Metabolic Syndrome Risk Forecasting on Elderly with ML Techniques *

Elias Dritsas, Sotiris Alexiou, and Konstantinos Moustakas

Department of Electrical and Computer Engineering, University of Patras, 26504 Rion, Greece
{dritsase, salexiou}@ceid.upatras.gr
moustakas@ece.upatras.gr
http://www.vvr.ece.upatras.gr/en/

Abstract. Metabolic syndrome is a disorder that affects the overall function of the human body. It is manifested by elevated levels of cholesterol and triglycerides, a significant reduction in energy levels, weight gain with visceral fat deposition in the abdomen, and menstrual disorders while increasing the risk of cardiovascular disease, autoimmune diseases and diabetes. A public dataset is exploited to evaluate the metabolic syndrome (MetS) occurrence risk in the elderly using Machine Learning (ML) techniques concerning Accuracy, Recall and Area Under Curve (AUC). The stacking method achieved the best performance. Finally, our purpose is to identify subjects at risk and promote earlier intervention to avoid the future development of MetS.

Keywords: Metabolic Syndrome · Risk Prediction · Machine Learning.

1 Introduction

MetS can be described as a scourge of the modern age, associated with a sedentary lifestyle and poor diet. The rates of people with metabolic syndrome are constantly increasing in the western world. The effects of this increase have already begun to show with the rise of type 2 diabetes from a young age [22]. It is a disease that has no specific symptoms. Hence, its diagnosis is made through laboratory indicators. Insulin resistance (a state of decreased activity and sensitivity, accompanied by increased insulin secretion) is a key feature of the metabolic syndrome that causes a set of symptoms that may not be immediately apparent as being related to the disorder. Under normal conditions, the human body breaks down food into glucose. Insulin is the hormone secreted by the pancreas that helps glucose pass from the blood into the cells [3, 19, 13].

In people with insulin resistance, the body secretes more and more insulin, which leads to the appearance of [17]:

* This work has been supported by the European Union's H2020 research and innovation programme GATEKEEPER under grant agreement No 857223, SC1-FA-DTS-2018-2020 Smart living homes-whole interventions demonstrator for people at health and social risks.

- Hypertension which is characterized by high systolic above 130mmHg or low diastolic blood pressure above 85mmHg.
- Elevated blood sugar levels, namely, fasting glucose above 100 mg/dl.
- Low levels of good cholesterol HDL less than 40mg /dl for men and 50 mg/dl for women.
- Elevated triglycerides with values greater than 150 mg/dl.
- Waist fat deposition with development of central obesity and increased visceral fat: waist circumference over 102 cm in men and over 88 cm in women.

In recent years, more and more research works have scientifically seen that sleep remains a major and valuable aspect of human life as, among else, it regulates the proper functioning of the human body. Hence, its absence may have serious consequences. According to [23, 15, 21], the "incorrect" sleep patterns (short duration <7 hours and long duration >9 hours) can also affect the risk of developing metabolic syndrome.

Based on the above, a big challenge in the healthcare field is the early forecasting of various chronic conditions, such as diabetes (as classification [12, 8] or regression task for continuous glucose prediction [1, 5]), hypertension [7], high cholesterol [11, 9], COPD [4], CVDs [6], stroke [10] etc. Similar to other conditions, several research studies have been conducted for MetS using ML models. In [2], decision tree (DT) was selected for MetS features selection and data classification. This model was evaluated considering the Youden index, Positive Predicted Value (PPV), Negative Predicted Value (NPV), Sensitivity and Specificity. The key idea in this work was to derive possible rules for the determination of MetS that could enhance its diagnosis. In [14], besides to decision tree, the authors applied the support vector machine (SVM) method to predict MetS. Sensitivity, specificity and accuracy were employed to assess their behavior. In [16], the authors investigated the performance of ML methods using data sampling techniques to generate balanced training sets in order to identify dependencies between diabetes mellitus and metabolic syndrome. For this purpose, they applied DT and Naïve Bayes. In [24], SVM, DT, random forest (RF), artificial neural network (ANN), principal component analysis (PCA) and association analysis (AA) are applied for the modelling and construction of predictive models for metabolic syndrome characterization. In [25], the XGBoost model is the best performing in terms of AUC, Accuracy, Precision, F1-score, Specificity and F2-score.

An ongoing comparative study of various ML techniques is the main contribution of this work. Moreover, the current research models will be integrated into the AI services of the GATEKEEPER [1] system, which aims to improve the independence and ability overtime of the elderly and provide information to professionals to support their decision for implementing personalised prevention and intervention plans (lifestyle changes).

The rest of this paper is organized as follows. Section 2 describes the features of the dataset which are used by experts as diagnostic criteria of MetS. Section 3 presents the methods for data balancing and feature importance ranking. Also,

[1] https://www.gatekeeper-project.eu/

Section 4 presents the evaluation of the ML models and Section 5 summarizes the paper.

2 Dataset Description

Our research was based on a dataset from Kaggle. From this dataset, we focused on participants who are over 50 years old. The number of participants is 396, and all the attributes (11 as input to ML models and 1 for target class) are described as follows

- Age (years): This feature refers to the age of a person who (in this study) is over 50 years old.
- Gender: This feature refers person's gender. The number of men is 193 while the number of women is 203.
- Marital: This feature represents the marital status of the participants and has 5 categories(Widowed, Married, Single, Divorced and Separated).
- WaistCirc -WC (cm): It is the measurement taken around the abdomen at the level of the umbilicus.
- BMI (Kg/m^2): This feature captures the body mass index of a person.
- Albuminuria (mg/g): This feature represents the person's urine albumin level and is categorized as normal to mildly increased ($< 30mg/g$), moderate increased ($30 - 300mg/g$ - microalbumin) and severe increased ($\geq 300mg/g$ - macroalbumin) [26].
- UrAlbCr: It captures the urine albumin to creatinine ratio. Its values define the level of albuminuria, as it is explained in the previous feature.
- UricAcid (mg/dL): Uric acid is a chemical created when the body breaks down purines.
- Blood Glucose: This feature captures the person's blood glucose level.
- HDL (mg/dL): High-density lipoprotein absorbs cholesterol and carries it back to the liver.
- Triglycerides (mg/dL): Triglycerides are a type of fat (lipid) found in human blood.
- MetS: This feature represents if a person has Metabolic Syndrome or not.

In Table 1, we summarize the statistical characteristics of the numerical features in the dataset. The participants are older than 50 years, and their maximum age is 80 years. In the current data, the number of participants who have been diagnosed with MetS is approximately similar in the age groups 55-59 (29), 60-64 (32), 65-69 (35) and 70-64 (32), while about two times greater (57) is the number of participants in the age group of older than 75 years. Moreover, the distribution of patients with MetS and albuminuria severity level is shown in Table 2.

Most of the patients have urine albumin to creatinine ratio in the normal to mild range and even less in the moderate class. In the third class, only patients older than 75 years have occurred. As in [18], here, it is also verified that the prevalence of both metabolic syndrome and albuminuria increases with age.

Features	Min	Max	Mean ± std
Age	50	80	67.23 ± 9.36
BMI	15.7	59.2	29.97 ± 5.93
WaistCirc	66.4	145.6	102.92 ± 14.86
UrAlbCr	1.87	338.54	19.98 ± 42.09
Uric Acid	1.9	9.9	5.69 ± 1.45
Blood Glucose	73	327	113.09 ± 33.
HDL	25	108	55.08 ± 15.89
Triglycerides	41	560	139.29 ± 81.4

Table 1. Statistical Characteristics

	level 0	level 1	level 2
50-54	16	3	0
55-59	25	3	0
60-64	30	1	0
65-69	20	1	0
70-74	26	4	0
75+	39	21	2

Table 2. Patients with MetS and Albuminuria level per age group

In this study, 27 women and 24 men suffer from MetS with the simultaneous presence of waist circumference above 88 for females and 102 for males, triglycerides above 150 and HDL lower than 50 and 40, respectively. In Table 3, we see MetS patients distribution for each criterion separately. Also, in the dataset, there are 162 MetS patients with glucose levels above 100. These patients are mainly distributed in the overweight (50) and obese (96) classes. In Table 4, we present the prevalence of MetS in overweight and obesity classes when the waist circumference criterion is also satisfied.

WC	**No**	**Yes**
> 102 **(male)**	34	84
> 88 **(female)**	71	91
HDL	**No**	**Yes**
< 40 **(male)**	5	43
< 50 **(female)**	6	42
Triglycerides	**No**	**Yes**
> 150	12	112

Table 3. WaistCirc, HDL, Triglycerides per class

	No	**Yes**		**No**	**Yes**
Overweight	**48**	**50**	**Obese I**	**29**	**69**
WC > 102	20	22	**WC** > 102	10	39
WC > 88	28	28	**WC** > 88	19	30
Obese II	**8**	**28**	**Obese III**	**4**	**16**
WC > 102	2	15	**WC** > 102	0	8
WC > 88	6	13	**WC** > 88	4	8

Table 4. WaistCirc vs BMI classes

3 Data preprocessing and Feature Importance

In this study, random oversampling has been applied to produce a balanced dataset. For the training of the ML-based models, all features were kept, except for race and income. Based on the relevant literature, we focus only on the most important risk factors for metabolic syndrome. To estimate the importance of an attribute x, we employed Gain Ratio (GR) method. The Gain Ratio of an attribute x is calculated as $GR(x) = \frac{IG(x)}{H(x)}$, where $IG(x)$, $H(x)$ capture the Information Gain and entropy of x, respectively [20]. The entropy of x is defined as $H(x) = -\sum_i P(x_i) log_2(P(x_i))$ where $P(x_i)$ captures the probability to have the value x_i by considering all values of an attribute. In the balanced dataset the features' importance and the related weight are as follows: Triglycerides (0.2972), BMI (0.1795), WaistCirc (0.1698) BloodGlucose (0.1526), HDL (0.1065), Age (0.0845), UricAcid (0.0825), UrAlbCr (0.0519), Albuminuria (0.0357), Marital (0.0105) and Sex (0.0000186).

	Logistic Regression		**SVM (Linear)**		**MultiLayer Perceptron**		**Random Forest**		**Stacking**	
	Yes	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**
Accuracy	0.783		0.785		0.888		0.909		0.919	
Recall	0.773	0.793	0.785	0.785	0.899	0.879	0.919	0.899	0.924	0.914
AUC	0.891	0.891	0.785	0.785	0.927	0.927	0.974	0.974	0.971	0.971

Table 5. Machine Learning Models Performance

4 Performance Evaluation of ML models

In this section, the performance of several ML models is evaluated in the WEKA environment using 10-cross validation on the balanced dataset. Logistic Regression (LR), Support Vector Machine (SVM), MultiLayer Perceptron (MLP), Random Forests (RFs) and a Stacking ensemble (using as base classifiers the previous models while as a meta classifier the LR) was applied.

The results in Table 5 indicate that LR and (linear) SVM models present similar satisfactory accuracy and recall 78.5%. The LR model demonstrated higher AUC, and it can discriminate the prevalence of MetS with a higher probability than SVM in populations similar to the dataset. An even higher performance demonstrated RF (as a single classifier) and the superior outcomes were acquired by the stacking method. Stacking performed best concerning the accuracy and recall metrics with a bit (0.3%) lower AUC. Finally, this model is considered powerful for the personalized risk assessment of the MetS in the context of the GATEKEEPER system.

5 Conclusions

In this research work, a publicly available dataset was considered to examine the order of importance of specific risk factors on MetS, aiming at risk prediction in older people living at home. A limitation of the current study is that in the features set was not available the blood pressure in relation to metabolic syndrome. Several ML methods were assessed and the Stacking method was found to yield the best prediction performance against the single classifiers. The results of the stacking method presented consistently high accuracy (0.919), recall (0.919) and AUC (0.971), a fact that seems promising for the discrimination ability of the model regarding possible subjects with MetS.

In future work, we aim to extend the Machine Learning framework through the use of Deep Learning methods by applying Long-Short-term-Memory (LSTM) algorithm and Convolutional Neural Networks (CNN) based on the same dataset and comparing the results concerning the aforementioned metrics.

References

1. Alexiou, S., Dritsas, E., Kocsis, O., Moustakas, K., Fakotakis, N.: An approach for personalized continuous glucose prediction with regression trees. In: 2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM). pp. 1–6. IEEE (2021)
2. Babič, F., Majnarić, L., Lukáčová, A., Paralič, J., Holzinger, A.: On patient's characteristics extraction for metabolic syndrome diagnosis: predictive modelling based on machine learning. In: International Conference on Information Technology in Bio-and Medical Informatics. pp. 118–132. Springer (2014)
3. Basciano, H., Federico, L., Adeli, K.: Fructose, insulin resistance, and metabolic dyslipidemia. Nutrition & metabolism **2**(1), 1–14 (2005)
4. Dritsas, E., Alexiou, S., Moustakas, K.: Copd severity prediction in elderly with ml techniques. In: Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments. pp. 185–189 (2022)
5. Dritsas, E., Alexiou, S., Konstantoulas, I., Moustakas, K.: Short-term glucose prediction based on oral glucose tolerance test values. In: International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF. vol. 5, pp. 249–255 (2022)
6. Dritsas., E., Alexiou., S., Moustakas., K.: Cardiovascular disease risk prediction with supervised machine learning techniques. In: Proceedings of the 8th International Conference on Information and Communication Technologies for Ageing Well and e-Health - ICT4AWE,. pp. 315–321. INSTICC, SciTePress (2022)
7. Dritsas, E., Fazakis, N., Kocsis, O., Fakotakis, N., Moustakas, K.: Long-term hypertension risk prediction with ml techniques in elsa database. In: International Conference on Learning and Intelligent Optimization. pp. 113–120. Springer (2021)
8. Dritsas, E., Trigka, M.: Data-driven machine-learning methods for diabetes risk prediction. Sensors **22**(14), 5304 (2022)
9. Dritsas, E., Trigka, M.: Machine learning methods for hypercholesterolemia long-term risk prediction. Sensors **22**(14), 5365 (2022)
10. Dritsas, E., Trigka, M.: Stroke risk prediction with machine learning techniques. Sensors **22**(13), 4670 (2022)

11. Fazakis, N., Dritsas, E., Kocsis, O., Fakotakis, N., Moustakas, K.: Long-term cholesterol risk prediction with machine learning techniques in elsa database. In: International Joint Conference on Computational Intelligence (IJCCI). pp. 445–450. SCIPTRESS (2021)
12. Fazakis, N., Kocsis, O., Dritsas, E., Alexiou, S., Fakotakis, N., Moustakas, K.: Machine learning tools for long-term type 2 diabetes risk prediction. IEEE Access **9**, 103737–103757 (2021)
13. Freeman, A.M., Pennings, N.: Insulin resistance. StatPearls [Internet] (2021)
14. Karimi-Alavijeh, F., Jalili, S., Sadeghi, M.: Predicting metabolic syndrome using decision tree and support vector machine methods. ARYA atherosclerosis **12**(3), 146 (2016)
15. Konstantoulas, I., Kocsis, O., Dritsas, E., Fakotakis, N., Moustakas, K.: Sleep quality monitoring with human assisted corrections. In: International Joint Conference on Computational Intelligence (IJCCI). pp. 435–444. SCIPTRESS (2021)
16. Perveen, S., Shahbaz, M., Keshavjee, K., Guergachi, A.: Metabolic syndrome and development of diabetes mellitus: predictive modeling based on machine learning techniques. IEEE Access **7**, 1365–1375 (2018)
17. Raikou, V.D., Gavriil, S.: Metabolic syndrome and chronic renal disease. Diseases **6**(1), 12 (2018)
18. Shih, H.M., Chuang, S.M., Lee, C.C., Liu, S.C., Tsai, M.C.: Addition of metabolic syndrome to albuminuria provides a new risk stratification model for diabetic kidney disease progression in elderly patients. Scientific reports **10**(1), 1–9 (2020)
19. Tappy, L., Lê, K.A.: Metabolic effects of fructose and the worldwide increase in obesity. Physiological reviews (2010)
20. Trabelsi, M., Meddouri, N., Maddouri, M.: A new feature selection method for nominal classifier based on formal concept analysis. Procedia Computer Science **112**, 186–194 (2017)
21. Troxel, W.M., Buysse, D.J., Matthews, K.A., Kip, K.E., Strollo, P.J., Hall, M., Drumheller, O., Reis, S.E.: Sleep symptoms predict the development of the metabolic syndrome. Sleep **33**(12), 1633–1640 (2010)
22. Vollenweider, P., Eckardstein, A.v., Widmann, C.: Hdls, diabetes, and metabolic syndrome. High Density Lipoproteins pp. 405–421 (2015)
23. Wolk, R., Somers, V.K.: Sleep and the metabolic syndrome. Experimental Physiology **92**(1), 67–78 (2007)
24. Worachartcheewan, A., Schaduangrat, N., Prachayasittikul, V., Nantasenamat, C.: Data mining for the identification of metabolic syndrome status. EXCLI journal **17**, 72 (2018)
25. Yang, H., Yu, B., OUYang, P., Li, X., Lai, X., Zhang, G., Zhang, H.: Machine learning-aided risk prediction for metabolic syndrome based on 3 years study. Scientific Reports **12**(1), 1–11 (2022)
26. Zhang, A., Li, M., Qiu, J., Sun, J., Su, Y., Cai, S., Bao, Q., Cheng, B., Ma, S., Zhang, Y., et al.: The relationship between urinary albumin to creatinine ratio and all-cause mortality in the elderly population in the chinese community: a 10-year follow-up study. BMC nephrology **23**(1), 1–10 (2022)