Automatic Diagnosis Metabolic Syndrome via a $K-$Nearest Neighbour Classifier

Omar Behadada
Biomedical Engineering Laboratory
Faculty of technology
University of Tlemcen, Algeria
Email: o_behadada@mail.univ-tlemcen.dz

Meryem Abi-Ayad
Biology department
University of Tlemcen, Algeria
Email: m.abi.ayad.2007@gmail.com

Marcello Trovati
Department of Computer Science
Edge Hill University
Ormskirk, UK
Email: Marcello.Trovati@edgehill.ac.uk

Georgios Kontonatsios
Department of Computer Science
Edge Hill University
Ormskirk, UK
Email: georgios.kontonatsios@edgehill.ac.uk

Chikh MA
Biomedical Engineering Laboratory
Faculty of technology
University of Tlemcen, Algeria
Email: mea_chikh@mail.univ-tlemcen.dz

***Abstract*—In this paper, we investigate the automatic diagnosis of patients with metabolic syndrome, i.e., a common metabolic disorder and a risk factor for the development of cardiovascular diseases and type 2 diabetes. Specifically, we employ the $K-$Nearest neighbour (KNN) classifier, a supervised machine learning algorithm to learn to discriminate between patients with metabolic syndrome and healthy individuals. To aid accurate identification of the metabolic syndrome we extract different physiological parameters (age, BMI, level of glucose in the blood etc) that are subsequently used as features in the KNN classifier. For evaluation, we apply the proposed k-NN algorithm against two baseline machine learning classifiers, namely Nave Bayes and an artificial Neural Network, on a manually curated dataset of 64 individuals. The results that we obtained demonstrate that the $K-$NN classifier improves upon the performance of the baseline methods and it can thus facilitate robust and automatic diagnosis of patients with metabolic syndrome. Finally, we perform feature analysis to determine potential significant correlations between different physiological parameters and the prevalence of the metabolic syndrome.**

***Keywords*-$K-$Nearest Neighbour Classifier; Data Mining; Knowledge Extraction;Big Data.**

I. Introduction

The metabolic syndrome (MetS) is defined as the cluster of metabolic abnormalities (e.g., obesity, hypertension, glucose intolerance) that constitute high risk factors for the developments of cardiovascular diseases and type 2 diabetes mellitus [13]. Over the past decade, MetS has been rapidly increasing worldwide affecting both adults and children and thus pauses a major clinical and public health challenge [14], [15].

MetS definition appeared for the first time about 25 years ago when this risk factors clustering and its association with insulin resistance suggested the investigators the existence of a unique pathophysiological condition [7]. In order to provide uniformity in the description of this phenomenon, different diagnostic criteria have been proposed for MetS. Firstly defined by The World Health Organisation in 1998 [8], many international agencies and organisations subsequently proposed various definitions, among which the most widely used are: the Third Report of the National Cholesterol Education Program Expert Panel on Detection, Evaluation and Treatment of High Blood Cholesterol in Adults (NCEP-ATPIII) (National Cholesterol Education Program 2002), the International Diabetes Federation (IDF) (International Diabetes Federation 2006), and the harmonising criteria of the International Diabetes Federation and American Heart Association/National Heart, Lung and Blood Institute (AHA/NHLBI) [9].

Whilst MetS pauses a major clinical and public health challenge, studies report that a large number of patients that meet the criteria of MetS remains largely undiagnosed. As an example, Helminen et al. [18], showed that the manual diagnosis of MetS by general practitioners in Finland achieves a sensitivity of 0.31 and a specificity of 0.73. Computer-aided techniques that employ machine learning methods to learn to identify patients with MetS can thus facilitate a more reliable diagnosis of the metabolic syndrome. Existing approaches for the automatic identification of MetS proposed different machine learning methods including Decision Trees [16], Artificial Neural Networks [19] and Tree Regression [20]. Jose M. Bioucas-Dias and Antonio Plaza [1] proposed an algorithm that uses Multinomial Logistic Regression (MLR) to find the posterior class probability which is aided by a semi supervised segmentation [2], [3].
In this paper we propose $K-$Nearest Neighbour (KNN) as classifier of metabolic Syndrome (MetS), in order to create a robust and accurate knowledge-based system, which provides a crucial insight into MetS diagnosis from a variety of information sources. This paper is arranged as follows. In the rest of this section, we discuss the relevant

Table I
EVALUATION DATA

	Normal	MetS	Male	Female
Class	14	50	19	45
	21.88%	78.12%	29.69%	70.31%

medical background.

In Section II, data preparation and correlation studies are discussed, and in Section III an overview of $K-$Nearest Neighbour (KNN) based classification is presented, which is subsequently investigated and assessed in Section IV. Finally, Section V concludes the paper and discusses future research directions.

II. DATA PREPARATION

A total of 50 metabolic syndrome patients were identified at EPSP-CHU Tlemcen (Algeria), as part of [**REF to PhD thesis of Miss Meryem Abi-Ayad**]. All subjects received physical examination and anthropometric measure at diabetes centre (EPSP Tlemcen) and assessed by a questionnaire. Patients with cardiovascular, renal, hepatic or thyroid diseases were excluded from the study. The patients age ranged between 40 to 69 (average age: 59.78 ± 9.7), Sexes repartition (24.49%Men and 74.51% Women). All subjects have metabolic syndrome according the International Diabetes Federation (IDF, 2005) definition. Patients' waist circumference was $\geq$ 94 cm (men) and $\geq$ 80cm (women). A total of 14 volunteering patients in apparent good health were taken as control in the same range age as the patients, (19% Males and 45% Females) in the period between July 2015 and September 2015.

A. Feature Selection

The descriptors utilised in the feature selection process are as follows

- Age: age of patients;
- ALAT: alanine amino transferase;
- BMI: body mass index;
- CHDL: Cholesterol high density lipoprotein;
- Creatinine: degradation product of creatinine phosphate in the muscle;
- CT: total cholesterol;
- CLDL: cholesterol low density lipoprotein;
- Glycaemia: level of glucose in the blood;
- Proteins: biological macromolecule formed of one or more polypeptide chains;
- Taille: height;
- TourDeTaille: waist circumference;
- Uree: An organic compound with the chemical formula CO (NH2) 2;
- TGT: total triglyceride;
- ASAT: aspartate amino transferase, and finally

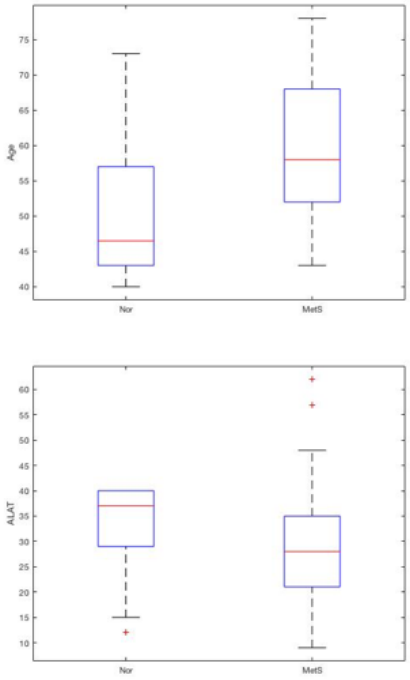

Figure 1. The distribution of features Age and ALAT in different classes

- Poids: weight.

B. Data Visualisation

Classifying two classes: Normal (Nor) and Metabolic Syndrome (MetS) with only one feature, namely Age ALAT, BMI, etc. is, in general, a complex task due to the lack of threshold for each class. Figures 1 o 7 depict such difficulty in distinguishing between classes according of each feature separately. Note that in Figures 4, 5 and 6 it is possible to separate between Nor class and MetS, which is a promising aspect.

C. Correlation Study

The aim of this section is to assess the existence of any relationship between the different features. Figures 7 and 8 show the distribution of samples. In particular, the histograms indicate the presence of a linear correlation. In order to evaluate and assess the relationships between all features and the output, we have investigated the corresponding correlations properties. In particular, the calculation of the R parameters is analysed individually, as well as within the corresponding classes. This has enabled to deduce that the mean feature maximising the relationships within the corresponding classes are Age, CHDL, Glycaemia and TourDeTaille. This confirms that these two parameters are very important in arrhythmias detection, also suggesting the presence of correlation between and we can confirm that there are correlation between some features such as BMI

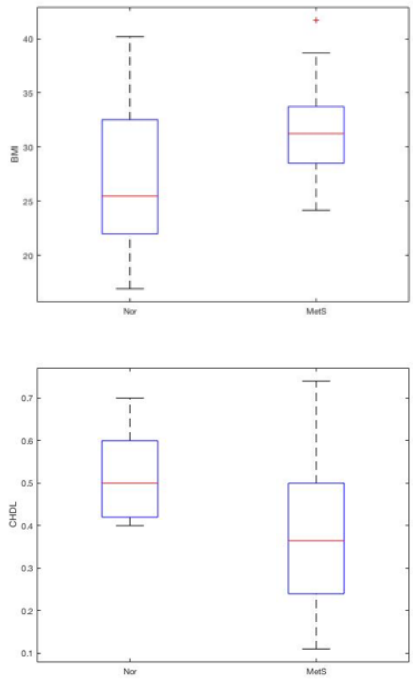

Figure 2. The distribution of features BMI and CHDL in different classes

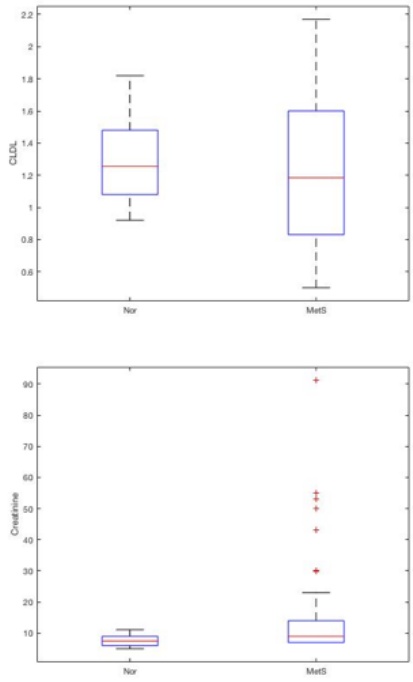

Figure 3. The distribution of features CLDL and Creatine in different classes

/Age, BMI/ Poids (weight), BMI/ TourDeTaille, which is justified due to the fact they depend on the body state.

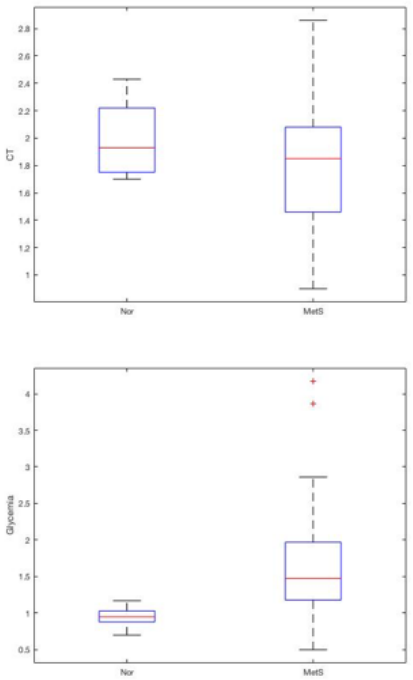

Figure 4. The distribution of features CT and Glycaemia in different classes

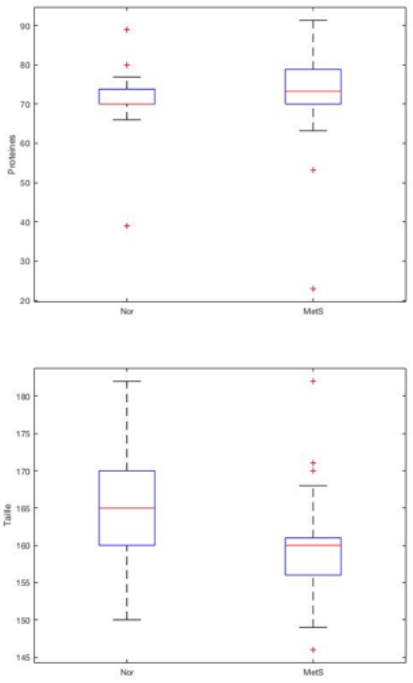

Figure 5. The distribution of features Proteins and Taille in different classes

We also found a strong relationship between CT / CLDL

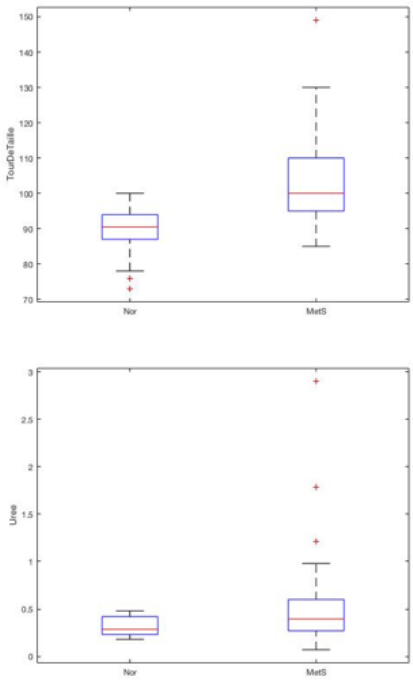

Figure 6. The distribution of features TourDeTaille and Uree in different classes

,which is expected since CT is the addition of CLDL and other component. In conclusion, we can claim that there are some features, which are relevant to the diagnosis of MetS like BMI, Age, CHDL, Glycaemia and TourDeTaille.

III. K-Nearest Neighbour (KNN)

KNN is one of the important non-parameter algorithms [10] and it is a supervised learning algorithm, where classification rules are generated by the training samples without any additional data. Loosely speaking, the training phase is defined by vectors in a multidimensional feature space with a class label attached to each of them. On the other hand, the classification phase is based on an unlabelled vector classified by identifying the most frequent label among the K nearest training samples. This enables the identification of the categories the data are likely to belong.
The best value of K depends on the data itself. In general, large values of K have an impact in the reduction of noise during the classification process. More specifically, as discussed in [11], the classification of a sample X via the KNN algorithm includes the following steps :

- Consider $C_1, C_2, \ldots, C_j$ training categories so that, after feature reduction, they become an $m-$dimension feature vector;
- Let the sample X be the feature vector of the form $(X_1, X_2, \ldots X_m)$;

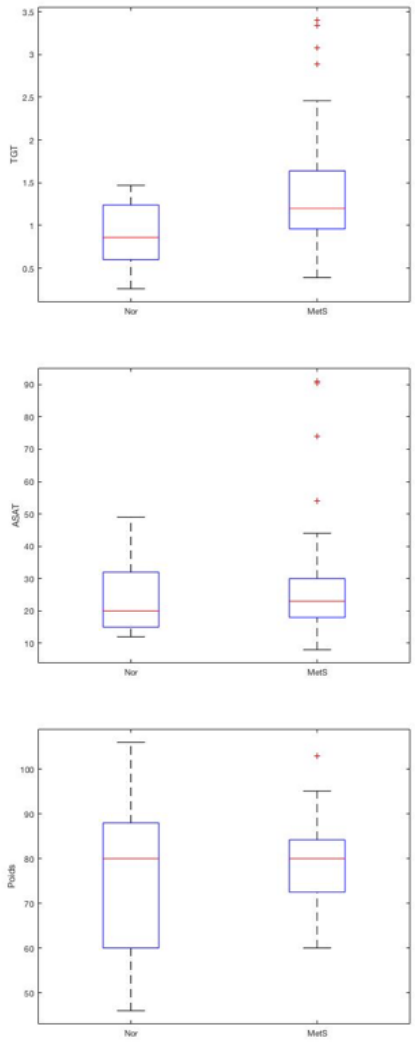

Figure 7. The distribution of features TGT,ASAT and Poids in different classes

- The similarities between all training samples and X are evaluated. As an example, for a sample $d_i = (d_{i1}, d_{i2}, \ldots, d_{im})$, the similarity SIM$(X, d_i)$ is defined as follows

$$\text{SIM}(X, d_i) = \frac{\sum_{j=1}^{m} X_j d_{ij}}{\sqrt{\left(\sum_{j=1}^{m} X_j\right)^2}\sqrt{\left(\sum_{j=1}^{m} d_{ij}\right)^2}} \quad (1)$$

Subsequently, consider K samples which are larger than the values of SIM(X, d_i), for $i = 1, 2, \ldots, N$, and regard them as a KNN collection of X. The probability of X to belong to each category is [12]:

$$P(X, C_j) = \sum_{d} \text{SIM}(X, d_i) y(d_i, C_j), \quad (2)$$

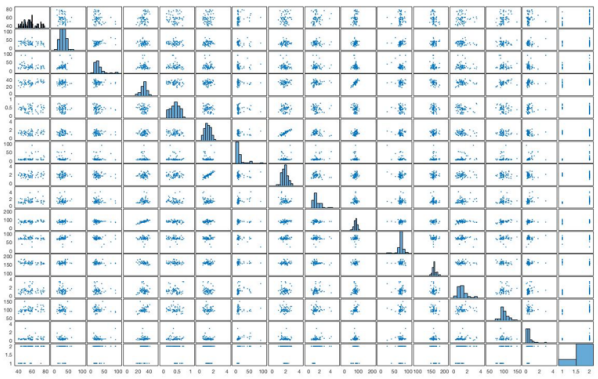

Figure 8. The matrix of correlation

	Age	ALAT	ASAT	BMI	CHDL	CLDL	Creatinine	CT	Glycemie	Poids	Proteines	Taille	TGT	TourDeTaille	Uree	Class
Age	1	-0.2355	-0.2357	0.3587	-0.0120	0.1006	0.1753	0.0962	0.1784	0.2059	0.0163	-0.3367	0.0004	0.2774	0.1191	**0.4013**
ALAT	-0.2355	1	0.4297	-0.034	-0.0579	0.1172	-0.0819	0.1138	-0.0495	0.1133	0.1375	0.2169	-0.1474	-0.0074	0.0909	-0.1426
ASAT	-0.2357	0.4297	1	-0.0004	-0.0742	-0.0691	0.0617	-0.1152	-0.0527	-0.1205	-0.1068	-0.1218	-0.0924	-0.0841	0.2758	0.0975
BMI	**0.3587**	-0.03447	-0.0004	1	-0.1780	0.1588	-0.0916	0.1475	0.1531	**0.8354**	0.0549	**0.4504**	0.0653	0.4427	0.0045	0.3651
CHDL	-0.0120	-0.0579	-0.0742	-0.1780	1	-0.1869	-0.0595	0.0289	-0.0925	-0.1898	0.0049	-0.0766	0.0310	-0.1342	0.0151	**-0.3422**
CLDL	0.1006	0.1172	-0.0691	0.1588	-0.1869	1	-0.0571	0.8802	-0.2374	0.2258	0.0122	0.0909	-0.0848	-0.0197	-0.1215	-0.1074
Creatinine	0.1753	-0.0819	0.0617	-0.0916	0.0595	0.0571	1	0.0478	0.3164	-0.0550	-0.4266	0.0537	0.2753	-0.0275	0.2527	0.2204
CT	0.0962	0.1138	-0.1152	0.1475	0.0289	**0.8802**	0.0478	1	-0.1734	0.2455	-0.0326	0.0868	0.0650	0.0267	-0.1058	-0.1376
Glycemie	0.1784	-0.0495	-0.0527	0.1531	-0.0925	-0.2374	0.3164	-0.1734	1	0.0431	0.06213	-0.1945	0.2402	0.0831	0.2623	**0.4070**
Poids	0.2059	0.1133	-0.1205	**0.8354**	-0.1898	0.2258	-0.0550	0.2455	0.0431	1	0.0426	0.0434	0.0835	**0.4235**	0.0175	0.1796
Proteines	0.0163	0.1375	-0.1068	0.0549	0.0049	0.0122	-0.4266	-0.0326	0.0621	0.0426	1	0.0247	-0.1570	0.0791	0.2126	0.1137
Taille	-0.3367	0.2169	-0.1218	0.4504	0.0766	0.0909	0.0537	0.0868	-0.1945	0.0434	0.0247	1	-0.0230	-0.0413	0.0246	-0.2801
TGT	0.0004	-0.1474	-0.0924	0.0653	-0.0310	-0.0848	0.2753	0.0655	0.2402	0.0835	-0.1570	-0.0230	1	0.1969	0.0987	0.2791
TourDeTaille	**0.2774**	-0.0074	-0.0841	0.4427	-0.1342	-0.0197	-0.0275	0.0267	0.0831	0.4235	0.0791	-0.0413	0.1969	1	-0.0754	**0.4561**
Uree	0.1191	0.09099	0.2758	0.0045	0.0151	-0.1215	0.2527	-0.1058	0.2623	0.0175	0.2126	0.0246	0.0987	-0.07545	1	0.2055
Class	0.4013	-0.1426	0.0975	0.3651	-0.3422	-0.1074	0.2204	-0.1376	0.4070	0.1796	0.1137	-0.2801	0.2791	0.45613	0.2055	1

Figure 9. Correlation coefficients

where $y(d_i, C_j)$ is a category attribute function, which satisfies

$$y(d_i, C_j) = \begin{cases} 1, & \text{if } d_i \in C_j \\ 0, & \text{if } d_i \notin C_j \end{cases} \quad (3)$$

The above equations allow to assign the sample X to be the category with the largest $P(X, C_j)$.

IV. Results and Evaluation

In this paper, we have chosen a $K-$Nearest Neighbour approach, which was tested on the database described above, and the results are represented in Table 4 with a comparison with two others algorithms artificial Neuronal network (ANN) and Naive Bayes classifier (RBF). Figure 10 shows the main properties and performances criteria. In particular, all the parameters depicted in Figure 10 are defined as follows:

- CorrectRate : Correctly Classified Samples / Classified Samples
- ErrorRate: Incorrectly Classified Samples / Classified Samples
- LastCorrectRate : the following equation applies only to samples considered the last time the classifier performance object was updated. This is Correctly Classified Samples / Classified Samples
- LastErrorRate : the following equation applies only to samples considered the last time the classifier performance object was updated, which is Incorrectly Classified Samples / Classified Samples
- InconclusiveRate : Nonclassified Samples / Total Number of Samples
- ClassifiedRate : Classified Samples / Total Number of Samples
- Sensitivity : Correctly Classified Positive Samples / True Positive Samples
- Specificity : Correctly Classified Negative Samples / True Negative Samples
- PositivePredictiveValue : Correctly Classified Positive Samples / Positive Classified Samples
- NegativePredictiveValue : Correctly Classified Negative Samples / Negative Classified Samples
- PositiveLikelihood: Sensitivity / (1 - Specificity)
- NegativeLikelihood: (1 - Sensitivity) / Specificity
- Prevalence: True Positive Samples / Total Number of Samples.

Performance criteria	ANN	RBF	Kmn
CorrectRate	0.9688	0.6562	1
ErrorRate	0.0312	0.3438	0
LastCorrectRate	0.9688	0.6562	1
LastErrorRate	0.0312	0.3438	0
InconclusiveRate	0	0	0
ClassifiedRate	1	1	1
Sensitivity	0.9286	0.3750	1
Specificity	0.9800	0.9375	1
PositivePredictiveValue	0.9286	0.8571	1
NegativePredictiveValue	0.9800	0.6000	1
PositiveLikelihood	46.4286	6	NaN
NegativeLikelihood	0.0729	0.6667	0
Prevalence	0.2188	0.5000	0.2188

Figure 10. Comparison of performance of classifiers

In this work, we have adapted the algorithm of KNN algorithms. A variety of observations have been carried out and we have obtained the classification rate of 100%[1], which is a clear indication that our proposed method is suitable to this type of data for ANN just 96.88% and only 65.62%. Furthermore, it suggests that it successfully addresses the challenge of the classification of metabolic syndrome, with a sensitivity of 100%[2]. We also note that our approach improves the criterion of transparency and interpretability of the process, by the simplicity of implementation of the algorithms. Furthermore, the readability of the results has also been enhanced, which is an important aspect of the interpretation process carried out by cardiologists expert. As a consequence, it is clear that our approach provides an advantage over other methods of classifications.

[1]Need to check this!
[2]Also need to check this!

V. Conclusion

In this paper we have presented a classifier based on $K-$Nearest Neighbour (KNN). Currently KNN offers a major advantage in the classification due to their simplicity. In the medical field, experts need automatic diagnostic support to facilitate and justify their decisions, which tends to lack in several techniques cited in the literature in particular neural networks. The method presented in this paper offers physicians an explicit knowledge base on probability acquired from a medical database. The contribution in the classification process is demonstrated by an accuracy of 100%, which is a very good result compared with others methods. Furthermore, our method offers more flexibility and transparency in the system of detection. In future research, we are aiming to investigate the integration of the approach presented in this paper with fuzzy partition rules [5], to provide an efficient and scalable tool in arrhythmia detection.

Acknowledgment

The authors thank Miss Meryem Abi-Ayad for providing essential annotation information, as well as for their scientific comments regarding this work and to give us access to part of her data.

References

[1] Li, Jun, Jose M. Bioucas-Dias, and Antonio Plaza. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. Geoscience and Remote Sensing, iEEE Transactions on 48.11 (2010): 4085-4098.

[2] Bohning, Dankmar. Multinomial logistic regression algorithm. Annals of the institute of Statistical Mathematics 44.1 (1992): 197-200.

[3] Camps-Valls, Gustavo, and Lorenzo Bruzzone. "Kernel-based methods for hyperspectral image classification." Geoscience and Remote SenSing, iEEE Transactions on 43.6 (2005): 1351-1362.

[4] Behadada O, Trovati M, Chikh MA and Bessis N Big data-based extraction of fuzzy partition rules for heart arrhythmia detection: a semi-automated approach Concurrency and Computation: Practice and Experience, 2015

[5] Ginevra Biino and All, Dissecting metabolic syndrome components: data from an epidemiologic survey in a genetic isolate; Biino et al. SpringerPlus (2015) 4:324 DOI 10.1186/s40064-015-1049-9

[6] Kaur J (2014) A comprehensive review on metabolic syndrome. Cardiol Res Pract 2014:943162. doi:10.1155/2014/943162

[7] Meigs JB, Tracy RP (2000) Invited commentary: insulin resistance syndrome? Syndrome X? Multiple metabolic syndrome? A syndrome at all? Factor analysis reveals patterns in the fabric of correlated metabolic risk factors. Am J Epidemiol 152:908-912. doi:10.1093/aje/152.10.908

[8] Alberti KG, Zimmet PZ (1998) Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. Diabet Med 15:539-553. doi:10.1002/(SICI)1096-9136(199807)15:7¡539::AID-DIA668¿3.0.CO;2-S

[9] Alberti KGMM, Eckel RH, Grundy SM, Zimmet PZ, Cleeman JI, Donato K et al (2009) Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International. Circulation 120:1640-1645.doi:10.1161/CIRCULATIONAHA.109.192644

[10] Belur V. Dasarathy, Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques, Mc Graw-Hill Computer Science Series, IEEE Computer Society Press, Las Alamitos, California, pp. 217-224, 1991.

[11] Y. Lihua, D. Qi, and G. Yanjun, Study on KNN Text Categorization Algorithm, Micro Computer Information, 21, pp. 269-271, 2006.

[12] Suguna1, and Dr. K. Thanushkodi An Improved $K-$Nearest Neighbor Classification Using Genetic Algorithm IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 2, July 2010ISSN (Online): 1694-0784 ISSN (Print): 16940814

[13] Eckel, Robert H., Scott M. Grundy, and Paul Z. Zimmet. The metabolic syndrome. The lancet 365.9468 (2005): 1415-1428.

[14] Heier, E. C., Meier, A., Julich-Haertel, H., Djudjaj, S., Rau, M., Tschernig, T., Lukacs-Kornek, V. (2017). Murine CD103+ dendritic cells protect against steatosis progression towards steatohepatitis. Journal of Hepatology.

[15] M. Blachier, H. Leleu, M. Peck-Radosavljevic, D.C. Valla, F. Roudot-Thoraval The burden of liver disease in Europe: a review of available epidemiological data J Hepatol, 58 (2013), pp. 593-608

[16] Worachartcheewan, A., Nantasenamat, C., Isarankura-Na-Ayudhya, C., Pidetcha, P., and Prachayasittikul, V. (2010). Identification of metabolic syndrome using decision tree analysis. Diabetes Research and Clinical Practice, 90(1), e15-e18. [

[17] Makrilakis, K., Liatis, S., Grammatikou, S., Perrea, D., Stathi, C., Tsiligros, P., and Katsilambros, N. (2011). Validation of the Finnish diabetes risk score (FINDRISC) questionnaire for screening for undiagnosed type 2 diabetes, dysglycaemia and the metabolic syndrome in Greece. Diabetes and metabolism, 37(2), 144-151.

[18] Helminen, E. E., Mntyselka, P., Nykanen, I., and Kumpusalo, E. (2009). Far from easy and accurate-detection of metabolic syndrome by general practitioners. BMC family practice, 10(1), 76.

[19] Ushida, Y., Kato, R., Niwa, K., Tanimura, D., Izawa, H., Yasui, K., and Murohara, T. (2012). Combinational risk factors of metabolic syndrome identified by fuzzy neural network analysis of health-check data. BMC medical informatics and decision making, 12(1), 80.

[20] De Kroon, M. L., Renders, C. M., Kuipers, E. C., van Wouwe, J. P., Van Buuren, S., De Jonge, G. A., and Hirasing, R. A. (2008). Identifying metabolic syndrome without blood tests in young adults - The Terneuzen Birth Cohort. The European Journal of Public