Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

A Classification Model for Predicting Fetus with down Syndrome – A Study from Turkey

Alptekin Durmuşoğlu, Memet Merhad Ay & Zeynep Didem Unutmaz Durmuşoğlu

APPLIED ARTIFICIAL INTELLIGENCE
2020, VOL. 34, NO. 12, 898–915
https://doi.org/10.1080/08839514.2020.1790246

A Classification Model for Predicting Fetus with down Syndrome – A Study from Turkey

Alptekin Durmuşoğlu[a], Memet Merhad Ay[b], and Zeynep Didem Unutmaz Durmuşoğlu[a]

[a]Department of Industrial Engineering, Gaziantep University, Gaziantep, Turkey; [b]Department of Industrial Engineering, Erciyes University, Kayseri, Turkey

ABSTRACT

The triple test is a screening test (blood test) used to calculate the probability of a pregnant woman having a fetus that has a chromosomal abnormality like Down Syndrome (DS). AFP (Alpha-Fetoprotein), hCG (Human Chorionic Gonadotropin), and uE3 (Unconjugated Estriol) values in the blood sample of pregnant women are computed and compared with the similar real records where the outputs (healthy fetus or a fetus with DS) are actually known. The likelihood of the indicators is used to calculate the probability of having a fetus with chromosomal abnormality like DS. However, high false positive rate of the triple test has been a problematic issue. One of the reasons of the high false positives is the differences in the norm values of indicators for the pregnant women from different geographical regions of a country. We use 81 patient records retrieved from Şahinbey Training and Research Hospital of Gaziantep University; Turkey. In our study, nine different classification algorithms were trained based on triple test indicators. Multilayer perceptron outperformed with 94.24% detection rate and 13% false positive rate. The multilayer perceptron can predict the outcome of triple test with a high level of accuracy and fewer patients are suggested for amniocentesis. This study is the first study using the MLP model for Turkish triple test data. Regional MLP models can eliminate the bias due to local biological differences.

Introduction

One of the most common chromosomal defect in fetuses is known to be Down Syndrome (DS or trisomy 21) (Shurtz, Brzezinski, and Frumkin 2016). DS does significantly impact both quality and length of life of individuals having this abnormality (Temming and Macones 2016) and their families. In this respect, it has been important for pregnant to be informed about DS existence before the birth. Various maternal serum biomarkers with or without ultrasonography measurement are commonly used for DS screening (Ökem et al. 2017). Measurement of maternal serum alphafetoprotein (AFP), human

CONTACT Alptekin Durmuşoğlu ✉ durmusoglu@gantep.edu.tr Department of Industrial Engineering, Gaziantep University, Gaziantep, Turkey

chorionic gonadotrophin (hCG), and unconjugated estriol (uE3) at the beginning of the second trimester of pregnancy (called as triple test) is a well-established screening test for DS (trisomy 21) (Witters et al. 2001). It is a blood test typically performed during the second trimester (Shurtz, Brzezinski, and Frumkin 2016) and these three markers are used in combination to modify the maternal age-related risk of DS (Founds 2014) and thus determine individual risk of fetal DS (Shaw, Chen, and Cheng 2013). In the calculation of associated risk, maternal age-related risk is multiplied by likelihood ratios, determined according to the deviation of the measured levels of three markers from the expected median values. Most of existing approaches use posterior probabilities based on the median and the standard deviation of the markers, or by using a suitable multivariate statistical approach (Neocleous, Nicolaides, and Schizas 2016). Screening results that are equal to or greater than 1:274 (the risk of a 35-year-old for fetal Down syndrome at the second trimester) are accepted as positive (Phillips et al. 1992).

On the other hand, noninvasive identification of fetuses with DS is a diagnostic challenge (Yagel et al. 1998). Even though there have been some modifications in the method of prenatal screening over the last few years to increase the accuracy of the used method (Kaur et al. 2013) the best detection rate was obtained with four maker test which is 65% for 5% false positive rate (Wald et al. 1994). If gestational age is included (obtained via ultrasound scan) then detection rate increases to 72% with the same false positive rate. An abnormal screening result usually ends with suggestion of chromosomal analysis of amniotic fluid (amniocentesis) which is considered as confirmatory follow-up test (Tamminga et al. 2016). Fatal loss risk of amniocentesis is not outweighed by the rate of adverse obstetrical outcome induced by amniocentesis (Muller et al. 2002). However there is still a risk associated with amniocentesis. Increasing the accuracy of detection rate of screening tests can be assistive to avoid risks associated with amniocentesis. Predictive classification with triple test data refers to the assignment of a particular unknown blood sample to a DS class based on its similarity to certain quantitative patterns from learning data set. This classification can be performed by a training predictive classifier, like a neural network classifier. Neural networks provide an effective and promising platform for medical data analysis and especially classification, since they allow us to solve rather complicated classification tasks (Autio, Juhola, and Laurikkala 2007).

It is expected to be useful to employ neural network classifiers while the detection rate is low to recognize DS by traditional statistical approaches. In this paper, we propose a specific type of neural networks (Paiva, Cardoso, and Pereira 2018), multilayer perceptron model, to predict risk of DS in a more accurate manner.

The remainder of this paper is organized as follows: In Section 2, a description of the data and the methodology employed is presented.

Section 3 shows the results obtained. Discussion and conclusion are presented in Sections 4.

Markers of Triple Test

Screening methods needs to be planned and organized to be applied as routine clinical practice. This process covers selection of markers which there is sufficient scientific evidence of efficacy, quantifying performance in terms of detection and false positive rates (Wald et al. 1997). There are different markers which have been proven to be important for prenatal screening. One of these markers is maternal serum alpha-fetoprotein. Lower level of maternal serum AFP values has been associated with the increased risk for DS. Almost 35% of Down's syndrome in fetuses can be identified by measuring maternal serum alpha-fetoprotein during the second trimester in the general population of pregnant women (Haddow et al. 1992). In the late 1980s, high levels of another serum marker, hCG were found to be associated with DS (Driggers and Seibert 2008). Later on, uE3 measurements made the test a triple test by the end of the 1990s (Shaw, Chen, and Cheng 2013; Tamminga et al. 2016). Recent case-control studies indicate that current detection rate can be approximately doubled by measuring serum levels of unconjugated estriol hCG, which are abnormally low and abnormally high, respectively, in women carrying fetuses affected by Down's syndrome (Haddow et al. 1992).

Maternal age, gestational age, gestational weight, and smoking have been other relevant factors affecting the value of the tree markers and the associated risk. Maternal age is the age of mother candidate at the beginning of pregnancy. It is well established that the risk for DS increases with maternal age (Harris, Reed, and Vora n.d.). By the maternal age of 40, the risk of delivering an affected term newborn with DS is 1% (Skrzypek and Hui 2017). On the other hand, as a screening test it has poor performance alone (Norton and Rink 2016). If maternal age is included, the gestational age also needs to be specified (Benn 2016). Gestational age is the total duration that a baby has been in the uterus. It can be calculated using the current date and the patients estimated date of delivery. To compare individual results, values for AFP, HCG, and u-E3 were expressed as a multiple of the medians (MOM) for gestational age (Bar-Hava et al. 2001).

Studies show that there is also a significant relationship between the marker levels and the weight. At the beginning it was found that heavier pregnant women have lower median values of AFP due to larger blood volume (Crandall et al. 1983; Haddow et al. 1981). In another study, this relationship was obtained for all markers (Reynolds, Vranken, and Nueten 2006). Since weight is an important determinant of marker levels, obesity rises the risk of failure of noninvasive prenatal screening regardless of gestational age (Yared et al. 2016). On the other hand, maternal weight adjustments can be used to correct the related problems (Wald et al. 1996).

Serum marker levels may be different in women who smoke and who do not (Wald et al. 1997). Therefore, smoking habits should be taken into account for risk assessment (Engels et al. 2014). Smoking significantly reduces median levels of uE3 and hCG while increasing the AFP (Zhang et al. 2011).

Multilayer Perceptron Models

Multilayer perceptron (MLP) is a feed forward artificial neural network model (Brasil, de Azevedo, and Barreto 2001) which provides the linkage between the sets of input data and a set of outputs (Aye and Heyns 2015). The neurons in MLP are interconnected in a one-way and one-directional fashion (Alameer et al. 2019). A classical MLP model has three kind of layers: an input layer, one or several hidden layers, and one output layer (Bienvenido-Huertas et al. 2019b; West and West 2000). Particularly, each unit from one layer is connected with all the units from the following layer. The hidden layer processes and transmits the input information to the output layer (Orhan, Hekim, and Ozer 2011). The value of the response predicted by the model corresponds to the output of the neuron of the last layer. The output value is simply the sum of the values of the neurons of the previous layers which are weighted by synaptic weights and by using activation, transference, and propagation functions (Bienvenido-Huertas et al. 2019a). MLP learns the complexity of the data and optimizes the weights to minimize classification error (Mulongo et al. 2019).

The MLPs are one of the well-known and widely applied artificial neural networks architectures with their capacities of universal approximation. We preferred the MLP neural network for this study since it produces highly accurate results particularly in problems requiring classification, recognition and generalization (Avuçlu and Başçiftçi 2018; Mukherjee 2018). MLPs have also certain advantages to map nonlinear relationships in the data (Güler et al. 1998). Since the multilayer perceptron is used as a classifier, the attributes of the model will be our network inputs while the network outputs is the actual classes defined for the problem (Setsirichok et al. 2012).

Methods

Data Acquisition and Preprocessing

The data used in this study was obtained from Şahinbey Training and Research Hospital of Gaziantep University, which is managed by the antenatal care unit for the years between 2010 and 2016.

Our purpose and data retrieval was approved by the local committee of ethics (Gaziantep University). The patient records and data were gathered from different departments of the hospital such as obstetrics and gynecology clinic, biochemistry laboratory, and molecular genetics laboratory of the hospital. Maternal serum samples that had AFP, hCG, uE3 levels, and maternal age, were taken from the triple screening test results saved by the biochemistry laboratory. Since, the patients with higher risk of having a fetus with DS, are forwarded to amniocentesis, records of corresponding triple test results were matched by accessing the amniocentesis report of each patient from the molecular genetics laboratory.

At the beginning, we have obtained data of 6340 patients who had applied to obstetrics and gynecology clinic due to routine control of a fetus. 324 of these patients aborted, and 2815 of them were routed to have amniocentesis. To fulfill the main purpose of this analysis, a patient data must be complete (triple test and amniocentesis or birth should be performed at the same hospital to have the full record) to be considered in this study. We have checked each of patients who gives birth healthy/with DS and who had the amniocentesis by a patient number, file number and patient names. The remaining incomplete patient data was removed from the data set. The removed patients indicate that they have not visited the same hospital for all of the examinations during their pregnancy and births. After the removal, there were 81 full records which have the whole data including the genetic disorder status of babies. Seventy-six of them had no genetic disorder and 5 of them had trisomy 21 (DS).

Balancing of Imbalanced Data

Imbalanced data typically refer to a problem regarding the unequal representation of classes. In our data set we have 5 records associated with Down syndrome and the remaining 76 as unaffected. Our minority class members are the records labeled with DS and our main objective is to predict the minority class in a high accuracy. However, the most of the well-known data mining algorithms becomes unattractive in case of imbalance in the data sets, as the distribution of the data sets is not taken into consideration when these algorithms are designed (Han, Wang, and Mao 2005). Specifically, the generalization performance of MLPs trained with the unbalanced training subsets will be quite poor (Daqi, Chunxia, and Yunfan 2007). Many techniques have been developed to tackle the problem of imbalanced training sets. One of the widely applied balancing methodologies has been SMOTE. It considers each member from the minority class and generates new synthetic members along the lines between it and some of randomly selected its k nearest neighbors from the minority class (Abidine et al. 2014; Maciejewski and Stefanowski 2011).

SMOTE is widely applied since it can create new instances rather than replicate the existing instances (Jeatrakul, Wong, and Fung 2010). We have also applied SMOTE algorithm to increase number of DS cases in our data set. The algorithm (Chawla et al. 2002) that is outlined below was implemented to our data set. As described, synthetic data records were generated by calculating the difference between minority sample and its nearest neighbors(He et al. 2018). Subsequently, this difference was multiplied by a random number between 0 and 1, and was added to the feature vector under consideration. Thereby, random data points along the line segment between two specific features were generated. Oversampling amount (S) is a system parameter where different receiver operating characteristics (ROC) curves can be generated. Area under the ROC curves is used to measure the performance of the classification problems. While ROC is a probability curve that indicates how much model is capable of distinguishing between classes, the higher area shows a better model that is predicting each of the class values. Therefore, an area of 1.0 represents a perfect accuracy for the given classification problem.

Algorithm SMOTE (M, S, k)

```
Input:
Number of individuals in minority class: M;
Amount of SMOTE: S%;
Number of nearest neighbors: k
Output: (S/100)* M synthetic minority class members
if S < 100 (randomize the minority class members-SMOTE only a random percent of them)
then Randomize the M minority class members
    M = (S/100)*M
    S = 100
end if
S = (int)(S/100)(SMOTE amount is assumed to be in integral multiples of 100)
k = Number of nearest neighbors
attrs = Number of attributes
Sample [][]: array for original minority class members
newind: number of synthetic samples generated, initialized to 0
Synthetic [][]:array for synthetic members(*Compute k nearest neighbors for each minority class sample only.*)
for i←1 to M
    Compute k nearest neighbors for i, and save the indices in the nnarray
    Populate (S, i, nnarray)
endfor
Populate (S, i, nnarray)(generate the synthetic members)
while S = 0
    Let's select a random number between 1 and k, call it nn. This step picks one of the k nearest neighbors of i.
```

```
        for attr←1 to attrs
            Compute: dif = Sample[nnarray[nn]][attr]−Sample[i][attr]
            Compute: gap = a random value between 0 and 1
            Synthetic[newind][attr] = Sample[i][attr]+gap*dif
        endfor
        newind++
        S = S − 1
    endwhile
    return (*End of Populate.*)
End of Pseudo-Code.
```

In this respect, we over-sampled our data set at 100%, 200%, 300%, 400%, 500%, 600%, 700%, 800%, 900%, 1000%, 1100%, 1200%, 1300%, 1400%, 1500%, and 2000% of original size by using SMOTE algorithm. Amount of SMOTE (S) and its corresponding area under ROC curves was calculated as given at Table 1. The best ROC area was found at the amount of %900 SMOTE (0.897). For this setting, number of records labeled as DS increased to 50.

Algorithms Used in Classifying Data

We have implemented several different well-known algorithms (Lin, Ke, and Tsai 2017) to classify the fetus as with/without DS. The summary description of the classifiers is as follows.

ZeroR

The zero-rules classifier (0-R) is a classifier which assigns class of each sample member to the class with highest prior probability (Pota et al. 2015). Therefore, all instances are labeled with the mean (for a numeric class) or the mode (for a nominal class) of the dataset (Rani and Jyothi 2016). It is the most basic classification algorithm therefore; it is beneficial for determining a baseline performance to compare with other classifiers. In this study, we also use 0-R algorithm as the baseline model to make robust comparisons.

Table 1. Amount of SMOTE and area under the ROC curve.

Amount of SMOTE	Area under ROC Curve	Amount of SMOTE	Area under ROC Curve
0	0.207	*900**	*0.897*
100	0.424	1000	0.802
200	0.634	1100	0.838
300	0.765	1200	0.870
400	0.842	1300	0.848
500	0.783	1400	0.886
600	0.799	1500	0.843
700	0.825	2000	0.882
800	0.842		

** The best percentage is given by ROC Area value*

K-nearest Neighbors

It is a simple algorithm which is regarded as one of the top 10 algorithms in data mining (Wu et al. 2008). It classifies new cases based on a similarity measure. It is widely used for classification. The classification starts with linking of the training dataset onto a one-dimensional distance space based on the calculated similarities. Subsequently, the most dominant or mean of the labels of the k nearest neighbors are labeled (Ertuğrul and Tağluk 2017).

Bayesian Network

A Bayesian network (BN) is a visual model that shows the joint distribution of a set of random variables in a form of a directed acyclic graph (DAG) (Fareh 2019; Hwang, Boyle, and Banerjee 2019). Each node in a Bayesian network indicate propositional variables of interest and the links the informational or causal dependencies among the variables by calculating conditional probabilities of each node with its parents in the graph. The topology of a BN and the associated probabilistic relationships between variables are usually learned from data (Lin et al. 2019).

Naïve Bayesian

Naïve Bayesian (NB) algorithm is based on the Bayesian theorem with an assumption regarding the conditional independence of predictors (Diab and El Hindi 2017). The Naïve Bayes classifier intends to detect the class of data by a series of probabilistic values. The probability tests performed for the learning data and the new test data are activated according to the early obtained probability values and it is attempted to define which category of test data is given. In most of the time, real data sets fails to satisfy the condition of independence, despite the performance of the naïve Bayesian classifier is still very reasonable when compared to other classifiers (Wong 2012).

C4.5

C4.5 algorithm was proposed to overcome the limitations of the Iterative Dichotomiser 3 (ID3) algorithm. Mainly, all training patterns are fixed at root. These patterns are distributed based on features selected on an impurity function in recursive routine. Distribution lasts until all training patterns for a certain node is assigned to the similar class (Saeh et al. 2016). C4.5 uses an uncertainty (entropy) measure for a new split creation (Mantas, Abellán, and Castellano 2016).

Fisher Linear Discriminant Analysis (FLDA)

Although the method is simple, it produces good results in complex problems. FLDA is based on the search for a linear combination that best separates the variables between the two classes (targets). The method tests the ratio between within-group and between-group variance (Chen 2018). The ratio is calculated to express separability of the particular variable. The higher ratio value indicates the higher separability.

Logistic Regression

Logistic regression (LR) has been one of the widely used tools to solve classification problems, has continuously received excessive attention of both researchers (Zhang, Xu, and Zhang 2019). LR measures the relationship between the defined dependent variable and independent variables by assessing probabilities using a logistic function (Khairunnahar et al. 2019). It is a linear probabilistic classifier that provides a linkage between an input vector and multiple hyperplanes where each corresponds to an individual class (Asif, Majid, and Anwar 2019).

Sequential Minimal Optimization (SMO)

SMO is an optimization algorithm used to train a support vector machine (SVM) on a data set. SVM is capable of finding a solution to a nonlinear low dimensional classification problem by projecting it into high dimensional space by constructing an optimal separating hyperplane between the positive/negative classes with maximum margin (Hashmi et al. 2015). To define the maximum margin, it is necessary to maximize the width (w) of the margin. Also, "w" and "b" is found by solving the objective function, with using Quadratic Programming (QP). A solution of the QP problems is hard and it takes a long time. SMO can rapidly find a solution for the SVM QP problems without using extra matrix storage and numerical QP optimization steps at all.

Input Variables

The MLP model developed for this study employs three input variables that are AFP, hCG and uE3 to predict target class (as DS and DS free). We have not used the weight and gestational week attributes due to multiple missing values in the records. The training phase adjusts the internal weights to get as close as possible to the known classes values.

Determining the Number of the Hidden Layer

Optimizing the number of hidden layer neurons for establishing Feedforward Neural Networks (FNN's) have been a difficult issue in the research area. On the other hand, the hidden neuron can affect the error on the nodes to which their output is linked (Sheela and Deepa 2013). In 2012, Hunter et al. (Hunter et al. 2012) developed a formula to be determine number of hidden layers in a proper NN architecture. This approach can be easily used in the absence of trial-and-error method and has generalization ability. The implemented formula for neural network is N-1, where N is number of input neurons. In this respect, we have used two hidden layers in our model.

The Applied Multilayer Perceptron Model

The Waikato Environment for Knowledge Analysis (WEKA) software tool (Hall et al. 2009) was used to construct the classification models and to develop the Multilayer Perceptron Model. In the output, there are four sigmoid nodes. As illustrated in Figure 1, node 0 and node 1 are output nodes and node 2 and node 3 are hidden nodes.

Results

This section presents the obtained results of the proposed MLP and the other classification models employed. A comparison is provided between the MLP model and other classifiers that predict the existence of DS. For evaluation purpose, a k-fold cross-validation technique was selected. In cross-validation, the data are randomly partitioned into k subsets or folds (Juez-Gil et al. 2019). In this evaluation approach, predictive model is trained k times; at each of

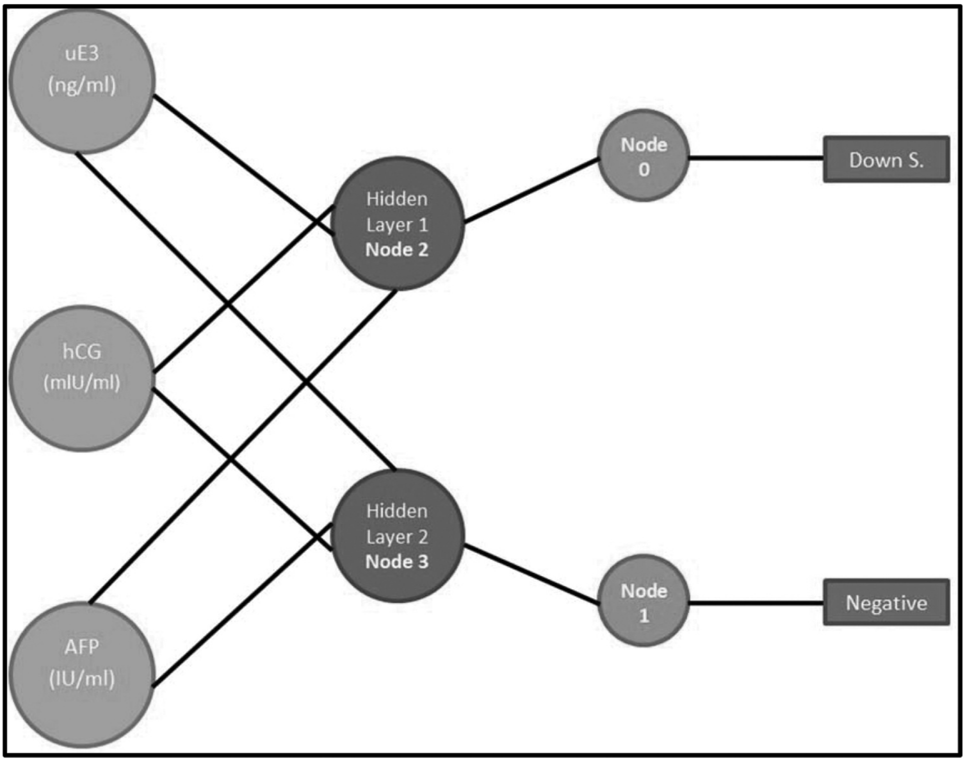

Figure 1. The structure of the proposed MLP model.

these training stages, one fold is used as the test set and the remaining k-1 folds as training set. Each fold is allowed to be "test set" exactly once. Thereby, it is avoided to use test data for training purposes. Thus k-fold cross validation provides a better generalization of the model (Bustillo et al. 2011). The repetitions of tests by cross-validation ensure that the prediction errors are not randomly good or bad and they are an average of multiple runs. In this research, we have implemented a 10-fold cross-validation approach and repeated the classifiers 10 times and therefore; a result provided is an average of 100 runs.

The result sheet of the multilayer is as given at Table 2. Number of correctly classified instances are 112 and correct classification rate is 88.89%. Kappa value (0.7726) shows the good agreement between predicted and observed instances. Mean absolute error value and relative absolute error values are satisfactory. ROC area values indicate that accuracy of the classifier is above 90%. RECALL is TP rate which means sensitivity, it shows how much of actual positives (DS) were predicted as positive. It is also named as detection rate and its value is 92% for this analysis.

Table 3 shows the average value of performance indicators (Percentage of Correctly Classified, Detection Rate, False Positive Rate, Area under ROC and F-Measure) which were calculated for each algorithm after 100 runs. For comparison purposes, we have selected the 0-R as the base classifier and the performance of the other classifiers were compared

Table 2. Results of the MLP model.

Sigmoid Node 0		Inputs	Weights			
		Threshold	3.782			
		Node 2	−8.761			
		Node 3	−11.124			
Sigmoid Node 1		**Inputs**	**Weights**			
		Threshold	−3.782			
		Node 2	8.761			
		Node 3	11.124			
Sigmoid Node 2		**Inputs**	**Weights**			
		Threshold	−17.226			
		AFP	−16.294			
		uE3	−3.310			
		hCG	−7.596			
Sigmoid Node 3		**Inputs**	**Weights**			
		Threshold	16.372			
		AFP	15.535			
		uE3	12.241			
		hCG	0.272			
10*10 Cross-Validation Summary						
Correctly classified instances			112 (88.89%)			
Kappa statistic			0.7726			
Mean absolute error			0.1534			
Relative absolute error			32.0168%			
Total number of instances			126			
	TP Rate	**FP Rate**	**Precision**	**F-Measure**	**ROC area**	**Class**
	0.920	0.132	0.821	0.868	0.910	**DS**
	0.868	0.180	0.943	0.904	0.910	**DS Free**
Weighted Avg.	0.889	0.100	0.895	0.890	0.910	

with 0-R. MLP has the highest correct classification rate (90%) among all classifiers considered.

SMO model has been the worst with 68.17% correct classification rate. Detection rate has been zero for the base classifier (0-R). It means that the base classifier did not predict any instances of DS class correctly. Therefore, all classifiers had a better result when compared to the 0-R. However, the best performance was obtained by Bayesian Network algorithm with 97%. The closest performance was of Naïve Bayes classifier.

False Positive Rate (FPR) indicates that a subject without a DS is misclassified as having DS aneuploidy. This might be costly in real life. The family is given a wrong information and there may be psychological results and may mislead family to an invasive test. Therefore the main objective is to minimize the FPR. As it can be seen from Table 3, MLP is the best performing classifier according to FPR measure. We can see that, the base classifier (0-R) has 0.0 FP rate while it classifies all instances as negative.

While ROC is a probability curve that indicates how much model is capable of distinguishing between classes, the higher area shows a better model that is predicting each of the class values. From the viewpoint of area under the ROC curves, Bayesian Network and Naïve Bayes classifiers has the best performance. Subsequently, MLP and k-NN classifiers have area of 0.93 which are also very close to 1 (perfect value).

In a statistical evaluation of classification, the F-measure is a harmonic mean of precision and recall. The perfect/best value of F-measure is 1 and the worst value is 0. Multilayer Perceptron had the best F-measure performance with 0.88 value.

Using several different performance criteria in a classification study, may not end up with a clear decision about the goodness of a classifier. In this respect, it is possible to weight each of criteria and summing them up to find a total performance score. We have negotiated with some experts from the

Table 3. The performance of the classification algorithms.

Classifier	% Correctly Classified	% Detection Rate (DR)	% False Positive Rate (FPR)	Area Under ROC	F Measure	Total Weighted Scores
Weights	*35%*	*50%*	*−40%*	*35%*	*30%*	*110,00*
ZeroR	*0,60*	*0,00*	*0,00*	*0,50*	*0,00*	*0,00*
k-NN	0,86	0,92	0,17	0,90	0,85	*87.8*
Bayesian network	0.86	0.97*	0.21	0.93*	0.85	*89.75*
Naïve Bayes	0.81	0.96	0.28	0.92	0.81	*83.15*
C4.5	0.82	0.81	0.17	0.85	0.78	*77.05*
FLDA	0.68	0.74	0.36	0.78	0.65	*54.7*
Logistic regression	0.68	0.55	0.24	0.77	0.56	*46.95*
MLP	0.90*	0.94	0.13*	0.92	0.88*	93.4*
SMO	0.68	0.81	0.41	0.7	0.67	54,00

** The best value among the alternatives*

Gaziantep University Hospital and decided to weight each of the performance criteria as follows. The weight assigned to correct classification rate was 35, and it was 50 for the detection rate. The detection rate has been considered as the most important criteria for the study. The ROC area and F-measure criteria were weighted with 35 and 30 respectively. While value of a FPR measure is a negative indicator of performance −40 was used as weighting factor for the corresponding performance.

We have used the equation 1, to calculate the total score of performance the classifiers. This equation takes the base model (0-R) into consideration to show the relative performance of classification algorithm.

$$\sum x_i = \sum (x_i - y_i) * w_i \tag{1}$$

where

x is the classifier,

y is the base classifier,

i is the criteria,

(x_i) is the value of the specified criterion of the particular classifier,

(y_i) is the value of specified criterion of the base classifier,

(w_i) is the weight of the specified criterion.

As a result, the best weighted total score was obtained by the Multilayer Perceptron algorithm with 94.24% detection rate and 13% false positive rate.

Discussion and Conclusion

Triple test has been a simple and affordable screening test to detect DS. However, the high false positive rate has been considerable issue for the test. Many patients are suggested to go for amniocentesis however most of them are found to be free of defect.

In this study, we have evaluated nine classifiers with dissimilar features related to DS occurrence. Using a total of three features obtained with the triple test, we obtained an AUC score of 0.92 using the MultiLayer Perceptron model. The MLP model produces promising results for estimating DS according to our findings. The present work suggests that MLP model can be considered an effective tool for the prediction of DS by using triple test values. Successful detection of the DS can decrease the number of patients routed to amniocentesis and thereby the physiological effects of waiting for a test result can be decreased for some patients. As a future implementation software devoted of risk calculation of pregnant women can be modified to calculate the risk by using MLP models.

ORCID

Alptekin Durmuşoğlu http://orcid.org/0000-0001-9800-5747
Memet Merhad Ay http://orcid.org/0000-0002-6892-7924
Zeynep Didem Unutmaz Durmuşoğlu http://orcid.org/0000-0001-7891-3764

References

Abidine, M., B. Fergani, M. Oussalah, and L. Fergani. 2014. A new classification strategy for human activity recognition using cost sensitive support vector machines for imbalanced data. *Kybernetes* 43 (8):1150–64. doi:10.1108/K-07-2014-0138.

Alameer, Z., M. A. Elaziz, A. A. Ewees, H. Ye, and Z. Jianhua. 2019. Forecasting gold price fluctuations using improved multilayer perceptron neural network and whale optimization algorithm. *Resources Policy* 61:250–60. doi:10.1016/j.resourpol.2019.02.014.

Asif, A., M. Majid, and S. M. Anwar. 2019. Human stress classification using EEG signals in response to music tracks. *Computers in Biology and Medicine* 107:182–96. doi:10.1016/j.compbiomed.2019.02.015.

Autio, L., M. Juhola, and J. Laurikkala. 2007. On the neural network classification of medical data and an endeavour to balance non-uniform data sets with artificial data extension. *Computers in Biology and Medicine* 37 (3):388–97. doi:10.1016/j.compbiomed.2006.05.001.

Avuçlu, E., and F. Başçiftçi. 2018. New approaches to determine age and gender in image processing techniques using multilayer perceptron neural network. *Applied Soft Computing* 70:157–68. doi:10.1016/j.asoc.2018.05.033.

Aye, S. A., and P. S. Heyns. 2015. Acoustic emission-based prognostics of slow rotating bearing using bayesian techniques under dependent and independent samples. *Applied Artificial Intelligence* 29 (6):563–96. doi:10.1080/08839514.2015.1038432.

Bar-Hava, I., M. Yitzhak, H. Krissi, M. Shohat, J. Shalev, B. Czitron, Z. Ben-Rafael, and R. Orvieto. 2001. Pregnancy: Triple-test screening in in vitro fertilization pregnancies. *Journal of Assisted Reproduction and Genetics* 18 (4):228–31. doi:10.1023/A:1009455912670.

Benn, P. 2016. Posttest risk calculation following positive noninvasive prenatal screening using cell-free DNA in maternal plasma. *American Journal of Obstetrics and Gynecology* 214 (6):676.e1-676.e7. doi:10.1016/j.ajog.2016.01.003.

Bienvenido-Huertas, D., A. Pérez-Fargallo, R. Alvarado-Amador, and C. Rubio-Bellido. 2019a. Influence of climate on the creation of multilayer perceptrons to analyse the risk of fuel poverty. *Energy and Buildings* 198:38–60. doi:10.1016/j.enbuild.2019.05.063.

Bienvenido-Huertas, D., C. Rubio-Bellido, J. L. Pérez-Ordóñez, and J. Moyano. 2019b. Optimizing the evaluation of thermal transmittance with the thermometric method using multilayer perceptrons. *Energy and Buildings* 198:395–411. doi:10.1016/j.enbuild.2019.06.040.

Brasil, L. M., F. M. de Azevedo, and J. M. Barreto. 2001. Hybrid expert system for decision supporting in the medical area: Complexity and cognitive computing. *International Journal of Medical Informatics* 63 (1):19–30. doi:10.1016/S1386-5056(01)00168-X.

Bustillo, A., E. Ukar, J. J. Rodriguez, and A. Lamikiz. 2011. Modelling of process parameters in laser polishing of steel components using ensembles of regression trees. *International Journal of Computer Integrated Manufacturing* 24 (8):735–47. doi:10.1080/0951192X.2011.574155.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16:321–57. doi:10.1613/jair.953.

Chen, B.-W. 2018. Incomplete data classification-fisher discriminant ratios versus welch discriminant ratios. *Future Generation Computer Systems*. doi:10.1016/j.future.2018.05.003.

Crandall, B. F., T. B. Lebherz, P. C. Schroth, and M. Matsumoto. 1983. Alpha-fetoprotein concentrations in maternal serum: Relation to race and body weight. *Clinical Chemistry* 29 (3):531–33. doi:10.1093/clinchem/29.3.531.

Daqi, G., L. Chunxia, and Y. Yunfan. 2007. Task decomposition and modular single-hidden-layer perceptron classifiers for multi-class learning problems. *Pattern Recognition* 40 (8):2226–36. doi:10.1016/j.patcog.2007.01.002.

Diab, D. M., and K. M. El Hindi. 2017. Using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification. *Applied Soft Computing* 54:183–99. doi:10.1016/j.asoc.2016.12.043.

Driggers, R. W., and D. C. Seibert. 2008. Prenatal screening: New guidelines, new challenges. *The Journal for Nurse Practitioners* 4 (5):351–56. doi:10.1016/j.nurpra.2008.03.003.

Engels, M. A. J., S. L. Bhola, J. W. R. Twisk, M. A. Blankenstein, and J. M. G. van Vugt. 2014. Evaluation of the introduction of the national Down syndrome screening program in the Netherlands: Age-related uptake of prenatal screening and invasive diagnostic testing. *European Journal of Obstetrics & Gynecology and Reproductive Biology* 174 (SupplementC):59–63. doi:10.1016/j.ejogrb.2013.12.009.

Ertuğrul, Ö. F., and M. E. Tağluk. 2017. A novel version of k nearest neighbor: Dependent nearest neighbor. *Applied Soft Computing* 55:480–90. doi:10.1016/j.asoc.2017.02.020.

Fareh, M. 2019. Modeling incomplete knowledge of semantic web using Bayesian networks. *Applied Artificial Intelligence* 33 (11):1022–34. doi:10.1080/08839514.2019.1661578.

Founds, S. 2014. Innovations in prenatal genetic testing beyond the fetal karyotype. *Nursing Outlook* 62 (3):212–18. doi:10.1016/j.outlook.2013.12.010.

Güler, E. Ç., B. Sankur, Y. P. Kahya, and S. Raudys. 1998. Visual classification of medical data using MLP mapping. *Computers in Biology and Medicine* 28 (3):275–87. doi:10.1016/S0010-4825(98)00010-9.

Haddow, J. E., E. M. Kloza, G. J. Knight, and D. E. Smith. 1981. Relation between maternal weight and serum alpha-fetoprotein concentration during the second trimester. *Clinical Chemistry* 27 (1):133–34. doi:10.1093/clinchem/27.1.133.

Haddow, J. E., G. E. Palomaki, G. J. Knight, J. Williams, A. Pulkkinen, J. A. Canick, D. N. Saller Jr, and G. B. Bowers. 1992. Prenatal screening for down's syndrome with use of maternal serum markers. *New England Journal of Medicine* 327 (9):588–93. doi:10.1056/NEJM199208273270902.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations Newsletter* 11 (1):10–18. doi:10.1145/1656274.1656278.

Han, H., W.-Y. Wang, and B.-H. Mao. 2005. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Advances in intelligent computing. Lecture notes in computer science*, ed. D.-S. Huang, X.-P. Zhang, and G.-B. Huang, 878–87. Berlin Heidelberg: Springer.

Harris, S., D. Reed, and N. L. Vora. n.d. Screening for fetal chromosomal and subchromosomal disorders. *Seminars in Fetal & Neonatal Medicine*. doi:10.1016/j.siny.2017.10.006.

Hashmi, S., S. M. Halawani, O. M. Barukab, and A. Ahmad. 2015. Model trees and sequential minimal optimization based support vector machine models for estimating minimum surface roughness value. *Applied Mathematical Modelling* 39 (3):1119–36. doi:10.1016/j.apm.2014.07.026.

He, L., R. A. Levine, A. J. Bohonak, J. Fan, and J. Stronach. 2018. Predictive analytics machinery for STEM student success studies. *Applied Artificial Intelligence* 32 (4):361–87. doi:10.1080/08839514.2018.1483121.

Hunter, D., H. Yu III, M. S. P. Kolbusz, and B. M. Wilamowski. 2012. Selection of proper neural network sizes and architectures-A comparative study. *IEEE Transactions on Industrial Informatics* 8 (2):228–40. doi:10.1109/TII.2012.2187914.

Hwang, S., L. N. Boyle, and A. G. Banerjee. 2019. Identifying characteristics that impact motor carrier safety using Bayesian networks. *Accident Analysis & Prevention* 128:40–45. doi:10.1016/j.aap.2019.03.004.

Jeatrakul, P., K. W. Wong, and C. C. Fung. 2010. Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. In *Neural information processing. Models and applications. Lecture notes in computer science*, ed. K. W. Wong, B. S. U. Mendis, and A. Bouzerdoum, 152–59. Berlin Heidelberg: Springer.

Juez-Gil, M., I. N. Erdakov, A. Bustillo, and D. Y. Pimenov. 2019. A regression-tree multilayer-perceptron hybrid strategy for the prediction of ore crushing-plate lifetimes. *Journal of Advanced Research* 18:173–84. doi:10.1016/j.jare.2019.03.008.

Kaur, G., J. Srivastav, S. Sharma, A. Huria, P. Goel, and B. S. Chavan. 2013. Maternal serum median levels of alpha-foetoprotein, human chorionic gonadotropin & unconjugated estriol in second trimester in pregnant women from north-west India. *The Indian Journal of Medical Research* 138 (1):83–88.

Khairunnahar, L., M. A. Hasib, R. H. B. Rezanur, M. R. Islam, and M. K. Hosain. 2019. Classification of malignant and benign tissue with logistic regression. *Informatics in Medicine Unlocked* 16:100189. doi:10.1016/j.imu.2019.100189.

Lin, S., Q. Zhang, F. Chen, L. Luo, L. Chen, and W. Zhang. 2019. Smooth Bayesian network model for the prediction of future high-cost patients with COPD. *International Journal of Medical Informatics* 126:147–55. doi:10.1016/j.ijmedinf.2019.03.017.

Lin, W.-C., S.-W. Ke, and C.-F. Tsai. 2017. Top 10 data mining techniques in business applications: A brief survey. *Kybernetes* 46 (7):1158–70. doi:10.1108/K-10-2016-0302.

Maciejewski, T., and J. Stefanowski, 2011. Local neighbourhood extension of SMOTE for mining imbalanced data. *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, Paris, 104–11, April. doi: 10.1109/CIDM.2011.5949434.

Mantas, C. J., J. Abellán, and J. G. Castellano. 2016. Analysis of Credal-C4.5 for classification in noisy domains. *Expert Systems with Applications* 61:314–26. doi:10.1016/j.eswa.2016.05.035.

Mukherjee, S. 2018. Malignant mesothelioma disease diagnosis using data mining techniques. *Applied Artificial Intelligence* 32 (3):293–308. doi:10.1080/08839514.2018.1451216.

Muller, F., D. Thibaud, F. Poloce, M.-C. Gelineau, M. Bernard, C. Brochet, C. Millet, J.-Y. Réal, and M. Dommergues. 2002. Risk of amniocentesis in women screened positive for Down syndrome with second trimester maternal serum markers. *Prenatal Diagnosis* 22 (11):1036–39. doi:10.1002/pd.449.

Mulongo, J., M. Atemkeng, T. Ansah-Narh, R. Rockefeller, G. M. Nguegnang, and M. A. Garuti. 2019. Anomaly detection in power generation plants using machine learning and neural networks. *Applied Artificial Intelligence* 1–16. doi:10.1080/08839514.2019.1691839.

Neocleous, A. C., K. H. Nicolaides, and C. N. Schizas. 2016. First trimester noninvasive prenatal diagnosis: A computational intelligence approach. *IEEE Journal of Biomedical and Health Informatics* 20 (5):1427–38. doi:10.1109/JBHI.2015.2462744.

Norton, M. E., and B. D. Rink. 2016. Changing indications for invasive testing in an era of improved screening. *Seminars in Perinatology* 40 (1):56–66. doi:10.1053/j.semperi.2015.11.008.

Ökem, Z. G., G. Örgül, B. T. Kasnakoglu, M. Çakar, and M. S. Beksaç. 2017. Economic analysis of prenatal screening strategies for Down syndrome in singleton pregnancies in Turkey. *European Journal of Obstetrics & Gynecology and Reproductive Biology* 219 (SupplementC):40–44. doi:10.1016/j.ejogrb.2017.09.025.

Orhan, U., M. Hekim, and M. Ozer. 2011. EEG signals classification using the K-means clustering and a multilayer perceptron neural network model. *Expert Systems with Applications* 38 (10):13475–81. doi:10.1016/j.eswa.2011.04.149.

Paiva, J. S., J. Cardoso, and T. Pereira. 2018. Supervised learning methods for pathological arterial pulse wave differentiation: A SVM and neural networks approach. *International Journal of Medical Informatics* 109:30–38. doi:10.1016/j.ijmedinf.2017.10.011.

Phillips, O. P., S. Elias, L. P. Shulman, R. N. Andersen, C. D. Morgan, and J. L. Simpson. 1992. Maternal serum screening for fetal down syndrome in women less than 35 years of age using alpha-fetoprotein, hCG, and unconjugated estriol: A prospective 2-year study. *Obstetrics and Gynecology* 80 (3):353.

Pota, M., E. Scalco, G. Sanguineti, G. M. Cattaneo, M. Esposito, and G. Rizzo. 2015. Early classification of parotid glands shrinkage in radiotherapy patients: A comparative study. *Biosystems Engineering* 138:77–89. doi:10.1016/j.biosystemseng.2015.06.007.

Rani, A. S., and S. Jyothi, 2016. Performance analysis of classification algorithms under different datasets. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, 1584–89, March 2.

Reynolds, T. M., G. Vranken, and J. V. Nueten. 2006. Weight correction of MoM values: Which method? *Journal of Clinical Pathology* 59 (7):753–58. doi:10.1136/jcp.2005.034280.

Saeh, I. S., M. W. Mustafa, Y. S. Mohammed, and M. Almaktar. 2016. Static security classification and evaluation classifier design in electric power grid with presence of PV power plants using C4.5. *Renewable and Sustainable Energy Reviews* 56:283–90. doi:10.1016/j.rser.2015.11.054.

Setsirichok, D., T. Piroonratana, W. Wongseree, T. Usavanarong, N. Paulkhaolarn, C. Kanjanakorn, M. Sirikong, C. Limwongse, and N. Chaiyaratana. 2012. Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening. *Biomedical Signal Processing and Control* 7 (2):202–12. doi:10.1016/j.bspc.2011.03.007.

Shaw, S. W. S., C.-P. Chen, and P.-J. Cheng. 2013. From Down syndrome screening to noninvasive prenatal testing: 20 years' experience in Taiwan. *Taiwanese Journal of Obstetrics & Gynecology* 52 (4):470–74. doi:10.1016/j.tjog.2013.10.003.

Sheela, K. G., and S. N. Deepa. 2013. Review on methods to fix number of hidden neurons in neural networks. *Mathematical Problems in Engineering* 2013:1–11. doi:10.1155/2013/425740.

Shurtz, I., A. Brzezinski, and A. Frumkin. 2016. The impact of financing of screening tests on utilization and outcomes: The case of amniocentesis. *Journal of Health Economics* 48 (SupplementC):61–73. doi:10.1016/j.jhealeco.2016.02.001.

Skrzypek, H., and L. Hui. 2017. Noninvasive prenatal testing for fetal aneuploidy and single gene disorders. *Best Practice & Research. Clinical Obstetrics & Gynaecology* 42 (SupplementC):26–38. doi:10.1016/j.bpobgyn.2017.02.007.

Tamminga, S., M. van Maarle, L. Henneman, C. B. M. Oudejans, M. C. Cornel, and E. A. Sistermans. 2016. Maternal plasma DNA and RNA sequencing for prenatal testing. In *Advances in clinical chemistry* 74, 63–102. Elsevier. doi:10.1016/bs.acc.2015.12.004.

Temming, L. A., and G. A. Macones. 2016. What is prenatal screening and why to do it? *Seminars in Perinatology* 40 (1):3–11. doi:10.1053/j.semperi.2015.11.002.

Wald, N. J., J. W. Densem, L. George, S. Muttukrishna, and P. G. Knight. 1996. Prenatal screening for Down's syndrome using inhibin-a as a serum marker. *Prenatal Diagnosis* 16 (2):143–53. doi:10.1002/(SICI)1097-0223(199602)16:2<143::AID-PD825>3.0.CO;2-F.

Wald, N. J., J. W. Densem, D. Smith, and G. G. Klee. 1994. Four-marker serum screening for Down's syndrome. *Prenatal Diagnosis* 14 (8):707–16. doi:10.1002/pd.1970140810.

Wald, N. J., A. Kennard, A. Hackshaw, and A. McGuire. 1997. Antenatal screening for Down's syndrome. *Journal of Medical Screening* 4 (4):181–246. doi:10.1177/096914139700400402.

West, D., and V. West. 2000. Improving diagnostic accuracy using a hierarchical neural network to model decision subtasks. *International Journal of Medical Informatics* 57 (1):41–55. doi:10.1016/S1386-5056(99)00059-3.

Witters, I., E. Legius, K. Devriendt, P. Moerman, D. V. Schoubroeck, A. V. Assche, and J.-P. Fryns. 2001. Pregnancy outcome and long term prognosis in 868 children born after second trimester amniocentesis for maternal serum positive triple test screening and normal prenatal karyotype. *Journal of Medical Genetics* 38 (5):336–38. doi:10.1136/jmg.38.5.336.

Wong, -T.-T. 2012. A hybrid discretization method for naïve Bayesian classifiers. *Pattern Recognition* 45 (6):2321–25. doi:10.1016/j.patcog.2011.12.014.

Wu, X., V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, et al. 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems* 14 (1):1–37. doi:10.1007/s10115-007-0114-2.

Yagel, S., E. Y. Anteby, D. Hochner-Celnikier, I. Ariel, T. Chaap, and Z. B. Neriah. 1998. The role of midtrimester targeted fetal organ screening combined with the "triple test" and maternal age in the diagnosis of trisomy 21: A retrospective study. *American Journal of Obstetrics and Gynecology* 178 (1, Part 1):40–44. doi:10.1016/S0002-9378(98)70623-4.

Yared, E., M. J. Dinsmoor, L. K. Endres, M. J. Vanden Berg, C. J. Maier Hoell, B. Lapin, and B. A. Plunkett. 2016. Obesity increases the risk of failure of noninvasive prenatal screening regardless of gestational age. *American Journal of Obstetrics and Gynecology* 215 (3):370.e1-370.e6. doi:10.1016/j.ajog.2016.03.018.

Zhang, C.-X., S. Xu, and J.-S. Zhang. 2019. A novel variational Bayesian method for variable selection in logistic regression models. *Computational Statistics & Data Analysis* 133:1–19. doi:10.1016/j.csda.2018.08.025.

Zhang, J., G. Lambert-Messerlian, G. E. Palomaki, and J. A. Canick. 2011. Impact of smoking on maternal serum markers and prenatal screening in the first and second trimesters. *Prenatal Diagnosis* 31 (6):583–88. doi:10.1002/pd.2755.