataset with a ratio of 70:30 for training and test set. We have got impressive results as we expected. The Table 2 shows the performance of those two algorithms trained and tested on this dataset.

Now let us see whether the answers for the research questions are obtained or not. The first question was about effect of birth weight on the development of physical and mental health. We have found that birth weight has a very big role to play in the development of physical and mental health. The answer was obtained by thorough study and discussion with doctors.

The second question was about the possibility of predicting birth weight of new born baby by using features of mother. Since, before birth the baby is totally dependent on mother for nutrition, hence, physical and mental status of the mother will have significant influence on the baby as well. The physical status of mother can be well understood by investigating some features of mother. So, it is clear that by analysing the features of mother, the nutritional status of the new born baby can be predicted.

The third question was about the correctness of the prediction of birth weight by machine learning techniques. To get the answer to this question, two machine learning algorithms have been trained and tested. It has been found that Random Forest outperforms Gaussian Naïve Bayes in terms of Accuracy, Precision, Recall and F1-Score. From Table 2, it is clear that machine learning techniques can predict the birth weight of new born with very good accuracy.

4.1 *Comparison with other's findings*

It has been found that limited studies have been carried out for prediction of birth weight of new born. Most of the studies are for the nutritional status of the child below five years of age. Different researchers have implemented different algorithms on different datasets and varied number of attributes. One of the researchers has obtained 95% of accuracy for Random Forest [3], another researcher has used Minimum Error Rate classifier and got an accuracy of 96.77% [6], another researcher got 80.372% of accuracy using Naïve Bayes classifier [1]. A researcher has claimed 100% accuracy using Artificial Neural Network [7]. In our study we have implemented two algorithms. We have got an accuracy of 86% for Gaussian Naïve Bayes and 100% for Random Forests.

5. Conclusion

Birth weight plays vital role in the development of physical and mental health. For a peaceful society, children with proper physical and mental development are essential. Early prediction of the nutritional status of new born can be a signal for the parent and doctors to take preventive and curative measures to help the baby to be healthy. In this study two machine learning techniques have been implemented using a small dataset. Both the techniques have shown impressive accuracy in prediction. As a future work, the dataset will be expanded and other state-of-the-art machine learning techniques like neural network, deep learning can also be implemented.

Acknowledgment

The data required for this work has been provided by Geramari MPHC, Gauripur, Dhubri, Assam, India and Kazigaon SD, Kokrajhar, Assam, India. Authors would like to offer heartfelt thanks to the authorities of both the organizations.

References

[1] A. Desiani, R. Primartha, M. Arhami, O. Orsalan, 2019. Naive Bayes classifier for infant weight prediction of hypertension mother. IOP Conf. Series: *Journal of Physics*: Conf. Series 1282 (2019) 012005, doi: 10.1088/1742-6596/1282/1/012005.

[2] S. Aruna, P. Sudha, 2016. An efficient identification of malnutrition with unsupervised classification using logical decision tree algorithm. *International Journal of Innovative Research in Computer and Communication Engineering*, 4(2), pp. 1311-1315, doi: 10.15680/IJIRCCE.2016.0402036.

[3] P. Ahmadi, H. A. Majd, S. Khodakarim, L. Tapak, N. Kariman, S. Safshekan, F. Pazhuheian, 2017. Prediction of low birth weight using Random Forest: A comparison with Logistic Regression. *Journal of Paramedical sciences* (JPS), 8(3), ISSN: 2008-4978, pp.36-43.

[4] A. Faruk, E. S. Kahyono, N. Eliyati, I. Arifieni, 2018. Prediction and Classification of Low Birth Weight Data using Machine Learning Techniques. *Indonesia Journal of Science & Technology*, 3(1), pp.18-28.

[5] D. Senthilkumar, S. Paulraj, 2015, March. Prediction of Low Birth Weight Infants and Its Risk Factors Using Data Mining Techniques. In Proceedings of the 2015 International Conference on Industrial Engineering and Operations Management, Dubai, United Arab Emirates (UAE).

[6] A. R. Yarlapati, S. R. dey, S. Saha, 2017. Early Prediction of LBW cases via Minimum Error Rate Classifier: A Statistical Machine Learning Approach, IEEE.

[7] M. Al-Shawwa, S. S. Abu-Naser, 2019. Predicting Birth weight Using Artificial Neural network. *International journal of Academic Health and Medical Research* (IJAHMR), ISSN: 2000-007X, 3(1), pp. 9-14.

[8] P. Watcharapasorn, N. Kurubanjerdjit, 2016. The surgical patient mortality rate prediction by machine learning algorithms. In IEEE 2016 13th International Joint Conference on Computer science and Software Engineering (JCSSE).

[9] S. N. Ariyadasa, L.K. Manasinghe, S. H. D. Senanayake, MGNAS Fernando, 2013. Knowledge extraction to mitigate child malnutrition in developing countries (Sri Lankan context). In IEEE 2013 4th International Conference on Intelligent Systems, Modeling and Simulation.

[10] D. Thangamani, P. Sudha, 2014. Identification of Malnutrition with use of Supervised Data mining Techniques –Decision Trees and Arti-

ficial Neural Networks. *International Journal Of Engineering And Computer Science*, 3(9), ISSN: 2319-7242, pp. 8236-8241.

[11] Dalal A. D C Dutta's Textbook of Gynaecology and Textbook of Obstetrics. *J Obstet Gynaecol India*. 2016;66(4):303–304. doi:10.1007/s13224-016-0848-4.

[12] S. Soumiya, H. Rajkumar, P. Amika, K. Nicholas, L. Avula, L. Thingnganing, . . .D. Lalit. 2019. The burden of child and maternal malnutrition and trends in its indicators in the states of India: the global burden of disease study 1990–2017.The Lancet Child & Adolescent Health, 3: 855 – 870.

[13] WHO. 2019. 2018 global nutrition report: Shining a light to spur action on nutrition. Retrieved from https://www.unicef.org/rosa/press-releases/2018-global-nutrition-report-reveals-malnutrition-unacceptably-high-and-affects (Last accessed 26 December2019)

[14] Ali Ünlü & Waqas Ahmed Malik (2008) Modified item tree analysis of inductive reasoning data, J*ournal of Interdisciplinary Mathematics,* 11:5, 641-652, DOI: 10.1080/09720502.2008.10700589.

[15] Jean-Charles Lamirel & David Reymond (2014) Automatic Websites Classification and Retrieval using Websites Communication Signatures, *COLLNET Journal of Scientometrics and Information Management,* 8:2, 293-310, DOI: 10.1080/09737766.2014.954855.

[16] Zakir Hussain and Malaya Dutta Borah, (2020). "Nutritional Status Prediction in Neonate Using Machine Learning Techniques: A Comparative Study". In: Bhattacharjee A., Borgohain S., Soni B., Verma G., Gao XZ. (eds). International Conference on Machine Learning, Image Processing, Network Security and Data Sciences - MIND 2020. Communications in Computer and Information Science, 1241. Springer, Singapore. Link: https://link.springer.com/chapter/10.1007%2F978-981-15-6318-8_7

[17] Mohan Kumar, Sunil Kumar Khatri & Masoud Mohammadian (2020) Breast cancer identification and prognosis with machine learning techniques - An elucidative review, *Journal of Interdisciplinary Mathematics,* 23:2, 503-521, DOI: 10.1080/09720502.2020.1731963.

[18] Samta Rani & Sarfaraz Masood (2020) Predicting congenital heart disease using machine learning techniques, *Journal of Discrete Mathematical Sciences and Cryptography,* 23:1, 293-303, DOI: 10.1080/09720529.2020.1721862.