ISSN (e): 2250 – 3005 || Volume, 09 || Issue, 3 || March– 2019 ||
International Journal of Computational Engineering Research (IJCER)

Performance Evaluation of ANN Classifier for Knowledge Discovery in Child Immunization Databases

[1]Arun Singh Bhadwal, [1]Sourabh Shastri, [1]Paramjit Kour, [1]Sachin Kumar, [1]Kuljeet Singh, [1]Monika Kumari, [2]Dr. Anand Sharma, [1]Prof. Vibhakar Mansotra
[1]*University of Jammu, Jammu and Kashmir*
[2]*Guru Kashi University, Talwandi Sabo, Punjab*
Corresponding Author: Arun Singh Bhadwal

ABSTRACT: *Knowledge Discovery in Databases (KDD) is a magnificent process of discovering informative patterns and knowledge from enormous amount of unorganized databases by using the techniques and algorithms of data mining and machine learning. In the present paper, the authors have made an attempt to discover novel knowledge from the child immunization dataset that has been collected from Health Management Information System (HMIS), a web portal facilitated by Ministry of Health and Family Welfare (MoHFW), Government of India. The data consists of diverse health indicators related to the immunization of children and it covers all districts of India. The Artificial Neural Network (ANN) classifier has been used to build a child immunization predictive model based on the available past data to categorize the districts of India into High Focus (HF) and Non-High Focus (NHF) districts. In addition to model building, various measurement methods have been used to evaluate the performance of the predictive model.*

Date of Submission: 28-03-2019 Date of Acceptance: 08-04-2019

I. INTRODUCTION

The various sectors like healthcare, banking, finance, marketing, insurance, education, transport, revenue etc. are generating the digital data at an exponential rate due to the unceasing development and advancement in database technology. In this research paper, the healthcare field has been selected for study in general with exclusive emphasis on the domain of child immunization as child immunization is a highly requisite process of administering vaccines to children for protecting them from infectious diseases.

Therefore it is propagated throughout the world with the conducting of events pertains to the immunization of children and their survival. It is observed that millions of children die every year because of vaccine preventable diseases (VPDs) while a disproportionate number of these children are found in developing countries and recent estimations of the world claim that approximately 34 million children are not completely immunized with more or less 98% of them residing in developing countries [1]. It is hence extremely momentous to provide vaccination to children on time in order to avoid children's suffering from vaccine preventable diseases. Almost all countries have formulated their vaccination schedule that includes vaccine, route, timing, number of doses etc. In India, Ministry of Health and Family Welfare (MoHFW), Government of India is regularly taking various initiatives to cover those children who are living in far flung areas and remain unvaccinated against vaccine preventable diseases. One of such initiative of MoHFW is NHM-HMIS digital database that contains the child immunization data from all over the country in the form of standard and analytical reports in addition to voluminous data of several other healthcare subfields including maternal health, family planning, patient services, blindness control programme, laboratory testing, stock position etc. It is not out of place to mention here that these gigantic datasets at our disposal are of no use until some informative knowledge is not extracted from them. The Knowledge Discovery Process (KDD) along with Data Mining and Machine Learning provides a new generation of computational models for providing powerful analytical solutions for the extraction of hidden informative patterns and knowledge from these expeditiously growing volumes of digital data, thereby facilitating decision making.

There are numerous computational models and techniques (such as classification, clustering, association, time series etc.) provided by Data Mining and Machine Learning to discover hidden knowledge from historical databases but the technique taken under consideration for the present research study is classification, a supervised learning technique of machine learning in order to build national model for child immunization data. The classification technique has numerous algorithms viz. C5.0, Naïve Bayes, Bayesian Tree, ANN, CHAID,

C&RT, QUEST and several others for building classifiers based on massive datasets. In this research paper, the authors have applied Artificial Neural Network (ANN) algorithm of classification on the enormous chunk of child immunization data. ANN is a computational model based on biological neural network that consists of three layers of nodes viz. input, hidden and output and has diverse characteristics including robustness, self-organization, adaptive learning, parallel processing, distributed storage and fault tolerance [2]. These qualities of ANN make it immensely powerful and useful in the field of knowledge discovery, data mining, machine learning and artificial intelligence. In this study, ANN has been used to build a predictive model based on the child immunization data of India.

II. REVIEW OF LITERATURE

S. Shastri et al. [3] have proposed a data mining based model for the classification of child immunization data of Jammu and Kashmir State, India into priority and non-priority districts using Naïve Bayes classification algorithm. They have collected data from NHM-HMIS web based portal, provided by the Ministry of Health and Family Welfare, Government of India. Another study conducted by M. Hemalatha and S. Meghala [4] reflected the effective use of decision tree and ANN to handle tremendous immunization data. They have applied the concepts of data mining classification using ANN on data of children with immunization details that would help to administer the health care. In another research, A. Meleko et al. [5] analyzed the data of 12-23 months old children belonging to Mizan Aman town and assessed the factors that affect child immunization and came to a conclusion that the education level of mother/caretaker, place of delivery, knowledge about vaccine and vaccine preventable diseases showed note worthy association with full child immunization.

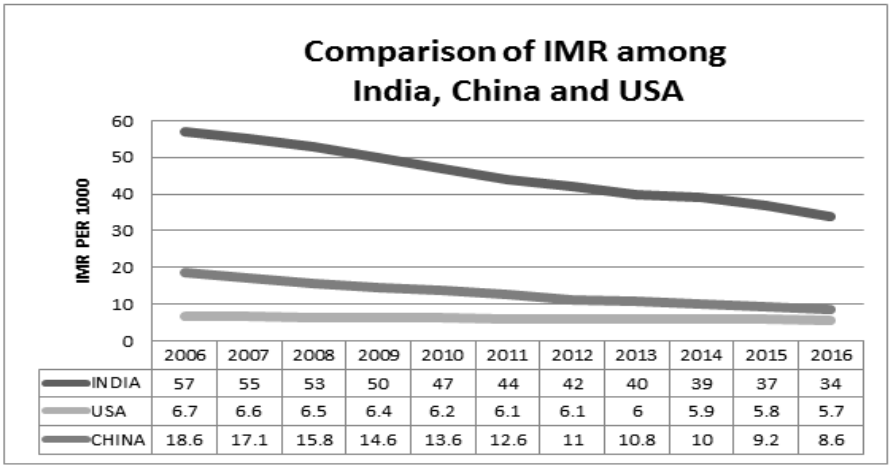
Another research done by P. Gaur [6] indicated the use of neural networks in data mining to discover proficient and informative patterns. R. Revathi and T. P. Senthilkumar [7] also examined the potential use of classification on vast volume of immunization data. Another researcher, G. P. Zhang [8] highlighted some of the most vital developments in neural network classification research especially posterior probability estimation, association between neural and conventional classifiers, feature selection, learning and generalization tradeoff in classification etc, K. Amarendra et al. [9] in their research paper laid emphasis on the detailed study of data mining using ANN. It is presumed in their study as a conclusion that the robustness, self-organizing, adaptive, parallel processing, distributed storage and high degree of fault tolerance, characteristics of neural make it so useful in the field of data mining. K. Streatfield et al. [10] illustrated that formal education of women or education related to immunization is important factor for child immunization by taking the data of two villages of western district of Yogyakarta, Indonesia.

M. Charles Arockiaraj [11] proffered that ANN is a good tool for data mining practitioners. Sonalkadu and S. Dhande [12] elaborated that advantages of ANN like affordability to the noise data, low error rate and rule extraction algorithm increase the use of ANN in data mining. S. B. Maind and P. Wankar [13] explained ANN, its working and training phases exhaustively. Furthermore, they described ANN as an analytical powerful model that is an alternate to the conventional techniques for model building. G. Tiwary [14] explained the importance of ANN for extracting symbolic rules from trained model by using Extraction of Symbolic Rules from ANNs (ESRNN). He further emphasized that this weight freezing is based on constructive and pruning algorithm that worked in three phases. Appropriative network architecture is determined by the first and second phase whereas in the third phase, symbolic rules are extracted using the frequently occurred pattern based rule extraction algorithm by examining the activation values of the hidden nodes.

III. PRESENT STATUS OF CHILD IMMUNIZATION

The immunization system of India was started in 1985 and it is one of the largest in the world covering about 2.7 crore children every year. In spite of being running from last more than 30 years, only 65% of the children in India received all vaccines during their first year of life [15]. According to the report of United Nation Interagency Group for Child Mortality Estimation (UNIGME), about 8,02,000 infant deaths occured in India in 2017, that is, a decline from 8,67,000 in 2016 which is about 7.5% of decline from 2016 to 2017. This achievement has been made possible by Government of India by showing great interest and running many successful schemes like Janani Suraksha Yojana (JSY), Janani Shishu Suraksha Karyakaram (JSSK), Mission Indradanush, Navjat Shishu Suraksha Karyakaram (NSSK) etc. On December 25, 2014, Mission Indradhanush was initiated by the Ministry of Health and Family Welfare (MoHFW) as a plan to propagate all over the country to immunize all unimmunized and partially immunized children and pregnant women by 2020 [16]. The program was started from April 2015 to July 2017, where the total number of children who were vaccinated was around 25.5 million and 6.9 million pregnant women were also vaccinated. This contributed to an increase of 6.7% in full immunization coverage with 7.9% in rural areas and 3.1% in urban areas after the first two phases [17]. The target of achieving 90% immunization in the country by 2020 is reduced to 2018 by launching Intensified Mission Indradanush (IMI) in October 2017.

Janani Shishu Suraksha Karyakarm (JSSK) was launched on 1st June 2011 and it provides free facilities to all women who opted government medical institutions. The objective of JSSK is to increase delivery at government institutions. According to the report of National Health Mission, JSSK benefits approximately more than 12 million women. Another important initiative was Mahila Arogya Samiti (MAS) that played a very important role for bringing awareness among women. MAS played a significant role in the massive campaign in the urban Odisha which helped the state to achieve an overall immunization rate of 98.2% and an urban immunization rate of 95.7%, thereby, making it one of the highest performing states in the country [18]. Diseases that are preventable by vaccines are still contributing about 25% of 10 million deaths occurring annually among children below 5 years of age [19]. Although the infant mortality rate of India declined with a very rapid rate but still infant deaths were reported highest in the world followed by Nigeria, Pakistan and the Democratic republic Congo. India is still far behind in terms of IMR from many countries including China and USA as shown in figure 1. From global point of view, India still needs more important and successful implementation steps for improving the condition of child immunization.



	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
INDIA	57	55	53	50	47	44	42	40	39	37	34
USA	6.7	6.6	6.5	6.4	6.2	6.1	6.1	6	5.9	5.8	5.7
CHINA	18.6	17.1	15.8	14.6	13.6	12.6	11	10.8	10	9.2	8.6

Figure 1: Comparison of IMR among India, China and USA

IV. DATA MINING

As due to unceasing development in the database technology, the amount of data in digital form is increasing at dramatic pace, so there is a need of tools and techniques in order to get some lore from this huge amount of data and these results will further help in decision making. Earlier, number of terms were used for the process of extracting useful patterns or trends from data like data mining, data extraction, information discovery, data archaeology and many more and then in 1989 a term KDD was coined for the purpose of knowledge discovery [20]. Both the terms Data mining and KDD are different but have a relationship with each other. The whole process of discovering knowledge from data is known as Knowledge Discovery or KDD. It comprises of number of components, starting from data selection to pattern interpretation whereas data mining is an essential component of KDD for applying algorithms to extract the patterns [20]. The other components of KDD are preprocessing, transformation and integration.

The data mining techniques can be mainly categorized into two categories on the basis of data i.e. predictive and descriptive mining. Predictive mining is also known as supervised learning that uses historical data to perform prediction. Classification, Time series analysis, Regression etc. are some techniques of predictive mining. On the other hand, Descriptive mining also known as unsupervised learning uses real world data in order to perform prediction by finding some kind of relationships in data. Clustering, association rules, summarization etc. are some techniques of descriptive data mining [3]. In the present era, these techniques of data mining mentioned supra are being used ubiquitously for discovering knowledge from huge databases [21].

V. CLASSIFICATION USING ARTIFICIAL NEURAL NETWORK (ANN)

Artificial Neural network (ANN) is a concept that has been derived from biology in computer engineering. The working of ANN is very much similar with human body neurons. ANNs are built out of dense set of simple interconnected units where each unit takes a number of real valued inputs (possibly the outputs of other units) and produces a single real–valued output (which may become the input to many other units) [21]. The main reason to study neural network is to understand the style of parallel computation which is truly inspired by the human brain that computes the results with a big parallel networks. It is very unlike from the way computation is done on schematic serial processors. It is very helpful in numerous situations like matching of handwriting but it is very futile on division of two large numbers just like human brain. Classification plays a pivotal role in the

field of research. Classification of districts into HF and NHF can be performed by the statistical classifiers models like Naïve Bayes, J45 and many more but due to instability and computational complexity in the results, their popularity is hampered. In the recent years, ANN with Multilayer Perceptron has been widely used for the purpose of data classification. Although we cannot compare statistical classifiers and ANN directly as ANN is a non-linear and model free method whereas statistical classifiers are linear and model based. In this paper, data has been divided into three partitions viz. training, test and validation. ANN with Multilayer Perceptron has been applied on training data to firstly learn the patterns of the data. The accuracy of the model has been evaluated on the basis of unseen data in test and validation partitions. The main goal of the final model is to classify the new data into HF and NHF based on the learning from the training data.

VI. DATA COLLECTION

The secondary data related to child immunization of 674 districts of India for the year 2016-17 has been collected from NHM-HMIS portal of Ministry of Health and Family Welfare (MoHFW), Government of India for this research work. The Artificial Neural Network (ANN) algorithm of classification has been applied on the aforementioned data of all the districts of India for the year 2016-2017. The total number of features in the dataset were 37 primarily but with the use of predictor importance, the most 10 important features were extracted for building the model as shown in figure 2. As the classification technique is a supervised learning so the dataset requires a class label having categorical data type for categorizing the input variables into these classes by building a classifier. The class label used for the study is Focus (FS) with two categorical values viz. High Focus (HF) and Non High Focus (NHF). The HF reflects weaker districts of the country that requires more attention in reference to primary health services as compared to NHF. An attempt has been made to build a predictive model using ANN that categorizes the districts of India into HF and NHF areas based on child immunization data that has been collected from NHM-HMIS repository.

S.no.	Description	Indicator
1.	Number of infants (0 to 11 month old) received OPV0(Birth Dose)	OPV0_0_11
2.	Number of children (more than 5 years old) given DT5	DT5_5Y
3.	Total number of children (9 to 11 months old) fully immunized(BCG+DPT123/Pentavalent123+OPV123+Measles) during the month	T_FI_9_11
4.	Number of Infants (0 to 11 months old) received DPT1 immunization	DPT1_0_11
5.	Total number of female children (9 to 11 months old) fully immunized (BCG+DPT123/Pentavalent123+OPV123+Measles) during the month	F_FI_9_11
6.	Number of Infants (0 to 11 months old) received Hepatitis-B3 immunization	HEPB3_0_11
7.	Number of Infants (0 to 11 months old) received DPT3 immunization	DPT3_0_11
8.	Number of Infants (more than 16 months old) received Measles immunization (Second Dose)	MEAS2_M16
9.	Number of Infants (more than 16 months old) received DPT Booster dose	DPT_B_M16
10.	Number of Infants (0 to 11 months old) received Pentavalent3 immunization	PENTA3_0_11

Figure 2: Ten Important Indicators

VII. NATIONAL MODEL FOR CHILD IMMUNIZATION DATA USING ANN

The multilayer preceptron has been applied on the most 10 important features for the clasification of districts into HF and NHF. Multilayer preceptron (MLP) has total eleven inputs and out of these eleven inputs, ten are the imprtant features selected by predictor importance and the eleventh one is bias. In this model as shown in figure 3, there is only single hidden layer having four neurons and one bias. The output layer has only one indicator i.e. FS which classifies the districts into HF and NHF. The MLP is activated by feeding the input layer

with the ten important features and a bias. Then the activation function has been activated in feedforward manner by multiplying each input with the corresponding weights associated with the connections through the entire network and classifies the districts either into HF or NHF.

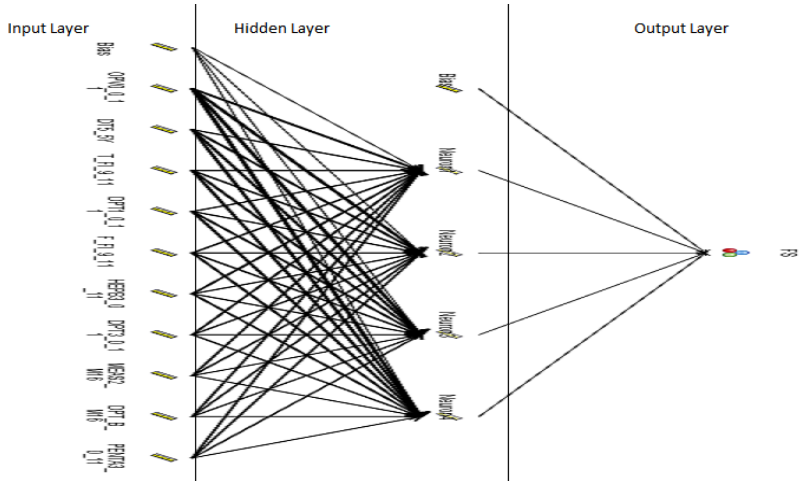


Figure 3: MLP model of ANN

VIII. NATIONAL MODEL EVALUATION

The most important step in any scientific study is to evaluate the performance of the model. The complete dataset of 674 records has been divided into 406, 124 and 144 for training, test and validation partitions respectively i.e. 70% for training, 20% for testing and 10% for validation. The training partition has been used to train the model whereas test and validation partitions have been used to evaluate the performance of the model. In the test partition, out of 124 records, 107 records have been correctly classified whereas 17 were incorrectly classified. Similarly, in validation partition, out of 144 records, 119 have been correctly classified whereas 25 were incorrectly classified. The performance of the present ANN can be measured by using various metrics viz. Accuracy, Precision, Recall, Specificity, AUC and Gini. The confusion matrix and ROC curves of the training, test and validation partitions are shown in figure 4 and 5 respectively.

Confusion Matrix		
Training Data	HF	NHF
HF	238 (tp)	18 (fn)
NHF	33 (fp)	117 (tn)
Testing Data		
HF	72 (tp)	7 (fn)
NHF	10 (fp)	35 (tn)
Validation Data		
HF	77 (tp)	8 (fn)
NHF	17 (fp)	42 (tn)

Figure 4: Confusion Matrix

The ROC (Receiver Operator Characteristic) curve is a graphical representative curve between x-axis and y-axis. On y-axis, sensitivity i.e. true positive rate has to be plotted and on x-axis, (1- sensitivity) i.e. false positive rate has to be plotted [3]. The range of area under curve is from 0 to 1. The AUC of 1 and 0 indicates the best and worst model respectively. Another important statistical measure is Gini that is also used for measuring the performance of the model. The value of Gini can be calculated as Gini=2*AUC-1. If the value of Gini is above 60% then model is considered as good model. The value of AUC can be classified as:

(a) 0.90-1 = excellent
(b) 0.80-0.90 = good
(c) 0.70-0.80 = fair
(d) 0.60-0.70 = poor
(e) 0.50-0.60 = fail [22]

The area under curve for training, testing and validation partitions are shown in figure 5 and the values of AUC and Gini in percentage are shown in the table 1. The value of AUC and Gini for testing partition is 0.921 and 0.841 respectively whereas for validation partition, AUC and Gini are 0.860 and 0.719 respectively. These values of AUC and Gini for testing and validation partitions show the good category of model.

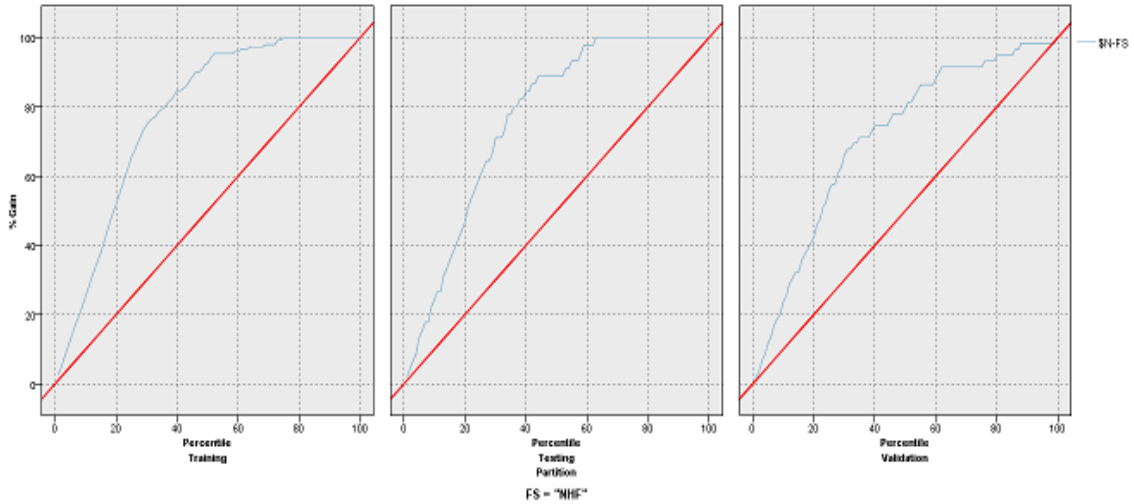


Figure 5: ROC for Training, Testing and Validation data

The overall accuracy of ANN model on the training data is 87.43%. The acceptability of accuracy depends upon the accuracy of test and validation data. The accuracy of ANN model on the unseen test data is 86.29% whereas the accuracy of unseen validation data is 82.63%. The typical outcome when test and validation data are passed through a model is that the accuracy drops by some amount. If accuracy drops by a large amount, it means that the model overfit the training data whereas if accuracy drops by only a small amount, it provides evidence that the model is reliable and will work well in the future on the new data. The small change of 1.14% and 4.80% in accuracy from training to test and validation data respectively indicates that the ANN model is validated. Additionally, the precision value of 87.80% and 81.91% for test and validation data, the recall value of 91.13% and 90.58% for test and validation data and the F-measure value of 89.44% and 86.03% for test and validation data shows that the existing model works well for the purpose of prediction.

Performance Matrix	**Training Data**	**Testing Data**	**Validation Data**
Accuracy	87.43%	86.29%	82.63%
Precision	87.82%	87.80%	81.91%
Recall	92.96%	91.13%	90.58%
Sensitivity (TPR)	92.96%	91.13%	90.58%
Specificity (TNR)	78.00%	77.77%	71.18%
F-Measure	90.32%	89.44%	86.03%
Kappa	72.46%	69.92%	63.25%
AUC	93.80%	92.10%	86.00%
Gini	87.60%	84.10%	71.90%

Table 1: Performance Matrix

IX. FINDINGS AND CONCLUSION

In this paper, Artificial Neural Network (ANN) technique has been used for classification of child immunization data of all districts of India into HF and NHF for the year 2016-2017. Initially, 37 features have been collected from the NHM-HMIS database but after applying the predictor importance technique, the important 10 features have been extracted for building the model. The whole dataset has been divided into three partitions viz. training, test and validation data. The various performance measures have been applied to check the validity of the model including accuracy, precision, recall, F-measure, kappa, AUC and Gini. The accuracy of test and validation partitions drops by only a small amount that means the model doesn't overfit the training data and provides evidence of reliability and stability and thus indicates that the ANN model is validated.

REFERENCES

[1]. Institute of Economic Growth [Online]. Available: http://162.144.90.128/IEGIndia/upload/pdf /wp283.pdf. [Assessed 16 August, 2018].

[2]. V. Sharma, S. Rai and A. Dev, "A Comprehensive Study of Artificial Neural Networks," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no.10, pp. 278-284, 2012.

[3]. S. Shastri et al., "Development of a Data Mining Based Model for Classification of Child Immunization Data," International Journal of Computational Engineering Research, vol. 08, no. 06, pp. 41-49, 2018.

[4]. M. Hemalatha and S. Megala, "Mining Techniques in Health Care: A Survey of Immunization," Journal of Theoretical and Applied Information Technology, vol. 25, no. 2, pp. 63-70, 2011.

[5]. A. Meleko, M. Geremew and F. Birhanu, "Assessment of Child Immunization Coverage and Associated Factors with Full Vaccination among Children Aged 12–23 Months at Mizan Aman Town, Bench Maji Zone, Southwest Ethiopia," International Journal of Pediatrics, vol. 2017, pp. 1-11, 2017.

[6]. P. Gaur, "Neural Networks in Data Mining," International Journal of Electronics and Computer Science Engineering, vol. 1, no. 3, pp. 1449-1453, 2014.

[7]. R. Revathi and T. P. Senthilkumar, "Mining Techniques in Health Care: A Survey of Immunization," International Journal of Computer Trends and Technology, vol. 10, no. 2, pp. 73-78, 2014.

[8]. G. P. Zhang, "Neural Networks for Classification: A Survey," IEEE Transactions on Systems, Man, and Cybernetics Part C Applications and Reviews, vol. 30, no. 4, pp. 451-462, 2000.

[9]. K. Amarendra, K. V. Lakshmi and K. V. Ramani, "Research on Data Mining using Neural Networks," International Journal of Computer Science & Informatics, vol. 2, no. 1, pp. 1-8, 2012.

[10]. K. Streatfield, M. Singarimbun and I. Diamond, "Maternal Education and Child Immunization," Demography, vol. 27, no.3, pp. 447-455, 1990.

[11]. M. C. Arockiaraj, "Applications of Neural Networks in Data Mining," International Journal of Engineering and Science, vol. 3, no. 1, pp. 08-11, 2013.

[12]. Sonalkadu and S. Dhande, "Effective Data Mining Through Neural Network," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 3, pp. 441-444, 2012.

[13]. S. B. Maind and P. Wankar, "Research Paper on Basic of Artificial Neural Network," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 2, no. 1, pp. 96-100, 2014.

[14]. G. Tewary, "Effective Data Mining for Proper Mining Classification using Neural Networks," International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 2, 2015.

[15]. Ministry of Health and Family Welfare [Online]. Available: https://mohfw.gov.in. [Accessed 13 January, 2019].

[16]. S. Shastri, A. Sharma and V. Mansotra, "Classification of Child Immunization Data using Bayesian Network," in Proceedings of 4th International Conference - Computing for Sustainable Global Development (INDIACom), 2017, pp. 1263-1268.

[17]. The BMJ [Online]. Available: https://www.bmj.com. [Accessed 23 January, 2019].

[18]. National Health Mission [Online]. Available:http://www.nhm.gov.in. [Accessed 13 January, 2019].

[19]. V. M. Vashishtha, "Status of immunization and need for intensification of routine immunization in India," Indian Pediatrics, vol. 49, pp. 357-361, 2012.

[20]. U. Fayyad, G. P. Shapiro and P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework," in Proceedings of KDD-96, 1996, pp. 82-88.

[21]. Big Data Made Simple [Online]. Available: http://bigdata-madesimple.com. [Accessed 15 August, 2018].

[22]. J. Devi and N. Sehgal, "A Technique for Improving Software Quality using Support Vector Machine," International Journal of Computer Sciences and Engineering, vol. 5, no. 6, pp. 100-105, 2017.

Arun Singh Bhadwal "Performance Evaluation of ANN Classifier for Knowledge Discovery in Child Immunization Databases" International Journal of Computational Engineering Research (IJCER), vol. 09, no. 3, 2019, pp 70-76