An Enhanced Decision Tree Ensemble Technique For Obesity Prediction

P. Kamakshi Priyaa, S. Sathyapriya, Dr. L. Arockiam

Abstract: Currently, obesity which is a non-communicable disease (NCD) is a serious health issue that leads to many life-threatening diseases like diabetes (Type 2), cancer and heart ailments. Obesity can be determined using various factors, namely age, weight, height and Body Mass Index (BMI). The techniques that are used for obesity prediction are extremely reliant on the BMI; however, the BMI cannot be applied universally. The proposed R ensemble based prediction model incorporating 13 variables has provided an average accuracy of 97.29%. The ensemble model leverages the Enhanced Decision Tree, Naïve Bayes and the Support Vector Machine (Radial SVM). The work has considered unique health parameters such as calories consumed, pregnancy status and bodybuilder which have proved to drastically improve the accuracy of the prediction model.

Index Terms: Ensemble Learning, Machine Learning, Non-Communicable disease, Obesity, Prediction.

1 INTRODUCTION

Obesity is becoming one of the most serious global health problem. Obesity is a term coined for the anatomical condition characterized by an excessive growth of body fat, individually the build-up of adipose tissue beneath the skin [1]. The adults and elderly people affected by obesity are at a higher risk to attract cardiovascular ailments [2] and the people who tend to be overweight tend to be affected by serious health issues in their latter part of life. Obesity is a prominent factor which gives rise to diseases like type 2 diabetes, high blood cholesterol, high blood pressure, heart ailments hypertension, dyslipidaemia and metabolic syndrome [3], [4]. The amount of people diagnosed with clinical obesity has increased greatly in recent years [5]. The psychological, physical and economic consequences of obesity has been discussed well in [6]. The intake of food items rich in saturated fat is directly associated with obesity [7], leading to recommendations to reduce body weight. The inference from studies shows that the process of weight loss has proven to make moderate change in reducing the risk factor among groups and provides a considerable amount of health benefits. Many people around the globe are ignorant and they are not conscious of maintaining a healthy body which reduces the risk of developing many diseases. Machine learning represents a broad array of different techniques, which can be broadly grouped based on the learning methods. The supervised methods include classifiers, the unsupervised methods include clustering and the semi-supervised methods include label propagation. The supervised machine learning technique has been incorporated in this research work.

- *P. Kamakshi Priyaa is pursuing her M.Phil. in Computer Science at St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India, E-mail: kamakshi1796@gmail.com. Her research area is IoT Data Analytics.*
- *S. Sathyapriya is doing her Ph.D. in Computer Science at St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India, E-mail: sathyapriya2822@gmail.com. Her research area is IoT Data Analytics.*
- *Dr. L. Arockiam is working as an Associate Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India, E-mail: larockiam@yahoo.co.in. His research interests are: Software Measurement, Cognitive Aspects in Programming, Data Mining, Mobile Networks, IoT and Cloud Computing.*

Machine learning approaches are being widely used to predict the risk of obesity by identifying causes and developing proactive strategies to prevent it. In the recent times many extensive factors are being considered as possible influencers of obesity.

2 RELATED WORKS

There are several research works which have been explored using data mining techniques in health and disease prediction [8], [9]. Machine learning and structural equation modelling techniques are being used to examine large amounts of data to identify patterns and relationships that would otherwise go undetected [10], [11]. Some studies have applied these methods to build predictive models to understand the obesity problem. The tendency of obesity was examined within selected 7 English medium school students and most of them consume the injurious fast food from the road side shop and get obese [12]. The author designed a cross-sectional study by using a statistical software SPSS. Meng et al., [13] proposed a system that compared the performance of logistic regression, artificial neural networks (ANNs) and decision tree models for predicting diabetes using common risk factors. The logistic regression model gave the highest accuracy and the ANN gave the lowest accuracy. The general attributes namely the Body Mass Index (BMI), age, physician supply, ethnicity and education are utilized for mining the clinical data to predict the heart disease. The BMI does not match the risk factor condition. Some of the risk factors like smoking and regular intake of sugar-sweeten beverage increased the obesity risk to a greater extent than other factors in the daily life [14], in which smoking was not a statistically significant risk factor in [15]. Verma et al., [16] introduced a Multi-layer Perceptron algorithm to achieve higher level prediction accuracy of 88.4%. The proposed hybrid model improves the class level accuracy for the coronary artery disease. Nur' Aina Daud et al., [17] proposed a system to predict obesity based on grocery data using J48 algorithm with an accuracy of 89.4118%.The dataset was based on three different types of data that were manually collected namely grocery data (food weight, quantity, description), demographic data (age, gender) and anthropometric data containing body height, weight and physical activity level. The work was limited to a compatibility of only one algorithm and also the calorie value was assigned based on approximation formula. Given the shortcomings of previous research, the present study aims to build an ensemble based machine learning model to predict obesity.

3 PROPOSED APPROACH

The proposed work aims to build an Ensemble based Obesity Prediction Model using machine learning approaches in mobile environment. The initial step is the data collection which is conjointly done using a mobile application and also through questionnaires distributed through Google Forms. The data collected from these dual sources are stored in the data storage repository. Data Preprocessing is done for feature selection using correlation analysis and to discard the missing values using the median value and the preprocessed data are stored in the data storage. The ensemble based model using Stacking method is built by analyzing the performance of 3 machine learning algorithms namely Enhanced Decision Tree, Naïve Bayes and Support Vector Machine (Radial SVM).Finally, obesity prediction is done by the top layer model which is the most efficient one and the performance of the model is evaluated. The architecture of the proposed system is shown in Fig. 1. A brief description of the various activities in the system are summarized in the following section.

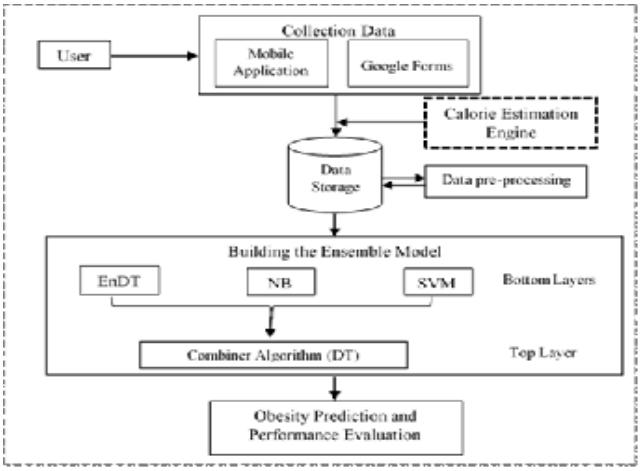

Fig. 1 *Proposed Architecture*

3.1 Data Collection

The data is collected from two sources namely through a mobile application and questionnaires distributed through Google Forms. Given below is a detailed overview of the process of data collection.

3.1.1 Mobile Application

An android based mobile application was developed using Android Studio which acts as the interface between the users and the system. The mobile application included the following set of attributes that can be split into 3 categories.

- Category A: contains demographic information including age and gender.
- Category B: contains the physical data like height, weight, waistline, physical activity level, and pregnancy status.
- Category C: contains the average calorie intake and frequency of processed food intake.

3.1.2 Questionnaire

The questionnaire was developed using Google Forms for the three categories of data as done through the mobile application. Unlike the mobile application which estimated the calories based on sensors, the questionnaire collected the calorie intake from the user. Based on the collected data, the following measures are calculated as

$$\text{BMI} = \frac{Weight\ in\ kg}{Height\ in\ m^2}$$

$$\text{WHfR} = \frac{Waist\ circumference\ (in\ cm)}{Height\ (in\ cm)}$$

The obesity index is calculated based on the values from Table 1. The average calorie intake was computed using a Deep learning based food image recognition system which is done in the previous work [18].

TABLE 1
BMI BASED OBESITY CHART

BMI (kg/m^2)	*Weight Classification*	*Obesity class*
<18.5	*Underweight*	-
18.5-24.9	*Normal*	-
25.0-29.9	*Overweight*	-
30.0-34.9	*Obese*	*Obesity class 1*
35.0-39.9	*Highly Obese*	*Obesity class 2*
>= 40	*Extremely Obese*	*Obesity class 3*

3.2 Data Storage

The data collected from both the sources were combined together and stored in .csv (Comma Separated Values) format. The collected data accounted to 324 records. The datatypes of the attributes were modified in order to suit the needs of the machine learning algorithm. The list of attributes and their description is provided in Table 2. These data were utilized to build the model.

TABLE 2
ATTRIBUTE AND DESCRIPTION

Attribute Name	*Description*
Gender	*Contains the gender of the person. Males are denoted as 0 and females as 1*
Height	*Numeric value that contains the height of an individual expressed in cm*
Weight	*The field contains the weight of the person expressed in kg*
Age	*The age of an individual*
Calorie_Intake	*The average amount of calories the user intakes per day expressed in kcal*
Waistline	*The circumference of the waistline was obtained in inches and was converted into cm.*
Pregnency	*The field determines whether a woman is pregnant which has values of 0 denoting no and 1 denoting yes.*
Physical_Activity_Level	*The attribute denotes the routine physical activity level of an individual which falls in either one of the three categories namely Extremely Active denoted as 0, Moderately Active denoted as 1 and Sedentary as 2.*
Processed_Food_Intake	*The field displays the frequency of processed food intake by the individual which was categorized as Rarely denoted as 0, Occasionally denoted as 1 and Frequently denoted as 2.*
BMI	*The BMI is calculated using the formula (weight (kg)/height (m^2) and the computed value is rounded to one decimal place*
Index	*The attribute denotes the BMI Index which is calculated from the standard BMI table which contains levels from 1 to 6*
WHfR	*The waist to height ratio is calculated by weight (cm)/height (cm)*
Obesity_Predict	*The target attribute which predicts whether or not an individual has obesity which is denoted as 0 for no and 1 for yes.*

3.3 Data Storage

Data Pre-processing is an important step in the data mining process. The data that is collected may contain out-of-range values, missing values. The reliability of the data is increased when these kinds of invalid data are handled efficiently else it may lead to undesirable outcomes and makes the knowledge discovery process a challenge. The proposed system's dataset contained some missing values namely the waist size and processed food intake which were pre-processed using the median technique which generated the probable value. The waist size was calculated using the standard charts depending upon the height and weight. Feature selection was also done using the Correlational Attribute Evaluator and the resulting pre-processed data is sent for further processing.

3.4 Building the Ensemble Model

Ensembling is a technique of collaboratively working on more than two similar or dissimilar types of algorithms which are called the base learners. The process is done to build a more robust system which incorporates the predictions from all the base learners into account while making the final decision. This makes the final decision more robust, accurate and unlikely to be biased. The proposed R ensemble model is built using the stacking technique with two levels. The base level is formed using three supervised machine learning based algorithms namely Enhanced decision tree, Naïve Bayes and Support Vector Machine (Radial SVM). The algorithms were chosen such that they satisfy the requirement criteria of Stacking technique which includes that the individual model should have average accuracy criteria and the predictions of the individual models have low correlation with the predictions of other models. The model was trained with a p value of 0.50. A detailed description of the algorithms that were used to build the Ensemble Model are described below.

3.4.1 Enhanced Decision Tree

The decision tree model was chosen to be enhanced as the obesity dataset was highly compactable with the tree structure based algorithm and it was comprehensive and easy to interpret. The construction of a decision tree model involves a collection of decision nodes, connected by branches, extending downward from the root nodes until terminating in leaf nodes. The proposed algorithm provides an optimal enhancement of selecting the root node based on the correlation value of the attributes to the decision tree algorithm. The pseudocode of the modified algorithm is given below.

Pseudocode

Algorithm: DT_Learn (TT, TC, Can_Attr)
Input:
TT: set of training tuples
TC: the target class
Can_Attr: set of candidate attributes
Output:
A decision tree

Steps:

1) {
2) Create a root node RN //unlabeled node
3) if all the rows in TT are of the same target class T then
4) return RN as the leaf of the single node tree labeled with the class T;
5) if Can_Attr = empty (i.e. if no candidate attribute is present) then
6) return RN as the leaf of the single node tree labeled with the widely held value of the target in TT
7) Otherwise
8) {
9) Select the attribute Best_attri from Can_Attr that best classifies TT based on best positively correlated attribute
10) Set Best_Attri as the root node attribute
11) for each of the permitted value of Best_Attri, pv_i, //the tuples are partitioned and the subtrees are grown for each of the partition
12) {
13) Add a new branch below the root node that corresponds to Best_Attri=pv_i
14) let TT_{pvi} be the subset of TT that satisfies the outcome Best_Attri=pv_i
15) if TT_{pvi} = empty then

16) Add a leaf node which has the majority class in TT
to the root node RN;
17) else
18) add the node that is returned by DT_Learn (TT_{pvi},
TC, Can_Attr – {RN});
19) }
20) }
21) return (RN);
22) }

3.4.2 Naïve Bayes

The Naïve Bayes algorithm is simple yet an efficient probabilistic algorithm in classification technique which gets the probability value based on the frequency calculation and combinational values from the related collection. The algorithm assumes that the attributes are strongly independent. The inverse conditional probability is given by

$$P(Y|X_1\ldots\ldots X_n) = \frac{P(Y).P(X_1\ldots\ldots X_n|Y)}{P(X_1\ldots\ldots X_n)} \quad (1)$$

In equation (1), the variable Y is a class and the variable $X_1\ldots\ldots X_n$ denotes the classification characteristics. $P(Y|X_1\ldots\ldots X_n)$ is called the posterior probability and P(Y) is known as the prior probability. The conditional probability is computed as

$$P(X|Y = y) = \prod_{j=1}^{k} P(Xj|Y = y)$$

Pseudocode

Input:
Training dataset Tr,
$X = (X_1, X_2,\ldots\ldots X_n)$ //predictor variable value in the test dataset
Output:
A testing class dataset

Steps:
1) Read the training dataset T;
2) Calculate the mean value and the standard deviation value of the predictor variables in each class;
3) do
4) Calculate the probability of X_i using the gauss density equation in each class;
5) until
6) the probability of all the predictor variables (X_1, $X_2,\ldots\ldots X_n$)
has been calculated.
7) Calculate the likelihood of each class;
8) return the greatest likelihood;

3.4.3 Support Vector Machine

The support vector offer another approach to classify a multi-dimensional dataset in which the samples on the margin are called the support vectors. A SVM is a linear or non-linear classifier, which is mathematical function that distinguishes two different types of attributes.

Pseudocode

Input:
X and Y loaded with the training data, α ← 0 or α ← partially trained SVM model

Output:
SVM classifier model

Steps:
1) C<= any value
2) Repeat
3) for all {Xi,Yi}, {Xj,Yj} do
4) Optimize α_i and α_j
5) end for
6) until no changes in α or other resource constraint criteria met
7) Ensure to retain only the support vectors ($\alpha_i > 0$)

3.5 Obesity Prediction

The obesity prediction in the proposed model is not completely dependent on the BMI alone. The WHfR is used so as a bodybuilder who has a higher BMI will have a low WHfR value. A pregnant lady may gain 11-15 kgs during pregnancy and so this factor was also considered. The average calorie intake per day based on the gender was included. The physical activity level and the frequency of processed food intake were also considered. In order to predict the obesity of an individual, all the attributes were considered which has proved to increase the accuracy of the proposed prediction model.

4 RESULTS AND DISCUSSION

This section presents the results that have been deduced from the proposed system. The prediction results of the base layer model containing Decision Trees with an accuracy of 96.87%, Naïve Bayes with an accuracy of 93.84% and svmRadial with an accuracy of 95.75% are shown in Fig. 2. The correlation between the various machine learning algorithms was analyzed and the maximum correlation of 0.41 was observed between decision tree and svmRadial which is less than the maximum threshold of 0.75 and hence the proposed set of algorithms are valid to be used for stacking. The plot matrix of the correlation between the models is shown in Fig. 3.

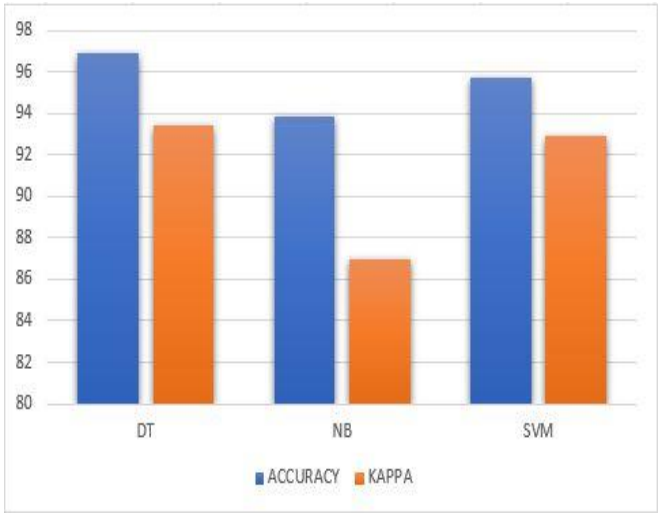

***Fig. 2** Comparison of Base Layer Accuracy*

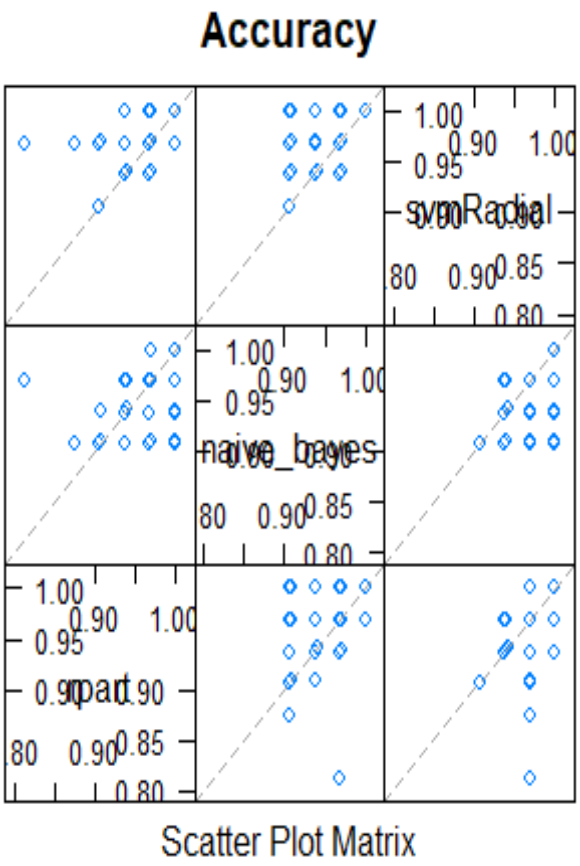

***Fig. 3**. Scatter Plot for model correlation*

The ensemble model with the decision tree as the top layer provided an increased accuracy of 97.29% which is an enhancement over the individual accuracy of 96.87%. A comparative analysis of the induvidual machine learning algorithm and the proposed ensemble model is shown in Fig. 4. The results prove that the proposed Ensemble model has improved the prediction accuracy.

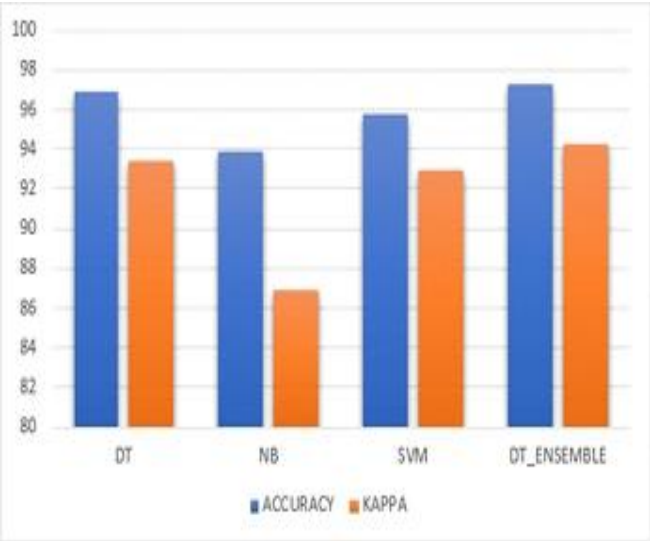

***Fig 4.** Comparative Analysis of Model Accuracy*

4.2 Performance Evaluation

The performance of the proposed system is compared with the related research area of obesity prediction and the proposed model has proven to have achieved a higher level of accuracy. Unlike the existing models which have only concentrated on the directly influential attributes of obesity, the proposed work has included extensive factors that have proven to have a positive correlation over the obesity level. The ensemble approach has proven to have a better prediction accuracy over the traditional algorithmic methods. The comparision of the previous research works is shown in Table 3.

TABLE 3

COMPARISON OF RELATED RESEARCH WORKS

Research Works	*Approach*	*Accuracy*
Zeyu et al. [19]	*Comparision of 4 Machine Learning Algorithms*	*88.82%*
Nur'Aina Daud et al. [17]	*Machine learning based J48 algorithm*	*89.41%*
Kapil et al. [20]	*Ensemble learning with three algorithms*	*89.68%*
Proposed work	*Ensemble learning with enhanced decision tree*	*97.29%*

The Table 3 shows that the proposed system has an improved accuracy over the previous related works. The comparative results show that ensemble models can be used to achieve a relatively higher accuracy than the individual algorithms.

5 CONCLUSION

The prediction of health conditions and diseases using machine learning techniques may be a challenging task but it increases the analytical accuracy and specificity. Data Analysis using machine learning techniques reduces the cost and time constraints involved. The proposed system has included the extensive factors of obesity which have proved to improve the prediction accuracy. The study has inferred that maintaining a healthy WHfR and an active physical activity level reduces the risk of obesity. The proposed ensemble model leveraging Enhanced Decision Tree, Naïve Bayes and Radial SVM provided an accuracy of 97.29%. The future work aims at including diverse factors for prediction and also increasing the levels of the ensemble model to further improve the accuracy of the model and also utilizing a larger dataset to analyze the model's efficiency.

6 REFERENCES

[1] N.L. Noor, N. Noordin, F.M. Saman and N.I. Teng, "Predicting Obesity from Grocery Data: A Conceptual Process Framework", In Proceedings of the 6th International Conference on Information and Communication Technology for the Muslim World (ICT4M), IEEE, pp. 286-291, 2016.

[2] M. Bouchonville, R. Armamento-Villareal, K. Shah, N. Napoli, D.R. Sinacore, C. Qualls and D.T. Villareal, "Weight loss, exercise or both and cardiometabolic risk factors in obese older adults: results of a randomized controlled trial", International Journal of Obesity, Vol. 38, No. 3, pp. 423, 2014.

[3] A. Dewan and M. Sharma, "Prediction of Heart Disease using a Hybrid Technique in Data Mining Classification", In Proceedings of the 2nd International Conference on

Computing for Sustainable Global Development (INDIACom), IEEE, pp. 704-706, 2015.

[4] J.D. Brown, J. Buscemi, V. Milsom, R. Malcolm and P.M. O'Neil, "Effects on Cardiovascular Risk Factors of Weight Losses Limited To 5–10%", Translational Behavioral Medicine, Vol. 6, No. 3, pp. 339-346, 2015.

[5] S. Mitra, Y. Qiu, H. Moss, K. Li and S.L. Pallickara, "Effective Integration of Geotagged, Ancilliary Longitudinal Survey Datasets to Improve Adulthood Obesity Predictive Models", In Proceedings of the International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), IEEE, pp. 1738-1746, 2018.

[6] J.A. Gazmararian, D. Frisvold, K. Zhang and J.P. Koplan, "Obesity is associated with an increase in pharmaceutical expenses among university employees", Journal of Obesity, 2015.

[7] R. Crescenzo, F. Bianco, A. Mazzoli, A. Giacco, R. Cancelliere, G. di Fabio, A. Zarrelli, G. Liverini and S. Iossa, "Fat quality influences the obesogenic effect of high fat diets", Nutrients, Vol. 7, No. 11, pp. 9475–9491, 2015, DOI: https://doi.org/10.3390/nu7115480.

[8] P. Repalli, "Prediction on diabetes using data mining approach", Oklahoma State University, 2011.

[9] N.A. Sundar, P.P. Latha and M.R. Chandra, "Performance analysis of classification data mining techniques over heart disease database", International Journal of Engineering Science & Advanced Technology, Vol. 2, No. 3, pp. 470-478, 2012.

[10] J. Yang, H. Gu, X. Jiang, Q. Huang, X. Hu and X. Shen, "Walking in the PPI network to identify the risky SNP of osteoporosis with decision tree algorithm", In Proceedings of the International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp. 1283-1287, 2016.

[11] T. Sivaranjani, "Comparative study on Obesity based on ID3 and KNN", International Journal of Advance Research in Computer Science and Management Studies, Vol. 2, No. 9, pp. 389-396, 2014.

[12] M.N. Rahman, S.A. Reza, M.A. Islam, A. Rahman and A.K. Nath, "Prevalence of obesity and overweight among English medium school children of Dhaka City in Bangladesh", Journal of Environmental Science and Natural Resources, Vol. 7, No. 1, pp. 63-67, 2014.

[13] X.H. Meng, Y.X. Huang, D.P. Rao, Q. Zhang and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors", The Kaohsiung Journal of Medical Sciences, Vol. 29, No. 2, pp. 93-99, 2013.

[14] A. Pochini, Y. Wu and G. Hu, "Data Mining for Lifestyle Risk Factors Associated with Overweight and Obesity among Adolescents", In Proceedings of the 3rd International Conference on Advanced Applied Informatics, IEEE, pp. 883-888, 2014.

[15] V. Silverwood, M. Blagojevic-Bucknall, C. Jinks, J.L. Jordan, J. Protheroe and K.P. Jordan, "Current evidence on risk factors for knee osteoarthritis in older adults: a systematic review and meta-analysis", Osteoarthritis and Cartilage, Vol. 23, No. 4, pp. 507-515, 2015.

[16] L. Verma, S. Srivastava and P.C. Negi, "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data", Journal of medical systems, Vol. 40, No. 7, pp. 178, 2016.

[17] N.L. Nur, S.A. Aljunid, N. Noordin and N.I. Teng, "Predictive Analytics: The Application of J48 Algorithm on Grocery Data to Predict Obesity", In Proceedings of the IEEE Conference on Big Data and Analytics (ICBDA), IEEE, pp. 1-6, 2018.

[18] P. Kamakshi Priyaa, S. Sathyapriya, L. Arockiam, "Nutrition Monitoring and Calorie Estimation Using Internet of Things (IoT)", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 8, No. 11, pp. 2669-2672, 2019, DOI:10.35940/ijitee.K2072.0981119.

[19] Z. Zheng and K. Ruggiero, "Using machine learning to predict obesity in high school students", In Proceedings of the International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp. 2132-2138, 2017.

[20] K. Jindal, N. Baliyan, P.S. Rana, "Obesity Prediction Using Ensemble Machine Learning Approaches", In Recent Findings in Intelligent Computing Techniques, Springer, pp. 355-362, 2018.