A Hybrid Approach Based on Machine Learning to Identify the Causes of Obesity

Anar Taghiyev[1], Adem Alpaslan Altun[1], Sona Caglar[2]

[1]*Department of Computer Engineering, Selcuk University, Konya, Turkey. (E-mail: anart@selcuk.edu.tr)*
[2]*Department of Health, Aksaray, Ministry of Health, Turkey*

Abstract: The obesity issue has international relevance in recent years and the study aim is to develop a hybrid classification model to identify the causes of obesity in the region of Turkey. In the period from March to November 2019, patient records retrieved from the database of Electronic Health Records (EHR) of Aksaray Sultanhani Family Health Center (ASFHC) were examined, and the questionnaire was conducted among the females aged 18 years and above. In the study, a two-stage hybrid model was used in order to better classify the collected data. The first-stage is the feature (i.e. best variable) selection while the second-stage is for classification. The performance of a proposed two-stage hybrid approach was compared with traditional single-stage classifiers: Decision Trees (DT) and Logistic Regression (LR) algorithms. In the study, the proposed hybrid system gives 91.4 of accuracy, which is better than other classifiers (i.e. 4.6 % higher than the performance of LR and 2.3 % higher than the performance of DT). Thus, the proposed hybrid system provides a more accurate classification of patients with obesity and a practical approach to estimating the factors affecting obesity.
In the future, we are going to research in detail the relationship between Type 2 Diabetes (T2D) and obesity in females.

Keywords: apache spark; machine learning; classification; hybrid intelligent systems; obesity.

1. INTRODUCTION

The objective of this study is to develop a two-stage classification model to better identify the causes of obesity among females aged 18 years and above in the region of Turkey. Because according to the World Health Organization (WHO), approximately 600 million (13%) adults aged 18 years and above are obese; at least 2.8 million people have been shown to die each year from overweight and obesity (Turkey: Ministry of Health, 2019; World Health Organization, 2019). In Turkey, the prevalence of obesity is 20.5% in adult men and 41% in women (Turkey: Ministry of Health, 2014). The prevalence of obesity in women is higher than in men. Physical inactivity is reported to account for 67.5 % of the prevalence of obesity (World Health Organization, 2019; Turkey: Ministry of Health, 2014). In this study, Body Mass Index (BMI) has been considered as a target/dependent variable. Because, the WHO utilizes BMI values to classify and measure obesity. BMI is a value obtained by the ratio of the body weight (kg) of the individual to the height square (m^2) as in (1).

$$BMI = \frac{Weight}{Height^2} \tag{1}$$

It also provides information about fat distribution in the body (Turkey: Ministry of Health, 2014). Accordingly, when BMI=25.0÷29.9 (kg/m^2), people are overweight, but when BMI≥30 (kg/m^2), people are considered obese. Obesity creates the condition for diseases such as cardiovascular diseases, neurological diseases, metabolic-hormonal complications (e.g., T2D, hyperinsulinemia, and hypertension) and cancer (Turkey: Ministry of Health, 2019). Therefore, obesity needs to be controlled and in this paper, we tried to investigate the factors causing obesity by using a two-stage hybrid model. There are many studies on the topic of obesity and most of them use the traditional (single-stage) classifiers. (Dugan et al., 2015) utilized RandomTree(RT), RandomForest (RF), J48, ID3, Naïve Bayes(NB), and Bayes for predicting children obesity with results: 85% of precision, and 89% of sensibility. (Ergün, 2009) used the single-stage algorithms such as LR or Neural Network (NN) for the classification of obesity disease. However, using single-step algorithms is difficult to achieve a more accurate classification.

In literature, some authors use two-stage methods from data mining (Adnan et al., 2010; Yang and Garibaldi, 2015) and machine learning (Murray et al., 2020; Ali et al., 2019a; Devi et al., 2020; Lin et al., 2019) for diseases classification and prediction. For instance, (Muhamad et al., 2012) have suggested a two-stage hybrid approach to predicting children obesity. The two-stage model consists of the combination of NB for prediction and Genetic Algorithm (GA) for parameter optimization and the hybrid model achieved the prediction accuracy of 75%.

Recently, (Devi et al., 2020) have proposed a two-stage hybrid approach, the combination of Farthest First (FF) clustering algorithm and the Sequential Minimal Optimization (SMO) classifier algorithms for diagnosing diabetes, where obesity was considered as the risk factor. A proposed approach achieved the classification accuracy of 99.4%. (Ni et al., 2020) have developed a two-stage hybrid

approach (MOGP-HMM): the first-stage is a multi-objective genetic programming (MOGP) algorithm to reduce the dimensions of data; the second-stage is a hidden Markov model (HMM), for predicting human physical activity status from lifelogging data. (Ramirez et al., 2020) have used a hybrid model based on NN and Fuzzy logic (FL) for 2-led cardiac arrhythmia classification and a proposed hybrid model achieved the classification accuracy of 93.80%. (Akgül et al., 2019) have suggested a hybrid approach (a combination of Artificial Neural Network (ANN) and (GA)) to improve the classification accuracy for diagnosis of Heart Disease. Experimental results for a hybrid classification model showed that the accuracy, precision, recall, and F-measures are 95.82%, 98.11%, 94.55%, and 96.30, respectively. (Ali et al., 2019b) have proposed a hybrid approach: $\chi 2$ statistical model is used to eliminate irrelevant features, while the optimally configured deep neural network (DNN) is searched by using exhaustive search strategy to improve the prediction accuracy for Heart Failure. A proposed two-stage model achieved the prediction accuracy of 93.3%. (Ali et al., 2019c) have also developed a two-stage (a feature-driven decision support system) method. In the first stage, $\chi 2$ statistical model is used to rank the commonly used features. Based on the $\chi 2$ test score, an optimal subset of features is searched using forward best-first search strategy. In the second stage, Gaussian NB classifier is used as a predictive model. The authors obtained the prediction accuracy of 93.33% using the proposed two-stage method. Furthermore, (Ali et al., 2019a) have suggested a hybrid approach for detection of Parkinson's disease (construction of an unbiased cascaded learning system based on feature selection and Adaptive Boosting Model). Experimental results for a hybrid classification model showed that the accuracy, sensitivity, specificity are 76.44%, 70.94%, and 81.94%, respectively. Moreover, a hybrid approach: a two-dimensional data selection method for sample and feature selection, proposed for early diagnosis of Parkinson's disease (Ali et al., 2019d) In the study, a proposed hybrid model achieved classification accuracy of 97.5% on training, and 100% on test datasets. In addition, (Ali et al., 2019e) have offered a hybrid approach (LDA-NN-GA) for the detection of Parkinson's disease (a proposed hybrid approach uses linear discriminant analysis (LDA) for dimensionality reduction and GA for hyperparameters optimization of NN, which is used as a predictive model). Authors were able to achieve the accuracy of 80% on training, and 82.14% on test dataset with a two-stage hybrid model.

Inspired by the various hybrid intelligent systems discussed above, in our study, we have also tried to go beyond single-stage classification models and to develop a hybrid classification model to identify the causes of obesity for a segment of the female population in Turkey. To our knowledge, none of the previous studies concerned the hybridization of the model to identify the causes of obesity. In the literature, only (Yang and Garibaldi, 2015) have also suggested a hybrid approach for the identification of risk factors but for heart disease. Their proposed hybrid approach is a combination of machine-learning methods with other NLP techniques.

In this study, we have used a combination of two classifiers such as DT and LR. We would like to emphasize that these two classifiers are complementary; i.e. DT handles linear interaction between independent and dependent features (subsets) good enough, but DT has problems with to handle linear relations between features (subsets). LR, on the contrary, handles linear relationships between features (subsets) good enough, but LR has problems with the interaction impacts between independent and dependent features (subsets). In principle, DT in the proposed hybrid approach splits the data into more effective features (subsets/variables) on which the LR is fit for every feature (subsets). Variable selection is performed by using the information value from the filter method (i.e. we pick feature with the highest gain). The final outputs of the proposed method are to get the odds ratio of more effective factors that causes of obesity, which is discussed in detail in this paper. A proposed hybrid method is carried out on the obesity dataset collected from ASFHC and the effectiveness of the model was compared with the performance of single-stage classifiers. In light of the outputs obtained from this research, factors causing obesity are revealed. Thus, the outputs presented in our article can contribute to the obesity-related planning of research centers, healthcare institutions, and managers, professional organizations.

The manuscript is organized as follows. Section 2 provides the materials and methods utilized for a proposed hybrid approach in detail. Validation and evaluation are shown in Section 3. Section 4 and Section 5 present the results and discussion of comparative experiments of the studies. Conclusion and future work information are given in Section 6.

2. MATERIALS AND METHODS

Firstly, we would like to note that to carry out the research; the necessary official permissions were obtained from the Ethics Committee of Faculty of Medicine at Selcuk University, as well as from Aksaray Provincial Directorate of Health and Family Health Center. After that, we proceeded to collect data. The data used in our study have been collected from ASFHC in Turkey. The main areas of the research process are shown in Fig 1.

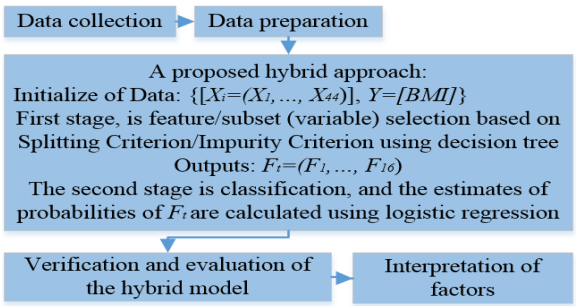

Fig. 1. Main stages of this reseacrh process.

2.1. Data Collection

The females who applied to ASFHC from 15.03.2019 to 01.11.2019, and accepted our study were surveyed (interviewed) by the researcher, the results of the blood test

were collected, and the hybrid approach was applied to identify the causes of obesity. The study involved four physicians and four health personnel working in ASFHC. We analysed the EHR database (Figueroa and Flores, 2016) and found the following information: the total number of applications was 22789; the number of females' applicants was 12306. The total number of lab records of applicants was 43312; the number of lab records for females was 31029; the total number of patients being examined was 5656, but only 651 of them are females aged 18 years and above who have undergone laboratory blood tests. Only 500 out of 651 females had been interviewed, who contacted the ASFHC and agreed to participate in this study.

2.2. Data Description and Preparation

In the course of the study, we have created a new database, which covers data retrieved from EHR (*DB1*) (Figueroa and Flores, 2016), as well as data collected from questionnaires (*DB2*). The females who participated in the study were asked a questionnaire consisting of 37 items. The contents of the questionnaire and the results of the patients' blood tests were presented in Fig 2 (a) and (b).

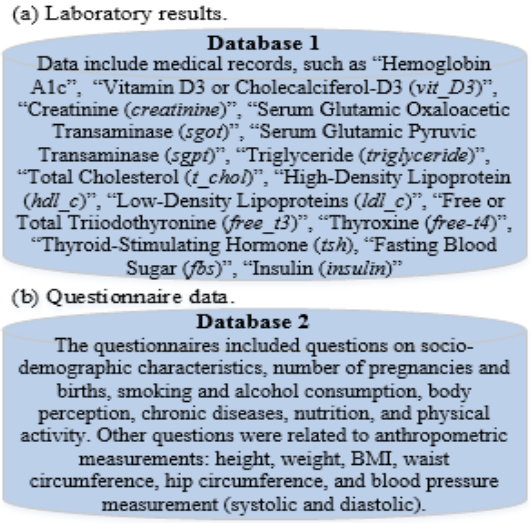

Fig. 2. Content of laboratory and questionnaire data.

Afterwards, the original dataset (O_{DB}) with 56 attributes and 500 instances has been created for the study. Since the hybrid approach is based on supervised training, the data were studied one by one, taking into account the opinion of the specialist doctor to determine the filtering strategy (Shi et al., 2019). As expected, the original dataset contained missing and noisy data. Attributes with many missing values proved useless for analysis. Consequently, these missing and noisy attributes (11 in total) were not investigated because they represented 50% of the total instance size in the dataset.

In addition, as we mentioned in the introduction, the presence of obesity was estimated according to BMI (kg/m^2) and in the current paper, it was considered as a target/dependent variable. The dependent variable (BMI values among females aged 18 years and above) was categorized into two groups:

1. When the *BMI* of 30 kg/m² or higher, female is considered "*OBESE*", i.e., the output will [0.0].
2. When the *BMI* of 29.9 kg/m² or less, female is classified as "*NONOBESE*", i.e., the output will [1.0];

Thus, after grouping the values of the attributes (see Table 1) and performing the necessary initial filtration used in the study, the analysis started on the training dataset ("O_{Tr}") with 45 attributes and 325 instances.

The filtration method was also used in the first stage of the hybrid model to feature selection by using the information value (i.e. we pick feature with the highest gain), while in the second stage these features were classified to identify the causes of obesity. Consequently, all this information has been provided in detail in the following section.

2.3. Description of the Proposed Hybrid Approach

In this study, the proposed two-stage hybrid approach was used to identify the factors causing obesity: in the first stage, we try to build trees to choose the best feature with the highest information gain (more effective variables-F_t) and in the next stage, LR applied to F_t-variables (See Fig. 3). DT is a good prognostic model, distinguished by their efficiency and clarity due to their simplicity (Hu, 2011; Khraisat et al., 2020). DT uses a process in which data are recursively broken down into smaller, cleaner subset, while repeatedly applying the search in the partitions of possible branches and choosing the optimal terminal node based on impurity criterion/splitting criterion (i.e., based on rules). Partitioning begins with the root node, which iteratively defines the most appropriate criteria for splitting, dividing the data into two groups: terminal and splitting nodes (Kumar and Nirmalkumar, 2019).

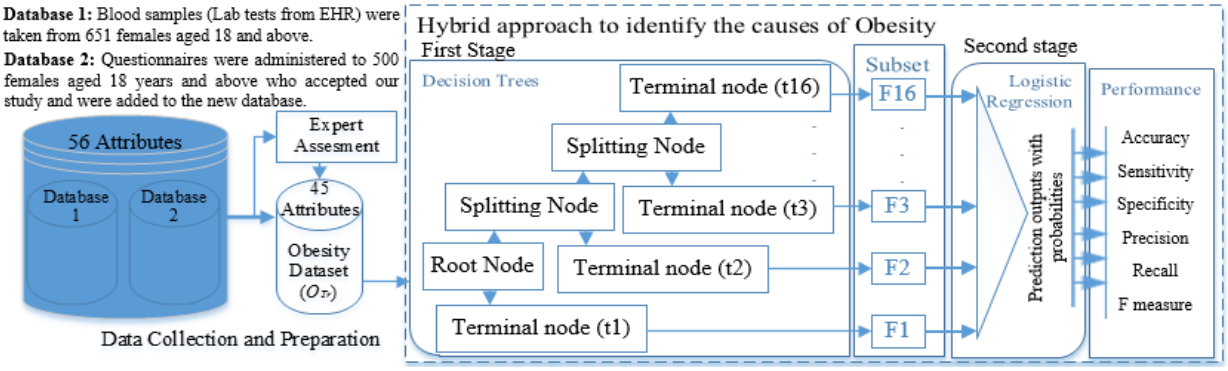

Fig. 3. A hybrid approach framework for determining the causes of obesity.

Table 1. List of attributes and descriptions.

Original Attributes	Description of values
age (year)	Patient's age grouped into: "[18-40)"; "[40-65)"; and "[65-90)"
fam_type	Family types of patients: 1- "*nuclear*"; 2- "*extended*"; 3- "*composite*"
mar_stat	Marital status of patients: 1- "*married*"; 2- "*single*"; 3- "*divorced*"; 4- "*widow*"
edu_stat	Educational status of patients, that is for Partner 1: 1- "*illiterate*"; 2- "*literate*"; 3- "*primary school*"; 4- "*secondary school*"; 5- "*high school*"; 6- "*undergraduate*"; 7- "*post-graduate*"
part_edu_stat	Educational status of Partner 2 (that is husband): 1- "*illiterate*"; 2- "*literate*"; 3- "*primary school*"; 4- "*secondary school*"; 5- "*high school*"; 6- "*undergraduate*"; 7- "*post-graduate*"
tjob	Type of job grouped into: 1- "*not work*"; 2- "*worker*"; 3- "*officer*"; 4- "*self-employed*"; 5- "*retired*"; 6- "*others*"
num_per	Number of family members: [1-4) that is "*4.0*"; [4-7) that is "*7.0*"; [7-20) that is "*20.0*"
smoke	Indicates if the patient smokes cigarettes: 1- "*yes*" or 2- "*no*"
drink_alc	Indicates whether a patient consumes alcohol: 1- "*no*" or 2- "*yes*"
chr_dis	Chronic diseases: 1- "*no*"; 2- "*yes*" or "*diseases*"
rec_diet	Compliance with the doctor's recommended diet: 1- "*no*" or 2- "*yes*"
pre_ac	Indicates if there was enough problem to prevent patients from doing exercise: 1- "*yes*", 2- "*no*"
transp	Vehicles that are often used to get home or anywhere else: 1- "*car*"; 2- "*pedestrian*"; 3- "*motorcycle*"; 4- "*bus*"; 5- "*others*"
comp	Indicates if there is a computer at home/office: 1- "*no*" or 2- "*yes*"
tspent_tech (hour)	The time spent in front of the computer, TV&phone is grouped: (0-1], that is "*1.0*"; (1-4], that is "*4.0*"; (4-12], that is "*12.0*"
diet_cha_hab	The issue of whether patients have used any diet to change eating and drinking habits at any point in their life has been grouped into two parts: 1- "*no*" or 2- "*yes*".
eng_phy_ac	Patients who were engaged in physical activity were identified: 1- "*no*" or 2- "*yes*"
dietexe_cha_hab	Grouping whether *patients have used diet & exercises to change bad habits*: 1- "*no*"; 2- "*yes*"; 3- "*sometimes*"; 4- "*others*"
ap_diet_preg	Have patients *applied diet to change eating and drinking habits for weight loss during pregnancy and after childbirth* (except during the postpartum period): 1- "*no*" or 2- "*yes*"
exe_dur_preg	Patients who did (or not) any *exercises during pregnancy & after childbirth*: 1- "*no*" or 2- "*yes*".
preg	Separation of patients who had a pregnancy: 1- "*no*" or 2- "*yes*"
num_preg	Total number of pregnancies: "*0*"; [1-4) that is "*4.0*"; [4-7) that is "*7.0*"; [7-20), that is "*20.0*"
had_childbirth	Patients who have given birth (or never given birth): 1- "*no*" or 2- "*yes*"
num_births	Total number of births: "*0*"; [1-4) that is "*4.0*"; [4-7) that is "*7.0*"; [7-15), that is "*15.0*"
weight_op	Patients' opinions on their own weight were identified: "*thin*"; "*norm*"; "*slightly fat*"; "*fat*"; "*very fat*"
ob_dur_preg	Patients who have been *diagnosed (or not) with obesity during pregnancy and after childbirth*: 1- "*no*" or 2- "*yes*"
weight (kg)	Patient's weights: "[40-70)"; "[70-100) and "[100-150)"
height (cm)	Patient's height range: "[135-160)" and "[160-185)"
waist_circ (cm)	Waist circumference range: <80, that is "*norm*"; [80-88), that is "*risk*"; ≥88, that is very "*very risk*"
hip_circ (ration)	Hip circumference (ration): 0.80 or lower, that is "*low*"; 0.81-0.85, that is "*moderate*"; 0.86 or higher, that is "*high*"
systolic (mmHg)	Systolic blood pressure grouped into "*low*" [70-90); "*ideal*" [90-120); "*pre-high*" [120-140); "*high*" (≥140)
diastolic (mmHg)	Diastolic blood pressure grouped into "*low*" [40-60); "*ideal*" [60-80); "*pre-high*" [80-90); "*high*" (≥90)
vit_D3 (ug/L)	Cholecalciferol-D3 (25-OH-Vitamin D3) grouped into "*low*" (<20); "*norm*" [20-60]; "*high*" (>60)
creatinine (mg/dL)	Creatinine grouped into "*low*" (<0.40); "*norm*" [0.40-1.20]; "*high*" (>1.20)
sgot (U/L)	Serum Glutamic Oxaloacetic Transaminase (*sgot*) grouped into: "*low*" (<5); "*norm*" [5-40]; "*high*" (>40)
sgpt (U/L)	Serum Glutamic Pyruvic Transaminase (*sgpt*) grouped into: "*low*" (<5); "*norm*" [5-45]; "*high*" (>45)
triglyceride(mg/dL)	Triglyceride grouped into "*low*" (<50); "*norm*" [50-200]; "*high*" (>200)
t_chol (mg/dL)	Total Cholesterol (*t_chol*) grouped into "*low*" (<140); "*norm*" [140-200]; "*high*" (>200)
hdl_c (mg/dL)	High-Density Lipoprotein (*hdl_c*) grouped into "*low*" (<40); "*norm*" [40-70]; "*high*" (>70)
ldl_c (mg/dL)	Low-Density Lipoproteins (*ldl_c*) grouped into "*low*" (<70); "*norm*" [70-130]; "*high*" (>130)
free_t3 (pg/mL)	Free or Total Triiodothyronine (*free_t3*) grouped into "*low*" (<3.40); "*norm*" [3.40-8.00]; "*high*"

	(>8.00)
free_t4 (ng/dL)	Thyroxine (*free_t4*) grouped into "*low*" (<7.6); "*norm*" [7.6-17]; "*high*" (>17)
tsh (μIU/mL)	Thyroid-Stimulating Hormone (*tsh*) grouped into "*low*" (<0.40); "*norm*" [0.40-5.60]; "*high*" (>5.60)
fbs (mg/dL)	Fasting blood sugar (*fbs*) grouped into: "*low*" (<70); "*norm*" [70-99]; "*pre-diabetes*" [100-125]; "*diabetes*" (≥126)
BMI (kg/m^2)	(If BMI value is 30 or higher), [0.0] mapped to the "*OBESE*"; (If BMI value is 29.9 or less), [1.0] mapped to the "*NONOBESE*"

Thus, in the first stage, a multi-way tree is created that finds the feature (variable selection by filter method) that will maximize the value of information *(Gain*-Information Gain*)* using the entropy *("E"* as in (2)) of an impurity criterion. In this way, we identify more valuable and effective variables (features), which are important for the further stage of the study.

$$E(X) = \sum_{i=1}^{Y} - P(Y) \log_2 P(Y), \ (i=1,2,...,n) \quad (2)$$

Here, "*E(X)"*-is the dimension of the number of doubtful/uncertainty in the splitting set-"*X"* (that is "*E(X)"*-characterizes the set-"*X"*). "*X"*-is the splitting set (variables/subsets) for "*E(X)"*-entropy. "*Y"*-is the set of classes (i.e. dependent variable BMI{*OBESE/NONOBESE*}) in "*X"*. "*P(Y)"*-is the proportion of the number of instances in cases of *OBESE/NONOBESE*) to number of instances in splitting set-"*X"*. If all instances in splitting set-"*X"* are same class of *OBESE* or *NONOBESE*, it means that data perfectly identified (i.e. *E(X)=0*).

Gain(F) is a measure of the difference in entropy from before to after the set-"*X"* is split on an attribute (feature/variable) "*F"*. In other words, how much doubtful/uncertainty in "*X"* was reduced after splitting set-"*X"* on "*F"* and can be expressed as the following (3).

$$Gain(F, X) = E(X) - \sum_{t \in T} P(t)E(t) \quad (3)$$

Here, "*T"*-is external nodes (i.e. the subsets created from splitting set-X), "*P(t)"*-is the proportion of the number of instances in "*t"*-leaf of tree to the number of instances in set-"*X"*. "*E(t")*-is the entropy of "*t"*. This process lasts until further separation is not possible or acceptable with a set of splitting (i.e., child) nodes, and as a result, reach the target terminal node (this is the leaves of the tree). Thus, the training obesity dataset ("O_{Tr}") with 45 attributes and 325 instances is divided into parts (partitions) like the structure of tree. It consists of many external nodes "*T*" (i.e., terminal), in splitting set-"*X"* of dataset *(O_{Tr})* into subsets (attributes/variables) "F_t" where a terminal node ($t_1,t_2,...,t_{16}$) in the tree represents every subset (4):

$$Splitting\ set\ X = \bigcup_{t \in T} F_t;\ \forall\ t \neq t' : F_t \cap F_{t'} = \emptyset \quad (4)$$

Separation gradually leads to more complex models and therefore to the over fitting risk, which has a direct impact on performance of model. There are two ways to prevent the complexity of the model: the first is by introducing a series of hyperparameters that regulate the separation process; the second is robust to post hoc pruning of weights or nodes that frequently occur in the datasets (Kumar and Nirmalkumar, 2019). In this study, pruning was used as a method to simplify the model in order to reduce tree size by removing partitions (subspace) of features (in the DT) that provide little power to classify obesity factors. In addition to the above, there are several other parameters that impact simplify the tree model: (a) the minimum leaf size parameter, (b) the parameter of the criterion for stopping the separation process.

In the hybrid approach, the parameter ranges of the tree were the same as the traditional single-stage DT (Khraisat et al., 2020). In the machine-learning framework of our study, the second-stage of hybrid approach, is a classification. The LR was fitted to each selected variables F_i=(F_1, ..., F_{16}) to identify the causes of obesity (See Table 2). LR is a valuable standalone technique used in medicine (Taghiyev et al., 2019; De Melo, 2016), because (a) in LR probabilities and predictions are assessed directly that makes them more understandable than "black box" methods, and (b) logistic regression model provides exact and reliable outcomes in comparative research for issues of classification.

Table 2. List of Outputs (Features).

Code	Features	Indicators
F1	age (year)	[40-65)
F2	fam_type	extended
F3	mar_stat	widow
F4	edu_stat	high school; undergraduate
F5	part_edu_stat	secondary school; illiterate
F6	eng_phy_ac	yes
F7	dietexe_cha_hab	no; yes; sometimes
F8	num_preg	[4-7)
F9	num_births	[1-4)
F10	weight_op	very fat; fat; thin; norm
F11	weight (kg)	[70-100)
F12	height (cm)	[160-185)
F13	waist_circ (cm)	very risk; risk; norm
F14	systolic (mmHg)	pre-high; high; ideal; low
F15	diastolic (mmHg)	pre-high; ideal; high; low
F16	fbs (mg/dL)	pre-high; ideal; high; low

Thus, we have a instance of *n* observations of the pair (F_i,Y_i), *i*=1,2,…,*n*, where "F_i" is the value of the independent variable for *i*-th patient. In order to simplify notation, we use the *P(i)=E(Y|F$_i$)* to represent the logistic distribution formulation. It can be linearized by using the transformations, even if it is nonlinear and will be as in (5):

$$P_i = \frac{1}{1 + e^{-Z_i}} \quad (5)$$

Here, Z-is an input to the function that is the linear combination of variables and their regression coefficient.

e=2.72-is a base of natural log. If P_i is the case's *OBESE* probability, so (1-P_i) is the case's *NONOBESE* probability and the probability estimate output will lie in the range of 0&1. By dividing the case's *OBESE* probability to the case's *NONOBESE* probability, function (6) is received:

$$\frac{P_i}{1-P_i} = e^{Z_i} \tag{6}$$

We get the formula (7), if apply natural logarithm to both sides:

$$\ln\left(\frac{P_i}{1-P_i}\right) = \ln\left(e^{Z_i}\right) \Rightarrow L_i = \ln\left(\frac{P_i}{1-P_i}\right) = Z_i \Rightarrow$$
$$\Rightarrow Z_i = \beta_0 + \beta_1 F_{i1} + \beta_2 F_{i2} + \ldots + \beta_{16} F_{i16} \tag{7}$$

L is called the logit model, with the variables ($(F_{i1},F_{i2},\ldots,F_{i16})$=[$age_i$, fam_type_i, mar_stat_i, edu_stat_i, $part\_edu\_stat_i$, $eng\_phy\_ac_i$, $dietexe\_cha\_hab_i$, num_preg_i, num_births_i, $weight\_op_i$, $weight_i$, $height_i$, $waist\_circ_i$, $systolic_i$, $diastolic_i$, fbs_i]) and model β-parameters ($(\beta_0,\beta_1,\beta_2,\ldots,\beta_{16})$ are the regression coefficients for the corresponding variables (k-th values of features)).

Thus, a hybrid approach based on the Spark Machine-Learning library (Spark MLlib) was used to identify the causes of obesity (Meng et al., 2015; Guler, 2015). The proposed hybrid model gives better results than other single-stage classifiers (e.g. DT and LR) because it has mechanism for selecting variables and it selects significant variables. The procedure of variable selection has many benefits: (a) The classifier provides good performance of classification model when training on more effective and valuable variables than a model based on a complex set of missing or noisy data. (b)The choice of independent variables (feature) leads to a laconic model, as the number of variables decreases. (c)Models with minor independent variables are generally more effective in practice.

3. VALIDATION AND EVALUATION

In this study, the obesity problem was examined in women classified as *OBESE* and *NONOBES*. A validation known as holdout was used for this paper and obesity dataset ("O_{DB}" with 500 instances) divided into 2 parts: (a) Training (65%-"O_{Tr}" with 325 instances) and (b) Validation set (35%-"O_{Ts}" with 175 instances) (see Fig. 4).

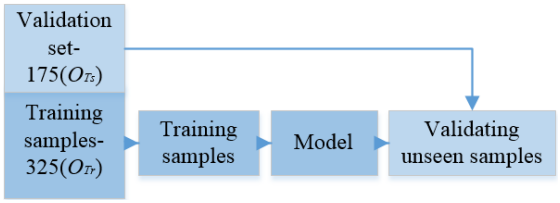

Fig. 4. All stages of validation.

The classification performance of the hybrid method was evaluated on the validation set (unseen instances) using the following measurements as in (8), (9), (10), (11), and (12), confusion matrix and compared with conventional single-stage algorithms:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$\text{Specificity} = \frac{TN}{FP+TN} \tag{9}$$

$$\text{Sensitivity} = \text{Recall} = \frac{TP}{TP+FN} \tag{10}$$

$$\text{Precision (Positive Predictive Value)} = \frac{TP}{TP+FP} \tag{11}$$

$$F_{measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

4. RESULTS

4.1. Experimental Evaluations

This study implemented on the virtual cluster environment, which installed on hardware: IntelXeon W-2145 CPU-3.70GHz. A proposed approach was performed using the Scala language (Scala version 2.11.8) in the Apache Spark (spark-2.3.1-bin-hadoop2.7.tgz) machine-learning environment (Meng et al., 2015). We used a holdout validation method. The first-stage of the proposed hybrid approach creates tree to determine more effective features-F_t that better classified in the second-stage. The outcomes of probability (prediction) of the factors that causes of obesity are shown in Fig. 5: *OBESE* outcome is [0.9215944726052586, 0.0784055273947414], 0.0], but *NONOBESE* outcome is [0.0523862837394741, 0.9476137162605259], 1.0].

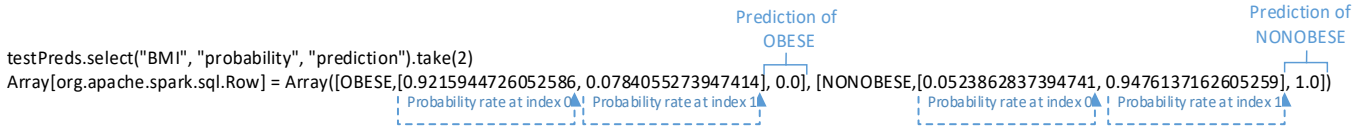

Fig. 5. Probability and prediction of OBESE and NONOBESE cases.

The probability is a vector, where 92.1594% indicates the probability index of "0" and 7.8406% indicates the probability index of "1". Thus, the higher prediction value (92.1594%) of the "0" index indicates that the *OBESE* value [0.0] was obtained on the output. In the other instance, 5.2386% indicates the probability index of "0" and 94.7614% indicates the probability index of "1" and the higher prediction value (94.7614%) "1" index, means that the *NONOBESE* value [1.0] was obtained on the output.

4.2. Performance of the Classification Model

First, the hybrid model is trained in the "O_{Tr}" (training dataset with 325 instances) then it is applied to the "O_{Ts}" (validation set with 175 instances). The evaluation metrics of

the model obtained and compared to the traditional standalone algorithms of DT and LR (see Fig. 6).

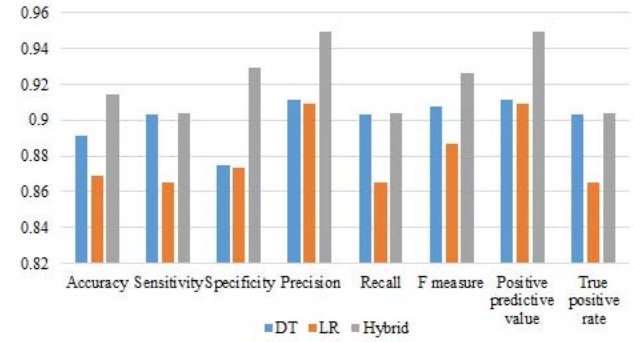

Fig. 6. Comparison of performance of classifiers.

Spark ML packages were used to evaluate the performance of the hybrid classification model, and the results obtained from the "O_{Ts}"-validation set are shown in Table 3 & Table 4.

Table 3. The confusion matrix of hybrid model.

		Predicted label	
		OBESE	**NONOBESE**
True label	**OBESE**	TP=94.0	FN=10.0
	NONOBESE	FP=5.0	TN=66.0

As shown in Table 4, the highest performance was achieved with a hybrid approach using machine learning and data mining methods. It can be observed that the performance of the hybrid model has improved compared to the evaluation metrics for supervised algorithms, such as stand-alone (single-stage) DT and LR.

Machine learning and data mining methods use odds ratio (probability coefficients) to identify the factors that cause obesity and can be used to comment on how influential factors are (see Fig. 7) (Koliopoulos et al., 2015). Obesity status was found in about half of the women interviewed. According to Table 5 and Fig. 7, coefficients greater than 1 were investigated as the factors affecting obesity. Among socio-demographic features with coefficients greater than 1: odds ratio of "*age=[40-65)*", "*fam_type=extended*", "*mar_stat=widow*", "*part_edu_stat=secondaryschool*", "*part_edu_stat=illiterate*", and "*num_preg=[4-7)*" increases 1.0989, 1.1910, 1.3341, 1.3517, 1.4886 and 1.225 times, respectively.

Table 4. Evaluation metrics of classifiers.

Metrics	**Classifiers**		
	OBESE label		
Performance metrics	DT	LR	Hybrid
Accuracy	0.8914	0.8685	0.9142
Sensitivity	0.9029	0.8653	0.9038
Specificity	0.8750	0.8732	0.9295
Precision	0.9117	0.9090	0.9494
Recall	0.9029	0.8653	0.9038
F measure	0.9073	0.8866	0.9261
Positive predictive value	0.9117	0.9090	0.9494
True positive rate	0.9029	0.8653	0.9038

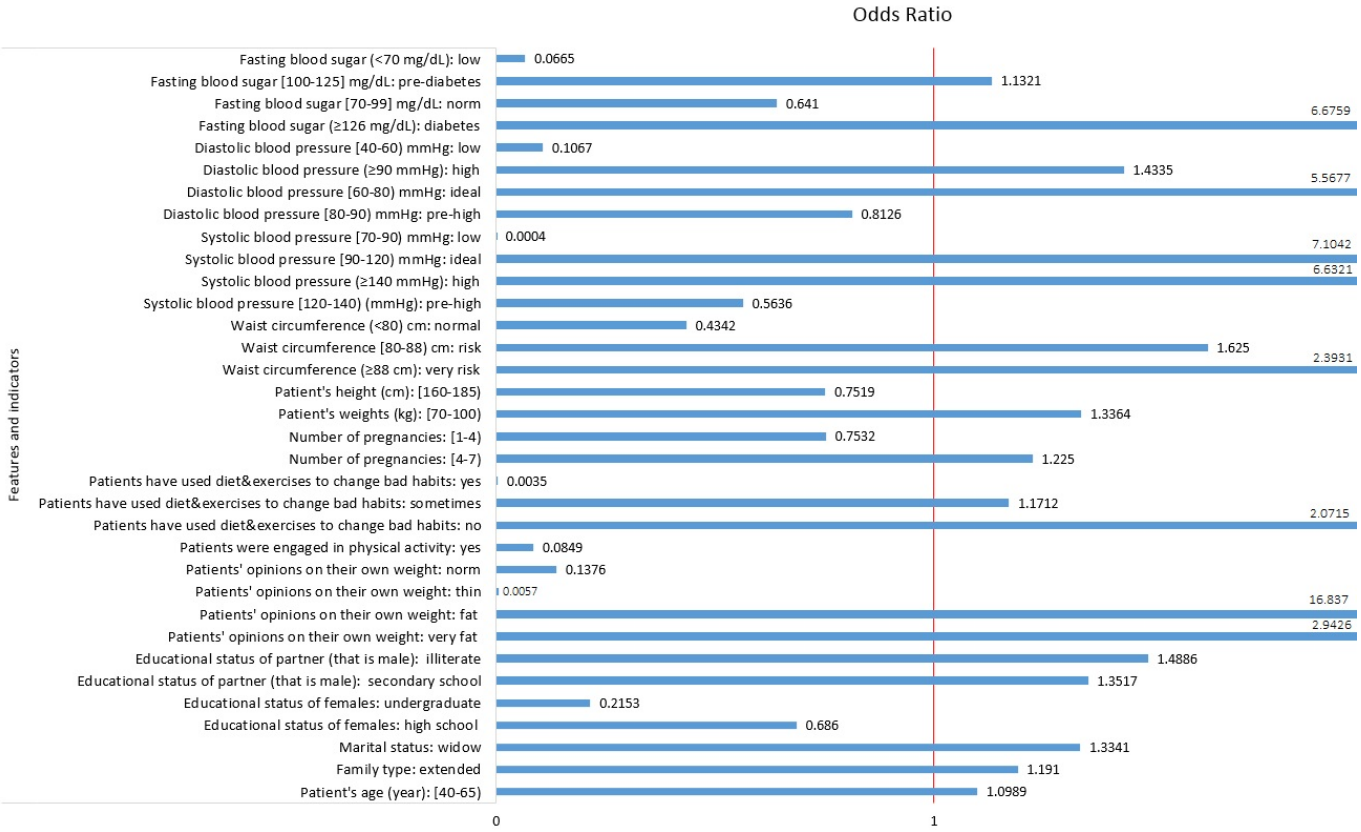

Fig. 7. Odds ratio of the factors that cause obesity.

Obesity increases about 1.2 times, in a woman who was four or more times pregnant. Odds ratio of "*weight_op=veryfat*", "*weight_op=fat*", "*dietexe_cha_hab= no*", "*dietexe_cha_hab=sometimes*","*weight*=[70-100)", "*waist_circ=veryrisk*", and "*waist_circ=risk*" of variables such as body perception, diet and activity, and anthropometric measurements increases 2.9426, 16.837, 2.0715, 1.1712, 1.3364, 2.3931, 1.625 times, respectively. In females with body perception of "*fat* and *very fat*", the risk of obesity increases about 17 and 3 times, respectively. Women who sometimes follow or do not follow a diet and exercise to change their bad habits increase the risk of obesity 2 and 1.2 times, respectively. In women with a weight of 70-100 kg, the risk of obesity increases 1.2 times, while in women with a waist circumference of 80-88 cm and ≥88 cm, the risk of obesity increases 1.6 and 2.4 times, respectively. Odds ratio of "*systolic=high*", "*diastolic=high*" variables, i.e., blood pressure measurements, were defined as 6.6321 and 1.4335 respectively.

Table 5. The outcomes (odds ratio).

Features	Indicators	Odds Ratio
age (year)	[40-65)	1.0989
fam_type	extended	1.1910
mar_stat	widow	1.3341
edu_stat	high school	0.6860
	undergraduate	0.2153
part_edu_stat	secondary school	1.3517
	illiterate	1.4886
weight_op	very fat	2.9426
	fat	16.837
	thin	0.0057
	norm	0.1376
eng_phy_ac	yes	0.0849
dietexe_cha_hab	no	2.0715
	sometimes	1.1712
	yes	0.0035
num_preg	[4-7)	1.2250
num_births	[1-4)	0.7532
weight (kg)	[70-100)	1.3364
height (cm)	[160-185)	0.7519
waist_circ (cm)	very risk (≥88 cm)	2.3931
	risk [80-88)	1.6250
	norm (<80)	0.4342
systolic (mmHg)	pre-high [120-140)	0.5636
	high (≥140)	6.6321
	ideal [90-120)	7.1042
	low [70-90)	0.0004
diastolic (mmHg)	pre-high [80-90)	0.8126
	ideal [60-80)	5.5677
	high (≥90)	1.4335
	low [40-60)	0.1067
fbs (mg/dL)	diabetes (≥126)	6.6759
	norm [70-99]	0.6410
	pre-diabetes [100-125]	1.1321
	low (<70)	0.0665

It follows that on systolic blood pressure of 140 mmHg and over and diastolic blood pressure of 90 mmHg and over, the risk of obesity increases 6.6 and 1.4 times, respectively. Odds ratio of "*systolic=ideal*" and "*diastolic=ideal*" variables are 7.1042 and 5.5677, respectively. In women with systolic blood pressure in the range of [90-120) mmHg and diastolic blood pressure in the range of [60-80) mmHg, the risk of obesity increases 7 and 5.6 times, respectively. Odds ratio of "*fbs=diabetes*" and "*fbs=prediabetes*" increases 6.6759 and 1.1321 times, respectively.

It means that the risk of obesity has increased by 1.1 times when the fasting blood glucose level was between 100-125 mg/dL, and the risk of obesity has increased by 6.7 times when the fasting blood glucose level was 126 mg/dL or higher. Odds ratio for "*eng_phy_ac=yes*" and "*dietexe_cha_hab=yes*" are 0.0849 and 0.0035, respectively. Odds ratio that is less than 1 indicates that features are effective in preventing obesity. For instance, odds ratio for "*edu_stat=highschool*", "*edu_stat= undergraduate*", "*waist_circ=norm*", "*weight_op=thin*", "*weight_op= norm*", "*height=[160-185)*", "*fbs=low*", "*fbs=norm*", "*systolic=prehigh*", "*diastolic=prehigh*", "*systolic=low*" and "*diastolic=low*" are 0.686, 0.2153, 0.4342, 0.0057, 0.1376, 0.7519, 0.0665, 0.641, 0.5636, 0.8126, 0.0004 and 0.1067, respectively. It means that females with high education and bachelor's degree, normal waist circumference, weight (thin or normal), height in the range [160-185), low blood presure and fast blood glucose level of 99 mg/dL and less were protected from obesity (see Fig. 7). Odds ratio of "*num_births=[1-4)*" was 0.7532, i.e. obesity prevention was observed in females with fewer than four births.

5. DISCUSSION

Data science should be more technologically advanced, to allow data engineers and doctors to use machine-learning methods to transform raw data into useful information and facts to achieve best solutions to healthcare problems. In literature, we may find that by combining several relevant algorithms one can obtain better results. For instance, (Akgül et al., 2019) presented a hybrid approach using ANN and GA and obtained similar results. (Ramírez et al., 2020) applied a hybrid system: NN&FL for 2-lead cardiac arrhythmia classification and obtained similar results, classification rate of 90.3% by using a cross-validation method. (Devi et al., 2020) proposed a hybrid model based on Farthest First and SMO algorithms for diagnosing diabetes mellitus with improving accuracy. (Khraisat et al., 2020) utilized a hybrid; the Stacking Ensemble of C5 Decision Tree Classifier&One-Class Support Vector Machine, but lower results were obtained. (De Melo, 2016) employed LR by Kaizen programming in a hybrid approach for automatic feature construction for breast cancer detection. Unlike the studies above, the proposed hybrid system gives 91.4 of accuracy, which is better than other classifiers.

In our study, the risk of obesity increases for women with socio-demographic characteristics, such as women aged [40-65], whose husbands have secondary education or no education, extended family, widows, and women who have had pregnancies 4 and over (See Fig. 7). Other similar study

shows that women's age, marital status, family structure, number of pregnancies and educational status are effective on obesity (Eunji et al., 2019; Fuentes et al., 2019; Pekcan et al., 2017). In this case, slowing down basal metabolism and the transition to a sedentary lifestyle may be effective with age. The risk factors of obesity, such as the low level of education of the partner (i.e. male), marital status, the family structure may be explained by psychosocial and cultural reasons. The effect of pregnancy on obesity may be due to the inability to lose weight gained during pregnancy. Besides, the fact that women may behave unconsciously about nutrition during pregnancy and lactation periods or their knowledge is insufficient may be effective in the increasing mild overweight or obesity. The risk of obesity increases in women with body perception, such as fat or very fat, as well as the risk of obesity increases in women who sometimes follow or do not follow, a diet and exercise to change their bad habits (See Fig.7), and it may be explained by a lack of knowledge and awareness of being obese. Other similar study gives the same results. (Pekcan et al., 2017).

In our study, the high blood pressure increases the risk of obesity (See Fig.7). Similar results were obtained in another study. (Nurdiantami, et al., 2018) High systolic and diastolic blood pressure and lack of treatment with antihypertensive drugs increase the risk of cardiovascular disease in overweight women. Systolic blood pressure in the range of [90-120) mmHg and diastolic blood pressure in the range of [60-80) mmHg, the risk of obesity increases 7 and 5.6 times, respectively. Obese females could control systolic and diastolic blood pressure with antihypertensive drugs (Murray et al., 2020), but the risk of obesity is high in these females. When the obesity problem is solved, the blood pressure may return to normal spontaneously. In our study, the risk of obesity increases when fasting blood glucose rises (See Fig.7) and similar results were obtained in another study (Campbell et al., 2019; Canan and Sazi, 2006). Obesity was observed in about half of the women participating in the study and most of them were physically inactive for socio-cultural reasons. In this study, physical activity and diet were found to have significant protective effects in obese women (See Fig. 7.) (Eunji et al., 2019; Wang et al., 2012).

Females with high education and bachelor's degree, normal waist circumference, weight (thin or normal), height in the range of [160-185), low blood pressure, low number of births, and fast blood glucose level of 99 mg/dL and less have a positive effect on obesity prevention (See Fig. 7); and similar results were obtained in another study (Kim et al., 2019; Nurdiantami et al., 2018; Pekcan et al., 2017). This may be explained by an increase in the level of consciousness of women with an increase in the level of education in society, and in parallel with this increase in personal health awareness. Since women eat healthily, their weight, waist circumference, blood pressure, and blood sugar levels are on normal levels, so we may think that they are protected from obesity.

6. CONCLUSION

The study presents outcomes on the main factors affecting obesity and discusses the implications of these factors. The proposed hybrid system provides a more practical approach that has yielded excellent and accurate outcomes (91.4% accuracy, 90.4% sensitivity, and 92.9% specificity) than LR or DT separately. In the fist-stage, the model is ability to select variables-F_t (more effective) by DT and in the second-stage, the variables-F_t are investigated more profound by using LR to account for specific group characteristics that would otherwise remain unknown. This approach provides a conceptually simple, effective, and accurate model.

Table 5 and Fig.7 present the factors that most affect the women who participated in the study. Approximately half of the females participating in the study were obese, and the causes of the problem were investigated. The risk of obesity increases with increasing age of women, the number of pregnancies, blood pressure, body weight, and blood glucose. There is also an increased risk of obesity in widowed women, in extended families, women with body perception such as fat and very fat, and women with a partner who has low education. Woman with high education or bachelor's degree, "*norm*" waist circumference, weight ("*thin*" or "*norm*"), height *[160-185)* (cm), physical activity and diet, low blood pressure, number of births above four, and fast blood glucose level of 99 mg/dL or less were protected from obesity.

In line with these results, women should be provided with training and counseling on the risks of obesity and possible health problems. Women should be informed about the weight that they should gain during pregnancy, how to lose weight after pregnancy, and the recommendations should be related to nutrition and physical activity. In order to protect, maintain and improve health, it is important to control the factors affecting obesity and to prepare health education programs for primary healthcare professionals. Primary healthcare professionals should be better informed on these issues.

In the future, our dataset may be used to study in detail the relationship between T2D and obesity (Albahli, 2020; Campbell, et al., 2019).

ACKNOWLEDGEMENTS

A. TAGHIYEV, A. A. ALTUN and S. CAGLAR thank the Selcuk University Scientific Research Projects Coordination Office (No: 18201154), Ethics Committee of Medicine faculty at Selcuk University, Aksaray Provincial Health Directorate and Aksaray Sultanhani Family Health Center for their support (No. 66472688-000-1966; No. 66472688-773.03).

REFERENCES

Adnan M., et al., (2010). A survey on utilization of data mining for childhood obesity prediction, *8th Asia-Pacific Symp.Infor.&Telecom.Tech.,* Kuching, pp. 1-6.

Akgül M., et al., (2019). Diagnosis of Heart Disease Using an Intelligent Method: A Hybrid ANN–GA Approach, *Advances in Intel.Sys.&Comp.,* 1029:1250-1257. DOI:10.1007/978-3-030-23756-1_147

Albahli S., (2020). Type 2 Machine Learning: An Effective Hybrid Prediction Model for Early Type 2 Diabetes

Detection, *J.Med.Imag. &Health Inf.,* 10(5):1069-1075. DOI: 10.1166/jmihi.2020.3000

Ali L., et al., (2019a). Reliable Parkinson's Disease Detection by Analyzing Handwritten Drawings: Construction of an Unbiased Cascaded Learning System Based on Feature Selection and Adaptive Boosting Model, in *IEEE Access*, vol.7, pp.116480-116489. DOI: 10.1109/ACCESS.2019.2932037

Ali L., et al., (2019b). An Automated Diagnostic System for Heart Disease Prediction Based on $\chi 2$ Statistical Model and Optimally Configured Deep Neural Network, in *IEEE Access*, vol.7, pp. 34938-34945, 2019. DOI: 10.1109/ACCESS.2019.2904800

Ali L., et al., (2019c). A Feature-Driven Decision Support System for Heart Failure Prediction Based on $\chi 2$ Statistical Model and Gaussian Naive Bayes, *Computational and Mathematical Methods in Medicine*, vol.2019, ArticleID 6314328, 8 pages. DOI: 10.1155/2019/6314328

Ali L, et al., (2019d). Early diagnosis of Parkinson's disease from multiple voice recordings by simultaneous sample and feature selection, *Expert Systems with Applications*, Vol.137, pp.22-28. DOI: 10.1016/j.eswa.2019.06.052

Ali L, et al., (2019e). Automated Detection of Parkinson's Disease Based on Multiple Types of Sustained Phonations Using Linear Discriminant Analysis and Genetically Optimized Neural Network, in *IEEE J.Tran.Eng. in Health&Med.*, vol.7, pp. 1-10. DOI: 10.1109/JTEHM.2019.2940900

Campbell, J.A, et al., (2019). Prevalence of diabetes, prediabetes, and obesity in the indigenous kuna population of Panamá. *J.Racial&Eth.Heal.Disp.*6, pp.743-751. DOI: 10.1007/s40615-019-00573-0

Canan E, Sazi I, (2006). Comparison of the obesity risk and related factors in employed and unemployed (housewife) premenopausal urban women, *Diabetes Research and Clinical Practice,* Vol.72, Issue 2, pp-190-196, DOI: 10.1016/j.diabres.2005.10.010

De Melo V.V., (2016). Breast cancer detection with logistic regression improved by features constructed by Kaizen programming in a hybrid approach, *2016 IEEE Cong.on Evol.Comp.*, 16-23. DOI: 10.1109/CEC.2016.7743773

Devi R.D.H., Bai A., Nagarajan N., (2020). A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms, *Obesity Medicine*, vol.17. DOI: 10.1016/j.obmed.2019.100152

Dugan TM., et al., (2015). Machine learning techniques for prediction of early childhood obesity*, App.clin.infor.,* 6(3):506-520. DOI: 10.4338/ACI-2015-03-RA-0036

Ergün U., (2009). The Classification of Obesity Disease in Logistic Regression and Neural Network Methods, *J.Med Syst.,* 33:67. DOI: 10.1007/s10916-008-9165-5

Eunji C, et al., (2019). Socioeconomic inequalities in obesity among Korean women aged 19-79 years: the 2016 Korean Study of Women's Health-Related Issues. *Epid.&health* vol.41: DOI:10.4178/epih.e2019005

Figueroa RL., and Flores CA., (2016). Extracting information from electronic medical records to identify the obesity status of a patient based on comorbidities and bodyweight measures, *J.Med.Syst.,* 40(8): 191. DOI: 10.1007/s10916-016-0548-8

Fuentes S, et al., (2019). Psycho-social factors related to obesity and their associations with socioeconomic characteristics: the RECORD study. *Eat Weight Disord.* DOI: 10.1007/s40519-018-00638-9

Guler M., (2015). Machine Learning with Spark. In: *Big Data Analytics with Spark, Apress,* Berkeley, 153-205. DOI: 10.1007/978-1-4842-0964-6_8

Hu R., (2011). Medical Data Mining Based on Decision Tree Algorithm, *J.Comp.&Inform.Sci.,* Vol.4, no.5. DOI: 10.5539/cis.v4n5p14

Khraisat A., et al., (2020). Hybrid Intrusion Detection System Based on the Stacking Ensemble of C5 Decision Tree Classifier and One Class Support Vector Machine, *Electronics* 9(1)173. DOI: 10.3390/electronics9010173

Kim D, et al, (2019). Factors affecting obesity and waist circumference among US adults. *Preventing chronic disease*, 16, E02. DOI: 10.5888/pcd16.180220

Koliopoulos A., et al., (2015). A Parallel Distributed Weka Framework for Big Data Mining Using Spark, *2015 IEEE Inter.Congress on Big Data,* pp.9-16. DOI: 10.1109/BigDataCongress.2015.12

Kumar N.S., Nirmalkumar P.A., (2019). Robust Decision Support System for Wireless Healthcare Based on Hybrid Prediction Algorithm, *J.Med Syst* 43, (170). DOI: 10.1007/s10916-019-1304-7

Lin C.J., et al., (2019). Supervised and Reinforcement Group-based Hybrid Learning Algorithms for TSK-type Fuzzy Cerebellar Model Articulation Controller, *J. of Control Engineering & Applied Infor.,* vol. 21, no 2, pp.11-21.

Meng X., et al., (2015). Mllib: Machine learning in apache spark, *J.Mach.Learn.Res.,* vol.17, pp.1-7. https://arxiv.org/abs/1505.06807

Muhamad A.M.H.B, et al., (2012). A hybrid approach using Naive Bayes and Genetic Algorithm for childhood obesity prediction, *2012 Inter.Conf.Comp.&Infor.Sci.,* pp.281-285. DOI: 10.1109/ICCISci.2012.6297254

Murray P.M., et al., (2020). Forecasting Ontario Oncology Drug Expenditures: A Hybrid Approach to Improving Accuracy, *Ap.Health Econ.Health Pol.,* 18, pp.127-137. DOI: 10.1007/s40258-019-00533-z

Ni J., et al., (2020). A hybrid model for predicting human physical activity status from lifelogging data, *Eur.J.Ope. Res.,* 281(3):532-542. DOI: 10.1016/j.ejor.2019.05.035

Nurdiantami Y, et al., (2018). Association of general and central obesity with hypertension, *Clin Nutr. Aug*; 37(4): pp.1259-1263. DOI: 10.1016/j.clnu.2017.05.012

Pekcan AG, et al. (2017). Population based study of obesity in Turkey: results of the Turkey utrition and health survey-2010, *Prog. inNutrition*; 19(3): 248-256.

Ramírez E., et al., (2020). Hybrid Model Based on Neural Networks and Fuzzy Logic for 2-Lead Cardiac Arrhythmia Classification, *Hyb.Int.Sys.Con,Pat. Rec.&Med.,* vol 827, pp. 193-217. DOI: 10.1007/978-3-030-34135-0_14

Shi P., et al., (2019). A hybrid model using LSTM and decision tree for mortality prediction and its application in provider performance evaluation, *IEEE Int.Con.BD*, 2773-2781. DOI: 10.1109/BigData47090.2019.9005958

Taghiyev A., Altun A.A., et al., (2019). A Machine Learning Framework to Identify the Causes of HbA1c in Patients With Type 2 Diabetes Mellitus, *Journal of Control Engineering and Applied Informatics,* vol. 21, no 2, pp.34-42. DOI: 10.6084/m9.figshare.11980896.v1

Turkey: Ministry of Health, (2019). Turkish public health institution. Department of Obesity, *Diabetes and Metabolic Diseases,* [Online] Available: http://www.saglik.gov.tr/ [Accessed: 2019].

Turkey: Ministry of Health, (2014). Public Health Agency of Turkey, *Turkey Healthy Nutrition and Active Life Program 2014-2017, Pub.no.773,* [Online] Available: https://hsgm.saglik.gov.tr/depo/birimler/saglikli-beslenme-hareketli-hayat-db/Yayinlar/programlar/hareketli-hayat-programi-2014-2017.pdf [Accessed: 2019].

Wang H., et al., (2012). Epidemiology of general obesity, abdominal obesity and related risk factors in urban adults from 33 communities of northeast china: the CHPSNE study. *BMC Pub.Health* 12, 967. DOI: 10.1186/1471-2458-12-967

World Health Organization, (2019). *Health topics, Obesity,* [Online] Available: http://www.who.int/topics/obesity/en/ [Accessed: 2019].

Yang H., Garibaldi JM, (2015). A hybrid model for automatic identification of risk factors for heart disease, *J.of Biomedical Informatics,* vol.58, Supplement, 2015, pp.S171-S182. DOI: 10.1016/j.jbi.2015.09.006n