International Journal of Emerging Engineering Research and Technology
Volume 2, Issue 8, August 2014, PP 155-163
ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

Machine Learning Techniques for Diabetes Identification

K. Srinivasa Reddy[1], N.Rajasekhar[2]

[1]Associate Professor, Department of Computer Science and Engineering BVRIT HYDERABAD College of Engineering for Women Hyderabad, India

[2]Assistant Professor, VNR VJIET, Hyderabad, India

Abstract: *Diabetes Mellitus, which in simple term known as Diabetes. It is due to metabolic disorders that is due to reduction of blood glucose level in the body. There are stages in diabetes based on the severity. So, it is an emergency task to identify diabetes at a very early stage to decrease the severity of the problem. By considering all the complications many research studies are done so as to solve the problem in an effective way. As most of the methods include Pima Indians Diabetes data set we have also implemented our paper by using the same but with different algorithm. At present we have diabetes based on Support Vector Machine (SVM), decision trees, PCA, etc., this algorithm chosen by us are related to subpart of Deep Learning named as Machine Learning. We have used data pre-processing techniques and classified the data which is well trained in a network. Through this paper we expect to get the maximum accuracy for predicting the diabetes detection.*

Keywords: *Diabetes, Machine Leaning, Data pre- processing.*

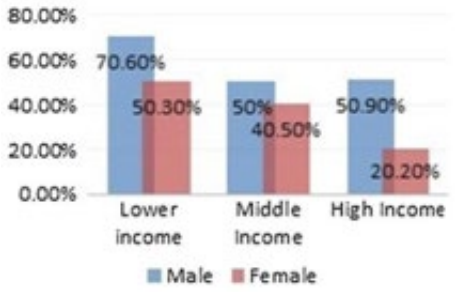
1. INTRODUCTION

Machine Learning at its most elementary is the rehearsal of using algorithms to analyze data, acquire from it, and then style a determination or prediction around something in the world. So instead of hand-coding software customs with a definite set of guidelines to achieve a particular task, the machine is "trained" by using very large totals of data and algorithms which give it the capacity to study how to complete the task.

Some techniques like Deep learning and Artificial Neural Networks (ANNs) are of highly capable tools of AI for solving highly complex problems, and these can also be developed and be leveraged in coming days. When we want an intelligent system similar to a robot so that it can perform as per our instructions and May also be able to hear the decision from any dialogue kind of clinical expert system, etc. something that is essential is the processing of a Natural Language.

Diabetes is majorly a combination of the metabolic disorders or a chronic disease. In such scenario the blood glucose present in the body of a person leads to suffer from a much extended equal of it that can either if production of the in sulin is insufficient, or may be the body's cells are unable to respond properly to the produced insulin. The persistent hyperglycemia of the diabetes is precisely connected to brokenness, long-haul harm and also due to miscarriage of various organs, majorly the eyes, nerves, kidneys, heart, and also veins.

The goal to be achieved through this research is to utilize the significant features, and then to model a prediction algorithm by making use of Machine learning. Later to get an optimal classifier which can give very closest result when comparing to that of the clinical outcomes. Here, the proposed methodology aims to focus majorly on the selection of the attributes which can ail in the early identification of Diabetes Miletus by using Predictive analysis. Hence, the result presents an algorithm related to decision tree and also Random forest is proven to be shown the specificity of about 98.20% and 98.00% which is treated as highest, respectively and holds the best for diabetic data analysis. The Naïve Bayesian has stated its outcome with its best accuracy of about 82.30%. Finally, the it generalizes selection of the optimal features from the dataset so as to improve classification accuracy.



2. LITERATURE SURVEY

In this paper we strived to implement a better system that could give the best results out of all the available systems. During this process we have come across many types of existing systems and have gone through various algorithm, since we are in growing technology things can be made much simpler by making use of advanced technologies. For example: in the past we have learnt only C, Java as the better languages to build software programs but now the world has changed completely with automation.

Thereby we are in search of such environments where we can create a robust platform for our problem statement of diabetes identification. We do have many implementations currently but those could not reach our imaginations. In this context we could explain the things that we have learnt by referring to various articles proposed by different researchers.

[11] As diabetes being a chronic condition in the body it is essential to create and then a model is to be predicted so that we can easily identify the condition of the body related to diabetes, keeping this in mind Pima Indians set is being proposed to conduct further implementations.

Diabetes is mainly of two types. Namely type1 and type 2. This is precisely based on the complexity of the problem. It depends on glucose levels and metabolism of the person [4].

Type 1 diabetes can be defined as the scenario in which the cells of pancreas that produce insulin are damaged. It is an auto immune condition. This may be due to genes sometimes.

Type 2 diabetes insulin is produced by pancreas but the problem is, it is not used by the body appropriately. [2] This type of diabetes can be seen in teens and adults. Of course, it is considered to be common and can be screened and diagnosed.

There are various tests available in present day medical field. According to a survey, almost 90% people are suffering with this type. Based on income of a country percentage of diabetic patients is shown in below figure.

According to a survey by Neha sharma, [10] the early identification and also screening is expected to play a key role for ineffectual prevention of the diabetes. The learning of such procedure starts with a survey or the data, as an example, the exact occurrences, or the commands, in an order so as to gaze for such type of patterns in data and then to make decisions satisfactorily.

Through an International journal of computer science and wireless security they said that “using tongue images innovative application can be developed and it is easy to use for diabetic deduction.”

In some journals, tongue images are used to detect diabetes. For this, MATLAB is used for diabetic identification. In this scenario both the diabetic and non-diabetic patients are diagnosed and then outputs are predicted.

[9] Patients suffering with type 2 Diabetes have larger area covered with yellow fur, thick fur and also bluish tongue than those of group under control. Also, an expressively higher portion of the patients having diabetics for long-term and those having yellow fur for the short-term was being noticed.

According to certain studies, [8] it is said that even eyes can lead to diabetes. This can be explained like if the retina’s blood vessels are damaged it causes diabetes which is usually called as Diabetic Retinopathy. In such scenario it can also lead to blindness.

J.Tuomileh to, [1] in this they considered both male and female samples, they took about 600 such samples and proposed a relationship for data mining for an efficient classification. They used the

methods of data mining so that the clinical data of diabetes can be classified and then prediction required that is whether the patient is suffering with diabetes or is not suffering is performed. For this they have presented a system which gives the training data for the data feature analysis to be performed and next the contrast of the classification algorithm, followed by selecting classifier and then an enhanced algorithm for classification is to be applied and have brought out an evaluation which is being compared to the training data. In this, they used an algorithm named C4.5 and the classification rate obtained through it is about91%.

Kanika, [12] they concluded that their proposed method gave more accuracy as it is based on blood vessels and its area and perimeter resulted in motivational outputs.

K.C. Tan, [6] in this they talked about the short filtering method that can remove the undesirable features that are present before the classification procedure starts, as a wrapper method is related for selecting certain optimal features by the classification algorithm. This Wrapper method is expected to give a higher accuracy of classification. But the drawback with this approach of wrapper is that it requires a longer runtime and this is because of ML algorithm that is chosen has to be worked iteratively to search for the subsets of attributes.

[5] Gentic Algorithms and Back Propagation Networks are proven to be given more accuracy when hybrid than implementing alone.GA-BPN works great with high performance and accuracy.

Swapna, G [3] in this they mentioned that diabetes can be identified by giving heart rate variability as input. Through this they estimated 97% accuracy.

The Deep Learning Techniques are broadly used to solve the problems more efficiently. In this paper they would be the techniques of Deep Learning, machine learning algorithms are implemented to get the best results of the available solutions. It involves robust environment and efficient programming language.

3. PROPOSED METHODOLOGY

To identify diabetes, we have used python language and its modules to implement machine learning algorithms. To achieve our requirement, this paper is built using jupyter notebook environment in collaboration with Anaconda.

Typically, people after age 20 are supposed to be affected by diabetes. [7] According to statistics of WHO, worldwide diabetes has become common among the adults after 18 years of age has then increased about 8.5% in the year 2014. Its existence rate has been increasing irrespective of the income levels of a country. It also becomes as a cause to the certain illnesses likewise blindness, cholesterol, kidney failure, heart diseases. It is the fact that the deaths because of diabetes and high blood glucose are rising. Hence, prediction of being suffered by diabetes must be done at an initial stage so that it would help patients to balance their sugar level.

It is well known that for the predictive analysis data mining approach is very well suited. As our papert deals with such predictive analysis problem that approach is considered for diabetes identification. At the end we have measured and also improved the performance by making use of feature selection and by selection of training set.

3.1. Exploratory Data Analysis

The sample dataset must be chosen and then it is separated as two datasets namely a training and a test dataset. Here, the major target is to get a feature subset which can produce classification with higher accuracy. The selection of features task is a significant problem in the knowledge discovery. So, after the selection of all the required features, we apply a classification algorithm so as to make a model for classification. Finally, model is practically used on our test set so that we can predict risk of diabetes. Proposed work is shown in the Fig.

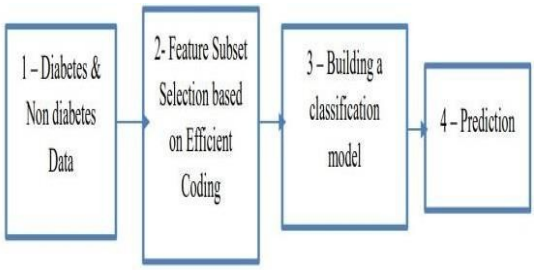


Fig. *Proposed Architecture*

3.2. Sample and Sampling Design

The sample for the experimental setup chosen is a Pima Indian Diabetes dataset. It is an open source in UCI repository. This set comprises of certain records including diabetic patients and non-diabetic. This possesses eight attributes and also class attributes.

768 instances can be seen in our data set. In dataset all the patients are of female category who are Pima Indians and their age is above 21 years. The features or attributes of the dataset are shown in below Table below.

Sample and Sampling Design for Diabetes Mellitus Prediction

Attribute_id	Attribute_name	Attribute_description
A1	Pregnant Times	Number of times pregnant
A2	Plasma Glucose	Plasma glucose concentration
A3	Diastolic BP	Diastolic Blood Pressure (mm Hg)
A4	Skin Thickness	Triceps Skin Fold Thickness (mm)
A5	Serum Insulin	2- Hour Serum Insulin (U/ml)
A6	BMI	Body Mass Index
A7	Pedigree	Diabetes Pedigree Function
A8	Age	Age in numbers
A9	Class Variable	Zero or One

The dataset will be classified by making use of recursive partitioning algorithm by which a model has to be built. Here, 70% of records are training set and the rest is test set.

Here, the data cannot be directly accessed so this is converted to CSV file. This will be learnt later after completion of data modelling.

In this paper, we will follow a set of steps so as to achieve the goal. The data is to be pre-processed, then classified and finally prediction is done during implementation. The major aspect is to model data effectively so that we can get efficient outcome.

3.3. Data Modelling

For Diabetic identification it includes a series of modules as explained in below context. It involves data pre-processing, classification and prediction.

3.4. Data Pre-processing

Data pre-processing not only involves the data cleaning but also handling of missing values, dimensionality reduction, handling of inconsistent data, selection of features, etc. Since it is already mentioned, the dataset of experiment chosen for this paper contains various attributes, such as pregnant Times, plasma Glucose, Age, Diastolic BP, Skin Thickness, Serum Insulin, BMI, Class Variable.

For an effective decision-making information, the gain feature selection method has been used. With a classification goal it will select the features that are highly correlated**.**

3.5. Classification

Classification module classifies samples into different groups. It is a supervised learning technique. Here, for getting accuracy, specificity, sensitivity and also precision of the Neural Networks. So, we need to give training dataset and then test the dataset as input then it gives the accuracy in percentage.

3.6. Prediction

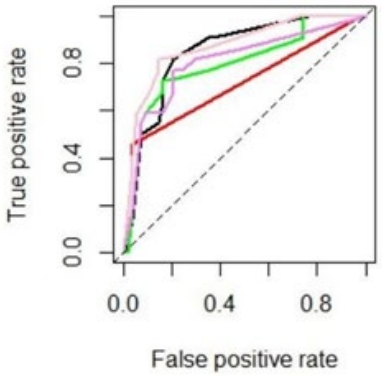
Prediction for this model is POSITIVIE or NEGATIVE. Neural Networks Classification algorithm is used for prediction.

In this paper, the measures of performance algorithm are done by using three types of performance metrics. They are the accuracy, the sensitivity and the specificity. We can calculate this by using confusion matrix In this matrix it represents all count of the True Positives (TP), the False Negatives (FN), the False Positives (FP), the True Negatives (TN).

Confusion matrix

Actual vs Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

The below figure depicts the relation of true positive rate and false positive rate that can be achieved through data classification.



The formulae to calculate performance metrics are shown in below equations. The accuracy is considered as a statistical measure which calculates that how better a classification test of binary can be able to recognize or may be ignores the condition correctly. The sensitivity is described as a recall otherwise as a true positive rate. Proportion of all positives result that can be correctly recognized as the positives are also measured. Then we have another measure for performance named specificity. This is also called as a true negative rate in this it measures proportion of all the negatives which are appropriately recognized as the negatives.

$$Accuracy = (TP+TN)/TP+FP+TN+FN$$

$$Sensitivity=TP/(TP+FN)$$

$$Specificity=TN/(TN+FP)$$

The ‘r’ part of the algorithm creates certain rules during formation of the binary trees as shown in below figure.

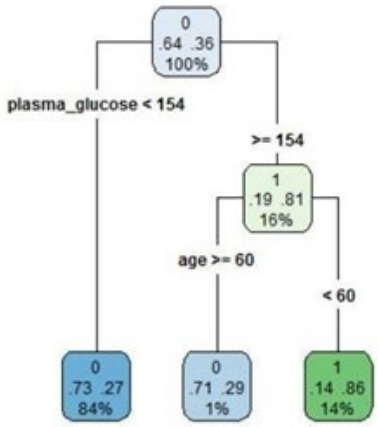


Fig. *Sample classification tree*

From the Fig.3.9, we can draw a point that is most important to be noted. This tree gives us the classification that is firstly depended on the plasma-glucose and then based on age the probability of occurrence of diabetes is further predicted.

However, to proceed further we have make sure that our input file has to be converted as discussed earlier. This can be achieved by the following.

3.7. Input as CSV File

Reading the data from CSV (comma separated values) is the basic necessity in the field of Data Science. Often, the data we get is obtained from the various sources that can get exported to the CSV format. So that, these can also be used by various other systems. The Pandas library has various features by using which we can not only read the CSV file in full but also in parts for a certain selected group of rows and columns.

The CSV file can be described as a text file and in this all the values available in columns are being separated by using a comma.

In our environment, that is in jupyter notebook the following line of code is used to extract the file and access its data throughout the work.

Import pandas as pd data=pd.read_csv(input data)

preg	plas	pres	skin	insu	mass	pedi	age	class
6	148	72	35	0	33.6	0.627	50	tested_positive
1	85	66	29	0	26.6	0.351	31	tested_negative
8	183	64	0	0	23.3	0.672	32	tested_positive
1	89	66	23	94	28.1	0.167	21	tested_negative
0	137	40	35	168	43.1	2.288	33	tested_positive
5	116	74	0	0	25.6	0.201	30	tested_negative
3	78	50	32	88	31	0.248	26	tested_positive
10	115	0	0	0	35.3	0.134	29	tested_negative
2	197	70	45	543	30.5	0.158	53	tested_positive
8	125	96	0	0	0	0.232	54	tested_positive
4	110	92	0	0	37.6	0.191	30	tested_negative
10	168	74	0	0	38	0.537	34	tested_positive
10	139	80	0	0	27.1	1.441	57	tested_negative
1	189	60	23	846	30.1	0.398	59	tested_positive
5	166	72	19	175	25.8	0.587	51	tested_positive
7	100	0	0	0	30	0.484	32	tested_positive
0	118	84	47	230	45.8	0.551	31	tested_positive
7	107	74	0	0	29.6	0.254	31	tested_positive
1	103	30	38	83	43.3	0.183	33	tested_negative
1	115	70	30	96	34.6	0.529	32	tested_positive
3	126	88	41	235	39.3	0.704	27	tested_negative
8	99	84	0	0	35.4	0.388	50	tested_negative

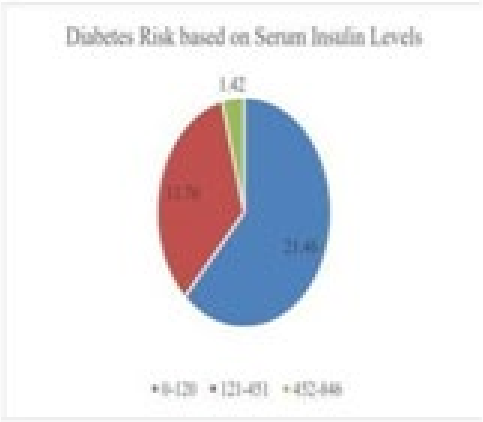
Diabetes input dataset

For our understanding we have class labels as tested_positive and tested_negative but these cannot be understood by the system. So to solve this problem we will be converting tested_positive and tested_negative as 0 and 1 respectively.

4. RESULTS

The preliminary analysis states the resulting visions of the data. Here, the given dataset comprises of patients belonging to female category and their ages are between 21-81.

Plasma glucose levels and serum insulin levels can be measured as a factor to diabetes risk is shown in the following figures.



The chosen dataset is to be divided into two categories. That is a training set and a test set after which the results are to be evaluated. Feature subset selection is being applied so as to improve the accuracy of the result. This subset selection mainly focuses on the identification of an attribute subset that can improve the accuracy of classification. Attributes which have resulted in highest accuracy are represented in the Table.

Performance Measures by Highest Value

Attribute	Accuracy	Sensitivity	Specificity
A1,A2,A3,A6	79.08%	90.56%	56.26%
A1,A2,A3,A4,A6	79.08%	90.56%	56.26%
A1,A2,A3,A4,A5,A6	79.08%	90.56%	56.26%
A2,A4,A5,A6,A8	79.08%	87.42%	62.5%
A1,A2,A3,A7	78.66%	88.05%	59.99%
A1,A2,A3,A4,A5,A6,A8	78.66%	87.42%	61.25%
A1,A2,A3,A4,A5,A8	78.24%	84.28%	66.25%

The feature subset selection is being soothed by eliminating attributes in sequence from dataset. The measures of performance are tabulated in Table.

Table. *Performance Measures by attribute selection*

Removed attribute	Accuracy	Sensitivity	Specificity
Full attribute set	77.82%	86.16	61.25
A8	74.48%	76.73	70.00
A7,A8	79.08%	90.56	56.26
A6,A7,A8	77.41%	88.05%	56.25%
A5,A6,A7,A8	76.14%	87.42%	53.73%
A4,A5,A6,A7,A8	74.06%	88.68%	44.99%
A3,A4,A5,A6,A7,A8	76.15%	94.96%	38.75%
A2,A3,A4,A5,A6,A7,A8	70.29%	86.79%	37.50%

The results have shown the highest accuracy at a point when attributes named pedigree and the other one named age are both removed from the attribute set. The choosing of the training as well as the test data sets do throw an impact on performance in addition to the selection of attributes of the algorithm. Here, the algorithm gives higher accuracy for increasement on data set is imposed above 85 percent. The accuracy of algorithm when the relation between the training set and the test data set is being varied. The renaming of attribute sets is as shown below.

$$\text{Set 1} : \{ A1, A2, A3, A6 \}, 2 : \{ A1, A2, A3, A4, A6 \}, 3: \{ A1, A2, A3, A4,$$

$$A5, A6 \}, 4: \{ A2, A4, A5, A6, A8 \}, 5: \{ A1, A2, A3, A7 \}, 6 : \{ A1, A2, A3,$$

$$A4, A5, A6, A8 \}, 7: \{ A1, A2, A3, A5, A8 \}$$

The accuracy of model rises for many attribute sets as long as the size of training set is enlarged. The dissimilarity of the sensitivity and specificity is represented in below figures.

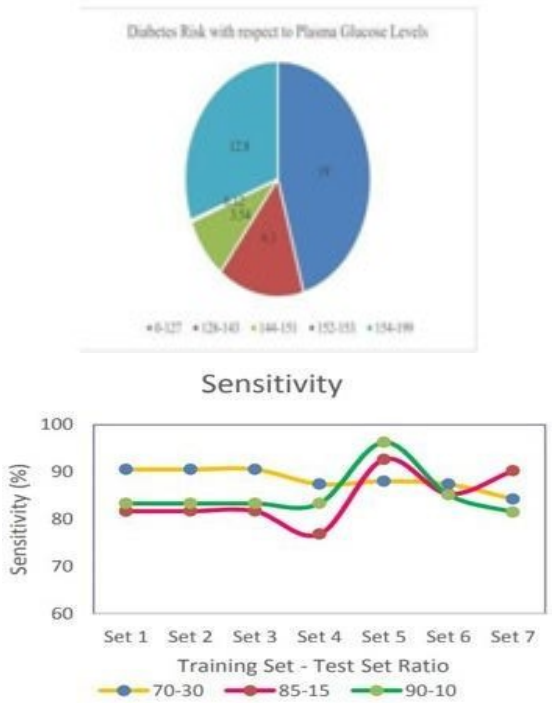
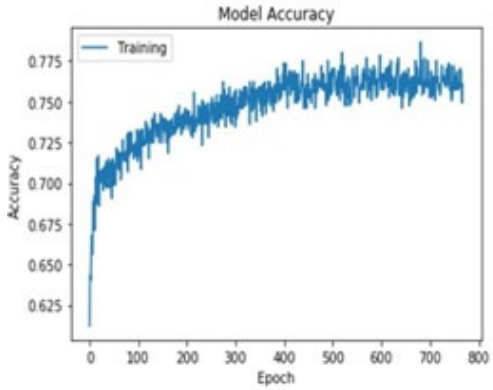


Fig. *Sensitivity for varying training-test set ratio of classification model*

Fig. *Specificity for varying training-test set ratio of classification model*

The accuracy of the proposed work is graphically represented below.



Graphical representation of accuracy of the model

5. CONCLUSION

This paper represents the way of creating a classification model by making use of a recursive partitioning algorithm which then implements this model for chosen dataset so as to classify patient's data whether they are diabetic or not. It is helpful in prediction of risk of being affected by diabetes on another dataset. Finally, the performance our model is verified by making use of the performance measures like accuracy, specificity and sensitivity. By changing the size of our training dataset, performance of the algorithm is being enhanced by using feature subset selection.

REFERENCES

[1] J.Tuomilehto, "Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance", Pubmed, vol. 344, no. 18, May 2001, pp.1343-1350.

[2] Kemal Polat, Salih Gunes, and Ahmet Arslan, "A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine," Expert Systemswith Applications, vol. 34. 1, January. 2008, pp. 482-487.

[3] Kayaer K and Yildirim T, "Medical diagnosis on Pima Indian diabetesusing general regression neural networks," Proceedings of the international conference on artificial neural networks and neural informationprocessing, 2003, pp. 181-184.

[4] Jack W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R.S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," Proc. Annu. Symp. Comput. Appl. Med. Care,November 9. 1988, pp. 261-265.

[5] Karegowda A. G., Manjunath A. S. and Jayaram M. A., "Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pimaIndians diabetes," International Journal onSoft Computing, vol. 2. 2, 2011, pp. 15-23.

[6] Carpenter G. A. and Markuzon N., "ARTMAP-IC and medical diagnosis:Instance counting and inconsistent cases," Neural Networks, vol. 11. 2,1998, pp. 323-336.

[7] Wold S., Esbensen K. and Geladi P., "Principal component analysis,"Chemometrics and intelligent laboratory systems, vol. 2. 1-3, 1987, pp.37-52.

[8] Balakrishnama S. and Ganapathiraju A., "Linear discriminant analysis-abrief tutorial," Institute for Signal and information Processing, vol. 18,1998.

[9] Deng L. and Yu D., "Deep learning: methods and applications," Foundations and Trends in Signal Processing, vol. 7. 3-4, 2014, pp. 197-387.

[10] Lee H., "Tutorial on deep learning and applications," NIPS 2010Workshop on Deep Learning and Unsupervised Feature Learning, 2010.

[11] Safavian S. R. and Landgrebe D., "A survey of decision tree classifier methodology," IEEE transactions on systems, man, and cybernetics, vol.21. 3, 1991, pp. 660-674.

[12] Suykens J. A. K. and Vandewalle J., "Least squares support vectormachine classifiers," Neural processing letters, vol. 9. 3, 1999, pp. 293-300